Title

Advancing Neural Granger Causality and Penalization Techniques

Author

Nguyen, Huy Khanh

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Advancing Neural Granger Causality and Penalization Techniques

A thesis submitted in partial satisfaction

of the requirements for the degree

Masters of Science in Statistics

by

Huy Khanh Nguyen

2024

ABSTRACT OF THE THESIS

Advancing Neural Granger Causality and Penalization Techniques

by

Huy Khanh Nguyen
Masters of Science in Statistics
University of California, Los Angeles, 2024
Professor George Michailidis, Chair

Thanks to its interpretability and absence of experimental data, Granger causality has become one of the most powerful tools in causal discovery. Focusing on temporal data, Granger causality evaluates where one time series can be predictive (Granger-causal) of another time series. However, there are several assumptions in traditional Granger causality: no unknown confounder, stationarity, and linearity. Linearity is especially challenging since linear time processes don't happen often in real life. Unlike stationarity which has established methods such as logarithmic transformation, resolving linearity is a much more difficult task. To capture these complex relationships, researchers have proposed many methods: additive nonlinear model, kernel space, and more. To improve interpretability, Structured Neural Networks is particularly powerful. A crucial component of the method is the regularization (penalty) as it helps with lag selection and ensures related features are dealt with together. This thesis first reviews a number of nonlinear Granger causality methods- additive nonlinear Granger causality, independent innovation analysis (IIA), kernel Granger causality, nonlinear permuted Granger causality- then focuses on neural Granger causality. It also explores different penalties including hierarchical lasso, group sparse group lasso, and elastic net which we illustrate on a data set on missing migrants.

The thesis of Huy Khanh Nguyen is approved.

Ying Nian Wu

Frederic Paik Schoenberg

George Michailidis, Committee Chair

University of California, Los Angeles

2024

To my family, friends, and mentors

your guidance and support made me who I am today.

Table of Contents

List of Figures

Acknowledgements

The majority of my codes came from the Git Neural-GC by Dr. Ian Covert of Stanford University with some changes. This is directly from Neural Granger Causality (Tank et al, 2021). The link to the GitHub is here: https://github.com/iancovert/Neural-GC.

I want to extend my deep appreciation to Dr. George Michailidis whose instruction and patience has been instrumental for this thesis.

I am also grateful to my mentor Brandon Thoma (M.S Statistics Class of 2021) in guiding me with formatting and general advice.

Chapter 1

Introduction

1.1 Traditional Granger Causality

Conceived by Nobel laureate Clive Granger in 1969, Granger causality has been pivotal in times series analysis. Despite the rapid advancement of new causal methods, Granger causality is still very popular. The main reason is that experiments are often lengthy, costly, or impossible to conduct due to ethical issues. Granger causality, by relying on observational data, is very applicable to many fields. Granger causality is also interpretable and flexible, which is very important for policy making and modifying the model for robust results.
Despite the name, it is important to understand that Granger causality does not imply causal interaction among time series. Rather, Granger causality evaluates if one time series can be useful in predicting another time series. As defined by Shojaie and Fox (2021), a time series y will Granger-cause another time series x if, by including the past values (history) of x, we can improve the prediction of y, compared to just the past values of y. We measure such improvement by reduction in variance (Shojaie & Fox, 2021):

$$var[x_t - P(x_t|H_{<t})] < var[x_t - P(x_t|H_{<t} \backslash y_{<t})]$$

Where $H_{<t}$ is all the historical values up to time $t-1$, $H_{<t} \backslash y_{<t}$ is the historical values excluding the those of $y_{<t}$, and $P(x_t|H_{<t})$ is the optimal prediction of x_t given $H_{<t}$ (Shojaie & Fox, 2021).

1.2 Vector Autoregressive (VAR)

A key framework for Granger causality is vector autoregressive (VAR). VAR is very important for several reasons. First, VAR can analyze interdependent time series simultaneously. Second, VAR captures temporal dynamics and significance for Granger causality by including lagged values and orders. Finally, VAR allows significance testing (most commonly F-test) to assess Granger causality.

We have the following definition for VAR (Shojaie & Fox, 2021): Let $x_t := (x_{1t}, x_{2t}, ..., x_{pt})^T$ be the values of x at t, d be the lag - the number of past time steps for predicting current value, $A^0, ..., A^d$ be the p x p lag matrices, e_t be the p-dimensional error. Now, to identify a linear model for Granger causality, we apply VAR for when A^0 is diagonal and structural vector autoregression (SVAR) for the general case:

$$A^0 x_t = \sum_{k=1}^{d} A^k x_{t-k} + e_t$$

One very crucial theorem for VAR is as followed (Shojaie & Fox, 2021):

Theorem 1 *Time series x_i is Granger-causal to time series x_j if and only if $A^k_{ij} \neq 0$ for $1 \leq k \leq d$.*

Despite being very rigid, VAR frameworks require many assumptions that are not applicable to real world data. First, the data generating process must be linear. Second, the time series must be stationary - statistical properties stay constant over time. Third, we assume there are no unknown confounders so there can be correct estimates and inferences. Finally, the time series are often assumed to be continuous.

Traditional VAR also has many weaknesses. One of which is the limited number of parameters for the model to execute in a reasonable time frame. This is not applicable to the complex relations in Granger causality. One way to mitigate this is factor-augmented VAR, which accounts for the exogenous variables in our model. Another way is to fit VAR

models for numerous endogenous variables (Shojaie & Fox, 2021). Here, we augment VAR loss function and utilize sparsity-inducing penalties. This turns many entries of matrix A^k to 0, and by theorem 1 we can look at non-zero entries to infer Granger causality. Following Shojaie and Fox (2021), let $\Omega(.)$ be the penalty on A^k for k = 1,..., d, T be the time series length, and λ be the tuning parameter. The least square loss function becomes:

$$\min_{A^1,...,A^d} \sum_{t=d+1}^{T} \|x_t - \sum_{k=1}^{d} A^k x_{t-k}\|_2^2 + \Omega(A^1, ..., A^d)$$

And to make Granger causality more interpretable, one common penalty choice is group lasso penalty (Shojaie & Fox, 2021):

$$\Omega(A^1, ..., A^d) = \lambda \sum_{i,j=1}^{p} \|(A_{ij}^1, ..., A_{ij}^d)\|_2$$

Finally, lag selection is another major weakness of VAR. Inappropriate lag lengths can lead to overfitting or missing Granger causal relationships. Truncating lasso penalty by Shojaie and Michailidis (2010) can assist in finding the optimal lag. From A^{k-1}, we will calculate weights ω to scale penalty $\Omega(.)$ for A^k. This will eventually shrink all coefficients of A^k to 0, with A^{k+1} only contains zero entries (Shojaie & Michailidis, 2010):

$$\Omega(A^1, ..., A^T) = \lambda \sum_{k=1}^{T} \omega^k \sum_{i,j=1}^{p} |A_{ij}^k|$$

Note that $\omega^1 = 1$. Otherwise, for $k > 1$, let $\mathbb{I}(.,.)$ be the convex indicator function (Shojaie & Fox, 2021):

$$\omega^k = \mathbb{I}(A^{k-1}, A : (T-k)\|A_0\| \geq p^2\beta)$$

1.3 Granger Causality with Discrete Values

With the discrete nature of missing migrant data, we need to redefine Granger causality to better fit the situation. Again following the work of Shojaie and Fox (2021), let g_i be the function specifying how the history of p time series map to a time series i. The definition of general VAR model now becomes:

$$x_{ti} = g_i(x_{<t1}, ..., x_{<tp}) + e_{ti}$$

From Shojaie and Fox (2021), we also adjust the theorem:

Theorem 2 *Time series x_j is non-Granger causal to time series x_i if and only if for all $(x_{<t1}, ..., x_{<tp})$ and $x'_{<tj} \neq x_{<tj}$, we have:*

$$g_i(x_{<t1}, ..., x_{<tj}, ..., x_{<tp}) = g_i(x_{<t1}, ..., x'_{<tj}, ..., x_{<tp})$$

Specifically for count data, we can analyze using integer-valued autoregressive (INAR) processes (Shojaie & Fox, 2021).

1.4 Neural Granger Causality for the Global Migrant Situation

For many migrants looking for safety and a better life, they will have to venture through very difficult routes like dense jungles, choppy seas, and barren desserts. Along the way, they may also have to deal with pirates, human traffickers, cartels, to name a few. With so many hazards, for many we never hear any news again. To make matters worse, their plights now become tools to divide governments and organizations. Thus, proper understanding of the situation is ever critical in this uncertain world. Otherwise, we cannot create effective

policies and drain the already limited resources.

As we know, the causes of these disappearances are often similar. An exodus stems likely stems from a combination of factors such as conflict, persecution, environmental degradation, political crisis,... Unsurprisingly, human trafficking worsens the dire situations day by day. Finally, environmental factors like harsh weather and terrain can exacerbate the risks for the migrants as well.

Neural Granger causality can help us understand the migrant situation. As a side note, our time series will be the aggregate counts of deaths and missing migrants by regions. Since these time series are discrete and sparse, neural Granger causality will be very helpful in many ways. First, the discrete nature requires a flexible method to capture the nuanced dynamics. Second, we can choose the right activation functions specifically for count time series. We will chose ReLu for this purpose mainly due to its speed. Finally and most importantly, it is very easy to interpret the time series dynamics with neural Granger causality. However, we have to be careful with our results and final claims. Even with its focus on interpretability, neural Granger causality can still be very difficult to make proper conclusions. That said, the method can still shed lights for us to resolve the migrant crisis.

Chapter 2

Background

As emphasized, the linearity of VAR greatly limited the applications of traditional Granger causality. Below are some predecessors of Neural Granger Causality that address certain aspects of nonlinearity.

2.1 Additive Nonlinear Granger Causality

2.1.1 Additive Non-linear Model

First, we establish what is an additive nonlinear time series model. Chu and Glymour (2008) provided this definition: Let $X_t = \{X_{t1}, ..., X_{tp}\}$ a p-dimensional observed time series, U_t a q-dimensional unobserved time series, ϵ_t a p-dimensional error like above. Then, they denoted $\{X\}_t = \{..., X_1, ..., X_T, ...\}$. For $\{X\}_t$ to be generated from a lag T additive non-linear model, Chu and Glymour (2008) outlined 4 conditions. They state the main condition is that for i = 1,..., p:

$$X_{ti} = \sum_{1 \le j \le p, j \ne i} c_{ji} X_{tj} + \sum_{1 \le k \le p, 1 \le l \le T} f_{kil}(X_{t-l,k}) + \sum_{m=1}^{q} b_{mi} U_{tm} + \epsilon_{ti}$$

Here, c_{ji}, b_{mi} are constants and f_{kil} are smooth univariate functions. We also have two of the many ways to think about causality here. Chu and Glymour (2008) stated X_{tj} is causal to X_{ti} if and only if $c_{ji} \neq 0$. Chu and Glymour (2008) also stated $X_{t-l,k}$ is causal to X_{ti} if and only if $f_{jil} \neq 0$.

2.1.2 Conditional Independence Test

One great advantage of additive time series models is their resistance to the curse of dimensionality. This is due to the inherent nature of additive regression method. As a result, we can efficiently test conditional independence for nonlinear series to see if potential causal links are indeed valid. Chu and Glymour (2008) provided the following: Let X_t^1, X_t^2 be two entries of X_t, X_t^c be any subset of $X_t \backslash \{X_t^1, X_t^2\}$, X_t^d be any subset of $X_t \backslash \{X_t^1\}$, $X^l = \{X_{t-T}, ..., X_{t-1}\}$, $X_{t-i,j} \in X^l$, and $X^e = X^l \backslash \{X_{t-i,j}\}$. Here, Chu and Glymour (2008) emphasized two important results. First, X_t^1 is independent of $X_{t-i,j}$ if and only if the conditional expectation of X_t^1 on X_t^d and X^l is constant in x_{lj}. Second, X_t^1 is independent of X_t^2 conditional on X_t^c and X^l if and only if the conditional expectation of X_t^1 on X_t^2,X_t^c, X^l is constant in x_t^2 the conditional value of X_t^2.

Most importantly, Chu and Glymour (2008) asserted **additive model regression** as a consistent estimator of conditional expectations, which combines with the above findings make conditional independence test possible.

2.1.3 Causal Graphs

A causal graph represents the causal relationships among observed or unobserved variables. Next, a directed acyclic graph (DAG) is the most common way to demonstrate the causal graph. In the case of our causal graph, the vertices and directed edges represent variables and causal link between variables, respectively. Focusing on additive non-linear, causal graphs expand our understanding of the procedure, whether capturing causal relationships, finding causal paths and confounders, or simplifying analysis for causal effect estimation.

We now provide details for causal graph in our case. Let $G =< V, E >$ be our causal graph, $e =< V_i, V_j >$ be edge of G (also indicates V_i is causal to V_j), V_m subset of V, G_m subgraph of G induced by V_m (Chu & Glymour, 2008). Next, they defined d-seperation and faithfulness. Vertices X, Y (random variables) are d-separated with respect to vertex set Z if and only if

between X and Y every undirected path contains a collider with no directed path into any vertex of Z or contains a non-collider that is a vertex of Z. Subsequently, a joint distribution on the vertices of a DAG is faithful if and only if all conditional independence relations (through d-separation property) applied to the DAG.

Chu and Glymour (2008) also underscored three important propositions. First, d-separation among the variables in X_t conditional on X_l in G_c the repetitive causal graph are the same as those in X_t in the subgraph of G_c induced by X_t. Second, let faithfulness be true. If there exists $X_{t-i,j} \in X^l$ such that X_t^2 and $X_{t-i,j}$ are independent conditional on X^e but not $X_{t-i,j}$ and X_t^1, then X_t^1 is not causal to X_t^2.

The third proposition is complicated, so we denoted some new terms. Let all the previous notations be unchanged, and X_t^d be the set of all observed contemporary direct causes of X_t^1. If faithfulness is true, then $X_{t-i,j}$ and X_t^1 are dependent on X_t^d and X^e if and only if one of these two are true. First, $X_{t-i,j}$ is causal to X_t^1. Otherwise, there exists path P between X_t^1, $X_{t-i,j}$ where the ordered set $< W_1, ..., W_m >$ of vertices between X_t^1, $X_{t-i,j}$ that satisfies the 5 causal and graph conditions outlined by Chu and Glymour (2008).

One last step before we go into causal inference procedure of additive non-linear models is its output named Partial Ancestral Graph (PAG). PAGs are a class represent the causal relationships among a set of variables, capturing both the observed and latent variables' interactions. In essence, PAG is an extension of DAG that encode conditional independence relations. In PAG, there is list of vertices (observed variables), 3 types of end points which combined form 4 types of edges representing causal relations between random variables.

2.1.4 Causal Inference Procedure

We will summarize the 3 main steps for the procedure. Note that There are more details and descriptions provided by Chu and Glymour (2008):

First Step - Identify contemporary causal relations: Determine if X_t^1 is independent of X_t^2 conditional on X_t^c and X^l through X_t^1, X_t^2, X_t^c. Then, treat these relations the same as

conditional independence relations between X_t^1, X_t^2 given X_t^c. To do this, generate a PAG denoted as π_t G for the contemporary causal structure among variables in X_t. From this, create set PCDC(X_t^1) of possible contemporaneous direct causes of X_t^1.

Second Step - Identify lagged causal relations: We create π_f with the vertex set $\{X_t, ..., X_{t-T}\}$ and π_t's edges. Determine if $X_t^1, X_{t-i,j}$ are independent given X^e and X_t^b through all of $X_t^1, X_{t-i,j}, X_t^b$. Finally, add the lagged causes (the edges) of each variable in X_t to π_f.

Third Step - Follow the conditions from Chu and Glymour (2008) to adjust contemporary PAG: Repeat edge orientation until no more change to π_f. Subsequently, apply causal inference algorithm FCI (or similar) to orient if still needed our contemporary PAG π_f.

2.1.5 Strengths and Shortcomings

Because its components are linear, non-linear additive causal inference's biggest strength is interpretability. In addition, non-linear additive causal inference is also less complex to other methods which is very important as time series data can be very large.

The primary limitation of non-linear additive causal inference is its reliance on additive forms. This may not capture more intricate or multiplicative interactions between variables. Our data is count time series, so the additive non-linear model itself is not very applicable. Finally, additive nonlinear Granger causality assumes temporal independence. Such assumption can lead to numerous hidden errors. As a result, we need to estimate errors for more arbitrary non-linear VAR (NVAR). This brings us to the next topic: Independent Innovation Analysis.

