MODEL AGNOSTIC METHODS FOR MULTIVARIATE TIME SERIES WITH ARBITRARY DEPENDENCE

Noah David Gade

Stillwater, Oklahoma

Bachelor of Science, Oklahoma State University, 2017

Master of Science, Oklahoma State University, 2019

A Dissertation submitted to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Doctor of Philosophy

Department of Statistics

University of Virginia

May 2024

Dr. Jordan Rodu, Chair

Dr. Leland Farmer

Dr. Xiwei Tang

Dr. Shan Yu

Model Agnostic Methods for Multivariate Time Series with Arbitrary Dependence

Noah David Gade

ABSTRACT

Nonlinear functional dependence in temporal data can be challenging to characterize. Linear methods often oversimplify the structure of data, and imposition of a rigid framework can be be limiting. A model-based method that is "close enough" usually works well, but specification of this functional form can be difficult, especially in dependent, multivariate data and when lacking a deep scientific understanding of the process. Representation learning methods, employing artificial neural networks, provide a model agnostic approach to capture even the most complicated and intricate interactions between covariates. These methods, if applied cautiously and when model recovery is not the main goal, can alleviate the burden of model specification and the difficulties that arise from model misspecification. They are applied to single and multiple change point detection, the adaptation of Granger causality to nonlinear functional dependence, and discussions of future research include strategic workarounds for the loss of covariate-specific and relational information.

Acknowledgments

This endeavor would not have been possible without the guidance of my advisor, Dr. Jordan Rodu. I am very thankful for his exceptional mentorship throughout my PhD program. Thanks should also go to committee members Dr. Leland Farmer, Dr. Xiwei Tang, and Dr. Shan Yu for their tremendous support and constructive comments in the dissertation process.

I am extremely grateful for my friends and colleagues Dr. Jesse E. Helman and Dr. Evan M. Bagley, and for their advice in navigating the challenges and obstacles of graduate work; they were immensely helpful in the day to day process. I'd like to extend sincere thanks to my office mates and friends Sydney Campbell and James Lee for their impromptu discussions and moral support, and acknowledge the encouragement of department faculty members Dr. Taylor Brown, Dr. Justin Weinstock, and Krista Varanyak. I'd also like to recognize the vision of my first academic boss, Dr. K. Darrell Berlin, and my first research supervisor, Dr. Ashlee N. Ford Versypt, who both knew my career aspirations before I did.

Lastly, words cannot express my gratitude to my family Dr. Mary N. Gade, Dr. David Gade, and Emma Gade for their constant belief and inspiration.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Text Abbreviations

-GL Group lasso penalty

-GSGL Group sparse group lasso penalty

-H Hierarchical lasso penalty

AMOC At most one change point

ANN Artificial neural network

AR Autoregressive model

ARCH Autoregressive conditional heteroskedasticity model

ARI Adjusted Rand index

ARIMA Autoregressive integrated moving average model

ARL Average run length

ARMA Autoregressive moving average model

CA1 Hippocampal cornu ammonis neurons

CCP Conceptor change point method

cLSTM Component-wise long short-term memory network

cMLP Component-wise multilayer perceptron

CNN Convolutional neural network

CUSUM Cumulative sum statistic / control chart

EDiv E-Divisive change point method

ESN Echo state network

FDR False discovery rate

FNN Feedforward neural network

FOR False omission rate

FPR False positive rate

FWER Family-wise error rate

GARCH Generalized autoregressive conditional heteroskedasticity model

GNS Gaussian white noise substitution in-sample model comparison

GVAR Generalized vector autoregressive model

HC Hippocampus

HMM Hidden Markov model

i.i.d. Independent and identically distributed

KCP Kernel change point method

kCUSUM Online kernel CUSUM method for change point detection

LeKVAR Learned kernel VAR model

LFP Local field potential

LRSM likelihood ratio scan method

LSM Liquid state machine

LSTM Long short-term memory network

MA Moving average model

MBB Moving block bootstrap

MCC Matthews correlation coefficient

MCCP Multiple conceptor change point method

MLP Multilayer perceptron

NP-MOJO Nonparametric moving sum procedure for detecting changes in the joint characteristic function

NPGC Nonlinear permuted Granger causality

NRMSE Normalized root mean squared error

PFC Prefrontal cortex

PPV Positive predictive value

R/U Restricted and unrestricted in-sample model comparison

REM Rapid eye movement sleep

RNN Recurrent neural network

SARIMA Seasonal autoregressive integrated moving average model

SBS1(2) Sparsified binary segmentation Type 1(2) method

ScanB Scan-B method for kernel change point detection

SCCP Sequential conceptor change point method

SMUCE Simultaneous multiscale change-point estimator

STAR Smooth transition autoregressive model

TAR Threshold autoregressive model

TCDF Temporal causal discovery framework

THAL Thalamus

TPR True positive rate

VAR Vector autoregressive model

Chapter 1

Introduction

Characterizing temporal dependence in data has long been sought from structural models, like the autoregressive (AR) and moving average (MA) models of the early 20th century (Hooker, 1901; Yule, 1909; Yule, 1921; Yule, 1927; Wold, 1938). The AR(γ) model examines a stationary time series y_t, $t = 1, \ldots, T$, as a function of past observations, as in Equation 1.1, where γ is the order of the model (the included number of lagged values in the functional form). The MA(γ) model writes a stationary time series as a function of past errors, $f(\varepsilon_{t-1}, \ldots, \varepsilon_{t-\gamma})$.

$$y_t = \beta_0 + \sum_{i=1}^{\gamma} \beta_i y_{t-i} + \varepsilon_t \tag{1.1}$$

These basic building blocks are combined for greater functionality in the ARMA, autoregressive integrated moving average (ARIMA), and seasonal ARIMA (SARIMA) models that allow for combined mechanisms of dependence, trends, and seasonal behavior in time series data (Box and Jenkins, 1970). Further iterations allow for inclusion of exogenous variables. Adaptation to multivariate time series data $\mathbf{y}_t \in \mathbb{R}^d$ generalizes the AR and ARMA forms to vector autoregressive (VAR) models, with form similar to Equation 1.1 where scalar coefficients are replaced by vector coefficients, that scale like the above methods to include compound mechanisms of dependence, trends, and exogenous variables (Sims, 1980).

An alternative approach examines temporal data in the frequency domain, where the variation is expressed in terms of regular, periodic components (Shumway and Stoffer, 2017). Like harmonic analysis, the data can be thought of as a composition of several sinusoids that constitute the overarching structure. Periodic variations of the sample data are characterized in a periodogram as an estimate of the spectral density (Schuster, 1898; Schuster, 1906a; Schuster, 1906b). Spectral analysis can relate to principal component analysis; for a stationary time series, the density can be interpreted as an approximation of the eigenvalues of the covariance matrix (Shumway and Stoffer, 2017).

Nonlinear modifications for serial correlation in variance led to the autoregressive conditional heteroskedasticity (ARCH) and the generalized ARCH (GARCH) models (Engle, 1982; Bollerslev, 1986). To account for regime switching behavior, Tong and Lim (1980) developed the threshold autoregressive (TAR) model, and this was extended by Chan and Tong (1986) with smooth transition autoregressive (STAR) models to allow for gradual movement between regimes. In the same vein, state space representations like Markov switching and hidden Markov (HMM) models look to capture regime switching behavior, or other behavior when a response is influenced by a latent state. These models apply when the Markov property is satisfied, where a state is only influenced by the previous state and independent of any long-term history, $\Pr\left(Y_t|Y_{<t} = \{y_{t-1}, y_{t-2}, \ldots\}\right) = \Pr\left(Y_t|Y_{t-1} = y_{t-1}\right)$ (Markov, 1906; Baum and Petrie, 1966). Higher-order Markov chains allow for extension of this dependence structure to states further in the past, and in a HMM, the process is governed by a discrete latent quantity, known as the "hidden" state (Baum and Petrie, 1966; Bishop, 2006). Kalman filtering and particle filtering allow for smoothing of the stochastic process, somewhat analogous to methods for the TAR above (Kalman, 1960; Del Moral, 1997;

J. S. Liu and R. Chen, 1998). Differential equation models prove useful tools to represent the dynamics of a time series, beginning with processes such as Brownian motion and increasing in complexity depending on the application; however, these require a comprehensive knowledge of the underlying scientific processes driving the data (Einstein et al., 1905; Langevin, 1908; Fokker, 1914; Planck, 1917; Schrödinger, 1926; Kolmogorov, 1931; Black and Scholes, 1973).

Despite the breadth of tools available, characterizing nonlinear functional dependence in temporal data can still be challenging. Linear methods, like many listed above, often oversimplify the structure of data. Imposition of a structure that is "close enough" usually works well, but specification of this functional form can be difficult, especially in multivariate data with elaborate dependence. Artificial neural networks provide model agnostic approaches to represent data; they can capture even the most complicated intricacies of interactions between covariates. These methods, if applied cautiously when model recovery is not the main goal, can alleviate the burden of model specification and the difficulties that arise from model misspecification.

1.1 Artificial Neural Networks

Artificial neural networks (ANN) consist of interconnected nodes, akin to neurons in the brain, that modify and relay information received from other nodes. These models have been adapted for use in a broad range of application areas including function approximation and prediction, classification, pattern recognition and regeneration, and data processing (Goodfellow, Bengio, and Courville, 2016). Despite this wide scope, the primary utility of ANNs remains focused on extracting relational information between covariates of interest, like some statement about the conditional

behavior of one variable given another (e.g., $\mathbb{E}[Y|X = x]$).

The emergence of ANNs began with efforts to mathematically model biological mechanisms, as in some early works by McCulloch and Pitts (1943), Widrow, Hoff, et al. (1960), Rosenblatt (1961), and Rumelhart, Hinton, and Williams (1986) (Bishop, 2006). Methods of computation and application have rapidly improved in the advent of the technological age, and network architectures have become more complex to prioritize certain aspects of behavior. Basic architectures in Section 1.1 are presented in the context of temporal data; several other frameworks exist for ANNs that incorporate unstructured data, or data with spatial components including geographic applications and image processing (Bishop, 2006; Goodfellow, Bengio, and Courville, 2016).

1.1.1 Artificial Network Architecture

Artificial neural networks can take the basic form in Equations 1.2 and 1.3, where $\mathbf{x}_t \in \mathbb{R}^p$ is a vector of inputs at time t, and $\mathbf{y}_t \in \mathbb{R}^d$ a vector of outputs.

$$\mathbf{h}_t = g\left(\mathbf{W^i x}_t + \mathbf{b}\right) \tag{1.2}$$

$$\mathbf{y}_t = \mathbf{W^o h}_t \tag{1.3}$$

The inputs to the ANN are multiplied by a matrix $\mathbf{W^i} \in \mathbb{R}^{N\times p}$, modified by a bias vector $\mathbf{b} \in \mathbb{R}^N$, and fed to an activation function g to form the network states $\mathbf{h}_t \in \mathbb{R}^N$. The network states (also called hidden states, hidden units, or reservoir states) are the individual neurons in the artificial network. The activation function g introduces nonlinear behavior; common choices are the rectified linear unit ($\text{ReLU}(x) = \max\{0, x\}$), the logistic sigmoid function ($\sigma(x) = [1 + e^{-x}]^{-1}$), and the hyperbolic tangent func-

tion ($\tanh(x)$). The network states are linearly mapped to the output with matrix $\mathbf{W^o} \in \mathbb{R}^{d \times N}$.

Constructed ANNs are characterized by their width and depth. Figure 1.1 displays a basic illustration of these properties. The width of a network N is the dimension

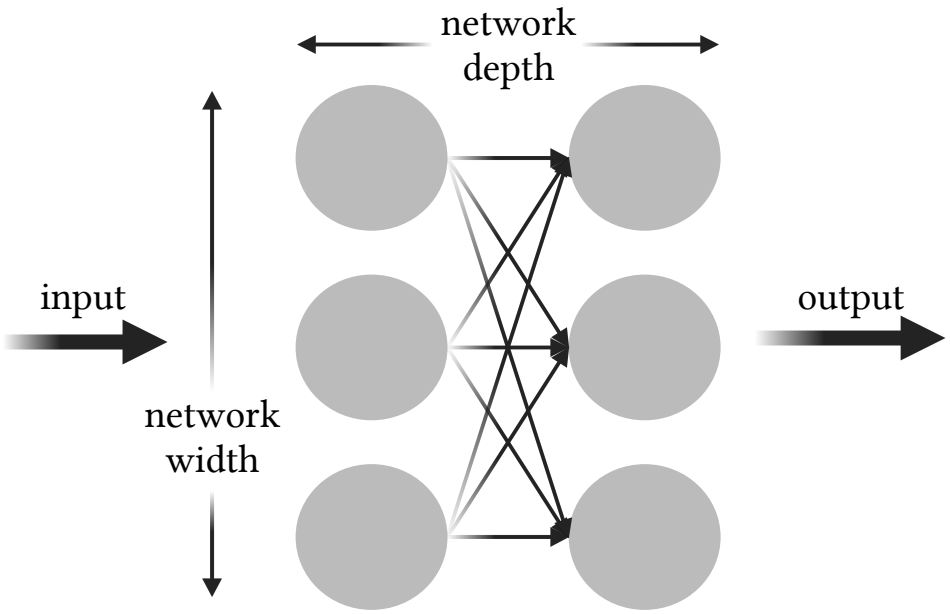


Figure 1.1: Simplified architecture of a basic artificial neural network.

of the network state vector $\mathbf{h}_t$, with wider networks containing a larger number of states. Network depth is fashioned from a repeated function composition of the form in Equation 1.2. In a deep network, each layer consists of one network state vector $\mathbf{h}_t^\ell \in \mathbb{R}^N$ for $\ell = 1, \ldots, L$, and the number of layers L constitutes the depth. Each subsequent layer receives the previous as input, like shown in Equation 1.5, and the final hidden layer in Equation 1.6 maps to the output in a linear fashion, identical to

Equation 1.3.

$$\mathbf{h}_t^1 = g_1\left(\mathbf{W^{1i}}\mathbf{x}_t + \mathbf{b^1}\right) \tag{1.4}$$

$$\mathbf{h}_t^\ell = g_\ell\left(\mathbf{W}^{\ell\mathbf{i}}\mathbf{h}_t^{\ell-1} + \mathbf{b}^\ell\right) \quad \text{for } \ell = 2, \ldots, L \tag{1.5}$$

$$\mathbf{y}_t = \mathbf{W^o}\mathbf{h}_t^L \tag{1.6}$$

Equations 1.2 and 1.3 represent the simplest ANN architecture, known as a feed-forward neural network (FNN). FNNs consist of a single layer of depth, and arbitrary width N, like shown in Figure 1.2. All information is "fed forward" from a specific input $\mathbf{x}_t$ and there is not any complicating structure present in the architecture.