2.2 Independent Innovation Analysis (IIA) for Granger Causality

2.2.1 Augmented NVAR model

Because NVAR uses past values to make future predictions, we also call NVAR a mixing model. On the other hand, independent innovation analysis (IIA) focuses on finding the inverse (demixing) of NVAR. The main goal of this demixing is to understand causal mechanism and hidden errors better.

To begin, Morioka et al (2020) denoted $x_t = [x_1(t), ..., x_n(t)]^T$ and $s_t = [s_1(t), ..., s_n(t)]^T$ as the observations and innovations at time t respectively. In its most basic form, the NVAR(1) model $f : R^{2n} \to R^n$ is: $x_t = f(x_{t-1}, s_t)$. As before, we assume temporal independence for the errors. From the formula, we can see unlike additive nonlinear VAR, NVAR allows nonlinear interaction between the observations and errors. Subsequently, the authors wanted to estimate innovations s only from observations x through NVAR. One thing to our advantage is that we can extend the model, algorithm, and theorem for NVAR(1) to NVAR(p) by replacing x_{t-1} with $[x_{t-1}, ..., x_{t-p}]$ (Morioka et al, 2020).

To make it possible for demixing, Morioka et al (2020) emphasized we need the invertible augmented NVAR model $\tilde{f} : R^{2n} \to R^n$ since NVAR f is not invertible. Note that we do not make any constraint on f (Morioka et al, 2020):

$$\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} = \tilde{f}\left(\begin{bmatrix} s_t \\ x_{t-1} \end{bmatrix}\right) = \begin{bmatrix} f(x_{t-1}, s_t) \\ x_{t-1} \end{bmatrix}$$

To start with demixing, let $g(x_t, x_{t-1}) \in R^n$ be the sub-space of the demixing model (Morioka et al, 2020). This sub-space shows the mapping from two temporally consecutive x_t, x_{t-1} to the innovation at time t. From Morioka et al (2020), the augmented demixing model for

innovation estimation is:

$$\begin{bmatrix} s_t \\ x_{t-1} \end{bmatrix} = \tilde{g}\left(\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}\right) = \begin{bmatrix} g(x_{t-1}, x_t) \\ x_{t-1} \end{bmatrix}$$

For NVAR, identifiability is an important condition. Identifiability makes sure a unique estimate for the model's parameters given sufficient data is possible. This is a requirement if we want to accurately understand nonlinear causal dynamics. Beyond that, identifiability makes statistical inference of causal relationships consistent, and provides reliable model evaluation.

To ensure identifiability for NVAR, Morioka et al (2020) required to make one important assumption. The distribution of innovations are time-dependent, and modulated through an auxiliary information about the error. Morioka et al (2020) denoted this auxiliary information at time t as random variable u_t. To be more specific, innovation s_i is dependent on m-dimensional random auxiliary variable u. However, s_i is conditionally independent of other innovations s_j. Finally, s_i has a univariate exponential family distribution conditioned on u formulated by Morioka et al (2020):

$$p(s|u) = \prod_{i=1}^{n} \frac{Q_i(s_i)}{z_i(u)} exp[\sum_{j=1}^{k} q_{ij}(s_i)\lambda_{ij}(u)]$$

Here, Q_i is the base measure, Z_i is the normalizing constant, k is the model order, q_ij is the sufficient statistics, and λ_{ij} is a scalar function of u per Morioka et al (2020). Now that we have the innovation model, we can talk about the learning frameworks.

2.2.2 Learning Frameworks

There are 3 frameworks to consider: **General Contrastive Learning Framework, Time-Contrastive Learning Framework, and Hidden Markov Model Framework**. Depending on the scenarios, they will be useful estimators of various causal structures.

General Contrastive Learning Framework (IIA-GCL): In IIA-GCL, the main goals are to train a feature extractor and a logistic regression classifier (Morioka et al, 2020). This classifier will distinguish a real dataset composed of the true observations of (x_t, x_{t-1}, u_t) from one where u is randomized. Morioka et al (2020) denote u^* the random value drawn from the distribution of u while also independent of x_t, x_{t-1}. They next write these two datasets as:

$$\tilde{x}_t = (x_t, x_{t-1}, u_t); \tilde{x}_t^* = (x_t, x_{t-1}, u^*)$$

Morioka et al (2020) also defined our estimator the nonlinear logistic regression system as:

$$r(\tilde{x}_t) = \sum_{i=1}^{n}\sum_{j=1}^{k} \psi_{ij}(h_i(x_t, x_{t-1}))\mu_{ij}(u_t) + \phi(x_{t-1}, u_t) + \alpha(u_t) + \beta(h_i(x_t, x_{t-1})) + \gamma(x_{t-1})$$

Where $\psi_{ij}, h_i, \mu_{ij}, \phi, \alpha, \beta, \gamma$ are scalar-valued functions. These functions are designed to match the difference of the log-pdfs of $x_t, x_{t-1}, u_t)$ in the two datasets (Morioka et al, 2020) and satisfy universal approximation capacity of Hornik et al (1989). We always take into account the innovation model outlined above. Finally, we will use neural network to learn these functions.

The main advantage of IIA-GCL is the certain identifiability of the innovations, up to a permutation and component-wise invertible nonlinearities Morioka et al (2020). This is because of the guaranteed convergence of the learning framework from the following theorem detailed by the above authors:

Theorem 3 *We have 6 assumptions to make: 1) The augmented model is invertible and (sufficiently) smooth; 2) The innovations are temporally independent, follow innovation model. Also, q_{ij} are twice differentiable; 3) Exist nk+1 distinct $u_0, ..., u_{nk}$ so nk × nk matrix $\boldsymbol{L} = (\lambda(u_1) - \lambda(u_0), ..., \lambda(u_{nk}) - \lambda(u_0))$ is invertible (note that $\lambda(u) = (\lambda_{11}(u), ..., \lambda_{nk}(u))^T$); 4) Regression system has universal approximation capability to distinguish $\tilde{x}$ and $\tilde{x}^*$; 5) Augmented function $\tilde{h}(x_t, x_{t-1}) = [h(x_t, x_{t-1}), x_{t-1}]$ is invertible; 6) ψ_{ij} are twice differentiable. Also for each i, exist $\theta \in R^k$ such that for all y if $\sum_{j=1}^{k} \psi_{ij}(y)\theta_j$ is constant, then $\theta = 0$. If*

all of these 6 assumptions are satisfied, then $h_i(x_t, x_{t-1})$ give the independent innovations, up to permutation and scalar (component-wise) invertible transformations. In short, $h(.,.)$ in the regression function provides a consistent estimator of our IIA method.

Time-Contrastive Learning Framework (IIA-TCL): If u_t is within a finite number of class $[1, T]$, IIA-TCL is our choice. IIA-TCL learns through a multinomial logistic regression (MLR) classifier using softmax function (Morioka et al, 2020):

$$p(u_t = \tau | x_t, x_{t-1}) = \frac{exp(\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij\tau})}{\sum_{l=1}^{T} exp(\sum_{i=1}^{n} \sum_{j=1}^{k} z_{ijl})}$$

$$z_{ijl} = \omega_{ijl} \psi_{ij}(h_i(x_t, x_{t-1})) + \phi(x_{t-1}, u_t = l) + b_l$$

With $\omega_{ij\tau}, b_\tau$ the class-specific weight and bias of MLR classifier respectively. Also, ψ_{ij}, h_i, ϕ is the same as IIA-GCL. The theorem for IIA-TCL now becomes (Morioka et al, 2020):

Theorem 4 *We have 7 assumptions similar to IIA-GCL with some additions and modifications: 3) u the auxiliary variable is an integer in $[1, T]$; 4) the nk × (T-1) modulation matrix $\boldsymbol{L} = (\lambda(2) - \lambda(1), ..., \lambda(T) - \lambda(1))$ is full rank row-wise (note that $\lambda(\tau) = (\lambda_{11}(u = \tau), ..., \lambda_{nk}(u = \tau))^T$); 5) MLR system has universal approximation capability to distinguish u_t (class label) from (x_t, x_{t-1}). If all of these 6 assumptions are satisfied, then $h_i(x_t, x_{t-1})$ give the independent innovations, up to permutation and scalar (component-wise) invertible transformations. In short, $h(.,.)$ in the regression function provides a consistent estimator of our IIA method.*

Hidden Markov Model Framework (IIA-HMM): When u is not observed, IIA-HMM is the method of choice. Morioka et al (2020) stated we assume the nonstationarity is described by hidden states following a discrete-time Markov model. To be more specific, the temporal structure for u is as followed. The auxiliary $u_t \in 1, ..., C$ shows a hidden random states at t. Next, Morioka et al (2020) described u_t by a Markov chain governed by a time-invariant transition-probability matrix $\mathbf{A} \in R^{C \times C}$. Here, they denote A_{ij} as the probability

of transitioning from state i to j.

With hidden Markov chain u_t, NVAR observation model which generates the innovations at time t has likelihood (Morioka et al, 2020):

$$p(x_0, ..., x_T; \mathbf{A}, \theta) = p(x_0) \prod_{t=1}^{T} |\mathbf{J}\tilde{g}(x_t, x_{t-1})| \times \sum_{u_1,...,u_T} \pi_{u_1} p(s_1|u_1; \theta) \prod_{t=2}^{T} \mathbf{A}_{u_{t-1},u_t} p(s_t|u_t; \theta)$$

We have λ the parameters of innovation model, g the demixing model with augmented $\tilde{g}$, $\pi = (\pi_1, ..., \pi_C)$ the stationary distribution of latent state u, $p(x_0)$ the marginal distribution of x_0, $\mathbf{J}\tilde{g}$ the Jacobian matrix of $\tilde{g}$, $\theta = \{\lambda, g\}$ (Morioka et al, 2020).

EM algorithm is often used to identify the innovations. E-step will try to find the optimal $(u_1, ..., u_T)$. M-step will parameters of the model so as to have MLE. The work for better understanding of identifiability for IIA-HMM is still in progress (Morioka et al, 2020).

2.2.3 Strengths and Shortcomings

In essence, IIA is a versatile method to uncover subtle or hidden causal connections. More importantly, the use of demixing models make causal mechanisms more interpretable.

On the other hand, EM algorithm and neural network can make IIA difficult to implement. There are numerous assumptions for each type of IIA. All of these can be a major issue with the variety in the real world. Finally, IIA may not be suitable for every types of nonlinear data. This is especially when facing with heteroskedasticity or non-independency. Consequently, kernel Granger causality is often more powerful since kernel methods are non-parametric and flexible.

2.3 Kernel Granger Causality

2.3.1 Connection with Information Theory

To understand how we can quantify kernel causal tests, we need to establish the relationship between kernel Granger causality and information theory. First, Schreiber (2008) defined transfer entropy as a measurement of the amount of information transferred from one system to another. It also indicates the flow of information between these systems. Now, consider two time processes X and Y. Denote the past k states of X and past l states of Y as x_t^k and y_t^l respectively. Let y_{n+1} the next state of Y, $T_{X\to Y}$ be the transfer entropy from X to Y given the past states of X and Y. From Marinazzo et al (2008) we have:

$$T_{X\to Y} = \sum p(y_{n+1}, y_t^l, x_t^k) log(\frac{p(y_{n+1}|y_t^l, x_t^k)}{p(y_{n+1}|y_t^l)})$$

With transfer entropy established, let us focus on time series. Marinazzo et al (2008) denoted time series $\{\xi_n\}_{n=1,...,N+m}$. For this time series, we can approximate using $p(\xi_n|\xi_{n-1}, ...\xi_{n-m}) = p(\xi_n|\xi_{n-1}, ...\xi_{n-m-1})$. This is the stationary Markov process of order m. Next, the authors denoted for i = 1, ..., N the N realizations of stochastic variables X and x as $X_i = (\xi_i, ..., \xi_{i+m-1})^T$ and $x_i = \xi_{i+m}$ respectively. Finally, they defined $R[f] = \int dXdx(x - f(X))^2 p(X, x)$ the risk functional. Marinazzo et al (2008) provided that the minimizer of $R[f]$- the best estimate of x given X- is $f^*(X) = \int dxp(x|X)x$. Similarly, they let $\{\eta_n\}_{n=1,...,N+m}$ be defined the same as $\{\xi_n\}_{n=1,...,N+m}$. So, $Y_i = (\eta_i, ..., \eta_{i+m-1})^T$. Marinazzo et al (2008) provided that the best estimate of x given X, Y is now $g^*(X, Y) = \int dxp(x|X, Y)x$. This also means Y does not provide any useful information to predict x. Now, the Markov property is $p(x|X) = p(x|X, Y)$. It then follows that (Marinazzo et al, 2008):

$$f^*(X) = g^*(X, Y)$$

From the formula for $T_{X\to Y}$, clearly then $T_{X\to Y} = 0$. In other words, if causality exists, $T_{X\to Y} \neq 0$.

2.3.2 Granger Causality Index

For traditional linear Granger causality, we evaluate causal connections based on established test statistics such as F-test. However, sample size, amount of parameters, and lack of cross validation can make these tests prone to overfitting. Thus, Marinazzo et al (2008) introduced a new estimator - filtered linear **Granger causality index**. We will outline the procedure for Granger causality index below.

We start with linear Granger causality. Note that the empirical risk $ER[f]$ is our choice instead $R[f]$ since $N < \infty$. To minimize $ER[f]$, consider $\alpha \in \{1, ..., m\}$. Marinazzo et al (2008) wrote $u_\alpha \in R^N$ as the vector generated by the samples of the α-th component of X with mean 0. Now, they denoted normalized $x = (x_1, .., x_N)^T$ with mean 0 and defined the linear regression of x_i versus X at X_i as $\tilde{x_i}$. Then, the space to look for the minimizer is the span of $u_1, ..., u_m$. They called this space H, meaning $\tilde{x} = \{\tilde{x_1}, ..., \tilde{x_m}\}$ is the projection of x on H. Marinazzo et al (2008) emphasized H coincides with the range of the N $\times$ N matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ consists of u_α as rows. Finally, Marinazzo et al (2008) stated the prediction error given X is $\epsilon_x = 1 - \tilde{x}^T\tilde{x}$. Subsequently, let $v_\alpha \in R^N$ be the vector generated by the samples of the α-th component of Y with mean 0. Follow previous steps, they denoted the space spans by $u_1, ..., u_m, v_1, ..., v_m$ as H' and the projection of x on H' as x'. Similarly, the authors let H' be the range of the N $\times$ N matrix $\mathbf{K'} = \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ consists of u_α and v_α as rows. they also stated that $\epsilon_{xy} = 1 - \tilde{x'}^T\tilde{x'}$. It is clear that $H' = H \oplus H^\perp$. From all of the above, for linear case Marinazzo et al (2008) defined Granger causality index as:

$$\delta(Y \to X) = \frac{\epsilon_x - \epsilon_{xy}}{\epsilon_x} = \frac{\|P^\perp x\|^2}{1 - \tilde{x}^T\tilde{x}}$$

Where P and $P^{\perp}$ are the projector on H and $H^{\perp}$ (Marinazzo et al, 2008). The authors asserted the orthonormal basis of $H^{\perp}$ is the eigenvectors of $\tilde{\mathbf{K}} = \mathbf{K}' - \mathbf{P}\mathbf{K}' - \mathbf{K}'\mathbf{P} + \mathbf{P}\mathbf{K}'\mathbf{P}$ with non-zero eigenvalues. They labeled these eigenvectors $t_1, ..., t_m$ and denote r_i the Pearson correlation between x and t_i. Overall, Marinazzo et al (2008) obtained: $\|P^{\perp}x\|^2 = \sum_{i=1}^{m} r_i^2$. However, we also have to filter out some r_i to prevent false causality by combining t test and Bonferroni correction. In the end, the filtered Granger causality index is (Marinazzo et al, 2008):

$$\delta(Y \to X) = \frac{\sum_{i'} r_i'^2}{1 - \tilde{x}^T \tilde{x}}$$

2.3.3 Extension to Kernel

We extend Granger causality index for nonlinear case. According to Marinazzo et al (2008), let k be kernel function: $k = \sum_a \lambda_a \Psi_a(X)\Psi_a(X')$ with λ the eigenvalues associated with the kernel function, $\Psi(.)$ the eigenfunction corresponding to the eigenvalue. Now, they stated the search space H is the range of the N x N Gram matrix $\mathbf{K}$ where $K_{ij} = k(X_i, X_j)$. Following the same definitions and steps for linear case, Marinazzo et al (2008) stated $\tilde{x}$ coincides with the linear regression of x in the feature space spanned by $\sqrt{\lambda_a}\Psi_a$.

Next, to predict x using both X and Y, Marinazzo et al (2008) constructed Z with samples $Z_i = (X_i Y_i)^T$, Gram matrix $\mathbf{K}'$ with $K'_{ij} = k(Z_i, Z_j)$. Subsequently, x' the regression value is same as the range of $\mathbf{K}'$. We can find filtered non-linear Granger causality index. We have two common kernel choices detailed below.

Inhomogeneous polynomial (IP): An IP kernel of order p is $k_p(X, X') = (1 + X^T X')^p$. In this case, Ψ_a are the monomials up to the p-th degree in the input variable (Marinazzo et al, 2008). The authors let $\tilde{\mathbf{K}} = \mathbf{K}' - \mathbf{P}\mathbf{K}' - \mathbf{K}'\mathbf{P} + \mathbf{P}\mathbf{K}'\mathbf{P}$. By choosing only the eigenvectors of $\tilde{\mathbf{K}}$ that meet the Bonferroni test, we obtain the filtered non-linear Granger causality index. For IP kernel, not only the index can help us evaluate causality without fear of overfitting, but also the degree of interaction (linear, quadratic,...). This is because our choice of order p can maximixe $\delta_F(X \to Y)$ (Marinazzo et al, 2008).

Gaussian: A Gaussian kernel is $k_\sigma(X, X') = exp(-\frac{(X-X')^T(X-X')}{2\sigma^2})$. One major distinction compared to linear or IP kernel is that $H \subseteq H'$ is not always true. We must adjust H accordingly. Marinazzo et al (2008) considered H the span of the eigenvectors of $\mathbf{K}$ whose eigenvalue is at least $\mu\lambda_{max}$. Here, μ is a very small number of our choice and λ_{max} is the largest eigenvalue of $\mathbf{K}$. We follow similar steps to assess and obtain $\mathbf{K'}$, but this time denote w the eigenvectors of $\mathbf{K'}$. Marinazzo et al (2008) also let ρ_i be eigenvalues of $\mathbf{K'}$ where ρ_i is at least $\mu\lambda'_{max}$ (λ'_{max} is the largest eigenvalue of $\mathbf{K'}$). Subsequently, they define $\mathbf{K}^*$:

$$\mathbf{K}^* = \sum_{i=1}^{m_2} \rho_i w_i w_i^T$$

Finally, redefine $\tilde{\mathbf{K}} = \mathbf{K}^* - \mathbf{P}\mathbf{K}^* - \mathbf{K}^*\mathbf{P} + \mathbf{P}\mathbf{K}^*\mathbf{P}$ and obtain filtered Granger causality index as before.

2.3.4 Strengths and Shortcomings

Kernel Granger causality is a step up in resolving non-linear Granger causality. The filtered Granger causality index can perform regardless of the degree of nonlinearity and thus not suffer from overfitting. Also, we can see that kernel Granger Causality is an extension of traditional linear Granger causality, so it is more interpretable than other advanced methods. However, the weakness of kernel Granger causality is in choosing the appropriate kernel. Thus, for some case such as count time series the method can be not as flexible. Kernel Granger causality may also suffer from overfitting and is sensitive to noises. Thus, we will need a method that can have robust causal tests and reduce overfitting. Nonlinear permuted Granger causality fits these criteria well.

2.4 Nonlinear Permuted Granger Causality

2.4.1 Redefining Granger Causality

We will first make some modifications for Granger causality to fit the work of Gade and Rodu (2023). Let x_t be our time series which we use to predict future of time series y_t. The authors denote $\mathcal{P}$ the optimal prediction function and $\mathcal{I}_{<t}$ the information prior to t. Finally, $X_{<t}$ is a matrix consisting of $x_t \in R^p$ prior to t. Now, they define Granger causality as we often understand: x_t is Granger-causal for y_t if $\text{Var}[y_t - \mathcal{P}(y_t|\mathcal{I}_{<t})] < \text{Var}[y_t - \mathcal{P}(y_t|\mathcal{I}_{<t} \backslash X_{<t})]$. From Gade and Rodu (2023), information isolation will happen when we screen through penalized variable selection. Thus, we will disregard the collective covariate set and likely up-weight contributions of the nonzero covariates. It is important then for us to redefine Granger causality to a permuted framework with out-of-sample testing (NPGC). This way, we can have both learning flexibility and improving identifiability.

Note that $\mathcal{P}(y_t|\mathcal{I}$ is hard to obtain. We will need to add an inherent explanatory covariate set $z_t \in R^q$. We redefine Granger causality - x_t is Granger-causal for y_t given z_t and history of response if (Gade & Rodu, 2023):

$$Var[y_t - \mathcal{P}(y_t|Y_{<t}, Z_{<t}, X_{<t})] < Var[y_t - \mathcal{P}(y_t|Y_{<t}, Z_{<t})]$$

In essence, this new definition means that if we reject the null hypothesis, x_t is truly causal to y_t as it provides information beyond that of z_t. Next, we also define conditional independence where z_t is an extension (Gade & Rodu, 2023): At a specific time t, x_t does not Granger cause y_t if and only if given $Y_{<t}$ then $Y_{<t+1} \perp X_{<t}$. One very important remark is that we cannot simply use this statement to conduct independent test because it may not be true for all time t.

Overall, these definitions matter because we have to understand the limitations of Granger causality. Often, we focus on the influence of individual variables which is affected by hidden

multiplicity. This is when we look too much into individual causal relationships and forget how a group of variables can collectively influence our prediction.

2.4.2 Out-of-sample Predictability

For many existing methods, Gade and Rodu (2023) emphasized that they rely on in-sample tests to assess causal relationships. The issue of this is that neural networks, under the right setting, can give approximations to any functional relationship between covariate sets. As such, we may connect two variables with no predictive relationship especially with overfitting situations where noises are mistaken for significant patterns.

To identify useful functional relationships, then, out-of-sample predictability is necessary. To accomplish this, the authors use a permutation structure which once more requires modifications to the definition of Granger causality. Permutation is very effective as it prevents dependence structure with response and maintain dependence among covariates (Gade & Rodu, 2023). Now, we have a null model and a restricted model. The authors denoted $\tilde{X}_{<t}$ as the copy of $X_{<t}$ with random permutations of rows (ie time). We now have two crucial definitions of conditional Granger causality (Gade & Rodu, 2023):

Definition 1: x_t is not conditional Granger causal for y_t given z_t and the history of $Y_{<t}$ if and only if

$$Y_{<t+1}|Y_{<t}, Z_{<t}, X_{<t} \overset{\mathrm{d}}{=} Y_{<t+1}|Y_{<t}, Z_{<t}, \tilde{X}_{<t}$$

Definition 2: x_t is not conditional Granger causal for y_t given z_t and the history of $Y_{<t}$ if

$$Var[y_t - \mathcal{P}(y_t|Y_{<t}, Z_{<t}, X_{<t})] < Var[y_t - \mathcal{P}(y_t|Y_{<t}, Z_{<t}, \tilde{X}_t)]$$

2.4.3 Structure

Gade and Rodu (2023) denoted ω as the set of possible realizations, and the data from ω as $(X, Y, X)_\omega$. Also, they denote $\omega_{obs} = \{1, ..., \varphi\}$ our set of observations.

We will describe the structure from Gade and Rodu (2023). First, select an appropriate γ lag values of y_t. This will give us representative history $Y_{lag} \in R^{(T+\gamma)\times\gamma d}$ (act as $Y_{<t}$ per our two definitions). We can only use the last T rows of X, Y, Z because the data length is finite $(T+\gamma)$. Because feed forward networks (FNNs) is used to capture nonlinear functional dependence, individual variables (columns) are standardized.

Next, create M permutations $\tilde{X}_m = \Pi_m X$ (row shuffling for X). For m = 1, $\Pi_1 = I$ i.e $\tilde{X}_1 = X$. This will be for our original model. Otherwise they will be for permuted model.

Next, generate predictor matrix $[1Y_{lag}Z\tilde{X}_m] \in R^{T\times(1+\gamma d+q+p)}$. Combine W^0, b^0 to create $W \in R^{(1+\gamma d+q+p)\times N}$ with conditions specified in Gade and Rodu (2023). With activation function g = tanh, the FNN model becomes (Gade & Rodu, 2023)::

$$H_m = tanh([1Y_{lag}Z\tilde{X}_m]W)$$

Next, the featurizations W_r (r=$1, ..., \mathfrak{R}$) have their behaviors aggregated to help with uncertainty. Now, the authors let $U_{m,r}$ be the variation in Y not captured by the functional relationship with the feature space $H_{m,r}$:

$$Y = H_{m,r}W^L_{m,r} + U_{m,r}$$

Next, Gade and Rodu (2023) let Θ_m be the covariance matrix for predicting Y given $Y_{lag}, Z, \tilde{X}_m$. Over ω all possible realization, they also let $\vartheta_m = tr(\Theta_m)$ be our out-of-sample variation parameter for estimation. Given realization $(X, Y, Z)_\omega$ and true functional form f, they define the specific covariance matrix of the prediction as $\sum_{m,\omega}$. Given a random featurization r, the estimate of $\sum_{m,\omega}$ is $S_{m,\omega,r}$ (Gade & Rodu, 2023).

Finally, we will use cross validation for each permutation to estimate ϑ_m.

2.4.4 Granger Causal Influence Test

We will detail **Granger Causal influence test** as our estimator for the influence. From Gade and Rodu (2023), we first choose sufficiently large featurization dimension N to linearize functional relationships. Then, split data to K sets for modelling and testing. For each set $k = 1, ..., K$, they state the number of observation is $T_k = \lfloor T/K \rfloor + 1(TmodK) \geq k$.
Again from Gade and Rodu (2023), let us generate $\mathfrak{R}$ FNNs and fix W_r for consistent error estimation. Now, let training data have subscript -k and R be the out-of-sample prediction residuals. We have then (Gade & Rodu, 2023):

$$\mathsf{R}_{m,\omega,r,k} = H_{m,\omega,r,k}(H'_{m,\omega,r,-k}H_{m,\omega,r,-k})^{-1}H'_{m,\omega,r,k}Y_{\omega,-k} - Y_{\omega,k}$$

Next, the authors denote $\hat{\vartheta}_m$ the estimate for the out-of-sample variation in prediction residuals. They measure the estimate as:

$$\hat{\vartheta}_m = \frac{1}{\varphi\mathfrak{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathfrak{R}}\sum_{k=1}^{K}\frac{1}{T_k}tr(\mathsf{R}'_{m,\omega,r,k}\mathsf{R}_{m,\omega,r,k})$$

Now, we approximate the null distribution of variation from each model m using the described permutation structure. Next, Gade and Rodu (2023) draw the variation estimates from:

$$\hat{\vartheta}_m \sim \hat{\mathcal{H}}(s) = (T!)^{-1}\sum_{i=1}^{T!}1\{\hat{\vartheta}_i \leq s\}$$

The empirical distribution from a size M subsample is (Gade & Rodu, 2023):

$$\hat{\mathcal{H}}(s) = \frac{1}{M}\sum_{m=1}^{M}1\{\hat{\vartheta}_m \leq s\}$$

In addition, they let $\hat{Q}_M = \hat{\mathcal{H}}(\hat{\vartheta}_1) = \frac{1}{M}\sum_{m=1}^{M}1\{\hat{\vartheta}_m \leq \hat{\vartheta}_1\}$. With this, given test level α, rejecting $H_0 : \hat{Q}_M \leq \alpha$ means X is Granger causal to Y conditional on Z and Y_{lag}.

2.4.5 Strengths and Shortcomings

Of all the methods in this paper, nonlinear permuted Granger causality (NPGC) is the most advanced and reliable because of its focus on out-of-sample predictability. It tackles specifically towards overfitting issues. This way, NPGC is more likely to have correct assessment of Granger causal effects. More so, the structure of NPGC means it prioritizes collective inference which has been disregarded previously.

For NPGC, its shortcoming is being computationally expensive. The permutation method may play a huge part in this. More importantly, the results are too advanced and convoluted for interpretations. While not as good in dealing with overfitting, our method of interest is one with strong interpretability: Neural Granger causality.

Chapter 3

Neural Granger Causality

3.1 Neural Networks Adaptation

Let the past of series i be $x_{<ti} = (..., x_{(t-2)i}, x_{(t-1)i})$ (Tank et al 2022). As before, the nonlinear autoregressive model is:

$$x_t = g(x_{<t1}, ..., x_{<tp}) + e_t$$

Where e_t is additive zero mean noise. Note that there are shortcomings of the original version. First, it is difficult to assign correct weights to have one series Granger-cause another series but not Granger-cause anything else. Second, it assumes we can apply the same past lags to every time series which rarely is the case in practice.

Thus, it is best to focus on how the past K lags are mapped to time series i (Tank et al 2022):

$$x_{ti} = g_i(x_{<t1}, ..., x_{<tp}) + e_{ti}$$

From Tank et al (2022), time series j is Granger non-causal to time series i if for all $x_{<t1}, ..., x_{<tp}$ and $x_{<tj} \neq x'_{<tj}$:

$$g_i(x_{<t1}, ..., x_{<tj}, ..., x_{<tp}) = g_i(x_{<t1}, ..., x'_{<tj}, ..., x_{<tp})$$

3.2 Multilayer Perceptron (MLP)

3.2.1 Definition

As the focus is on the components, we call the **method component-wise MLP (cMLP)**. According to Tank et al (2022), let MLP g_i have L-1 layers, $h_t^l \in R^H$ contain values at time t for m-dimensional l-th hidden layer, $W = \{W^1, ..., W^L\}$ be weights at each layer, $b = \{b^1, ..., b^L\}$ be biases at each layer. For the first layer, they let weight $W^1 \in R^{H \times pK}$ be $W^1 = \{W^{11}, ..., W^{1K}\}$. For $1 < l < L$, $W^l \in R^{H \times H}$. Also from the authors, $W^L \in R^H$, $b^l \in R^H$ for $1 < l < L$, and $b^L \in R$. Finally, let σ be the activation function of choice such as logistic or tanh. Now, at time t, for the first layer (Tank et al 2022):

$$h_t^1 = \sigma(\sum_{k=1}^{K} W^{1k} x_{t-k} + b^1)$$

Note that j does not Granger-cause i if for all k, $W_{:j}^{1k}$ is always a zero matrix. For the remaining L-2 layers (Tank et al 2022):

$$h_t^l = \sigma(W^l h_t^{l-1} + b^l)$$

The authors next denoted W^L the linear output decoder, h_t^L the hidden output from (L-1)-th layer. From them, our time series x_{ti} after passing through L-1 layers:

$$x_{ti} = g_i(x_{<t}) + e_{ti} = W^L h_t^{L-1} + b^L + e_{ti}$$

Now, from h_t^1 we know j does not Granger-cause i when for all k $x_{(t-k)j}$ has no impact on h_t^1 and x_{ti}. So, to select Granger causality our objective is (Tank et al 2022):

$$min_W \sum_{t=K}^{T} (x_{ti} - g_i(x_{(t-1):(t-K)}))^2 + \lambda \sum_{j=1}^{p} \Omega(W_{:j}^1)$$

Tank et al (2022) emphasized the penalty Ω will shrink $W^1_{:j} = (W^{11}_{:j}, ..., W^{1K}_{:j})$ to 0 and λ controls group sparsity. With that, there are three choices of penalties: **group lasso, group sparse group lasso, hierarchical group lasso**.

Group lasso will shrink equally the weights associated with lags for input series j (Tank et al 2022). It can only detect very few Granger causal connections estimations:

$$\Omega(W^1_{:j}) = \|W^1_{:j}\|_F$$

Group sparse group lasso can better help find lags with Granger causal effects. It can provide sparsity across groups and within each group (Tank et al 2022), and we control the tradeoff through $\alpha \in (0, 1)$:

$$\Omega(W^1_{:j}) = \alpha\|W^1_{:j}\|_F + (1-\alpha)\sum_{k=1}^{K}\|W^1_{:j}\|_2$$

Hierarchical group lasso can resolve Granger causality and lag order of interaction at the same time. The penalty is particularly powerful form large K as Granger causal relationships for higher lags are still accounted for (Tank et al 2022):

$$\Omega(W^1_{:j}) = \sum_{k=1}^{K}\|(W^{1k}_{:j}, ..., W^{1K}_{:j})\|_F$$

From Tank et al (2022), hierarchical group lasso can also do two things. First, it makes sure for each j, there exists lag k such that $W^{1k'}_{:j} = 0$ for $k' > k$ and $W^{1k'}_{:j} \neq 0$ otherwise. Second, for all k, the penalty also sets many $W^{1k}_{:j}$ columns to zero.