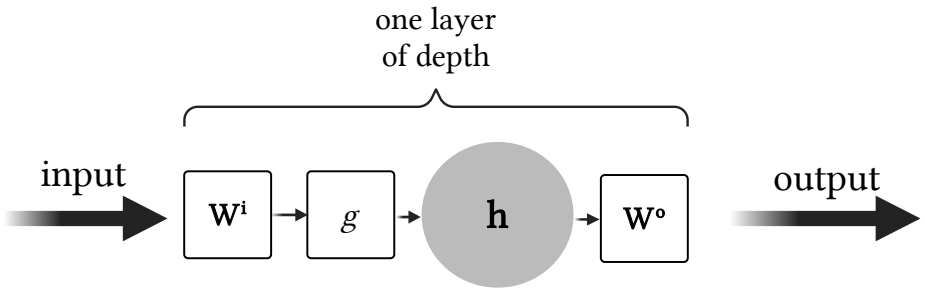


Figure 1.2: Simplified architecture of a feedforward neural network.

Multilayer perceptrons (MLP), extend the simple FNN to some level of network

depth. MLPs are constructed like shown in Equations 1.4 to 1.6. All information from a specific input $\mathbf{x}_t$ is still "fed forward" as in the FNN, but there are a greater number of intermediate steps (layers) between the input and the output $\mathbf{y}_t$.

Recurrent Neural Networks

Recurrent neural networks (RNN) are a type of ANN developed specifically for processing sequential data (Rumelhart, Hinton, and Williams, 1986; Goodfellow, Bengio, and Courville, 2016). Computation of the network state is modified to include dependence on the previous time point, usually through the last hidden state as in Figure 1.3. A generic, single layer RNN system can be described with Equations 1.7 and 1.8,

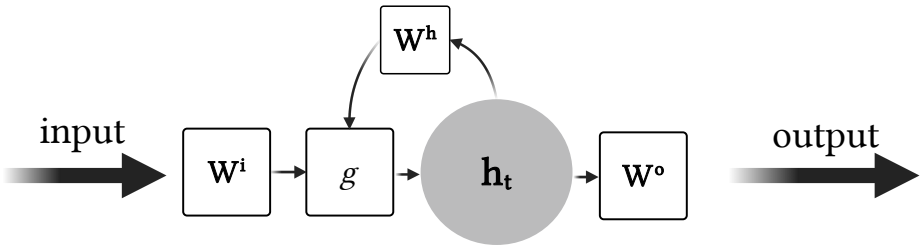


Figure 1.3: Simplified architecture of a recurrent neural network.

with straightforward extension to added depth similar to Equations 1.4 through 1.6.

$$\mathbf{h}_t = g\left(\mathbf{W}^{\mathbf{h}}\mathbf{h}_{t-1} + \mathbf{W}^{\mathbf{i}}\mathbf{x}_t + \mathbf{b}\right) \tag{1.7}$$

$$\mathbf{y}_t = \mathbf{W}^{\mathbf{o}}\mathbf{h}_t \tag{1.8}$$

In the architecture of RNNs, matching the other structures above, the parameter matrices $\mathbf{W}^{\mathbf{h}} \in \mathbb{R}^{N\times N}$, $\mathbf{W}^{\mathbf{i}}$, $\mathbf{W}^{\mathbf{o}}$, and vector $\mathbf{b}$ do not change depending on the time t in the sequence. This can be interpreted as a requirement for a stationary conditional distribution given the history of the sequence (Goodfellow, Bengio, and Courville, 2016).

Provided the parameters are set such that the RNN is stable, a recurrent network is able to capture temporal dependence that diminishes as time points become further separated. Stability is determined through the propagating gradient over several time points. A simple example follows that from Goodfellow, Bengio, and Courville (2016), where the function composition is assumed to lack the activation function g and resemble a linear transformation $\mathbf{h}_t = \mathbf{W}^{\mathbf{h}}\mathbf{h}_{t-1}$. This form has an alternative representation based on the values of the initial network state, $\mathbf{h}_t = \left(\mathbf{W}^{\mathbf{h}}\right)^t \mathbf{h}_0$. The eigenvalues of the matrix $\mathbf{W}^{\mathbf{h}}$ that are greater than one will produce an exploding product, while those that are less than one will vanish (Hochreiter, 1991; Bengio, Simard, and Frasconi, 1994). In the presence of a contractive activation function g, or a function that shrinks the $\mathcal{L}^2$-norm of the network state vector, the tipping point for stability grows larger than one (Goodfellow, Bengio, and Courville, 2016). The "vanishing gradient" problem of stable RNNs makes long-term dependence, as opposed to a recent memory of previous time steps, difficult to capture (Bengio, Simard, and Frasconi, 1994).

Methods to overcome the issue of capturing long-term dependence include adding skip dependence or incorporating longer delays in networks (T. Lin et al., 1996) and inserting "leaky" or "gated" units that accumulate information over longer durations (Mozer, 1991; El Hihi and Bengio, 1995). One such network that utilizes this architecture is the long short-term memory (LSTM) network of Hochreiter and Schmidhuber (1997), which contains dynamic self-loops that produce long running paths of dependence that change depending on the input sequence (Gers, Schmidhuber, and Cummins, 2000; Goodfellow, Bengio, and Courville, 2016).

Reservoir Computing

Parameters in RNNs, particularly as architectures increase in complexity to accommodate specific features of the input data, can be challenging and computationally expensive to learn. Reservoir computing circumvents this problem by defining the recurrent and input parameter matrices $\mathbf{W^h}$, $\mathbf{W^i}$, and $\mathbf{b}$ as fixed values, and only learning the output weights $\mathbf{W^o}$ (Lukoševičius and Jaeger, 2009). Reservoir computing networks transform the input sequence to a "reservoir" of features that capture distinct characteristics of the data (Goodfellow, Bengio, and Courville, 2016). The reservoir is a nonlinear, high-dimensional expansion of the original data, and features that are not linearly separable in the original input space can become separable in the reservoir (Lukoševičius, 2012). Echo state networks (ESN) of Jaeger (2001) and liquid state machines (LSM) of Maass, Natschläger, and Markram (2002) are two examples of this strategy with continuous and binary network states, respectively, where the output can be obtained from a simple linear mapping of this advantageous transformation of the data (Jaeger, 2002; Jaeger, 2007). The challenge is to obtain a rich enough representation of the data from the generated reservoir; Lukoševičius

(2012) and Yildiz, Jaeger, and Kiebel (2012) discuss standard practices and tricks to increase the likelihood of a rich representation.

In a recurrent network of the form shown in Equations 1.7 and 1.8, the input data $\mathbf{x}_t$ is commonly scaled such that most values are contained in a relevant domain of the activation function g, for example $[-1, 1]$ for the hyperbolic tangent. The size of the reservoir N (the dimension of the network state vector $\mathbf{h}_t$) should be as large as computationally affordable (Lukoševičius, 2012). Reservoir matrices $\mathbf{W^h}$, $\mathbf{W^i}$, and $\mathbf{b}$ are usually randomly generated, with individual matrix elements independent random Gaussian or uniform observations. The hidden state reservoir matrix $\mathbf{W^h}$ should be sparse to expedite computation, with an approximate fixed number of nonzero values in each row (Lukoševičius, 2012). This does not limit the capacity of the network, but reduces computational cost as the reservoir size increases. The input and bias matrices are often dense. The reservoir matrices are scaled to seek a suitable level of nonlinear behavior and control relative influence of the input and the previous network state. Scaling for $\mathbf{W^i}$ and $\mathbf{b}$ is performed on the standard deviation (limits) of the random Gaussian (uniform) observations. Higher scaling values will introduce a larger degree of nonlinearity in the network states as the activation function is pushed away from the zero neighborhood and toward the limiting values. Further fine-tuning of this process can allow for scaling of individual columns in the input $\mathbf{x}_t$ (Lukoševičius, 2012). Scaling for $\mathbf{W^h}$ is done through the spectral radius of the generated matrix. As in the example from Goodfellow, Bengio, and Courville (2016), if the largest eigenvalue is too large, the network states will become unstable. A stable network will eventually wash away ("forget") any previous states or initial conditions after a sufficient time length. In reservoir computing, this is also called the echo state property (Jaeger, 2007). A spectral radius less than one ensures the

echo state property in most cases, but often the tipping point for stability can be much larger than one (Lukoševičius, 2012; Yildiz, Jaeger, and Kiebel, 2012). The spectral radius directly influences how long the network retains information before it vanishes, and to maximize the quanitity of long-term information preserved, the spectral radius should be up against this threshold (Yildiz, Jaeger, and Kiebel, 2012). These network hyperparameters are evaluated based on the ability of the reservoir to capture the relationship between the input and output (evaluated with mean squared error) on training and validation sets. The hyperparameters can typically be chosen on a smaller reservoir than the selected size N and passed forward to the final system (Lukoševičius, 2012). Output weights $\mathbf{W^o}$ are the only learned values in the network; this is traditionally done with Ridge regression or some form of penalized regression, guarding against uncertainty due to overfitting or instability concerns (Lukoševičius, 2012). Lukoševičius (2012) also recommends averaging results from several generated reservoirs, like an ensemble of weak learners (Polikar, 2012).

Conceptors

Jaeger (2014) introduced the conceptor as a regularized identity mapping of reservoir states in a propagating ESN. A matrix $\mathbf{C}$ is added to the update equation that recognizes, controls, regenerates, and predicts state patterns in a dynamic reservoir (Jaeger, 2014; Jaeger, 2017). The matrix is placed at the front of Equation 1.7 to filter the network states in a manner associated with a specific pattern.

$$\mathbf{h}_t = g\left(\mathbf{W^h}\tilde{\mathbf{h}}_{t-1} + \mathbf{W^i}\mathbf{x}_t + \mathbf{b}\right) \tag{1.9}$$

$$\tilde{\mathbf{h}}_t = \mathbf{C}\mathbf{h}_t \tag{1.10}$$

$$\mathbf{y}_t = \mathbf{W^o}\tilde{\mathbf{h}}_t \tag{1.11}$$

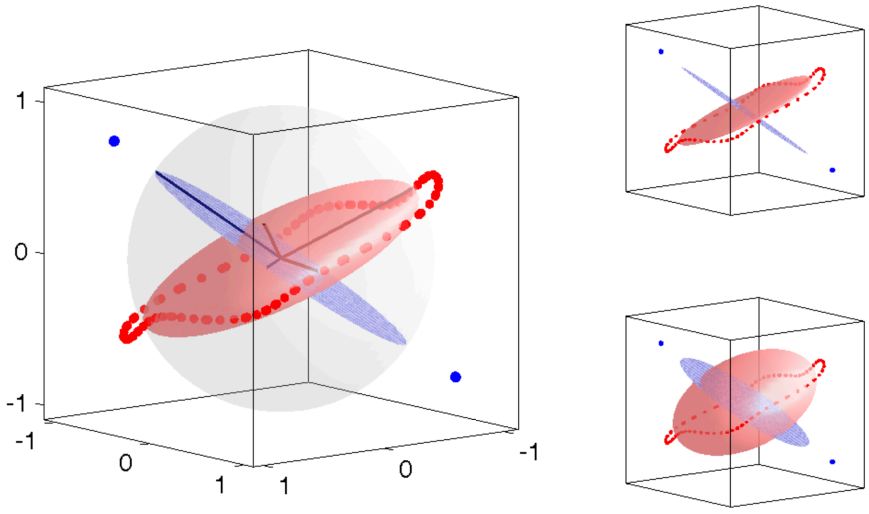
The conceptor matrix $\mathbf{C}$, defined in Equation 1.12, is computed from column space of the network states over a training window of data (of integer length T_{train}) that exhibits a pattern of interest. For states associated with the pattern from the training window, the conceptor matrix should leave the reservoir state unchanged and act like the identity, or $\tilde{\mathbf{h}}_t \approx \mathbf{h}_t$ (Jaeger, 2017). For states propagating in a fashion atypical to the pattern from the training window, the conceptor matrix will act like the null matrix and suppress the reservoir state (Jaeger, 2017).

Over the training window, $\mathbf{H}_{\text{train}} \in \mathbb{R}^{T_{\text{train}} \times N}$ compiles the network states as individual rows. The conceptor is a positive semidefinite matrix ($\mathbf{0} \preceq \mathbf{C} \preceq \mathbf{I}$) that forms an ellipsoid within the unit sphere. The principal axes of the ellipsoid are the eigenvectors of the reservoir state second moment matrix $\mathbf{R} = T_{\text{train}}^{-1} \mathbf{H}_{\text{train}}^{\top} \mathbf{H}_{\text{train}}$ that have been scaled by their eigenvalues.

$$\mathbf{C} \equiv \mathbf{R} \left(\mathbf{R} + \alpha^{-2} \mathbf{I} \right)^{-1} \tag{1.12}$$

A regularization parameter $\alpha \in (0, \infty)$, known as the aperture, influences the degree of dampening in the reservoir activity (Jaeger, 2014; Jaeger, 2017). Large apertures introduce small penalties, and the conceptor will tend to the identity with minimal activity dampening; smaller α shrinks the conceptor to the zero matrix. Figure 1.4 from Jaeger (2017) illustrates the geometry of the conceptor matrix, where the red and blue dots represent different three-dimensional state patters from two training windows of interest.

In many applications, conceptors are appealing for prediction and pattern regeneration tasks because it is possible to remove the input data from the update equation. Equation 1.9 is modified to $\mathbf{h}_t = g\left(\mathbf{W} \tilde{\mathbf{h}}_{t-1} + \mathbf{b} \right)$, where $\mathbf{W}$ is trained from



Left: The red and blue dots represent three-dimensional state patterns from two training windows of interest. Each associated conceptor matrix forms an ellipsoid within the gray unit sphere, and the principal axes are shown. *Right*: The effects of halving (*top*) and doubling (*bottom*) the aperture.

Figure 1.4: Geometric illustration of a conceptor matrix from Figure 2, Jaeger (2017).

the input and reservoir states, and the conceptor governs the dynamics of the system autonomously (Jaeger, 2014; Jaeger, 2017). For example, recognition of reservoir state patterns lends itself to classification problems where a series of known conceptor matrices are used to label sections of data (Bao et al., 2016). Control, regeneration, and prediction of reservoir state dynamics display the ability of the conceptor to learn patterns present in complex data (C. Kiefer, 2019; A. Zhang and Xu, 2020).