3.2.2 Optimization of Objectives

To interpret Granger non-causality, the input matrices columns must be exactly zero. As a result, proximal gradient descent is the algorithm of choice. In addition, we can incorporate line searches to guarantee convergence to a local minimum (Tank et al 2022). The authors denoted $\mathfrak{L} = \sum_{t=K}^{T}(x_{ti} - g_i(x_{<t}))^2$ the prediction loss, $prox_{\lambda\Omega}$ the proximal operator for

penalty Ω, $\gamma^(m)$ the step size. Starting with $W^{(0)}$, Tank et al (2022) updated network weights W as followed:

$$W^{(m+1)} = prox_{\gamma^(m)\lambda\Omega}(W^{(m)} - \gamma^{(m)}\nabla\mathfrak{L}(W^{(m)}))$$

Note that at higher levels the proximal steps for weights is the identity function (Tank et al 2022).

Specifically for group lasso, we use soft thresholding:

$$prox_{\gamma^{(m)}\lambda\Omega}(W^i_{:k}) = soft(W^i_{:k}, \gamma^{(m)}\lambda)$$

Denote $(x)_+ = max(0, x)$, we then have that (Tank et al 2022):

$$prox_{\gamma^{(m)}\lambda\Omega}(W^i_{:k}) = \Big(1 - \frac{\gamma^{(m)}\lambda}{\|W^1_{:j}\|_F}\Big)_+ W^1_{:k}$$

For step-by-step pseudocodes of proximal gradient descent as well as proximal steps with different penalties, we refer to the work of Tank et al (2022).

3.3 Recurrent Neural Network (RNN)

3.3.1 Definition

Even if we will not use it, component-wise RNN, specifically component-wise long short-term memory (cLSTM), can be another powerful option. We will only work with single-layer RNN since generalizations take similar forms. At time t, let h_t be H-dimensional hidden output. To update h_t, we may apply long short-term memory (LSTM). The details of LSTM formulas for h_t are in the paper of Tank et al (2022). LSTM works well for Granger causality as it can represent real nonlinear connections between time series which may also be long-range.

Now, x_{ti} the output for series i at time t is (Tank et al 2022):

$$x_{ti} = g_i(x_{<t}) + e_{ti} = W^2 h_t + e_{ti}$$

For details, they described the construction of input weight W^1, output weight W^2, and weight matrix W. To detect Granger causality using group lasso penalty, the objective is (Tank et al, 2022):

$$min_W \sum_{t=2}^{T} (x_{ti} - g_i(x_{<t}))^2 + \lambda \sum_{j=1}^{p} \|W^1_{:j}\|_2$$

3.3.2 Optimization of Objectives

Similar to cMLP, we also optimize the objective of cLSTM through proximal gradient descent. To compute the gradient, we use full or truncated back-propagation through time (BPTT) (Tank et al 2022). Again, the authors provided more pseudocodes of proximal gradient descent for objective optimization in the case of general group lasso penalty.

3.4 Strengths and Shortcomings

Both cMLP and cLSTM are powerful tools depending on our needs. For cLSTM, it can capture long range dependencies between time series which is very effective in sequence prediction. For cMLP, it is useful when our computational resources are limited. Overall, their main strengths are in interpreting complex causal connections. Such interpretability is possible thanks to flexible design architecture, inherent advantages of neural networks, and penalization schemes that ensures the network to represent non-causal relationships in the inputs' outgoing weights instead of other configurations (Tank et al 2022).

The main downside of NGC is the input parameter or penalty. Different combinations often lead to varied results, which in turn makes our conclusions unreliable. Predictability assessment can turn bad for a variety of reasons: overfitting, in-sample tests, individualistic

inferences,... Despite all this, the interpretability makes Neural Granger Causality the most useful method for our studies.

3.5 Additional Remarks

In the study, we will only consider cMLP over cLSTM for several reasons. cLSTM is much more computationally expensive. cMLP is simpler for modifications and additions so we can see how the different penalties perform. Also thanks to its simplicity cMLP can be more interpretable. Finally, the number of time points of 115 is quite small, so we can leave aside long-term dependencies for now.

Beside the penalties from Tank et al (2022), we can also use **elastic net penalty**. Unlike the above penalties, elastic net disregard the potential grouping of input features. In this case, the regularization term will be (Parikh & Boyd, 2014):

$$\lambda(||W||_1 + \frac{\nu}{2}||W||_2^2)$$

Where $\nu > 0$. From Parikh and Boyd (2014), the proximal gradient descent will have two steps: soft thresholding (for L_1), then multiplicative shrinkage (for L_2). Elastic net penalty can be more flexible, faster though may hinder interpretability. For our study, we will choose three penalties - group sparse group lasso, hierarchical lasso, and elastic net.

Chapter 4

Analysis and Results

4.1 Data Introduction, Cleaning, and Transformation

4.1.1 Introduction

Taken from Kaggle, our missing migrant dataset came from Kaggle user Nidula Elgiriyewithana. For more details, the reference section of our thesis provided the respective link. We credited that Elgiriyewithana (2023) scraped his data was from the Missing Migrants Project. It is an initiative of the International Organization for Migration (IOM) to record these tragic cases around the world. The main mission of the Missing Migrants Project is bring about more coordinated and united efforts among nations, an aspirations that we also hope to carry in this paper.

There are 13020 incidents (the rows) and 20 attributes (the columns) in the data set. We are only interested in the following 4 attributes: **Reported Month, Incident Year, Region of Origin, Total Number of Dead and Missing**. Here, we want to focus specifically on the origin of the dead and missing migrants. This way, we can better understand the dynamics among the origins of the migrants. Finally, there are 34 regions of origin for migrants. However, many of them only have a few time points. Thus, we will outline our data cleaning and transformation below.

4.1.2 Data Cleaning and Transformation

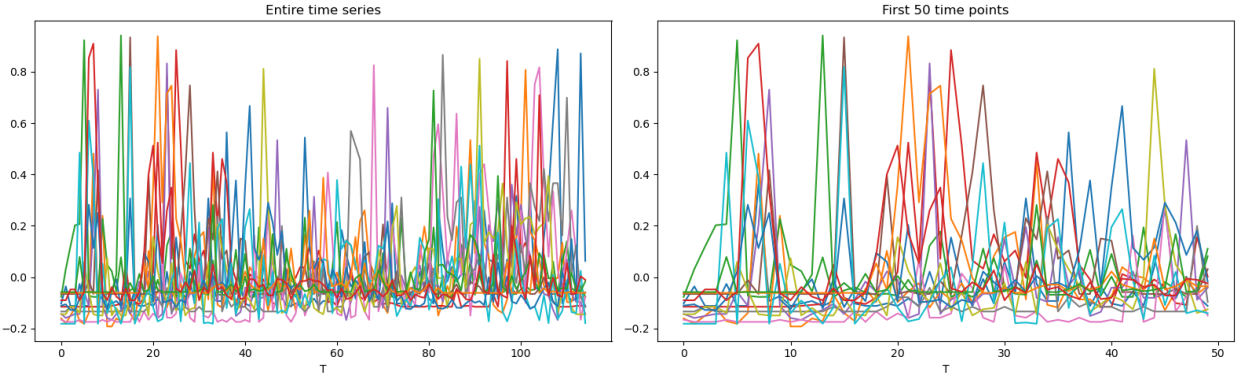
We will first combine Month and Year to a new variable called Date. We will next create time series based on the regions of origins. For each region, we sum the number of dead and missing migrants in each month-year pair of Date. To have sufficient data for analysis, each time series must have at least 50 non-zero counts. In the end, we have 14 such time series following this order: **Caribbean (1), Central America (2), Eastern Africa (3), Eastern Africa (P) (4), Latin America / Caribbean (P) (5), Mixed (6), Northern Africa (7), South America (8), Southern Asia (9), Sub-Saharan Africa (P) (10), Unknown (11), Western / Southern Asia (P) (12), Western Africa (13), Western Asia (14)**. Note that (P) stands for presumed, so we still keep these times separate from the confirmed region of origins. Subsequently, the maximum number of time points T is 115. For each time series, will fill empty time points with 0 assuming no dead or missing victims in that month and year.

When using the untransformed time series count data, the results showed the loss value remained high. Worse, the potential causal link plot looked very unsatisfactory. We believe we did not adjust the parameters enough. Through looking at Dr. Covert's Neural-GC NVAR simulation as well as trials and errors (standardization, min-max normalization, and unit length normalization), we would transform each time series with **mean normalization**:

$$\tilde{r} = \frac{r - \bar{r}}{max(r) - min(r)}$$

This way each time series will center around 0 and the effects of outliers are minimized. We now have the mean normalized count time series plots below:

Figure 4.1: Plots of the 14 Time Series when T = 115 (Left) and when T = 50 (Right)



4.2 Results

4.2.1 Algorithms and Parameters

We have two training algorithms to consider which are ISTA and GISTA. They are both designed to promote sparsity. Note that ADAM was considered but we eliminated it due to bad results. This is because ADAM does not promote sparsity. Unsurprisingly, ADAM converges too quickly when our data contains many zero values.

Iterative Shrinkage-Thresholding Algorithm (ISTA): To begin, ISTA finds the first loss through mean squared error and controls the weights through ridge regularization. Next, it performs gradient descent with backpropagation. Now, ISTA combines this with a learning rate to calculate the parameters. The sparsity of these parameters are encouraged through the proximal update from L_1 regularization. Then, as ISTA reaches the look back checkpoint after specific number of iterations, it promotes sparsity once more through a new smooth loss calculation and appropriate addition of non-smooth loss penalty. This checkpoint also ensures early stopping so we don't overfit.

Generalized Iterative Shrinkage-Thresholding Algorithm (GISTA): In order to improve ISTA especially after parameter update, GISTA employs line search mechanism. This can help the GISTA find a better learning rate and lead to earlier convergence. More over,

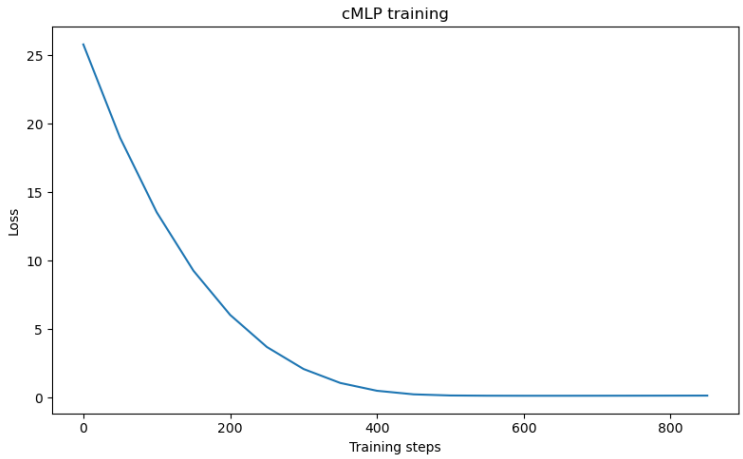
GISTA will handle the convergence of every single network. This means that for each component, the control and optimization will be refined and nuanced.
Coupled with hierarchical penalty being the most interpretable, we will mainly use Hierarchical-GISTA results. However, we may use other penalty-algorithm pairs to have a consensus when making key points.

4.2.2 ISTA

Elastic Net

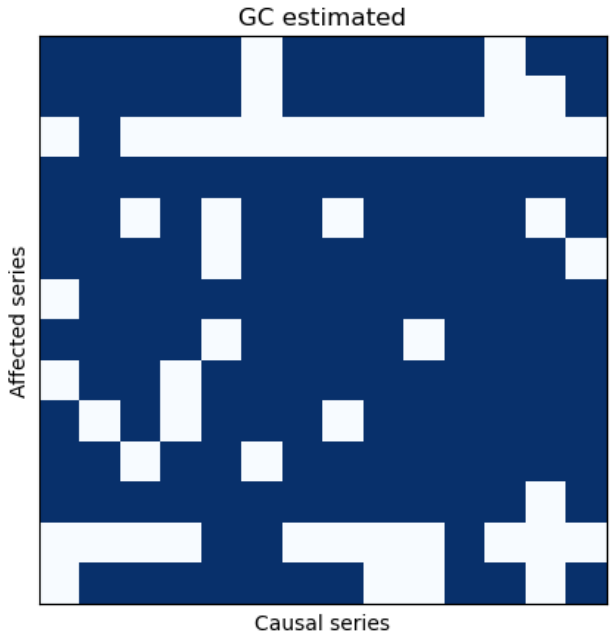
Since elastic net is the inferior penalty as outlined above, we will start with it. The loss function below shows rapid convergence with less than 1000 iteration needed. This is not a good sign as it indicates overfitting, extreme learning rate,... As such, we will not go too deeply into the Granger-causal relationships here and just consider it for consensus later.

Figure 4.2: Loss Function: Elastic Net - ISTA



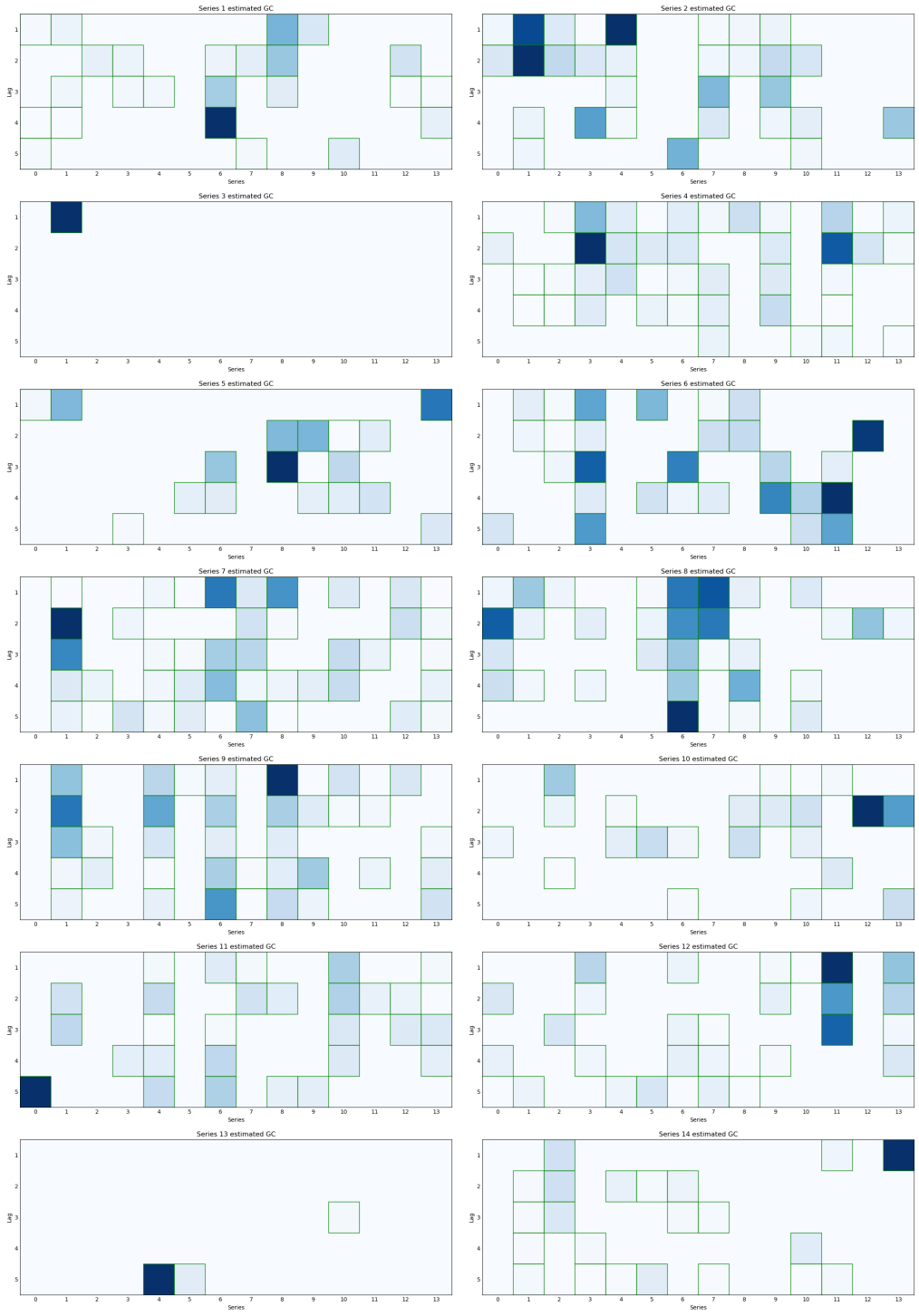
Below is figure 4.3 showing the potential causal relationships of time series. The column is our time series, and the row is the affected time series. For instance, last (14th) row, second column has a blue square. This means time series 2 (Central America) may Granger cause time series 14 (Western Asia). The results are very unrefined because the map indicate Granger causality almost everywhere.

Figure 4.3: General Granger Causality: Elastic Net - ISTA



Going into greater detail, figure 4.4 below will show Granger causal effects of each time series on others. The order of the time series are row by row. Similarly, the lag order of each time series is row by row in each plot. The lag order indicates how far back (in months) of information we used when predicting. For instance, the plot in the 4th row, 1st column is that of time series 7 (Northern Africa). Here, we look at the second column (Central America) and observe the green outer squares for all 5 lags. This means that for up to 5 months, time series 7 has predictive power (Granger-causal) over time series 2. The color of the inner squares indicate how large the Granger causal estimates are. For time series 7, we see the blue colors of row 2 and 3 in column 2. This means the data from two and three months ago play a huge role in predicting the "current" data of time series 2.

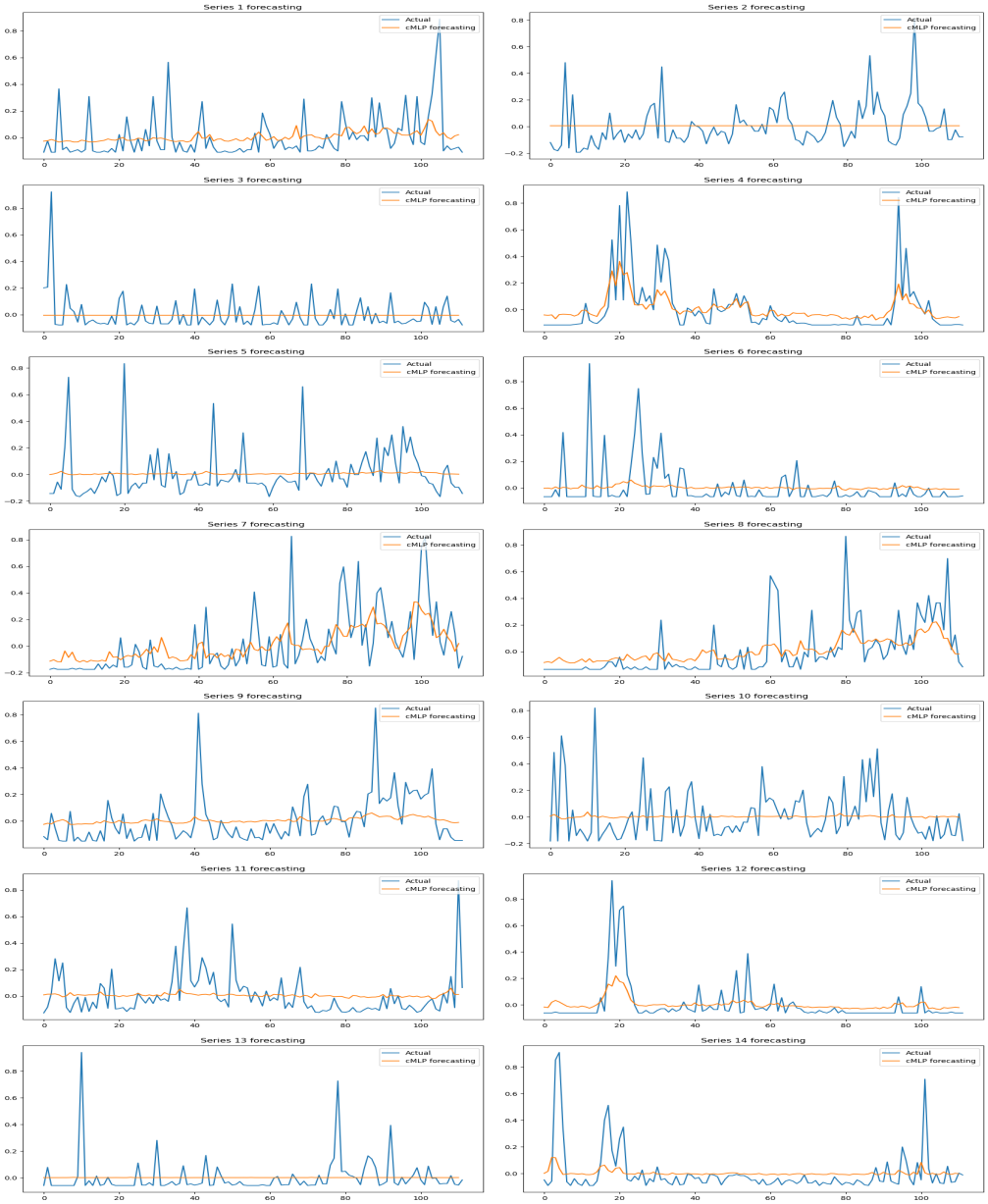
Figure 4.4: Granger Causality with Lag Order: Elastic Net - ISTA



Finally, we have forecasting plots for the time series in figure 4.5. The orange line shows the predicted time series using the Granger causal model, and the blue line is the true time series. We observe that many orange lines is almost horizontal, meaning elastic net is not a

good choice in forecasting.

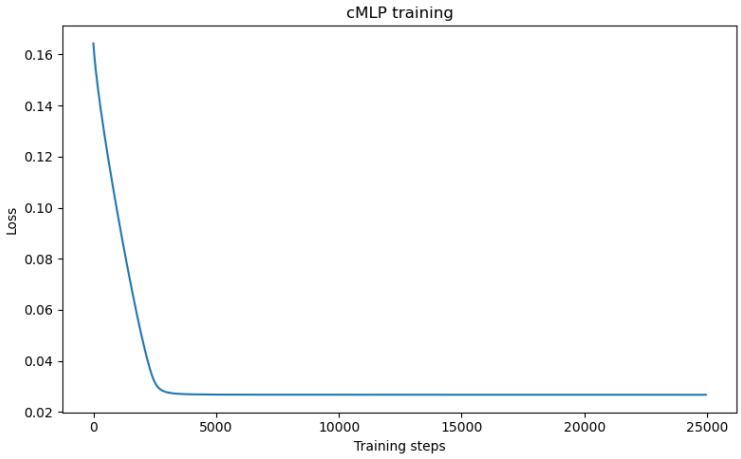
Figure 4.5: Forecasting Time Series: Elastic Net - ISTA



Group Sparse Group Lasso

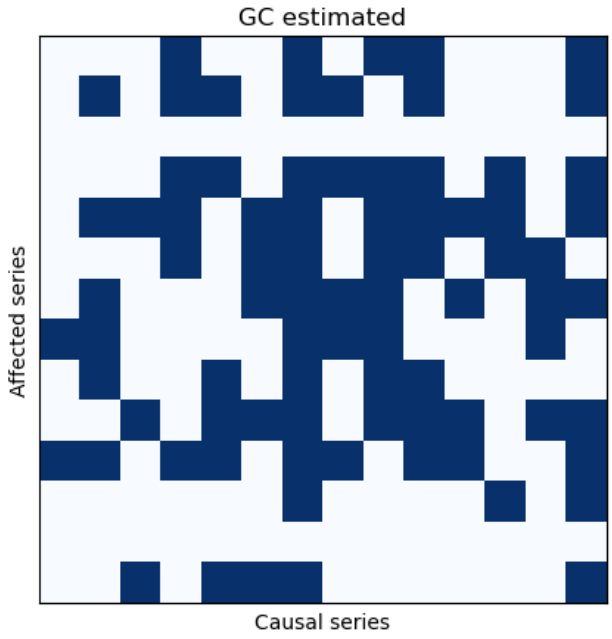
The loss function for GSGL is much better now as there are more iterations. This will allow for more exploration of solution space and improved learning rate.

Figure 4.6: Loss Function: GSGL - ISTA



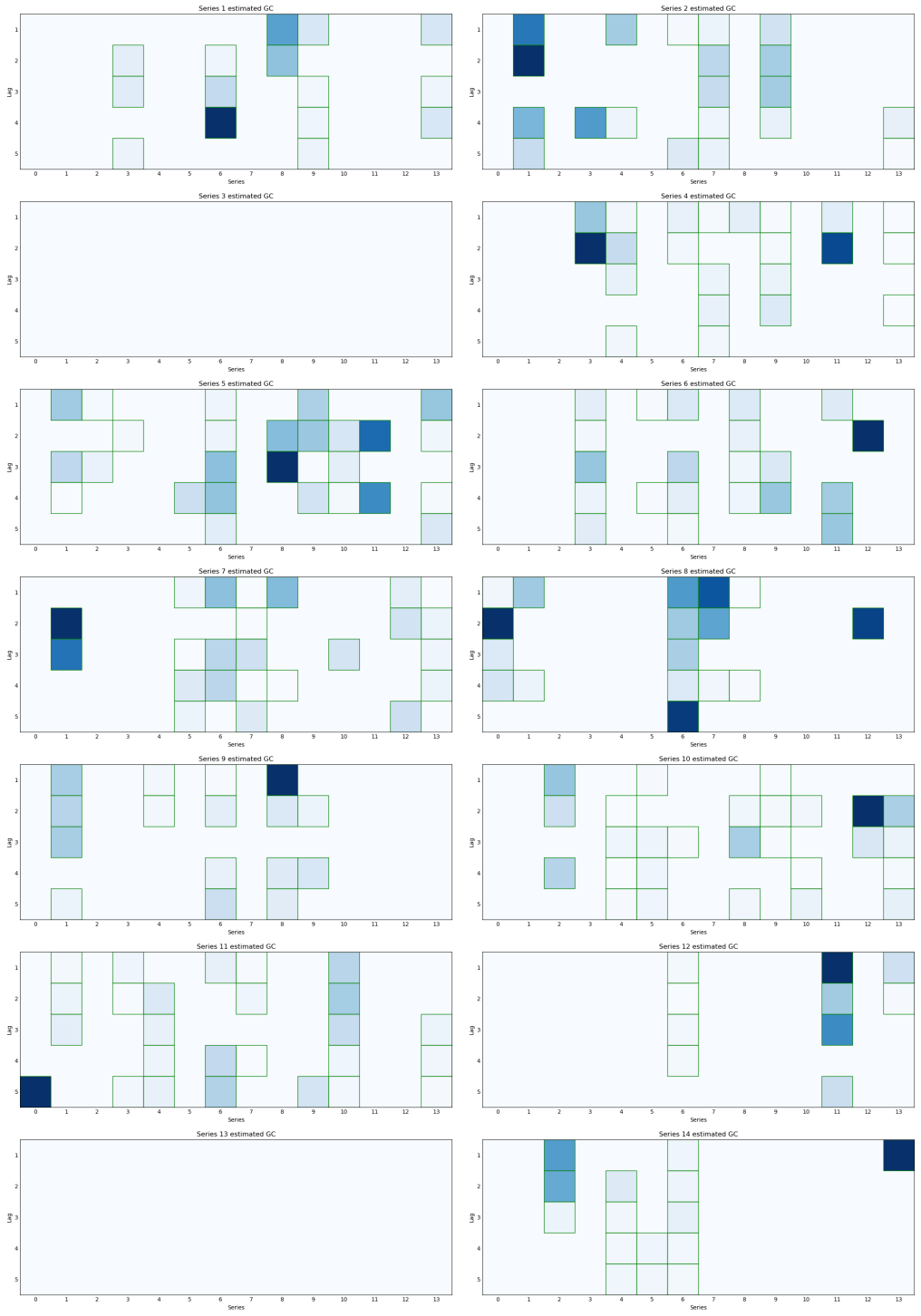
For figure 4.7, the Granger causal estimations are now more refined. That said, we only want to focus on the most influential time series. Thus, we will only focus on those that affected at least 7 series or half of all series. These time series are: 7 (Northern Africa), 9 (Southern Asia), 10 (Sub-Saharan Africa (P)), 14 (Western Asia).

Figure 4.7: General Granger Causality: GSGL - ISTA



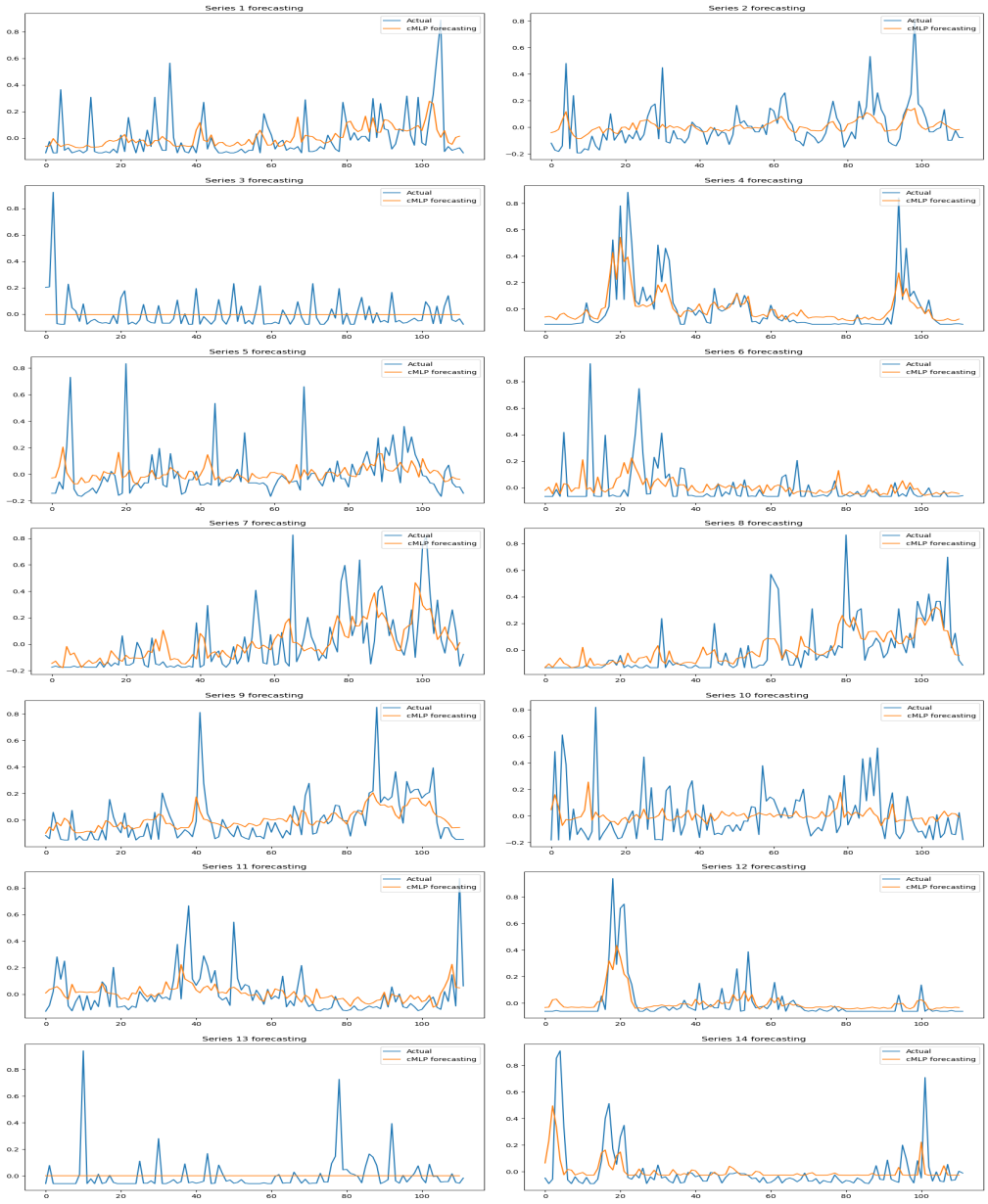
For figure 4.8, we can see that the Granger causal estimations are all over the lags. As such, we will not elaborate further and instead will rely on hierarchical penalty for interpretation.

Figure 4.8: Granger Causality with Lag Order: GSGL - ISTA



The forecasting plots are also better than those of elastic net. However, we can see horizontal forecast for series 3 (Eastern Africa) and 13 (Western Africa). They seem to have no connections with other regions of origin.