1.2 Representation Learning

While most artificial networks are focused on relating an input $\mathbf{x}_t$ to an output $\mathbf{y}_t$, they can also be used to uncover information about a single sequence of interest $\mathbf{y}_t$. ANNs can be thought of as advantageous functional transformations of a dataset to a higher-dimensional space where the nonlinear features and dependence structures become linearly separable (see Equation 1.13). Representation learning (also called feature learning) prioritizes the information encoded in the network states, and leverages these properties to extract information about the original sequence (Bengio, Courville, and Vincent, 2013; LeCun, Bengio, and Hinton, 2015; Goodfellow, Bengio, and Courville, 2016).

$$\Psi : \mathbf{Y} \in \mathbb{R}^{T \times d} \rightarrow \mathbf{H} \in \mathbb{R}^{T \times N} \text{ such that } N \gg d \tag{1.13}$$

This idea is not unique to artificial networks; it is also found in kernel methods and applied to problems in classification, natural language processing, signal processing, transfer learning, drug discovery, and genomics (Bishop, 2006; Bengio, Courville, and Vincent, 2013; LeCun, Bengio, and Hinton, 2015).

1.2.1 Universal Approximation Theorems

Universal approximation theorems, like those in Cybenko (1989) and Hornik (1991), reflect on the ability of an ANN to map the input sequence $\mathbf{x}_t$ to the output sequence $\mathbf{y}_t$ with any arbitrary degree of accuracy. From the context of representation learning, these demonstrate the quality of a representation of network states. If the time series of interest $\mathbf{y}_t$ can be approximated to an arbitrary degree of reliability, then the

reservoir states should provide a faithful representation and "linearize" the nonlinear features present in the data.

These universal approximation theorems hold in several cases. Denote any given function as f and some approximating artificial network $\mathcal{N}$. In the fixed depth and arbitrary width case, simple MLPs are universal function approximators provided the number of network states N is sufficiently large (Cybenko, 1989; Hornik, 1991; Barron, 1993). For a MLP with fixed depth $L \geq 1$, a sigmoidal activation function g, and any $\eta > 0$, there exists some width N such that $\|\mathcal{N} - f\| < \eta$ (Cybenko, 1989; Hornik, 1991). For bounded width and arbitrary depth, a similar result holds provided $\mathcal{N}$ has a sufficient number of layers and the width is above some minimum value (Lu et al., 2017). These results have been extended to fixed width, fixed depth networks by Maiorov and Pinkus (1999), Guliyev and Ismailov (2018a), and Guliyev and Ismailov (2018b).

With a sigmoidal activation function g, universal approximation theorems also apply to RNNs and ESNs of fixed depth and arbitrary width (Schäfer and Zimmerman, 2007; Grigoryeva and Ortega, 2018; Hart, Hook, and Dawes, 2020; Gonon and Ortega, 2021), and fixed width and arbitrary depth provided $N > p + d + \mathrm{C}$, where the input has dimension p, the output has dimension d, and the constant $\mathrm{C} = 3, 4$ depending on the activation function ($\tanh(x)$ or $\sigma(x)$, respectively) (Hoon Song et al., 2023).

1.3 Overview and Application

The methods discussed in the following chapters enlist universal approximation theorems as they apply to representation learning. For a single sequence $\mathbf{y}_t$, consider

the representation $\mathbf{h}_t$ created by an ANN leading to an estimate of the output $\hat{\mathbf{y}}_t = \mathbf{W}^o\mathbf{h}_t$. If $\|\hat{\mathbf{y}}_t - \mathbf{y}_t\|$ is arbitrarily small, one can argue $\mathbf{h}_t$ provides a faithful, higher-dimensional representation of the original data that allows the nonlinear features to become linearly separable with a convenient mapping. From this advantageous featurization, standard linear techniques and practices are more accessible, and the challenging problem of specifying a "close enough" functional form for nonlinear temporal dependence in multivariate time series is avoided.

Chapter 2 applies this perspective to the at most one change point problem in multivariate time series data with arbitrary dependence, Chapter 3 generalizes the first to target multiple change points and inches toward an interpretable structure, and Chapter 4 examines the implications of ANN representation learning on the notion of Granger causality.

Chapter 2

Change Point Detection with Conceptors

This chapter is adapted from the article *Change Point Detection with Conceptors* (Gade and Rodu, 2023a).

The offline change point identification problem is widely discussed in the literature, and used in a broad range of domains like signal processing, human activity monitoring, and finance (Truong, Oudre, and Vayatis, 2020). For time series data $\mathbf{y}_t \in \mathbb{R}^d$ with $t = 1, \ldots, T$, the goal is to retroactively identify points where the distribution changes. Despite the vast amount of work in this field, most of it is focused on identifying mean and variance changes, and the challenge of change point detection in the presence of nonlinear dependence is unresolved. Model based methods assume a known, rigid structure, and specification of a functional form to fit nonlinearities present in data can be difficult. Many nonparametric methods are only applicable to independent and identically distributed (i.i.d.) data, and are applied to cases where changes would be easily identified from visual inspection of a time series plot. Unless they target the aforementioned mean and variance scenarios, most nonparametric methods also have an uninterpretable or shrouded definition of the notion of "change." The contribution of this chapter is a model agnostic method for detecting change in multivariate and arbitrarily dependent nonlinear time series data. The con-

ceptor change point (CCP) methodology builds on prevalent ideas in statistics and representation learning literature. High-dimensional featurizations from ESNs, and a conceptor matrix from a specified and interpretable "baseline" state, allows for the flexibility of change detection in processes with elaborate dependence structures. The ESN controlled by a conceptor is defined in Equations 1.9 to 1.11 and Equation 1.12. This methodology should be used as a tool to suggest potential locations of interest in a dataset where traditional methods and inspection fail.

For this work, the framework is restricted to the at most one change point (AMOC) problem. The distribution functions for each vector in time $\mathbf{y}_1, \ldots, \mathbf{y}_\tau \sim \mathcal{F}_1$ and $\mathbf{y}_{\tau+1}, \ldots, \mathbf{y}_T \sim \mathcal{F}_T$ are compared with the hypotheses

$$\begin{aligned} H_0 &: \mathcal{F}_1 = \mathcal{F}_T \\ H_A &: \mathcal{F}_1 \neq \mathcal{F}_T, \end{aligned} \tag{2.1}$$

where $\mathcal{F}_1$ and $\mathcal{F}_T$ are unknown. Rejection of the null leads to the conclusion that a change took place immediately after time point τ. With sufficient spacing, the framework provides simple extensions to the sparse, multiple change point problem and the online, sequential change point problem.

2.1 Background

Change point detection methods can broadly be classified as sequential, agglomerative, or divisive. Sequential change point detection, like that of Lai (1995), best lends itself to the online problem where changes are identified in sequence while the data is observed. Online change point detection is not the main focus of this chapter.

Agglomerative change point methods begin by labelling each data point as a unique cluster, and proceed by strategically grouping adjacent clusters with an algorithmic criterion; Fryzlewicz (2018) and Matteson and James (2014) provide two examples of these criteria.

Divisive algorithms cluster a series of points and algorithmically search for breaks that best divide into chronological classifications. Many of these algorithms follow the binary segmentation approach, pioneered by Vostrikova (1981), and identify subsequent change points from each of the divided pieces. Penalized techniques like Ombao, Sachs, and W. Guo (2005), Lavielle and Teyssiere (2006), and Killick, Fearnhead, and Eckley (2012) are also prevalent and recent computational work by Haynes, Eckley, and Fearnhead (2017) and Tickle et al. (2020) build on these methods and aim to optimize the search process for efficient change point detection. Cumulative sum (CUSUM) type statistics provide the most common foundation to locate change points in ordered sequences (Picard, 1985; Gombay and Horváth, 1995; Gombay and Horváth, 1999; Cho and Fryzlewicz, 2012; Holmes and Kojadinovic, 2021; Kojadinovic and Verdier, 2021). Many modifications are in the form of the Kolmogorov-Smirnov and Cramer-von Mises criteria, including the self-normalization method of Shao and X. Zhang (2010). Most initial work in this field, like wild binary segmentation of Fryzlewicz (2014), applies only to univariate sequences, and several extensions of these methods to multivariate and high-dimensional applications build on the respective univariate versions (Cho and Fryzlewicz, 2015). Matteson and James (2014) propose a clustering algorithm based on a hierarchical divergence measure for the multivariate, multiple change point problem. This method imposes the strict assumption of i.i.d. data and is only asymptotically justified in the AMOC problem (Arlot, Celisse, and Harchaoui, 2019). Projection of the data into a high-dimensional space by Wang

and Samworth (2018), the kernel trick for change point estimation by Arlot, Celisse, and Harchaoui (2019), and Bayesian estimation like Cappello, O. H. M. Padilla, and Palacios (2023) also import the assumption of independent data. Dehling, Fried, et al. (2015) and Dehling, Vuk, and Wendler (2022) investigate relaxing the i.i.d. requirement by examining the effect of short-term dependence on large sample behavior, and Gerstenberger (2018) explores only the mean change problem for short-term dependence. Certain types of model based detection like Kirch, Muhsal, and Ombao (2015) also relax the i.i.d. requirement, but rely on strong parametric assumptions and impose a rigid structure onto the process. Characterizing nonlinear temporal dependence in change point problems is a challenging, relevant problem with relatively little progress.

2.2 Methodology

The hypothesis in Equation 2.1 is tested under the condition where at most one change point τ is present in the data. An initial training section of the time series is labelled as a "baseline" state from which changes are identified. It is assumed that any change takes place after this baseline window, and the initial distribution $\mathcal{F}_1$ produces a time series that is at least wide-sense cyclostationary. The second assumption ensures the training data covers a relevant range of the data and changes are not falsely identified from unrecognized network dynamics.

The selected training window (of integer length T_{train}) should be sufficiently long to capture the original dynamics of the time series and include representative values from the data prior to a suspected change. If the data exhibits a periodic or almost-periodic type structure, the training length should include at least one full cycle.

The training window need not be at the beginning of the time series; this can be generalized to take any section as a baseline state, where the method looks forward and backward for a potential change.

The proposed method involves three main steps. First, several ESNs are generated and the specified training window is used to select network parameters that satisfy an error tolerance $\varepsilon_{\text{train}}$, defined in terms of the normalized root mean square error (NRMSE) of the reservoir output. In Chapter 2, this is defined as

$$\text{NRMSE} = \left[(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2 \,/\, \left(\frac{1}{2}\text{Var}\,(\mathbf{y}_t) + \frac{1}{2}\text{Var}\,(\hat{\mathbf{y}}_t) \right) \right]^{1/2}. \tag{2.2}$$

The ESN featurizations serve as advantageous functional transformations to a high-dimensional domain where nonlinear relationships are "linearized" without imposing a rigid structure. For each ESN, a conceptor matrix is computed from the baseline that encapsulates information about the dynamics of the time series in that window. The input to the system and its relationship to the conceptor space is exploited; the conceptor records information about the baseline dynamics, and the relationship between the projected (filtered) and unprojected reservoir states is used to highlight differences in the mechanism of evolution. Second, a bisection technique is employed that estimates the most likely change point from these relative differences. Last, a moving block bootstrap is used to estimate the strength of evidence for a proposed change.

2.2.1 ESN Featurization

Along with an integer length for the baseline T_{train}, an integer length T_{wash} is specified and used to washout the initial conditions of a generated ESN reservoir, where

generally $T_{\text{wash}} < T_{\text{train}} \ll T$. A training error $0 < \varepsilon_{\text{train}} \ll 1$ influences hyperparameter settings; this error tolerance is the maximum allowable NRMSE between the conceptor governed ESN output and the original data. Define $T_0 = T_{\text{wash}} + T_{\text{train}}$, and the assumption of no distributional change applies to the time points where $t \leq T_0$, restricting the identification of any change point τ to the interval $[T_0 + 1, T - 1]$.

A reservoir size N is selected where $N \gg d$. A series of $r = 1, \ldots, \mathcal{R}$ ESNs are initialized by generating the matrices from Equation 1.9 like in Section 1.1.1. The input and bias matrices, $\mathbf{W^i}_r \in \mathbb{R}^{N \times d}$ and $\mathbf{b}_r \in \mathbb{R}^N$, are dense with independent random Gaussian realizations, and the variance is determined from a parameter grid such that the ESN output best fits the data as measured by NRMSE, like discussed in Section 1.1.1. Scaling of the input matrices is investigated in a range of 0.2 to 1.4, and scaling of the bias matrices in a range of 0.1 to 0.5 at a small reservoir size. Because the data is contained in a relevant domain prior to input into the ESNs, these grids remain constant for each dataset. Each $\mathbf{W^h}_r \in \mathbb{R}^{N \times N}$ is a sparse matrix with independent random Gaussian nonzero entries, and is scaled to a constant spectral radius of 0.8 to ensure the propagating reservoir states remain stable and wash out (or "forget") the information from any initial conditions. The chosen spectral radius is not at the limit of stability to capture the maximum amount of long-term dependence (Lukoševičius, 2012; Yildiz, Jaeger, and Kiebel, 2012). This is a balance between stability, introducing adequate temporal dependence, and ensuring the reservoir washout is not so long that valuable data is wasted. The full procedure for ESN initialization and scaling is found in Procedure A.1 of Appendix A.

The integer washout length T_{wash} can also be automatically selected from the data. It should be of sufficient length to allow the ESN to operate independently of the reservoir states at $t = 0$. Longer washout periods will remove relevant data from

the analysis. Unless specified, optimum selection is determined from the differences between reservoir states of ESNs with different initial values. After the differences are below a washout error tolerance $\varepsilon_{\text{wash}}$, the states have forgotten the initial condition; the network fitting process is not very sensitive to changes in this tolerance and a default value of 10^{-6} is used. This method is outlined in Procedure A.2 of Appendix A.

The reservoir size and aperture of the network are concurrently determined as the first values large enough to produce a NRMSE below $\varepsilon_{\text{train}}$. The parameter selection is outlined in Procedure A.3 of Appendix A, and the full ESN featurization procedure is given in Algorithm 2.1. Selection of $\varepsilon_{\text{train}}$ should take place prior to analysis and

Algorithm 2.1 ESN Featurization

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}

Outputs: ESN parameters; ESN network size N; aperture α; washout length T_{wash}

1: perform Procedure A.1 to obtain ESN scaling : $\{c_{\text{input}}, c_{\text{bias}}, \rho\}$

2: perform Procedure A.3 to obtain ESN reservoir size N and aperture α

return ESN scaling : $\{c_{\text{input}}, c_{\text{bias}}, \rho\}$, N, α, T_{wash}

not with iterations of the method on multiple parameter values. Smaller values may be more sensitive to learning variations in the noise component of the data, while larger values will extract general behavior. In Section 2.4, the effect of varying $\varepsilon_{\text{train}}$ is demonstrated, but the parameter should take a default of about 4 to 8% NRMSE unless compelling prior knowledge about the type of change sought suggests otherwise.

For each ESN featurization, the associated conceptor matrix is computed from the series of training time points after network washout, $t \in [T_{\text{wash}}+1, T_0]$. The network is propagated forward in time via Equations 1.9 and 1.10 with $\mathbf{C} = \mathbf{I}$, and the reservoir states $\tilde{\mathbf{h}}_t$ ($\mathbf{h}_t$) collected. Conceptor matrices $\mathbf{C}_r$ are obtained via Equation 1.12; then,

each ESN is run with $\mathbf{C}_r$ in place, like Equation 1.10, for all time points $t \geq T_{\text{wash}} + 1$.