Figure 4.9: Forecasting Time Series: GSGL - ISTA



Hierarchical Lasso

The loss function below is again what we expected.

Figure 4.10: Loss Function: Hierarchical - ISTA

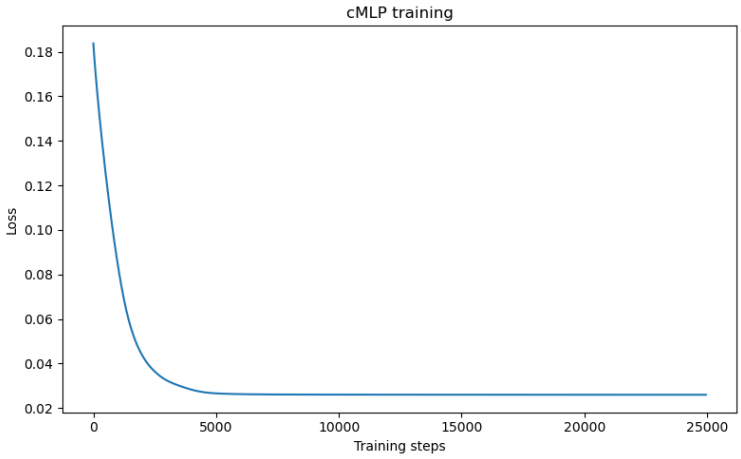


Figure 4.11 demonstrates hierarchical penalty is even more refined. The influential time series are: 7 (Mixed), 9 (Southern Asia), 10 (Sub-Saharan Africa (P)), 14 (Western Asia).

Figure 4.11: General Granger Causality: Hierarchical - ISTA

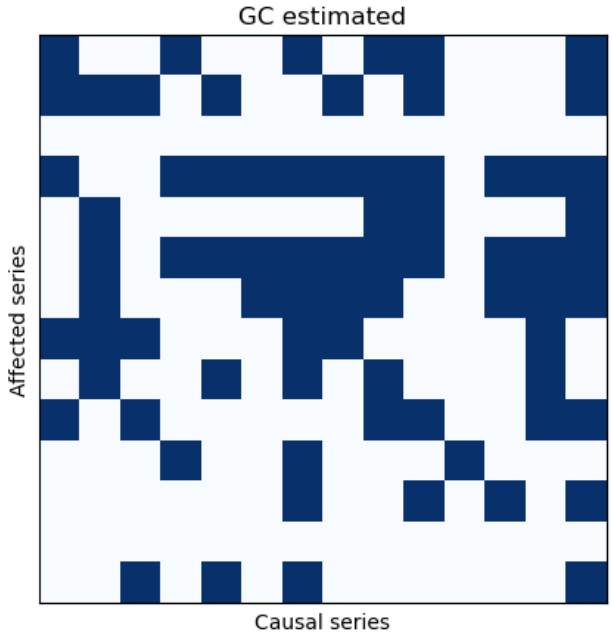
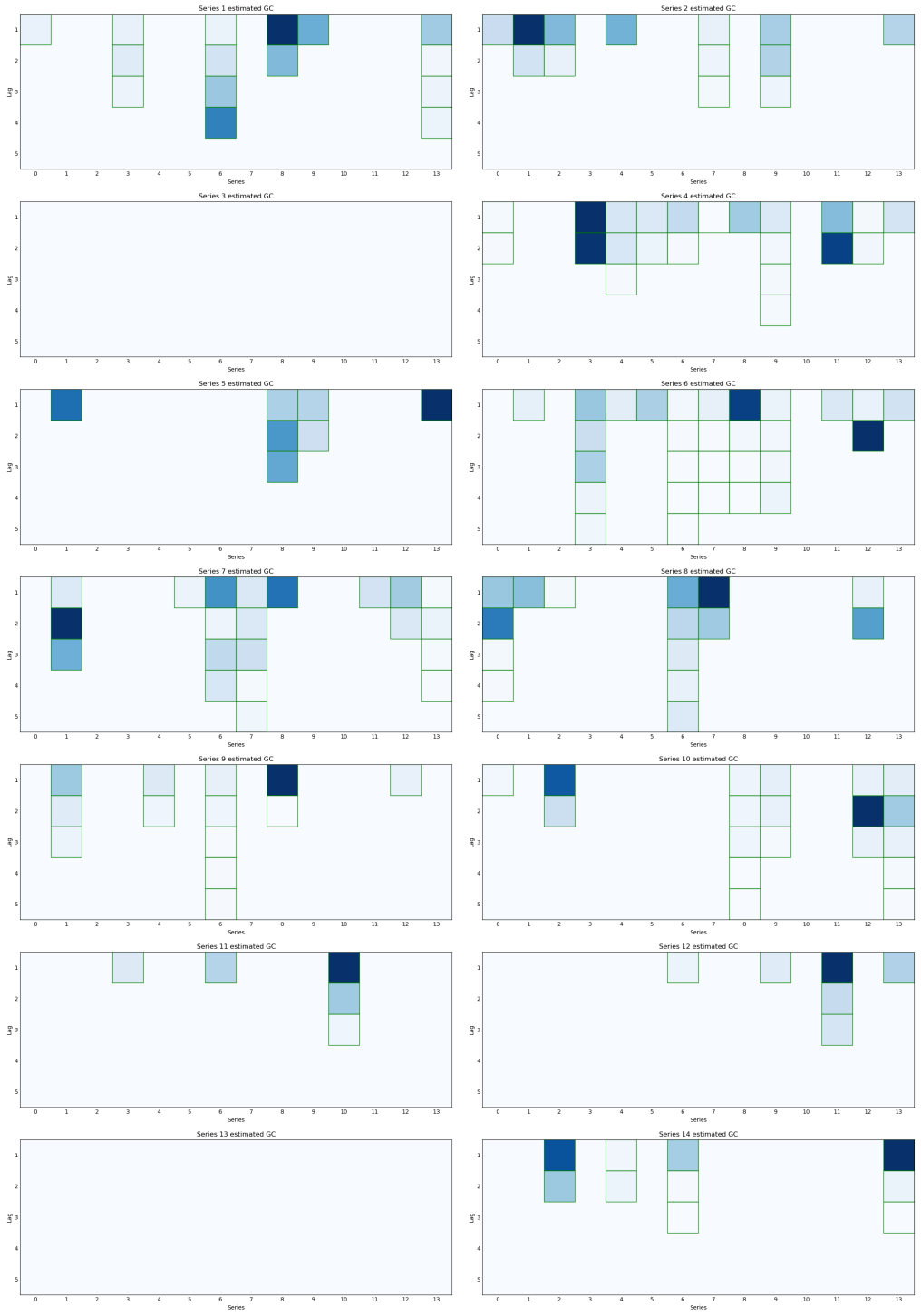


Figure 4.12 shows where hierarchical penalty truly shines. The lag order is clear now for interpretation. We will incorporate figure 4.12 with figure 4.21 of Hierarchical - GISTA for a more comprehensive interpretation in the Key Points section.

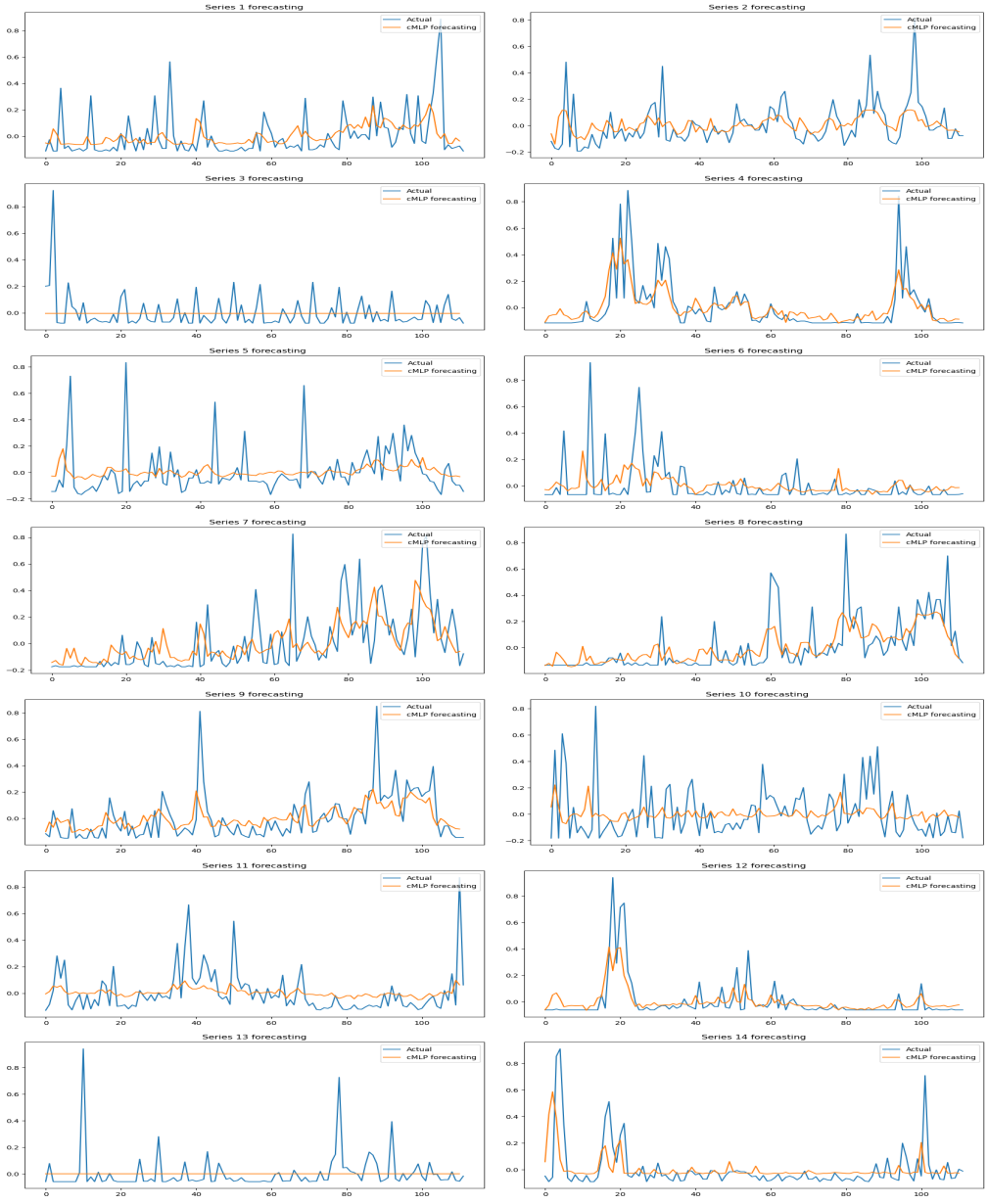
Figure 4.12: Granger Causality with Lag Order: Hierarchical - ISTA



Like GSGL, hierarchical penalty has decent forecast as shown in figure 4.13. Since hierarchical penalty is our choice, we want to make a comment. The model does not forecast very well when there are large peaks. This can be due to two reasons. First, we keep the parameters

the same as Dr. Covert's Neural-GC Git because forecasting is not our main goal. Second, mean normalization may not be enough or need further adjustments. These are the things we can do in the future.

Figure 4.13: Forecasting Time Series: Hierarchical - ISTA



4.2.3 GISTA

Elastic Net

Here, figure 4.14 shows some improvements. The influential time series are: 2 (Central America), 4 (Eastern Africa (P)), 6 (Mixed), 7 (Northern Africa), 8) (South America), 9 (Southern Asia), 12 (Western / Southern Asia (P)), 14 (Western Asia).

Figure 4.14: General Granger Causality: Elastic Net - GISTA

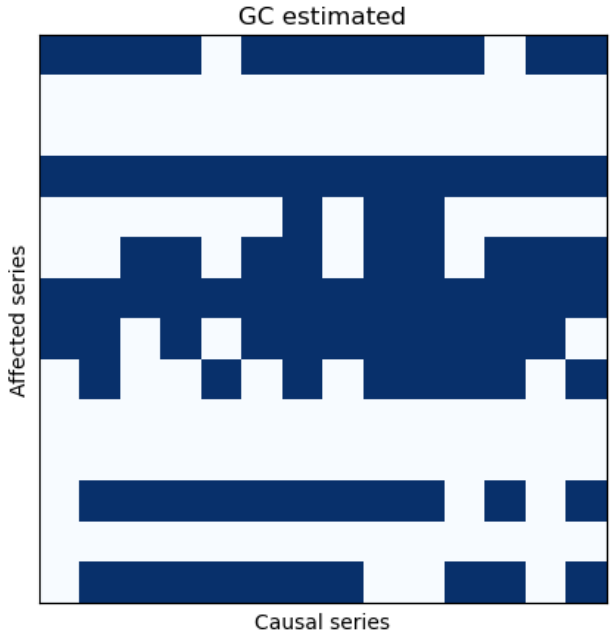


Figure 4.15 also shows substantial improvement compared to figure 4.4 (Elastic Net - ISTA). However, elastic net penalty is not our best option in general so we will not go further.