2.2.2 Change Point Proposal

The angles between the projected (filtered) reservoir states $\tilde{\mathbf{h}}_{r,t}$ and the unprojected reservoir states $\mathbf{h}_{r,t}$ are examined at each time t after the training period. The cosine similarities of these angles $s_{r,t}$, for each featurization r, form univariate sequences that quantify the proximity of the reservoir states to the space spanned by the corresponding conceptor matrix.

$$s_{r,t} = \frac{\tilde{\mathbf{h}}_{r,t}^{\top} \mathbf{h}_{r,t}}{\left|\left|\tilde{\mathbf{h}}_{r,t}\right|\right| \, ||\mathbf{h}_{r,t}||} = \frac{\mathbf{h}_{r,t}^{\top} \mathbf{C}_r^{\top} \mathbf{h}_{r,t}}{||\mathbf{C}_r \mathbf{h}_{r,t}|| \, ||\mathbf{h}_{r,t}||} \tag{2.3}$$

Values in each sequence $s_{r,t}$ are contained in the interval $[0, 1]$ because, by definition, each $\mathbf{C}_r$ is positive semidefinite. A similarity value can be interpreted as a measure of the strength of relationship between the reservoir state at time t and those in the period of training data. Values of zero imply the ESN is generating states orthogonal to the conceptor space, and those equal to one imply the ESN is generating states exactly in the conceptor space.

Other distance measurements may be used to quantify proximity to the conceptor space. Cosine similarity is a bounded, interpretable quantity that emphasizes angles further from zero at an increasing rate. The exact values of these similarities will vary and their absolute measure is not important; only relative differences are needed for comparison throughout the time series. Figures 2.6, A.2, and A.3 illustrate the relative nature of the similarity measure.

Because the networks are randomly generated, there is variation in the computed

conceptor matrices that extends to the cosine similarities across the ESNs. To extract a general behavior and reduce the dimension of the information, the average cosine similarity at each time point t is considered, $S_t = \mathcal{R}^{-1} \sum_{r=1}^{\mathcal{R}} s_{r,t}$. The aggregate cosine similarity sequence acts as an ensemble of weak learners from each generated ESN.

A proposed change point is selected using a modified CUSUM statistic that resembles the two-sample Kolmogorov-Smirnov distributional test (Smirnov, 1933). From the sequence S_t, empirical CDFs $\hat{\mathcal{F}}_{(T_0+1):t}(s)$ and $\hat{\mathcal{F}}_{(t+1):T}(s)$ are constructed by dividing the sample at each potential change point in the series.

$$\hat{\mathcal{F}}_{(T_0+1):t}(s) = \frac{1}{t - T_0} \sum_{i=T_0+1}^{t} \mathbf{1}\left\{S_i \leq s\right\} \tag{2.4}$$

$$\hat{\mathcal{F}}_{(t+1):T}(s) = \frac{1}{T - t} \sum_{i=t+1}^{T} \mathbf{1}\left\{S_i \leq s\right\} \tag{2.5}$$

A scaled statistic, like that used in Gombay and Horváth (1995) and O. H. M. Padilla, Y. Yu, et al. (2021), is computed at each observation, and the point of the maximum is identified as the most likely change point:

$$K = \max_t \frac{(t - T_0)(T - t)}{q(t)(T - T_0)^2} \sup_s \left| \hat{\mathcal{F}}_{(T_0+1):t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s) \right| \tag{2.6}$$

$$\hat{\tau} = \arg\max_t \frac{(t - T_0)(T - t)}{q(t)(T - T_0)^2} \sup_s \left| \hat{\mathcal{F}}_{(T_0+1):t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s) \right|. \tag{2.7}$$

The coefficient term in Equation 2.6 ensures the statistic converges in distribution under stationarity as $T \to \infty$ (Csörgő and Horváth, 1997; Gombay and Horváth, 1999; Holmes, Kojadinovic, and Quessy, 2013; Kojadinovic and Verdier, 2021). The $q(t)$ scaling function, with form shown in Equation 2.8, increases the sensitivity of the method near the edges of the sequence (Csörgő and Szyszkowicz, 1994a; Csörgő and Szyszkowicz, 1994b; Csörgő and Horváth, 1997; Csörgő, Horváth, and Szyszkowicz,

1997). Further properties of the statistic under the null are discussed in Section 2.3.

$$q(t) = \max\left\{\left(\frac{t-T_0}{T-T_0}\right)^{\nu}\left(1-\frac{t-T_0}{T-T_0}\right)^{\nu},\ \kappa\right\} \tag{2.8}$$

The statistic given in Equation 2.6 is similar to that discussed in Kojadinovic and Verdier (2021), where if $\nu = 1/2$ and κ is a small constant near zero, the mean and variance of the series of statistics remain approximately constant in the limit. Values $\nu = 1/2$ and $\kappa = 0.01$ are chosen for reasons explained in Section 2.3, and the full algorithmic process for proposing a change point is detailed in the Algorithm 2.2.

2.2.3 Moving Block Bootstrap

Estimating the null distribution of the statistic K from Section 2.2.2 may be of more importance than the identification of a potential change point. By nature, many change point detection problems are non-verifiable in real world applications. Thus, reliable change point algorithms should be robust in their ability to detect both change and the lack of change in a dataset.

The bootstrap method of Efron (1979) provides a foundation for inference based on repeated sampling. Several variations of the bootstrap for change point methodology are reviewed in Hušková (2004). The moving block bootstrap (MBB), developed by Kunsch (1989) and R. Y. Liu and Singh (1992), samples blocks of consecutive points to retain the dependence structure within each block. Extending the bootstrap to applications with dependent data structures, the MBB is able to asymptotically reproduce the underlying dependence structure (Lahiri, 2003). The block bootstrap is suggested by Dehling and Philipp (2002) for statistics of empirical processes, and Synowiecki (2007) describes its application to non-stationary data with periodic or

Algorithm 2.2 Change Point Proposal

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; ESN scaling: $\{c_{\text{input}}, c_{\text{bias}}, \rho\}$, N, α, T_{wash} from Algorithm 2.1

Outputs: most likely change point $\hat{\tau}$; statistic K

Default Parameters: number of featurizations $\mathcal{R} \leftarrow 100$

for r in $1 : \mathcal{R}$ **do**

  initialize $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$ where each element $\mathcal{N}(0,1)$, and $\mathbf{W^h}_r$ is sparse

  $\mathbf{W^i}_r \leftarrow c_{\text{input}} \mathbf{W^i}_r$; $\mathbf{b}_r \leftarrow c_{\text{bias}} \mathbf{b}_r$

  $\mathbf{W^h}_r \leftarrow \rho \mathbf{W^h}_r / \max\left\{\mathbf{v}^\top \mathbf{W^h}_r \mathbf{v} : ||\mathbf{v}|| = 1\right\}$

  $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$ for $t = 1, \ldots, T_{\text{wash}} + T_{\text{train}}$

  $\mathbf{C}_r \leftarrow \frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r \left(\frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r + \alpha^{-2} \mathbf{I}\right)^{-1}$

  where $\mathbf{H}_r = \left[\mathbf{h}_{r,T_{\text{wash}}+1} \ \cdots \ \mathbf{h}_{r,T_{\text{wash}}+T_{\text{train}}}\right]^\top$

  $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \tilde{\mathbf{h}}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$; $\tilde{\mathbf{h}}_{r,t} \leftarrow \mathbf{C}_r \mathbf{h}_{r,t}$ for $t = T_{\text{wash}}+1, \ldots, T$

  $s_{r,t} \leftarrow \frac{\tilde{\mathbf{h}}_{r,t}^\top \mathbf{h}_{r,t}}{\left|\left|\tilde{\mathbf{h}}_{r,t}\right|\right| ||\mathbf{h}_{r,t}||}$ for $t = T_{\text{wash}} + T_{\text{train}} + 1, \ldots, T$

end for

$S_t \leftarrow \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s_{r,t}$ for $t = T_{\text{wash}} + T_{\text{train}} + 1, \ldots, T$

for t in $(T_{\text{wash}} + T_{\text{train}} + 1) : (T - 1)$ **do**

  $\hat{\mathcal{F}}_{(T_{\text{wash}}+T_{\text{train}}+1):t}(s) \leftarrow \frac{1}{t - T_{\text{wash}} - T_{\text{train}}} \sum_{i=T_{\text{wash}}+T_{\text{train}}+1}^{t} \mathbf{1}\left\{S_i \le s\right\}$

  $\hat{\mathcal{F}}_{(t+1):T}(s) \leftarrow \frac{1}{T - t} \sum_{i=t+1}^{T} \mathbf{1}\left\{S_i \le s\right\}$

  $K_t \leftarrow \frac{(t - T_{\text{wash}} - T_{\text{train}})(T-t)}{q(t)(T - T_{\text{wash}} - T_{\text{train}})^2} \sup_s \left|\hat{\mathcal{F}}_{(T_{\text{wash}}+T_{\text{train}}+1):t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s)\right|$

end for

$K \leftarrow \max_t K_t$

$\hat{\tau} \leftarrow \arg\max_t K_t$

return $K, \hat{\tau}$, all $\mathbf{C}_r, \mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$

almost periodic behavior.

Potentially overlapping blocks of length L are selected and combined to form a bootstrapped time series of the original length T. When data are treated as i.i.d., as in Matteson and James (2014), the block length parameter reduces to a permutation of the time series with $L = 1$. With increasing block length, the estimated null distribution will exhibit less variation, and null values will tend closer to the statistic K. Choice of block length is discussed widely in literature, and the cross-validation like technique of Hall, Horowitz, and Jing (1995) allows for data driven selection (Bühlmann and Künsch, 1999; Lahiri, 2003; Politis and White, 2004; Lahiri, Furukawa, and Lee, 2007; Patton, Politis, and White, 2009). A large pilot block length is used to start the Hall, Horowitz, and Jing (1995) algorithm to ensure a long range dependence structure is considered, and the process is not iterated as convergence is not guaranteed (Lahiri, 2003). Adjusting the block length for the power and Type 1 error control trade-off should be considered if the researcher possesses some knowledge about the dependence structure of the data.

With the assumption of no change in the washout and training windows, the conceptor matrices do not need to be recomputed: in each bootstrapped series the points $t \leq T_0$ are identical to the original data. All remaining points $t \geq T_0 + 1$ are equally likely to be chosen as the beginning of a bootstrap interval of length L. To ensure equal inclusion probability, the series is wrapped such that a block near the end (a time point within L of T) will cycle back to the initial time considered, $T_0 + 1$. The process to generate bootstrapped data is shown in Procedure A.4 of Appendix A.

For each generated bootstrap time series $b = 1, \ldots, B$, the corresponding maximum statistic is computed as in Section 2.2.2, simulating an approximate null distribution.

A distribution quantile is estimated from the fraction of the B bootstrapped statistics that exceed the statistic from Equation 2.6, $p = B^{-1} \sum_{b=1}^{B} \mathbf{1}\{K_b > K\}$. The quantile estimate provides a notion of the strength of the evidence for a change point. The validity of applying the MBB to the AMOC problem is investigated through simulation and evaluation of Type 1 error control in Section 2.4 for a variety of data generating processes. For a dataset with no change point, the proportion of false rejections in $\mathcal{S}$ simulations is expected to be approximately $q\mathcal{S}$, where q is a predefined threshold of Type 1 error. The full algorithm for estimating a null distribution from the MBB is shown in Algorithm 2.3.

2.3 Theory

The hypothesis in Equation 2.1 is tested using the featurization and conceptor matrix as outlined in Section 2.2. Under the null hypothesis, the cosine similarity values S_t are expected to retain a consistent relative structure and fall close to one (*i.e.*, remain close to the space of the conceptor matrix). Under the alternative, changes in the relationship between the cosine similarity sequence and the space spanned by the conceptor matrix are observed. These changes, initiated by a shift in the data, are not strictly away from the conceptor space; a reduction in variation may lead to reservoir states that lie closer to the conceptor space.

For clarity, the time index is redefined to $T_0 = 0$ and T as the number of data points after washout and training. The assumption of any change after washout and training restricts the domain to $t > 0$. Suppose $S_t \sim \mathcal{F}_t(s)$ for all $s \in [0, 1]$, where each $\mathcal{F}_t$ is a defined distribution function. The hypothesis is reformulated in terms of

Algorithm 2.3 Null Distribution Estimate via Moving Block Bootstrap

Inputs: training window length T_{train}; T_{wash} from Algorithm 2.1; K, all $\mathbf{C}_r$, $\mathbf{W^i}_r$, $\mathbf{b}_r$, $\mathbf{W^h}_r$, $\mathcal{R}$ from Algorithm 2.2

Outputs: null distribution estimate of statistic K_b; quantile estimate at a defined Type 1 error threshold p

perform Procedure A.4 to obtain bootstrapped data $\mathbf{y}_{b,t}$ and B

for b in $1 : B$ **do**

  for r in $1 : \mathcal{R}$ **do**

    $\mathbf{h}_{b,r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \tilde{\mathbf{h}}_{b,r,t-1} + \mathbf{W^i}_r \mathbf{y}_{b,t} + \mathbf{b}_r\right)$; $\tilde{\mathbf{h}}_{b,r,t} \leftarrow \mathbf{C}_r \mathbf{h}_{b,r,t}$

    $s_{b,r,t} \leftarrow \frac{\tilde{\mathbf{h}}_{b,r,t}^{\top} \mathbf{h}_{b,r,t}}{||\tilde{\mathbf{h}}_{b,r,t}|| \, ||\mathbf{h}_{b,r,t}||}$ for $t = T_{\text{wash}} + T_{\text{train}} + 1, \ldots, T$

  end for

  $S_{b,t} \leftarrow \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s_{b,r,t}$ for $t = T_{\text{wash}} + T_{\text{train}} + 1, \ldots, T$

  for t in $(T_{\text{wash}} + T_{\text{train}} + 1) : (T - 1)$ **do**

    $\hat{\mathcal{F}}^{b}_{(T_{\text{wash}}+T_{\text{train}}+1):t}(s) \leftarrow \frac{1}{t - T_{\text{wash}} - T_{\text{train}}} \sum_{i=T_{\text{wash}}+T_{\text{train}}+1}^{t} \mathbf{1}\{S_{b,i} \leq s\}$

    $\hat{\mathcal{F}}^{b}_{(t+1):T}(s) \leftarrow \frac{1}{T - t} \sum_{i=t+1}^{T} \mathbf{1}\{S_{b,i} \leq s\}$

    $K_{b,t} \leftarrow \frac{(t - T_{\text{wash}} - T_{\text{train}})(T - t)}{q(t)(T - T_{\text{wash}} - T_{\text{train}})^{3/2}} \sup_{s} \left| \hat{\mathcal{F}}^{b}_{(T_{\text{wash}}+T_{\text{train}}+1):t}(s) - \hat{\mathcal{F}}^{b}_{(t+1):T}(s) \right|$

  end for

  $K_b \leftarrow \max_t K_{b,t}$

end for

$p \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{K_b > K\}$

return p, all K_b

these distribution functions.

$$
\begin{aligned}
&H_0 : \mathcal{F}_1(s) = \cdots = \mathcal{F}_T(s) \text{ for all } s \in [0,1] \\
&H_A : \exists\ \tau \in \mathbb{Z},\ 1 \leq \tau < T, \text{ such that } \forall\ s \in [0,1],\ \mathcal{F}_1(s) = \cdots = \mathcal{F}_\tau(s) \text{ and} \\
&\qquad \mathcal{F}_{\tau+1}(s) = \cdots = \mathcal{F}_T(s), \text{ and } \mathcal{F}_1(s_0) \neq \mathcal{F}_T(s_0) \text{ for some } s_0 \in [0,1]
\end{aligned} \tag{2.9}
$$

Define the corresponding empirical distribution functions at each point in the time series t, $\hat{\mathcal{F}}_{1:t}(s) = t^{-1} \sum_{i=1}^{t} \mathbf{1}\{S_i \leq s\}$ and $\hat{\mathcal{F}}_{(t+1):T}(s) = (T-t)^{-1} \sum_{i=t+1}^{T} \mathbf{1}\{S_i \leq s\}$. By the Glivenko-Cantelli theorem, these empirical CDFs are consistent estimators of the true distribution functions and they uniformly converge in the limit under stationarity and ergodicity (Tucker, 1959; H. Yu, 1993; Dehling and Philipp, 2002). Summarizing the data with the univariate sequence generated by the conceptor space S_t, rather than using the original multivariate time series, may also make investigation and verification of theoretical assumptions more accessible.

2.3.1 Limiting Distribution Under the Null Hypothesis

Asymptotic behavior of empirical processes for i.i.d. sequences stems from J. Kiefer (1972), that proved almost sure convergence to a Gaussian process. In this work, the $\mathcal{S}$-mixing definition of Berkes, Hörmann, and Schauer (2009) is used to obtain a similar result for the statistic in Equation 2.6.

Assume the sequence S_t is stationary under the null hypothesis, satisfied by construction in Equation 2.9, and can be represented as a shift process of i.i.d. random variables ε_t, $S_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$. Most stationary processes in practice admit a representation as a shift process, and this is causal due to the forward dynamics of the time

series (Berkes, Hörmann, and Schauer, 2009). These assumptions apply only to the sequence indicating the relationship of a given state to the baseline space spanned by the conceptor matrix; they are not imposed on the original data. In a time window without a change point present, assume S_t will arise from some common distribution. A process as $\mathcal{S}$-mixing if the two conditions in Definition 2.1 are satisfied.

Definition 2.1. A random process S_t is $\mathcal{S}$-mixing if:

(1) For any $t \in \mathbb{Z}$ and $m \in \mathbb{N}$, one can find a random variable S_{tm} such that $P(|S_t - S_{tm}| \geq \gamma_m) \leq \delta_m$ for some numerical sequences $\gamma_m \to 0$, $\delta_m \to 0$.

(2) For any disjoint intervals $\mathcal{I}_1, \ldots, \mathcal{I}_r$ of integers and any positive integers $m_1, \ldots, m_r$, the vectors $\{S_{jm_1}, j \in \mathcal{I}_1\}, \ldots, \{S_{jm_r}, j \in \mathcal{I}_r\}$ are independent provided the separation between $\mathcal{I}_1$ and $\mathcal{I}_r$ is greater than $m_1 + m_r$.

With mild assumptions on the function f, Berkes, Hörmann, and Schauer (2009) easily show the shift process representation for a general class of nonlinear processes. Construction of the approximating sequence S_{tm} is discussed via substitution, truncation, coupling, and smoothing techniques (Berkes, Hörmann, and Schauer, 2009). $\mathcal{S}$-mixing is not directly comparable to classical mixing conditions, like α-, β-, or ρ-mixing. The classical mixing definitions lead to clean and precise theoretical results, but verifying the required conditions can be challenging and their scope of application is limited (Berkes, Hörmann, and Schauer, 2009). $\mathcal{S}$-mixing relaxes these requirements to the existence of an approximating sequence that satisfies the above properties. Within the targeted class of shift processes, verification of assumptions is almost immediate, and the resulting strong approximation is used to derive the limiting distribution (Berkes, Hörmann, and Schauer, 2009).

Define the function $q : [0,1] \to (0,1)$ by $q(\delta) = \max\{\delta^{1/2}(1-\delta)^{1/2}, \kappa\}$ and some small $\kappa > 0$ resembling Equation 2.8 with $\nu = 1/2$.

Theorem 2.2. *Let S_t be a stationary sequence such that $\mathcal{F}(s) = P(S_1 \leq s)$ is Lipschitz continuous of order $C > 0$. Assume S_t is $\mathcal{S}$-mixing and that condition (1) of Definition 2.1 holds with $\gamma_m = m^{-AC}$, $\delta_m = m^{-A}$ for some $A > 4$. Under the null hypothesis for every $\kappa \in \left(0, \frac{1}{2}\right)$,*

$$\sqrt{T} \max_{1 \leq t < T} \frac{1}{q\left(\frac{t}{T}\right)} \left[\frac{t(T-t)}{T^2}\right] \sup_{s \in [0,1]} \left|\hat{\mathcal{F}}_{1:t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s)\right| \xrightarrow{D} \sup_{\delta \in [0,1]} \sup_{s \in [0,1]} |\mathcal{K}(s,\delta)| / q(\delta) \tag{2.10}$$

as $T \to \infty$, where $\mathcal{K}(s,\delta)$ is a Gaussian process with

$$\begin{aligned} &\mathbb{E}\left[\mathcal{K}(s,\delta)\right] = 0, \\ &\mathbb{E}\left[\mathcal{K}(s,\delta)\, \mathcal{K}(s',\delta')\right] = (\delta \wedge \delta')\, \Gamma(s,s'), \\ \textit{and} \quad &\Gamma(s,s') = \sum_{-\infty < t < \infty} \mathbb{E}\left[S_1(s) S_t(s')\right], \end{aligned} \tag{2.11}$$

and the limiting random variable is almost surely finite.

The mathematical exposition and proof is an extension of the independent case found in Csörgő and Horváth (1997), Theorem 2.6.1 and is shown in Appendix A. Theorem 2.2 implies the statistic K converges in probability to zero under the null hypothesis.

Stationarity of the average cosine similarities depends on the training window of data and adherence to the AMOC problem. With a well specified, sufficiently long training window such that a relevant range of the time series is covered, the reservoir will emit dynamics close to the conceptor space and the assumption is likely satisfied. With multiple changes present in a dataset, the stationarity assumption

may be violated. In practice, the data should be at least wide-sense cyclostationary, contain at most one change, and not exhibit some long run trend.

2.3.2 Consistent Change Point Estimation

The behavior of the change point estimate $\hat{\tau}$ is examined under the general class of alternatives given in the hypothesis of Equation 2.9. Under construction of the alternative, the average cosine similarity sequence divides into two stationary ergodic pieces on either side of a true change point τ represented by $\mathcal{F}_1(s)$ and $\mathcal{F}_T(s)$. When satisfying modest conditions, $\hat{\tau}$ is a consistent estimator of τ.

Theorem 2.3. *Suppose the sequence S_t, $1, \ldots, T$ divides into two stationary ergodic pieces on either side of the change point τ, and $\mathcal{F}_1(s_0) \neq \mathcal{F}_T(s_0)$ for some $s_0 \in [0,1]$. Then for every $\kappa \in \left(0, \frac{1}{2}\right)$, the change point estimate*

$$\hat{\tau} = \arg\max_{1 \leq t < T} \frac{1}{q(\frac{t}{T})} \left[\frac{t(T-t)}{T^2}\right] \sup_{s \in [0,1]} \left|\hat{\mathcal{F}}_{1:t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s)\right| \tag{2.12}$$

converges in probability to the true value τ under the domain restriction

$$\tau \in \left[\frac{T}{2} - \frac{T}{2}\sqrt{1-4\kappa^2}, \frac{T}{2} + \frac{T}{2}\sqrt{1-4\kappa^2}\right]. \tag{2.13}$$

The proof follows Theorem 2.1 from Newey and McFadden (1994) for consistency of extremeum estimators and can be found in Appendix A. Restricting a possible change point to the interval shown in Theorem 2.3 does not shrink the domain in practice if the chosen $\kappa < \sqrt{\frac{1}{4} - \frac{1}{4}\left(1 - \frac{2}{T}\right)^2}$.

2.4 Simulation Study

Performance of the CCP method is demonstrated through simulations restricted to the AMOC problem. The CCP method is compared to the e-divisive (EDiv) method of Matteson and James (2014) and the kernel change point (KCP) algorithm of Arlot, Celisse, and Harchaoui (2019), both via the `ecp` R package by James and Matteson (2015), along with the sparsified binary segmentation (SBS) methods of Cho and Fryzlewicz (2015), via the `sbs` R package by the same authors. Type 1 sparsified binary segmentation searches for changes in the center of the data, and Type 2 seeks other forms of distributional change (Cho and Fryzlewicz, 2015).

2.4.1 Simulation Settings

Simulated time series fall into the broad classes of VAR, periodic, Gaussian, and white noise processes. All simulated data $\mathbf{y}_t \in \mathbb{R}^2$, $t = 1, \ldots, T$, has length $T = 1000$ with a potential change located from $\tau = 181$ to $\tau = 999$. Table 2.1 summarizes the settings used for each method in the study. CCP requires specification of a training length of data with an associated training error tolerance. The washout length $T_{\text{wash}} = 60$ and

Table 2.1: Parameter settings for methods in simulation study.

Method	Settings
Conceptor Change Point (CCP)	$\varepsilon_{\text{train}} = 2, 4, 8, 16$
E-Divisive (EDiv)	$q = 0.05$
Kernel Change Point (KCP)	$\Pi = 1$, $C = 2$
Sparsified Binary Segmentation (SBS)	$q = 0.05$, Type $= 1, 2$

$\varepsilon_{\text{train}}$ is the error tolerance in % NRMSE, q the significance threshold, Π the maximum number of change points, C the KCP penalty scaling.

the training length $T_{\text{train}} = 120$ are fixed to ensure a constant window of estimation $\hat{\tau} \in [181, 999]$ for all compared methods. The error tolerance $\varepsilon_{\text{train}}$ varies from 2 to 16 percent of NRMSE. The EDiv, KCP, and SBS methods are restricted to the AMOC framework. For EDiv and SBS the estimate is the initial segmentation chosen by the algorithm. KCP accepts an input parameter to restrict to the AMOC problem. EDiv, SBS1, and SBS2 require a significance threshold for change point identification that is set to $q = 0.05$. The same value is used in the CCP method to set an upper threshold on the bootstrap null distribution. KCP requires specification of a penalty parameter for change point identification; Arlot, Celisse, and Harchaoui (2019) outline a procedure for selection of this parameter, and the suggested penalty scaling is used.

When a change in the data is present, the adjusted Rand index (ARI) of Hubert and Arabie (1985) compares the assignment of time points to the correct class, and the empirical CDF of the difference between the identified point and the true change point is computed. Given in Equation 2.14, with δ the fraction of the time series away from the true change point and $\mathcal{S}$ the total number of simulations for a selected setting, the empirical CDF shape in some neighborhood $\{\delta : 0 \leq \delta \leq T^* \ll T\}$ compares performance of the methods.

$$\hat{\mathcal{H}}_{\tau}(\delta) = \frac{1}{\mathcal{S}} \sum_{i=1}^{\mathcal{S}} \mathbf{1} \left\{ \frac{1}{T - T_{\text{wash}} - T_{\text{train}}} |\hat{\tau}_i - \tau_i| \leq \delta \right\} \tag{2.14}$$

Better performing methods will quickly increase to 1, and those that fail to identify an existing change point are evaluated as if it was placed at the end of the series, $\hat{\tau} = 1000$. When no change is present, $\hat{q}$ is defined as the observed Type 1 error and compared with the defined threshold q.

Tables 2.2 and 2.3 detail the data processes examined in the simulation study. For

each case, a no change point scenario is included where the initial data generating process held constant for the full time series. Each setting is indicated by a unique ID and repeated 300 times to create over 13000 simulated datasets.

Table 2.2: VAR and periodic simulation settings.

ID	Simulated Data	ID	Simulated Data
(1a); (2a)	$\rho = 0.5 \to 0.5$	(3a)	$\omega = 1 \to 0.5$
(1b); (2b)	$\rho = 0.5 \to 0.8$	(3b)	$\omega = 1 \to 0.8$
(1c); (2c)	$\rho = 0.8 \to 0.5$	(3c)	$\omega = 1 \to 1.2$
(1d); (2d)	$\rho = 0.8 \to 0.8$	(3d)	$\omega = 1 \to 1.5$
(1e); (2e)	$\rho = 0.5 \to$ NC	(3e)	$\omega = 1 \to$ NC
(1f); (2f)	$\rho = 0.8 \to$ NC		

VAR(1) + $\frac{1}{2}\mathcal{N}_2\left(\mathbf{0}_2, \mathbf{I}_2\right)$, VAR(2) + $\frac{1}{2}\mathcal{N}_2\left(\mathbf{0}_2, \mathbf{I}_2\right)$ spectral radius ρ change simulations, and periodic process frequency ω change simulations $\sin\left(\omega t \left\{+\omega\frac{\pi}{2}\right\}\right)\mathbf{1}_2 + \frac{1}{2}\mathcal{N}_2\left(\mathbf{0}_2, \mathbf{I}_2\right)$. All data $\mathbf{y}_t \in \mathbb{R}^2$, $t = 1, \ldots, T$, has length $T = 1000$ and the change point varies randomly $\tau \in [181, 999]$ or no change (NC). VAR(1) simulations are indicated by ID(1), VAR(2) by ID(2), and periodic by ID(3).

For VAR(γ) processes, the coefficient matrix is randomly generated to have a fixed spectral radius ρ (within a tolerance of 0.02). Change points from autoregressive processes with similar ρ may be more difficult to identify as they can exhibit similar dynamics. All VAR(γ) processes contain a white noise error term defined in Table 2.2. For periodic processes, the second dimension is shifted by of $\pi/2$ as noted by the braced parenthesis in Table 2.2, and all contain a white noise error term. The Ornstein-Uhlenbeck processes are defined by the stochastic differential equation $d\mathbf{x}_t = \theta\mathbf{x}_t dt + \lambda d\mathcal{W}_t$, where $\mathcal{W}_t$ denotes a two-dimensional Wiener process. The two-dimensional Ornstein-Uhlenbeck process is denoted $\mathcal{OU}_2(\Theta, \Lambda)$, where Θ is the 2×2 mean-reverting matrix and Λ is the 2×2 volatility matrix. Gaussian white noise

Table 2.3: Ornstein-Uhlenbeck and white noise simulation settings.