Figure 4.15: Granger Causality with Lag Order: Elastic Net - GISTA

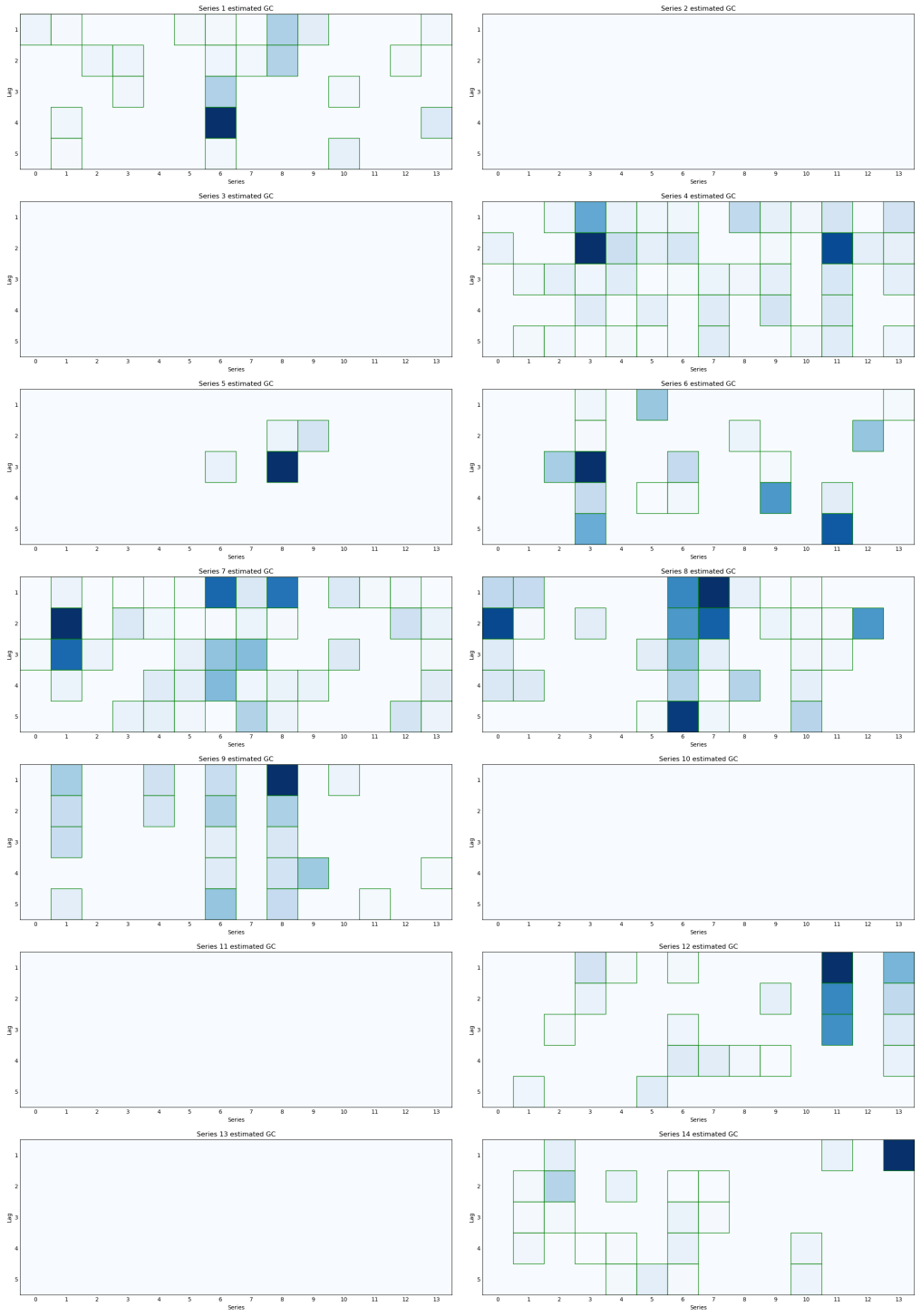
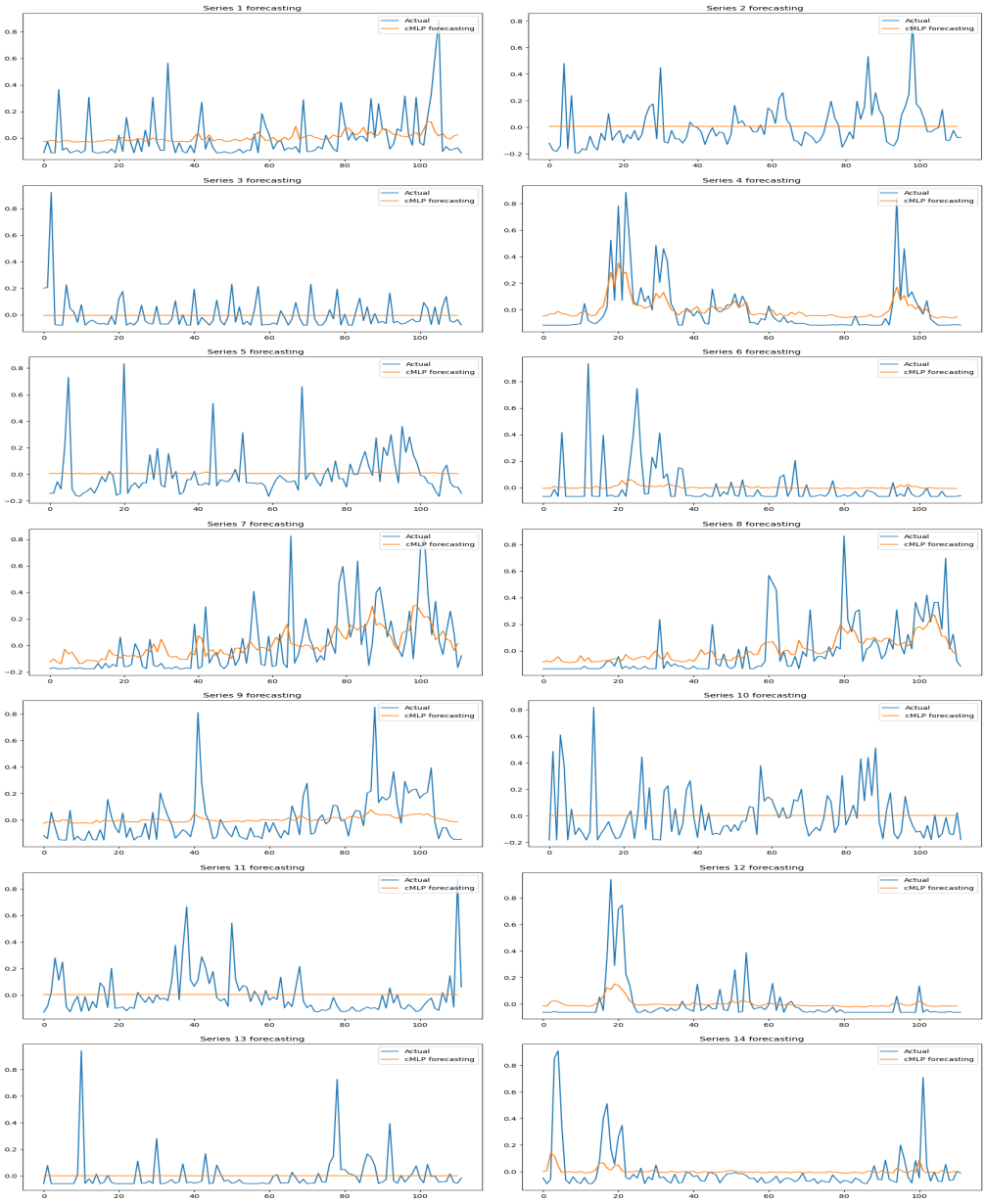


Figure 4.16 shows many horizontal forecast since many time series are not affected by others.

Figure 4.16: Forecasting Time Series: Elastic Net - GISTA



Group Sparse Group Lasso

Figure 4.17 indicates influential time series are: 7 (Mixed), 9 (Southern Asia), 10 (Sub-Saharan Africa (P)), 14 (Western Asia).

Figure 4.17: General Granger Causality: GSGL - GISTA

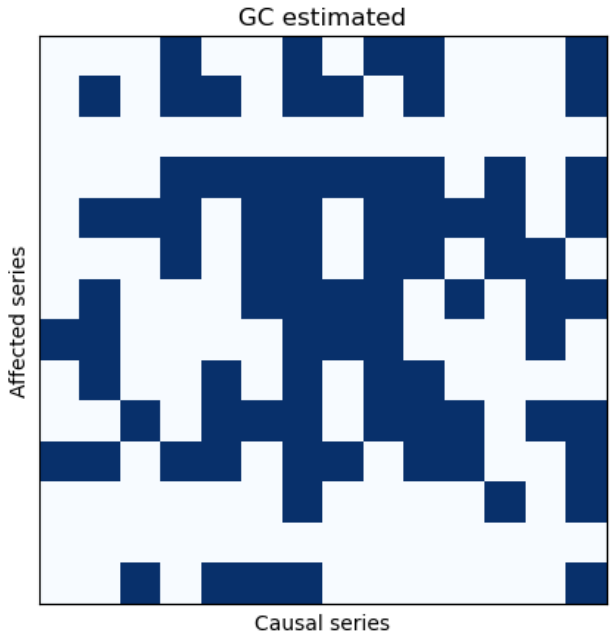


Figure 4.18 shows that for our study case, GISTA makes GSGL more interpretable. The lag order looks especially great for time series 4 (Eastern Africa (P)).

Figure 4.18: Granger Causality with Lag Order: GSGL - GISTA

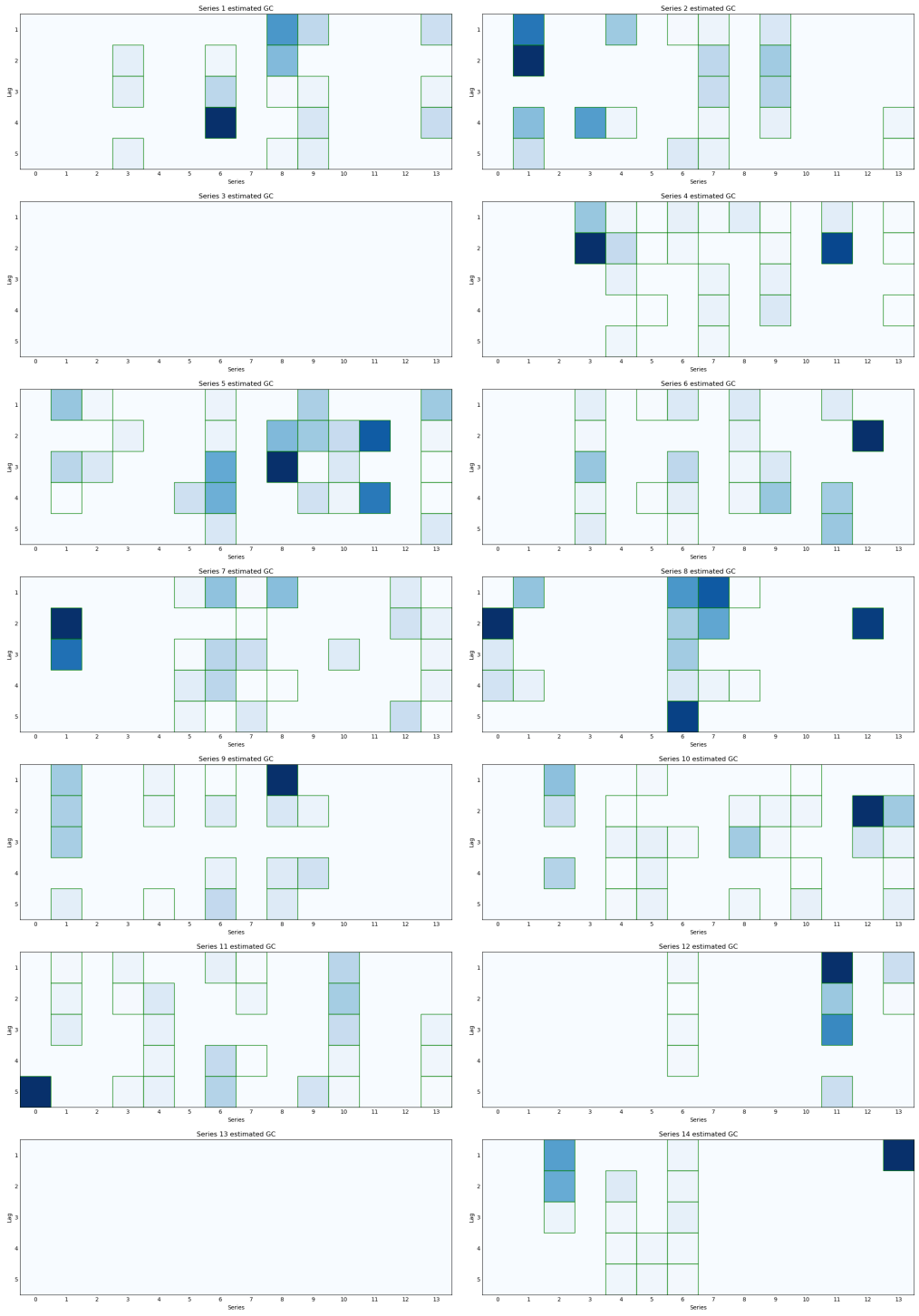
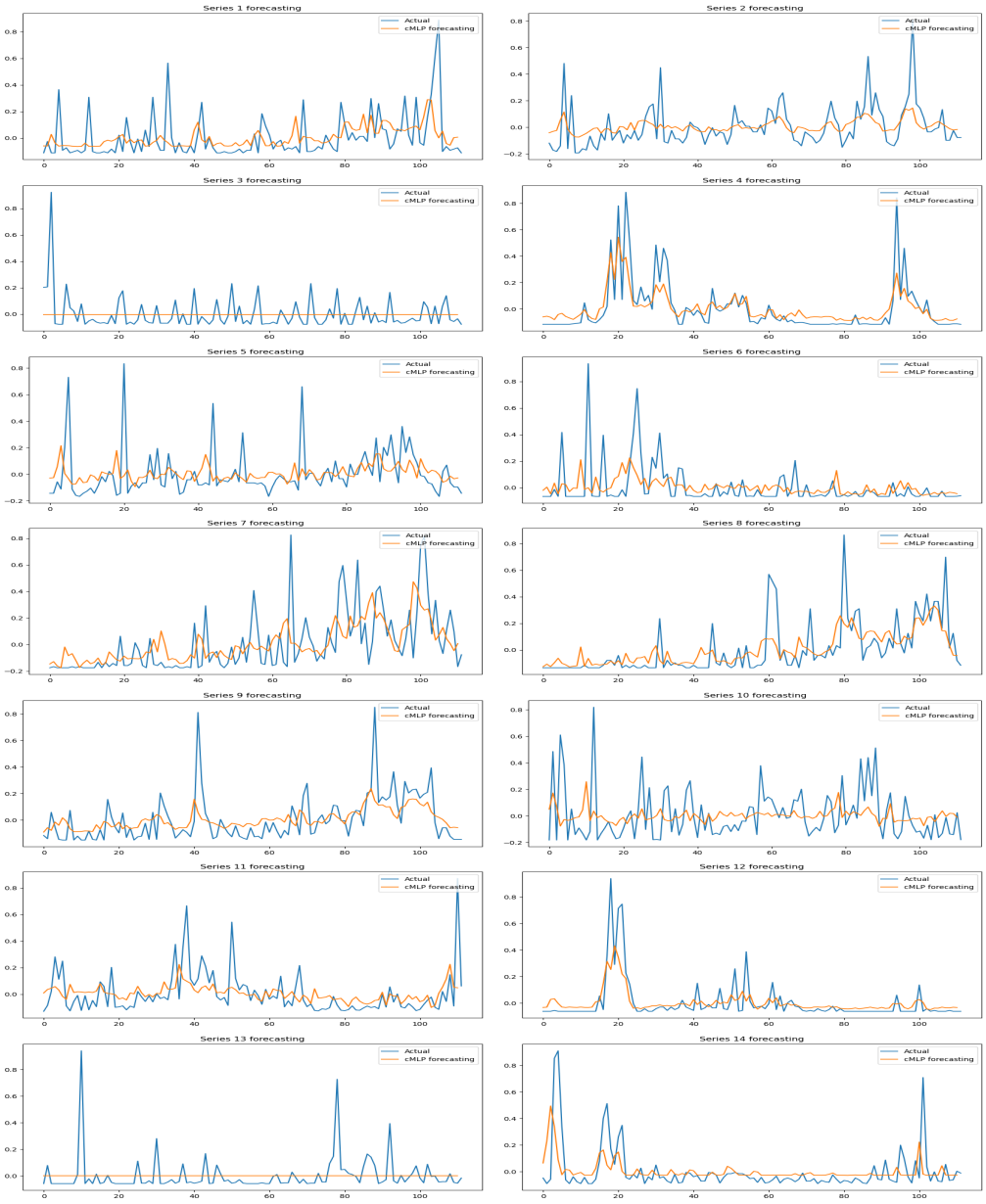


Figure 4.19 shows forecasting is slightly more complex since parameter tuning is more nuanced in GISTA.

Figure 4.19: Forecasting Time Series: GSGL - GISTA



Hierarchical Lasso

Figure 4.20 shows the influential time series are: 2 (Central America), 7 (Mixed), 9 (Southern Asia), 10 (Sub-Saharan Africa (P)), 14 (Western Asia). Now, after consulting with other results, we will examine all these 5 time series.

Figure 4.20: General Granger Causality: Hierarchical - GISTA

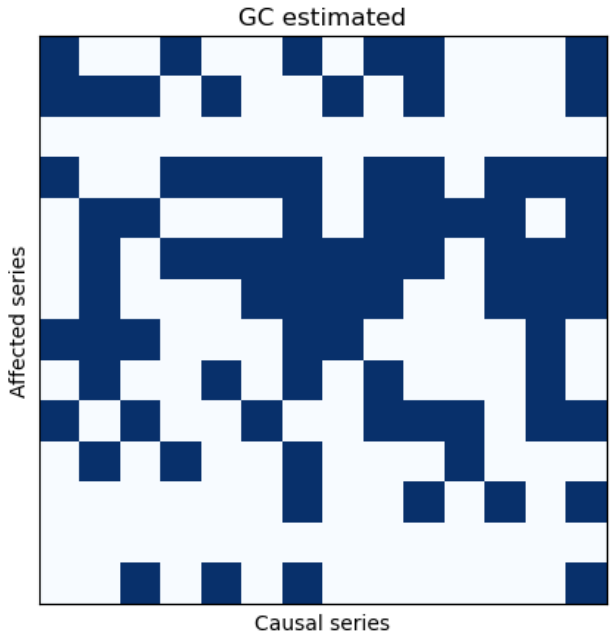
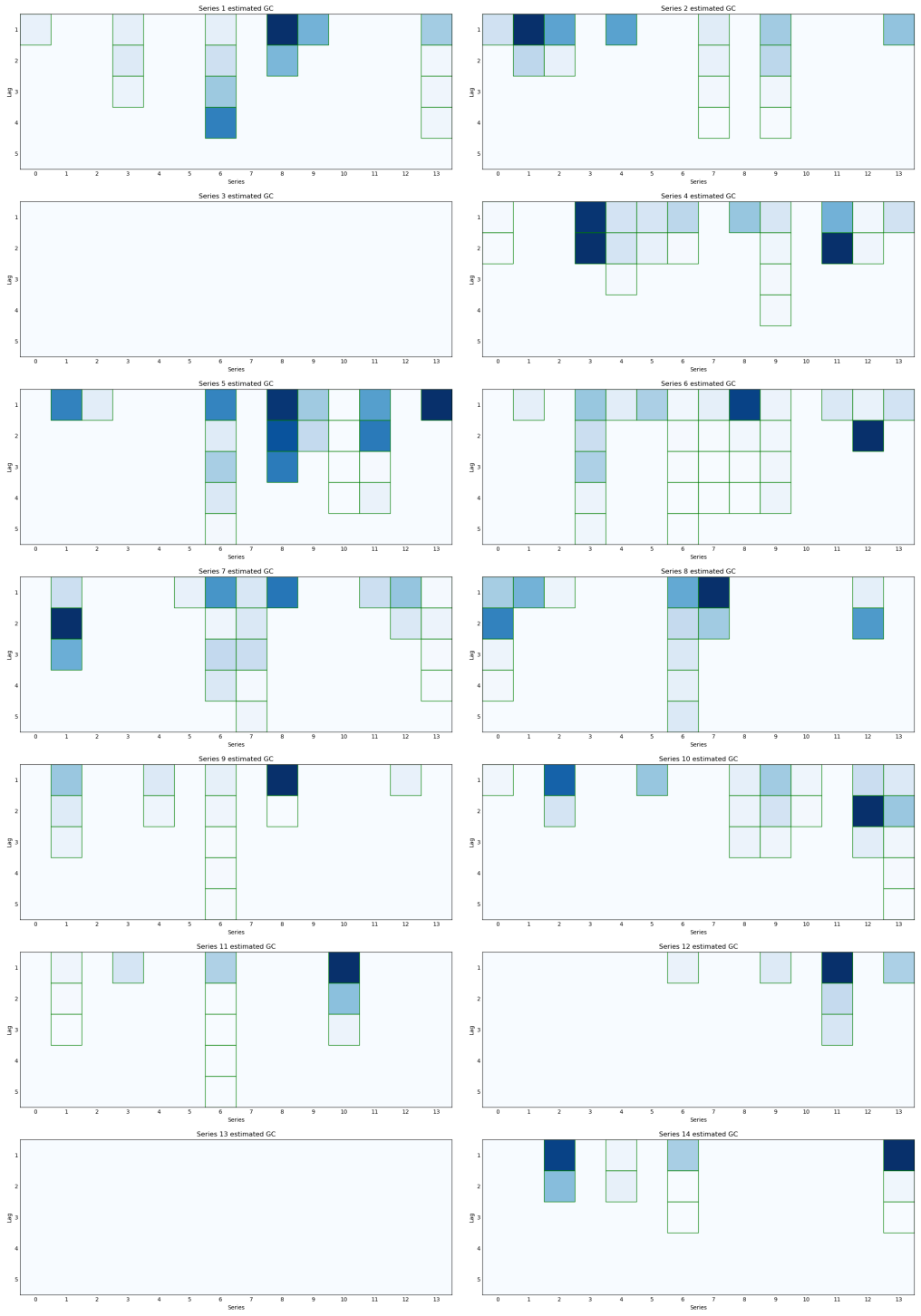


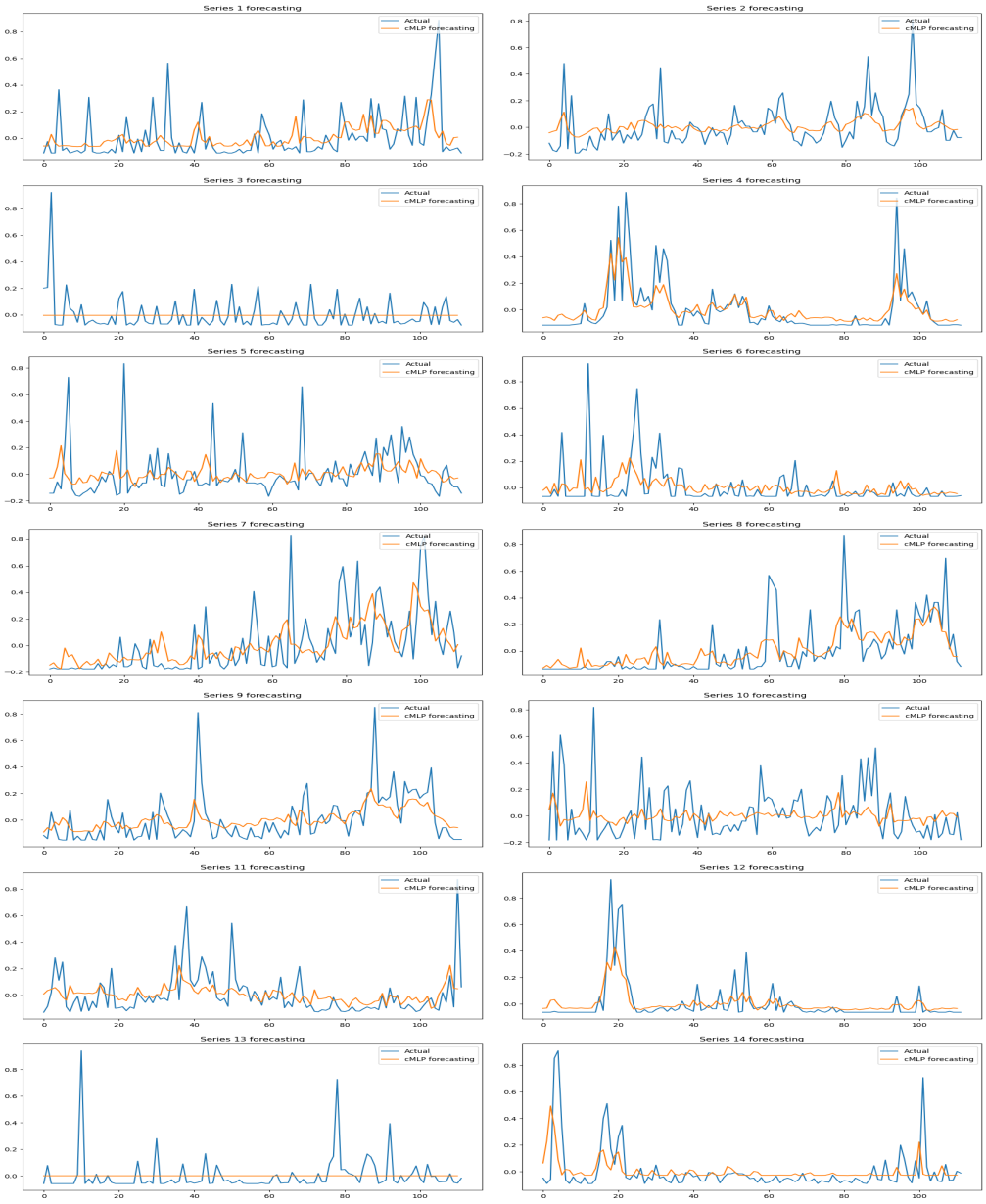
Figure 4.21 is very similar to figure 4.12 of Hierarchical - ISTA. For our 5 time series of interest, only 2 are slightly different: 2 (Central America) and 10 (Sub-Saharan Africa (P)). We will explore these results in greater details in the Key Points section.

Figure 4.21: Granger Causality with Lag Order: Hierarchical - GISTA



Finally, figure 4.22 shows the forecasting is as we expected. We already elaborated on the issue with time series peaks in Hierarchical - ISTA.

Figure 4.22: Forecasting Time Series: Hierarchical - GISTA



4.3 Key Points

In the end, our 5 influential time series are: **Central America (2), Northern Africa (7), Southern Asia (9), Sub-Saharan Africa (P) (10), Western Asia (14)**. We will explore each of them below. Note that we ignore the Granger causal effects of a time series on itself. Also, we will indicate the lag order and which lag has large Granger causal

estimate. One caveat is that our choice of blue which can be objective:

Central America (2) has predictive power over time series 1 (Caribbean) - order 1; 3 (Eastern Africa) - order 2, **lag 1**; 5 (Latin America / Caribbean (P)) - order 1, **lag 1**; 8 (South America) - **order 3 or 4**; 10 (Sub-Saharan Africa (P)) - **order 3 or 4**, **lags 1 and 2**; 14 (Western Asia) - order 1, **lag 1**.

Northern Africa (7) has predictive power over time series 2 (Central America) - order 3, **lag 2 and 3**; 6 (Mixed) - order 1; 8 (South America) - **order 5**; 9 (Southern Asia) - order 1, **lag 1**; 12 (Western / Southern Asia (P)) - order 1; 13 (Western Africa) - order 2, **lag 1 and 2**; 14 (Western Asia) - **order 4**.

Southern Asia (9) has predictive power over time series 2 (Central America) - order 3, **lag 1**; 5 (Latin America / Caribbean (P)) - order 2; 7 (Northern Africa) - **order 5**; 13 (Western Africa) - order 1.

Sub-Saharan Africa (P) (10) has predictive power over time series 1 (Caribbean) - order 1; 3 (Eastern Africa) - order 2, **lag 1**; 6 (Mixed) - order 1; 9 (Southern Asia) - **order 3 or 5**; 11 (Unknown) - order 2 or 3; 13 (Western Africa) - order 3, **lag 2**; 14 (Western Asia) - **order 5**.

Western Asia (14) has predictive power over time series 3 (Eastern Africa) - order 2, **lag 1 and 2**; 5 (Latin America / Caribbean (P)) - order 2; 7 (Northern Africa) - order 3.

Above, we highlighted affected time series with significant lags or having lag order of at least 4 since there are too many considerations. In addition, we use both Hierarchical - ISTA and Hierarchical - GISTA for lag order. With this, we will attempt to provide some possible reasons for these Granger causal links.

4.4 Possible Explanations

Now, we need to reiterate that Granger causality is not true causality. So, we can only say the time series of a region of origin may help predict that of another region. Additionally,

many of our explanations below are more hypotheses than concrete reasons. With that, we will divide our explanations into two categories: those that are intuitive and those that are otherwise.

Intuitive explanations: It makes sense to see the predictive power of Central America over South America, of Central America over Latin America / Caribbean (Presumed), of Northern Africa over Western Asia, of Northern Africa over Southern Asia, of Southern Asia over Northern Africa, of Northern Africa over Western Asia, of Sub-Saharan Africa (Presumed) over Eastern Africa, of Sub-Saharan Africa (Presumed) over Western Africa, of Sub-Saharan Africa (Presumed) over Western Asia, of Sub-Saharan Africa (Presumed) over Southern Asia, of Western Asia over Eastern Africa. These regions are either adjacent to one another or share the migration paths to North America and Europe.

Besides proximity, one likely reason is how these migration paths contain choke points. Here, many migrants have a higher chance of death and missing. For example, Schrank (2019) described the dangerous Mexican "La Bestia" train ride to reach the US border with cartels, extortionists, kidnappers, train accidents. With the vast desert of Mexico, they could die with no trace ever again. From what the lag orders showed, we believe South American migrants took 3 to 4 months to reach these points where tragedies struck. We assume the same for presumed Latin American / Caribbean victims who might sail to Mexico and traveled north.

Unprecedented crisis is another possibility. Africa and the Middle East currently face numerous civil wars, religious extremism, famine, and poverty. These hardships only exacerbated after the Arab Spring. Norman (2023) described the perils along migration journeys, from forced disappearance to sexual violence. Looking at time series Northern Africa and Southern Asia, we have another interesting finding. The lag order often depends on the distance between the two regions of origins. Like above, this may come from migrants reaching danger zones like the Sahara dessert or Mediterranean Sea. Finally, migrants faced no less danger even when they were not moving and staying in refugee camps. O'Callaghan (2021) high-

lighted Libya and Turkey were main jumping points to Europe. These countries, especially Libya after its civil war, could not cope with so many refugees. Hygiene, violence, disease, malnutrition,... all took a toll on migrants. Numerous significant lags of North Africa and Sub-Saharan Africa (P) for predictions seem to agree with this hypothesis: the migrants likely tried to reach the North African coast.

Non-intuitive explanations: Here, we are discussing the predictive power of Central America over Eastern Africa, of Central America over Sub-Saharan Africa (Presumed), of Central America over Western Asia, of Northern Africa over Central America, of Northern Africa over South America, of Southern Asia over Central America. These regions are oceans apart, yet the lag order and significant lags are about 2-3 months. Indeed, this is a new phenomenon where migrants outside of the Americas tried to reach the US through South America. Yates (2019) detailed an increasing volume of Asians and Africans through the Darien Gap in Panama. Paradoxically, the author emphasizes these long journeys are actually cheaper and less risky because of lax enforcement, easy plane travel with lenient visa policies, and fewer conflict zones.

Chapter 5

Conclusion

5.1 Summary

The interpretability of Neural Granger Causality by Tank et al (2021) is a major breakthrough for observational studies. We believe our migrant analysis effectively demonstrate that. The visualizations are very simple to make Granger causal connections. Also, it is straightforward to understand time series dynamics when using hierarchical penalty. In addition, neural Granger causality is flexible with so many options for parameters, training models, and penalties.

Thanks to our study, we learn a lot about migration connections and try to explain despite the caveats of Granger causality. We observe that many migrants had close regions of origins. This is unsurprising given that these regions face very similar and intertwined circumstances. Moreover, we discovered a recent issue which is cross-Atlantic migration.

From all of the above, we also want to provide some proposals. First, we have to ensure the stability of each and every country. Many of us agree that the migrant crisis is a worldwide issue, yet we often blame solely on specific governments for the instability of their nation. Whether such criticism is justified or not, we have to realize chaos from one place can spread widely. Our results indeed implied that. We believe only a serious international effort can bring peace and prosperity to each nation. For this to happen, dialogue and political unity is key. Next, we need our resources to concentrate on dangerous choke points such as the Darien Gap or the Mediterranean sea. Instead of providing resources to migrants only when they reach the borders, we should save lives in these critical junctions. Preventing the tragedies

of these migrants, even in these seemingly distant and remote areas, should be part of our strategy.

5.2 Discussion

Currently, our biggest shortcoming is the input parameters. The forecasting visualizations confirmed this. With time, we hope to improve these parameters likely through cross validation.

With more time, we may also want to analyse the Granger causal influence of more regions. Our choice of significant lags and lag orders seem arbitrary, but our paper is substantial already. Another interest we have is to use another attribute such as the region of incident or migration route. These attributes will be more applicable for our study. For now, we disregarded them since they did not have enough amount of time series. This would not have demonstrated the power of neural Granger causality.

If you want to see my new codes besides the main source (Dr. Ian Covert's Neural-GC Git), please contact me at huynguyen012016[at]gmail.com

References

Chu, T., & Glymour, C. (2008). Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, *9*(32), 967–991. Retrieved from http://jmlr.org/papers/v9/chu08a.html

Elgiriyewithana, N. (2023). *Global missing migrants dataset.* Kaggle. Retrieved from https://www.kaggle.com/datasets/ nelgiriyewithana/global- missing-migrants-dataset

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. Retrieved from https://www.sciencedirect.com/science/article/pii/0893608089900208 doi: https://doi.org/10.1016/0893-6080(89)90020-8

Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel method for nonlinear granger causality. *Physical review letters*, *100*(14), 144103.

Morioka, H., Hälvä, H., & Hyvarinen, A. (2021). Independent innovation analysis for nonlinear vector autoregressive process. In *International conference on artificial intelligence and statistics* (pp. 1549–1557).

Norman, K. (2023). *Migration and displacement in the arab world demands more equitable response.* Carnegie Endowment for International Peace. Retrieved from https://carnegieendowment.org/2023/05/03/ migration -and-displacement-in-arab-world- demands-more-equitable-response -pub-89520

O'Callaghan, L. (2021). *Migration in europe: How the crisis shifted shape in 2021.* The National News. Retrieved from https://www.thenationalnews.com /world/europe /2021 /12/29/migration-in-europe -how-the-crisis- shifted-shape-in-2021/

Parikh, N., & Boyd, S. (2014, jan). Proximal algorithms. , *1*(3), 127–239. Retrieved from

https://doi.org/10.1561/2400000003 doi: 10.1561/2400000003

Schrank, D. (2019). *Victims of "la bestia," mexico's notorious migrant train, learn to walk again.* Reuters. Retrieved from https://www.reuters.com/ article/ idUSKCN1VC1XL/

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, *85 2*, 461-4. Retrieved from https://api.semanticscholar.org/CorpusID:7411376

Shojaie, A., & Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, *9*, 289–319.

Shojaie, A., & Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, *97*(3), 519-538. Retrieved from https://doi.org/10.1093/biomet/asq038 doi: 10.1093/biomet/asq038

Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. B. (2021). Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4267–4279.

Yates, C. (2019). *Extracontinental migrants in latin america.* Migration Policy Institute. Retrieved from https://www.migrationpolicy.org /article/extracontinental -migrants-latin -america