ID	Simulated Data	ID	Simulated Data
(4a)	$\theta = 0.5 \to 0; \lambda = 0.5$	(5a)	$\mu = 0 \to 0.5$
(4b)	$\theta = 0.5 \to 1; \lambda = 0.5$	(5b)	$\mu = 0 \to 0.8$
(4c)	$\theta = 1 \to 0; \lambda = 0.5$	(5c)	$\mu = 0 \to 1$
(4d)	$\theta = 1 \to 0.5; \lambda = 0.5$	(5d)	$\sigma = 1 \to 0.5$
(4e)	$\theta = 0.5; \lambda = 0.5 \to 0.2$	(5e)	$\sigma = 1 \to 0.8$
(4f)	$\theta = 0.5; \lambda = 0.5 \to 0.8$	(5f)	$\sigma = 1 \to 1.2$
(4g)	$\theta = 0.5; \lambda = 0.5 \to 1$	(5g)	$\sigma = 1 \to 1.5$
(4h)	$\theta = 0.5; \lambda = 0.5 \to$ NC	(5h)	$\rho = 0 \to 0.8$
(4i)	$\theta = 1; \lambda = 0.5 \to$ NC	(5i)	$\mu, \rho = 0; \sigma = 1 \to$ NC

Ornstein-Uhlenbeck mean reverting θ and volatility λ change simulations $\mathcal{OU}_2\left(\gamma\mathbf{I}_2, \lambda^2\mathbf{I}_2\right)$ and white noise $\mathcal{N}_2\left(\mathbf{0}_2 + \mu\mathbf{1}_2, \sigma^2\mathbf{I}_2 + \rho\mathbf{J}_2\right)$ mean μ, variance σ, and covariance ρ change simulations, where $\mathbf{J}_2$ refers to the anti-diagonal matrix of ones. All data $\mathbf{y}_t \in \mathbb{R}^2$, $t = 1, \ldots, T$, has length $T = 1000$ and the change point varies randomly $\tau \in [181, 999]$ or no change (NC). Ornstein-Uhlenbeck simulations are indicated by ID(4) and white noise by ID(5).

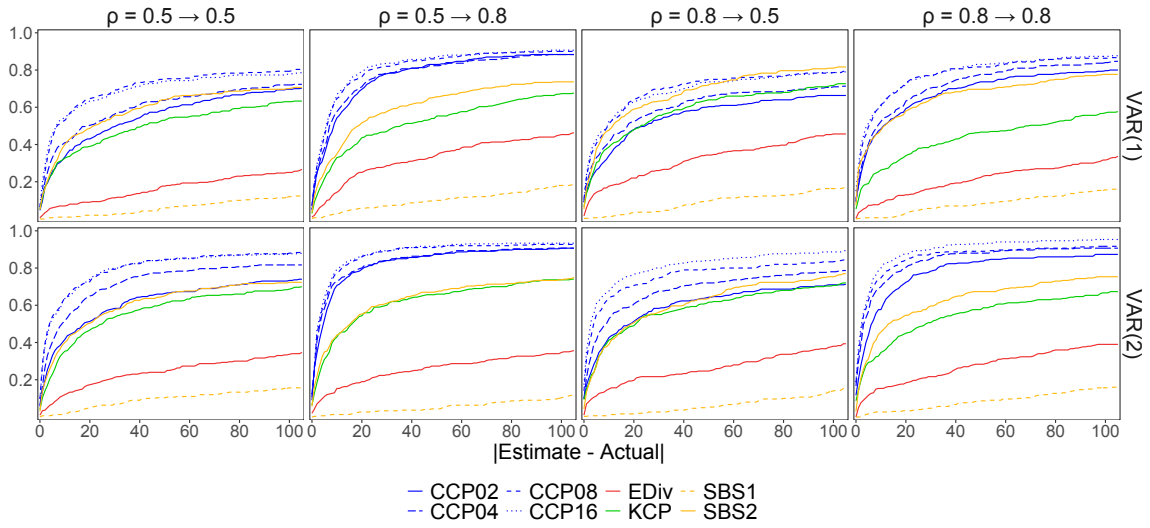
processes with mean, variance, and covariance shifts are included in the simulation to compare CCP methodology to existing methods in benchmark scenarios.

2.4.2 Simulation Results

Simulation results and figures for dependent processes are presented in this section. Tables of results, and tables and figures for white noise processes included for comparison to existing methods, are given in Appendix A.

For VAR(γ) processes with a change point present, the CCP method outperforms existing methodology. This advantage increases with more lagged values in the

dependence structure and when process transitions to a relatively large spectral radius. Figure 2.1 displays the graphical evaluation technique defined in Equation 2.14. Among the conceptor methods, the higher error tolerances (CCP08 and CCP16) pro-

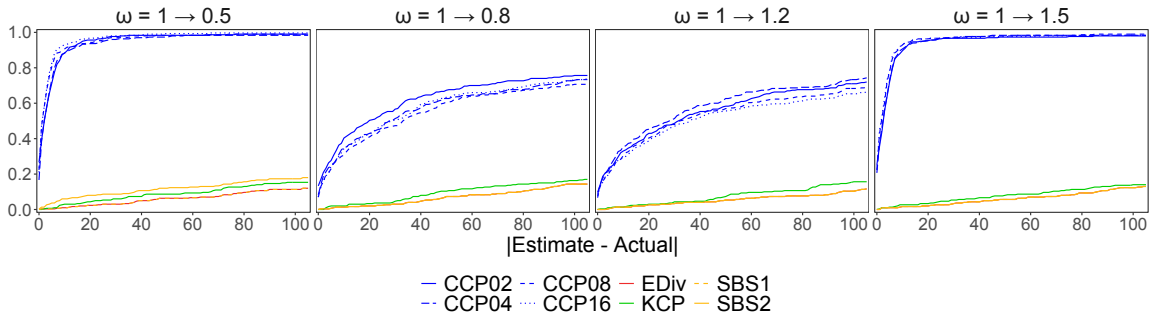


Fraction of identified points within error, VAR(γ) simulations with spectral radius change ρ, IDs (1a-d, 2a-d).

Figure 2.1: VAR(γ) simulation results.

vide a more general fit to the data, where smaller error tolerances (CCP02) produce networks with larger reservoirs that learn the minor deviations of the data. In the presence of noise, these minor deviations occlude the true signal, potentially leading to inconsistencies in the learned behavior. Caution should be taken when specifying the error tolerance in noisy data; moderate tolerances may perform better than small tolerances as they fit networks with constrained internal dynamics, placing more emphasis on a general signal. This phenomenon can be likened to an overfitting problem. The KCP and SBS2 methods are the closest existing methods to the conceptor performance. One major drawback of the SBS method is that the type of change point sought must be specified; the algorithm run with a Type 1 designation produces very low ARI scores for all VAR(γ) simulations.

Figure 2.2 shows the results for simulations where the underlying data is generated by a periodic process. The CCP method is able to reliably detect changes in the frequency of periodic processes when the deviation from an initial state is sufficiently large; existing methodology struggles with this class of processes. While methods in the frequency domain easily detect this type of change, methods that exist in the temporal domain often fail with periodic data. The CCP method in the temporal domain is able to capture this type of nonlinear dependence, as well as those more readily described by the time axis.

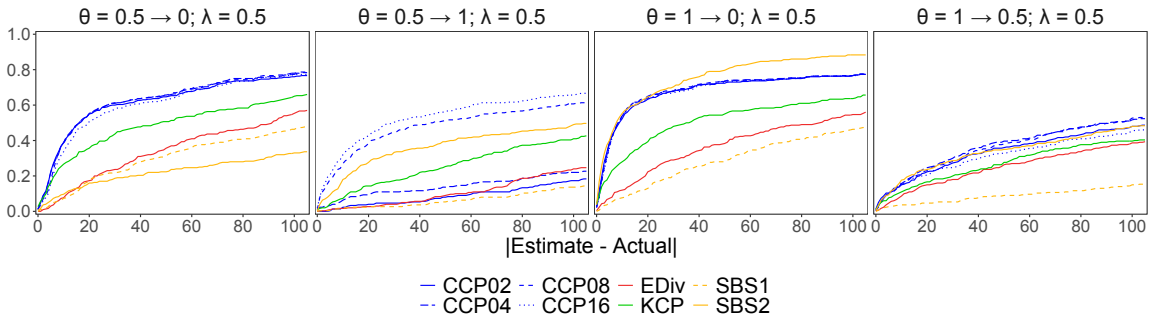


Fraction of identified points within error, periodic simulations with frequency change ω, IDs (3a-d).

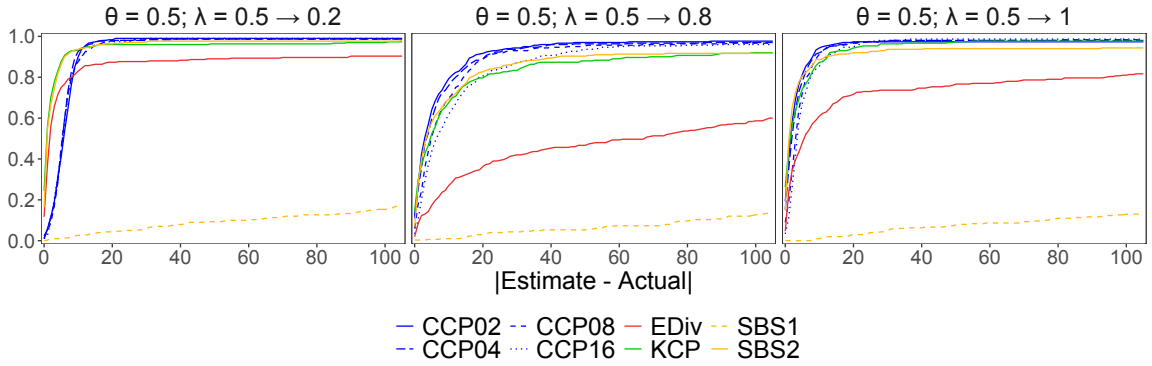
Figure 2.2: Periodic simulation results.

Figure 2.3 displays results for Ornstein-Uhlenbeck simulations with a mean reverting or volatility parameter change. The CCP method surpasses most other methods for detection in mean reverting parameter changes except in some cases when the data shifts to a random walk process (or the parameter goes to zero). The difficulties for all methods can likely be attributed to a relatively low signal to noise ratio. For the volatility, the CCP method is competitive with existing methodology. This behavior is also seen in white noise variance simulations (see Figure A.1 in Appendix A) that play to the strengths of the comparator methods. With a high signal to noise ratio, the CCP method is also competitive in detecting mean changes in white noise

processes.



(a) Fraction of identified points within error, Ornstein-Uhlenbeck simulations with mean reverting change θ, IDs (4a-d).

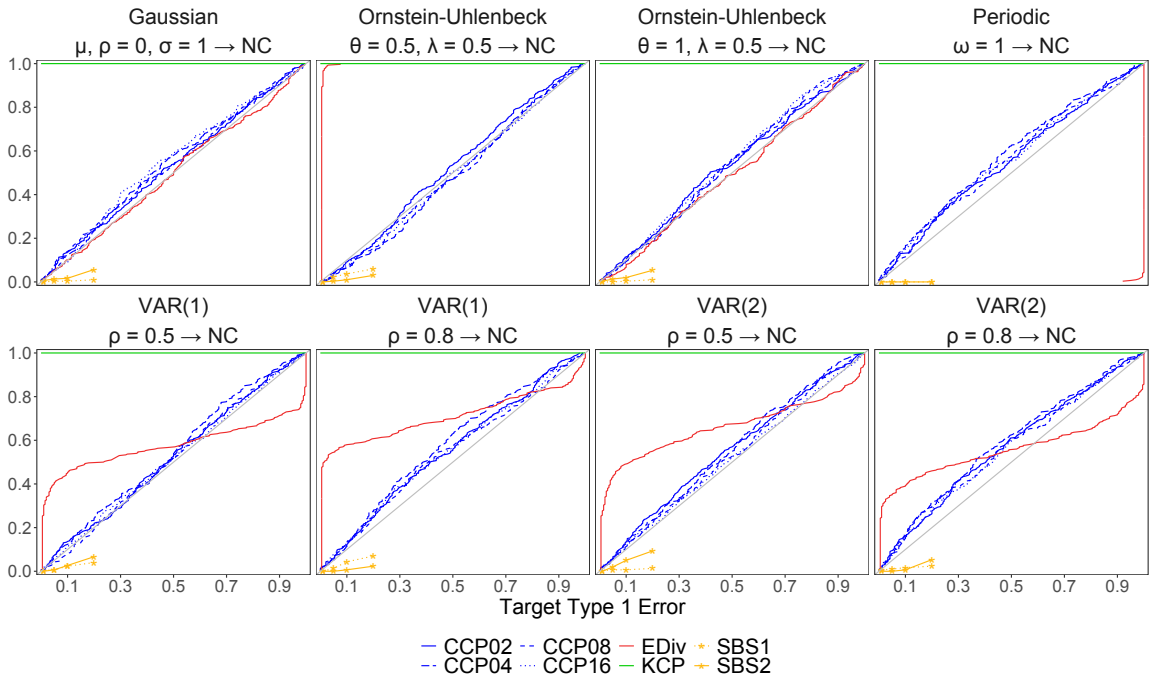


(b) Fraction of identified points within error, Ornstein-Uhlenbeck simulations with volatility change λ, IDs (4e-g).

Figure 2.3: Ornstein-Uhlenbeck simulation results.

To evaluate the validity of the moving block bootstrap in the CCP method, the Type 1 error of each method is observed when no change takes place in the time series. Control for false discovery of change points is as important as the correct identification of a change. Figure 2.4 shows the observed probability of erroneous detection for each method. SBS methods provide conservative error control, KCP methodology almost always flags a change point, and EDiv does not hold to a desired level for data that is not Gaussian white noise. SBS methods do not return a quantile estimate, but only a binary "present" or "not present" flag; to estimate coverage at

the points indicated, the method was run with different values of the threshold q. CCP methodology tracks along the uniform cdf with only slight undercoverage in periodic data and VAR data with large spectral radii.



Uniform cdf included for reference, IDs (1e-f, 2e-f, 3e, 4h-i, 5i).

Figure 2.4: Type 1 error control.

2.5 Application Study

The CCP method is applied to data from Varela and Wilson (2019). The authors record local field potential (LFP) up to 600Hz in the midline thalamus (THAL), medial prefrontal cortex (PFC), and the CA1 region of the hippocampus (HC) in rats experiencing bouts of non-REM sleep and wakefulness while they remained in a quiet, square-shaped enclosure (Varela and Wilson, 2019; Varela and Wilson, 2020). The data, obtained from the Collaborative Research in Computational Neuroscience

data sharing website, also includes determinations of sleep state (awake or non-REM sleep), as well as spiking, spindle, and sharp-wave ripple information over the course of the experiment (Varela and Wilson, 2019). Session 1 is selected (prior to exploration of a radial maze), and the data is filtered with a finite impulse response filter to focus on the delta band (1-4Hz), characterizing slow wave sleep. Finally, the data is downsampled to a frequency of 4Hz.

Three periods of transition identified by Varela and Wilson (2019) are isolated; each spans 100 seconds: sleep to wake (650 to 750s, change point at 740s), wake to sleep (740 to 840s, change point at 800s), and wake to sleep (1080 to 1180s, change point at 1150s). Relatively short windows are selected to satisfy the AMOC assumption; shorter time periods focus the methods to detecting the sleep state transition rather than other dynamic neural process changes almost certainly present in the data. The CCP, EDiv, SBS2, and KCP methods are evaluated on their ability to locate the change points in each transition period.

Approximately 10 seconds are reserved for reservoir washout, and 30 seconds are used for conceptor training with the CCP method. Change points are identified in the remaining one minute for all methods. Settings for methods are given in Table 2.1, and the CCP error tolerance is kept at $\varepsilon_{\text{train}} = 4\%$ NRMSE. Figure 2.5 presents the results from applying the methods to the LFP data. Methods that fail to identify a change point display as a vertical line at the far right edge of the figure.

Figure 2.6 displays the internal dynamics of the conceptor methodology applied to the first (top) time series segment in Figure 2.5. Figures A.2 and A.3 present similar visuals of the second and third segments, respectively. The top plot of Figure 2.6 displays the series of CUSUM-like statistics, with an estimated bootstrap null distribution on the right vertical axis and quantiles as horizontal lines on the plot. The

middle plot displays S_t with a relative vertical axis, as only comparative differences through the time series are sought. The bottom plot gives empirical CDFs of S_t over segmented windows of time; the plotted points display the difference between the empirical CDF of the specific window and the overall empirical CDF of the full time series. Shading in the middle and bottom plots represents the internal reservoir dynamics and their relationship to the conceptor space; blue refers to dynamics that are behaving similarly to the original conceptor space, and red indicates further away. The scale of color shading in the middle plot is tied to percentiles the sequence S_t, and in the bottom plot is a relative difference between empirical CDFs. Change points will be identified as a peak in the top plot and a color transition in the middle and bottom plots. Excessive undulation, a secondary peak, or multiple shifts may suggest violation of the AMOC assumption or a slow transition between states.

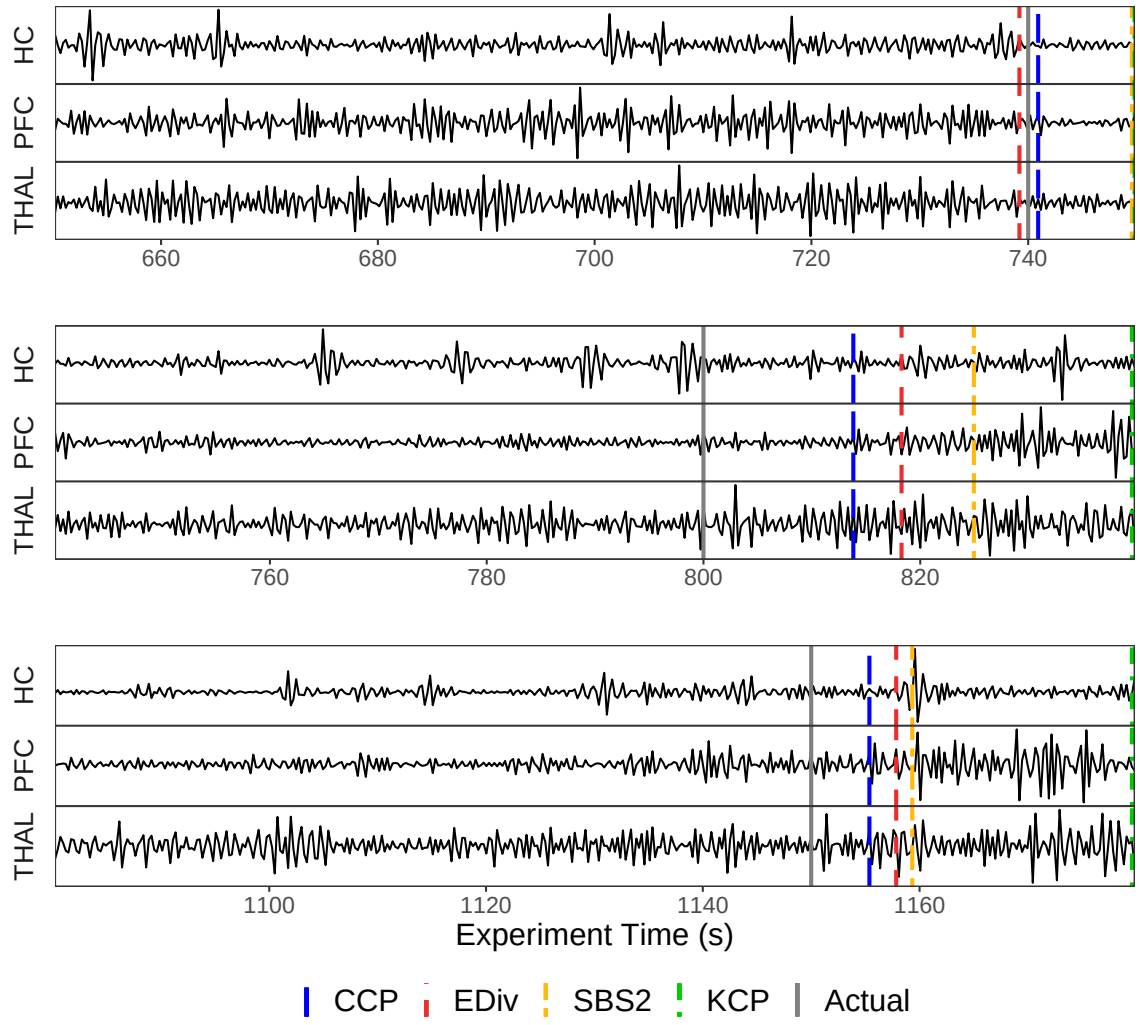
2.6 Discussion

The CCP method provides a model agnostic framework for addressing nonlinear temporal dependence in change point identification problems. This relaxes the common i.i.d. assumption of most existing methodology, and allows for flexible definition of a baseline state without the rigidity of an imposed structure. The method also alleviates the problem of specifying a functional nonlinear form, which can be challenging.

The ESN learns the characteristic dynamics of a training window, and the deviation from the conceptor space ix examined with a CUSUM-like statistic that consistently estimates the true change point under mild assumptions. The method is able to flag important locations for future scrutiny and provide guidance on the strength of evidence for a change point via the moving block bootstrap. CCP outperforms existing

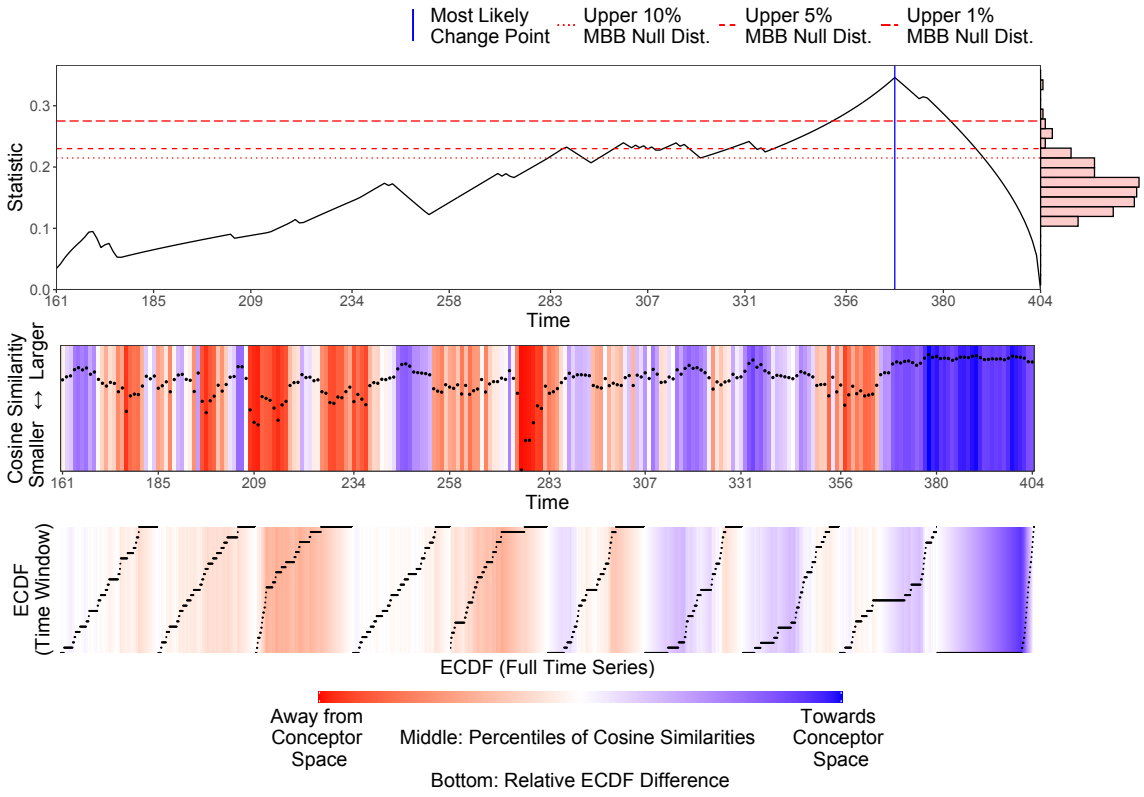
methods in temporally dependent and periodic processes, and is even competitive in i.i.d. processes with a change in variation or a high signal to noise ratio. In practice, the training window should be sufficiently long to capture representative variation of the original time series, and $\varepsilon_{\text{train}}$ left at a default value unless there is prior knowledge of the type of change sought. Assumptions include a baseline period of stable data generation, where the data is at least wide-sense cyclostationary, and stationarity of the obtained similarity sequence S_t on either side of at most one potential change point. Violation of the cosine similarity stationarity assumption will affect theoretical results, but does not diminish use of the method for investigative study of a dataset. Implementation of the method may require isolation of a time segment of interest, like the application in Section 2.5, so that the AMOC assumption is met. These segments must be identified by the researcher with prior knowledge of their data.

The framework in this chapter only applies to the AMOC problem, which can be limiting. Further, all covariate specific information regarding the nature of a change in the dataset is lost. Chapter 3 outlines a straightforward extension to the multiple change point problem provided changes are sparse and sufficiently spaced in a time series. After one identified change, a new baseline window is defined and the algorithm run consecutively for application in both offline and online problems. Future work can improve the data representation so that information pertaining to individual series in the data is preserved, making qualitative conclusions about change points more accessible.



Top: Sleep to Wake, *Middle & Bottom*: Wake to Sleep. Actual change points identified in Varela and Wilson (2019).

Figure 2.5: Estimated change points in LFP example.



Proposed Change Point: 368 (740.9s), Statistic = 0.346, MBB quantile = 0. *Top*: Identification of most likely change point from CUSUM-like statistics. Null bootstrap distribution included on the right vertical axis with estimated quantiles. *Middle*: Cosine similarities between conceptor and reservoir space over the time series. Shading represents percentiles of cosine similarities over the full time series. *Bottom*: Compares segment specific cosine similarity empirical CDFs to the full time series. Shading represents a relative difference of empirical CDFs.

Figure 2.6: CCP method visualization of Figure 2.5 (*top*).

Chapter 3

Multiple Change Point Detection with Conceptors

Extensions to the multiple change point problem are met by challenges beyond just estimating the potential locations. Methodology needs to address the propensity of a technique to overestimate (or underestimate) the number of potential breaks in a multivariate time series and perform separate inference on each of these breaks. With imperfect estimation, an evaluation technique should address the combined error of misplacement (distance away from a true change point) and mismatching (the number of estimated change points not corresponding to the number of true changes, also called annotation error in Truong, Oudre, and Vayatis (2020)). This chapter builds on the work of Chapter 2 and Gade and Rodu (2023a) to provide a framework for the online and offline multiple change point problems in arbitrarily dependent time series data, provided changes are sparse and sufficiently spaced.

3.1 Multiple Change Point Detection

The change point problem has amassed an extensive quantity of published literature and has been applied to problems in engineering, economics, biological sciences, signal processing, and genomics (Niu, Hao, and H. Zhang, 2016; Truong, Oudre, and

Vayatis, 2020; L. Xie et al., 2021). In both retrospective (offline) and sequential (online) analysis, detecting shifts in the underlying process is challenging with nonlinear temporal dependence.

Retrospective change point detection considers a complete section of a time series, $\mathbf{y}_t \in \mathbb{R}^d$ with $t = 1, \ldots, T$, and segments data into sections of consistent dynamics. Often, it is beneficial to reformulate the multiple change point problem into a series of AMOC problems by leveraging local information (Niu, Hao, and H. Zhang, 2016). Sequential detection, as in Lai (1995), aims to identify changes in a continuous data stream as they are observed, usually with a goal of prompt detection. A better result minimizes $T - \tau_j$, with τ_j a change point in the data. In both schemes, with $\boldsymbol{\tau}$ the set of true change points, it is common to include assumptions on sparsity of the total number of change points $n(\boldsymbol{\tau}) \ll T$ and spacing between two adjacent change points with a lower bound $|\tau_j - \tau_{j'}| \geq \gamma^*$ (Niu, Hao, and H. Zhang, 2016).

Ideally the set of estimated change points $\hat{\boldsymbol{\tau}}$ recovers all true change points $\boldsymbol{\tau}$, or $\hat{\tau}_j = \tau_j$ for $j = 1, \ldots, n(\boldsymbol{\tau})$. Denote $\mathcal{F}_t$ as the distribution of the multivariate time series at t. The hypothesis for the AMOC problem, considered in Chapter 2, is adjusted to the non-specific form shown in Equation 3.1.

$$\begin{aligned} H_0 &: \mathcal{F}_1 = \mathcal{F}_2 = \cdots = \mathcal{F}_{T-1} = \mathcal{F}_T \\ H_A &: \exists \text{ at least one } \tau_j \text{ such that } \mathcal{F}_{\tau_j} \neq \mathcal{F}_{\tau_j+1} \end{aligned} \tag{3.1}$$

This statement, as noted in Niu, Hao, and H. Zhang (2016), is too broad for any direct procedure. Specific forms relative to the sequential or retrospective nature of the methodology are shown in Sections 3.1.1 and 3.1.2.

The inherent repeated nature of the hypothesis framework introduces a multiple

testing problem that permits family-wise error rate (FWER) or false discovery rate (FDR) control depending on the tolerance of false detection (Benjamini and Hochberg, 1995). For conservative estimation schemes where false detection is undesirable, a FWER approach controlling $\mathbb{P}\left[(\#\text{ False Detections}) \geq 1\right] \leq q$ may be preferred, but a procedure looking for potential existence of breaks with limited downside of false alarm may utilize the FDR control of $\mathbb{E}\left[(\#\text{ False Detections})/n(\hat{\boldsymbol{\tau}})\right] \leq q$ for some suitable testing threshold q (Benjamini and Hochberg, 1995; H. Li, Munk, and Sieling, 2016). A secondary consideration may also be the definition of a "relevant" change. In a parametric setting driven by a parameter (or set of parameters) θ, a researcher may only be interested in changes larger than some magnitude $\|\theta_t - \theta_{t+1}\| > \Delta$ (Dette and Wied, 2015).

Much of the multiple change point literature focuses on the mean of a univariate sequence, relies on assumptions of temporal independence, or examines parameterized versions of the change point problem where likelihood ratio approaches are applicable (Niu, Hao, and H. Zhang, 2016; L. Xie et al., 2021). Simple data structures may not require complicated methodology, and sparse changes can often be discovered from visual inspection of a time series. In reality, especially when lacking a fundamental scientific understanding, the relevance of changes in a temporally dependent dataset may be much more complex, and parametric methods are rendered unreliable due to the multitude of possible functional forms (Gade and Rodu, 2023a; McGonigle and Cho, 2023).

The sequential conceptor change point (SCCP) and the multiple conceptor change point (MCCP) methods address several of these challenges. They harness the high-dimensional featurization and representation learning strategy from Gade and Rodu (2023a) and Chapter 2 to allow for multiple change point detection in multivari-

ate processes with arbitrary, potentially elaborate and nonlinear, dependence structures. With an interpretable "baseline" state defined from a training window of data, the methods allow for flexible change point detection beyond first and second order changes, or changes in parametric models that impose a rigid structure. Under assumptions on sparsity and spacing of potential change points, these tools can be used to suggest multiple locations of interest when traditional methods and visual inspection fail.

3.1.1 Online Multiple Change Point Detection

Online change point detection (sometimes also classified as statistical process control) examines time series data sequentially (potentially in real-time) with the goal of quick identification. The generic alternative in Equation 3.1 is adjusted to the more specific form in Equation 3.2, where a procedure steps forward in a recursive manner at each time point t.

$$\begin{aligned} H_0 &: \mathcal{F}_1 = \mathcal{F}_2 = \cdots = \mathcal{F}_{t-1} = \mathcal{F}_t \\ H_A &: \mathcal{F}_1 = \mathcal{F}_2 = \cdots = \mathcal{F}_{t-1} \neq \mathcal{F}_t \end{aligned} \tag{3.2}$$

After rejection of the null hypothesis, the time index resets and a method proceeds forward to the next change point. Equation 3.2 represents an ideal situation with immediate detection at time t; often it is necessary to incorporate a detection delay γ, and the resulting alternative considers $\mathcal{F}_1 = \cdots = \mathcal{F}_{t-\gamma} \neq \mathcal{F}_{t-\gamma+1} = \cdots = \mathcal{F}_t$.

Gösmann, Kley, and Dette (2021) makes the distinction between open-ended sequential detection, referring to a continuous observed stream of data where only data

points up to t are realized, and closed-ended detection where the entire dataset has been realized prior to the analysis (Gösmann, Kley, and Dette, 2021). Normalization schemes for a chosen statistic (such as the likelihood-based approaches of Dette and Gösmann (2020) and Gösmann, Kley, and Dette (2021)) may be different in each case, but successful detection in both requires sufficient spacing of consecutive change points such that the minimum separation is greater than the detection delay, $\gamma^* > \gamma$.

Research in this space stems from Page (1954) and Page (1955) and procedures from statistical process control (Shewhart, 1925; Shiryaev, 1963). For i.i.d. streams of data, the likelihood-based methods of Lorden (1971), Pollak (1985), Moustakides (1986), and Ritov (1990) attempt to optimize the sensitivity to change through the trade-off between false alarm and detection delay (L. Xie et al., 2021). These methods also inherently rely on parametric assumptions and assume the notion of change is clearly defined prior to observation of the data. Bai (1997a) and Bai (1997b) investigate least squares methods for parametric change point detection, relax the strict notion of independence among errors, and establish consistency of the resulting estimates. Relaxation of other individual assumptions include S. Zou, Fellouris, and Veeravalli (2017) and Rovatsos et al. (2017) for quickest change detection in transient processes where non-stationary pieces may be present, Tartakovsky, Nikiforov, and Basseville (2014) for change point methods using the generalized likelihood ratio approach in dependent data, and Fryzlewicz (2014) for online or offline change identification in dependent data from ARCH and GARCH models.

Dette and Gösmann (2020) uses approximately linear functionals of the empirical distribution function to extend the likelihood-based approach to dependent processes, but still requires a clear metric or significant underlying knowledge of the scientific process. Gösmann, Stoehr, et al. (2022) considers high-dimensional data and the

trade-off between controlling Type 1 error and the time to a false alarm (false detection), akin to FWER and FDR control (Benjamini and Hochberg, 1995). Generally, statistical process control accepts false detection and considers the relaxed false discovery error rate approach (H. Li, Munk, and Sieling, 2016; Gösmann, Stoehr, et al., 2022).

Lau, Tay, and Veeravalli (2019) examines quickest change point detection when the pre-change distribution is known, but this places a rather inflexible assumption on the process. Nonparametric methods look to relax the presumption of known properties of the data; O. H. M. Padilla, Athey, et al. (2019) develops a sequential, windowed Kolmogorov-Smirnov test (a maximal statistic in a backward-looking section of data) that accepts a user-defined error tolerance for false alarms. While relaxing the rigid, parametric form, this work reverts to the assumption of i.i.d. data (O. H. M. Padilla, Athey, et al., 2019). O. H. M. Padilla, Y. Yu, et al. (2021) and Yi Yu and Rinaldo (2023) develop sequential methods based on a CUSUM statistic, but only apply to univariate sequences. The kernel-based Scan B approach of S. Li et al. (2019), similar to the kernel method of Arlot, Celisse, and Harchaoui (2019), allows for a fully nonparametric change point detection framework, but still relies on the i.i.d. assumption and requires a relatively large fraction of reference data. Wei and Y. Xie (2023) adopts a similar technique with an online kernel-based method, and C. M. M. Padilla et al. (2023) relaxes the supposition of independent data by deriving a change point estimator under the conditions of α-mixing. The i.i.d. assumption is limiting, and α-mixing, while more generalizable than independence, still places restrictions on the observed data and the mechanism of a temporal process. In this work, SCCP addresses these concerns with a fully nonparametric method that allows for nonlinear interactions and does not limit the dependence structure to a specific

form. The goal of quickest change detection is temporarily relaxed to demonstrate the utility of the sequential methodology in capturing arbitrary, potentially nonlinear and complicated, dependence with a limited amount of training data. Extension of the method to optimize the detection delay is left for future research.

3.1.2 Offline Multiple Change Point Detection

Offline multiple change point detection considers a complete section of a time series where all relevant data has been observed (Truong, Oudre, and Vayatis, 2020). The generic hypothesis is adjusted to a similar form as Equation 3.2, but the sequential nature of testing is no longer inherent to the procedure. For each time point in the data $t = 1, \ldots, T-1$,

$$\begin{aligned} H_0 &: \mathcal{F}_t = \mathcal{F}_{t+1} \\ H_A &: \mathcal{F}_t \neq \mathcal{F}_{t+1}, \end{aligned} \tag{3.3}$$

or all points t are initially considered a potential change (Niu, Hao, and H. Zhang, 2016). Multiple testing of this nature permits FWER or FDR control, and H. Li, Munk, and Sieling (2016) examines the latter under the simple scenario of jump discontinuities in a piecewise constant function.

Penalized regression approaches are grounded in parametric methods where a change in a well-defined metric or parameter is sought. The simultaneous multiscale change-point estimator (SMUCE) of Frick, Munk, and Sieling (2014) examines changes in a piecewise constant function by estimating the mean vector via a constrained optimization problem on a local log-likelihood ratio test statistic. Extensions to less restrictive settings include generalization of the Gaussian regression model to

include heterogeneity, modifications of the piecewise function for shift processes (stationary datasets with a causal representation), and investigation of quantile regression for segmentation (Pein, Sieling, and Munk, 2016; H. Li, Q. Guo, and Munk, 2019; Dette, Eckle, and Vetter, 2020; Vanegas, Behr, and Munk, 2022). Other model-based methods, like the likelihood ratio scan method (LSRM) of Yau and Zhao (2015) for univariate data, allow for confidence regions of change, but impose a rigid structures making inference contingent on "close enough" specification of the functional form.

Complicated change point problems can sometimes be transformed to detecting mean changes in a univariate sequence (Niu, Hao, and H. Zhang, 2016), and M. Yu and X. Chen (2020) provides another extension of this framework to account for high-dimensional data. Derivation of this sequence based on a metric of interest, however, is not always straightforward. Nonparametric methods that don't require an explicit specification of a functional form can be more flexible, but may only be applicable to univariate data (Yau and Zhao, 2015; Haynes, Eckley, and Fearnhead, 2017; Haynes, Fearnhead, and Eckley, 2017; Korkas and Fryzlewicz, 2017; Messer, 2022) or assumptions that data is i.i.d. (Matteson and James, 2014; C. Zou et al., 2014; Cho and Fryzlewicz, 2015; Arlot, Celisse, and Harchaoui, 2019; S. Li et al., 2019). Rigid definition of the metric of interest (like a specific search for mean, variance, or correlation changes (Dette, W. Wu, and Zhou, 2019)) can also hinder existing methods in change point analysis when complicated, nonlinear interactions muddle the process.

McGonigle and Cho (2023) develops a method that allows for a flexible definition of change in multivariate time series and does not rely on assumptions of independence. The nonparametric moving sum procedure for detecting changes in the joint characteristic function (NP-MOJO) only requires that the data be piecewise station-

ary (McGonigle and Cho, 2023). The MCCP method targets these same scenarios and relaxes the requirements to include wide-sense cyclostationary processes.

3.2 Methodology

Methodology in this section divides the multiple change point problem sequentially into sets of AMOC problems where local information is leveraged. The SCCP and MCCP methods are extensions of the work presented in Chapter 2 and Gade and Rodu (2023a). The SCCP method steps forward in time to identify potential change points with a fixed detection delay, and the MCCP method aggregates information from both forward and backward-looking algorithmic processes.

3.2.1 Online Multiple Change Point Methodology

To test the hypothesis in Equation 3.2, SCCP begins with selection of an initial training window, of integer length T_{train}, near the beginning of the time series to serve as a "baseline" state from which changes are identified. Several ESNs are generated, with parameters chosen to best fit the data from the baseline window. The baseline state is padded from the left edge by an integer length T_{wash} that serves to wash out the influence of the initial zero-state reservoir conditions. The SCCP methodology builds on repeated iterations of an elementary sequential algorithm outlined below.

For a multivariate dataset $\mathbf{y}_t \in \mathbb{R}^d$, assume the initial distribution $\mathbf{y}_1 \sim \mathcal{F}_1$ produces a time series that is at least wide-sense cyclostationary, the first potential change takes place after $t = T_{\text{wash}} + T_{\text{train}}$, and the minimum spacing between consecutive change points is at least $\gamma^* = T_{\text{wash}} + T_{\text{train}} + 1$. The first assumption prevents

a long-term trend in the data and ensures that the specified training window covers a relevant domain of network states produced by the featurization from ESNs. The spacing assumption fixes the detection delay, $\gamma = T_{\text{wash}} + T_{\text{train}}$, and defines a new baseline window after change identification.

As in Chapter 2, generally $T_{\text{wash}} < T_{\text{train}}$ and define $T_0 = T_{\text{wash}} + T_{\text{train}}$. The assumption of no distributional change applies to time points where $t \leq T_0$. In open-ended change point detection, this restricts the domain of a potential change point $\tau_j \geq T_0 + 1$, and in a closed-ended procedure where the data has a defined endpoint T, changes are restricted to the domain $\tau_j \in [T_0 + 1, T - 1]$.

The elementary sequential method, like the CCP method in Chapter 2, has three main steps. First, several featurizations of the time series are generated from ESNs, and conceptor matrices computed to describe the dynamics of the baseline training window. The behavior of the time series relative to the baseline is summarized in a univariate cosine similarity sequence, and time points of interest are flagged using local maxima of a modified Kolmogorov-Smirnov distance statistic. Change points in the time series are evaluated with moving block bootstraps of the original data at the flagged local maxima.

ESN Featurization & Conceptor Computation

The ESN reservoir size is chosen near the allowable maximum $N = \lfloor 0.9 T_{\text{train}} \rfloor$, with the restriction that $N < T_{\text{train}}$ to avoid degeneracy of the system. A series of $r = 1, \ldots, \mathcal{R}$ ESNs are initialized by generating the matrices $\mathbf{W^h}_r$, $\mathbf{W^i}_r$, and $\mathbf{b}_r$ from Equation 1.9 as in Section 1.1.1 with additional details provided in Section 2.2.1. As in Chapter 2, each associated conceptor matrix is calculated (Equation 1.12) from

the series of training time points after network washout, $t \in [T_{\text{wash}} + 1, T_0]$, and the reservoir states are collected. For all time points $t \geq T_{\text{wash}} + 1$, each ESN propagates forward in time with $\mathbf{C}_r$, like Equation 1.10.

In this chapter, the reservoir size and the aperture $\alpha = 100$ are fixed for computational simplicity. Fixing N at a large value heeds the recommendation of Lukoševičius (2012) to produce a rich representation of the data in the reservoir. The aperture is fixed at a value small enough to obtain slight differences in the projected and unprojected states $\tilde{\mathbf{h}}_t$ and $\mathbf{h}_t$, respectively. Large values of the aperture force the conceptor to the identity matrix, resulting in $\tilde{\mathbf{h}}_t \approx \mathbf{h}_t$ (Jaeger, 2014; Jaeger, 2017). Procedure B.1, similar to Procedure A.1 and the first part of Algorithm 2.2, outlines the determination of parameters for the ESN featurization procedure and computes the conceptor matrices from the baseline training window of data.

Computation of the Kolmogorov Distance Statistic

A scaled Kolmogorov-Smirnov statistic, resembling Equation 2.6 in Chapter 2, is restricted to include only local information around a time point of interest. With a fixed detection delay $\gamma = T_0$, the first point examined as a potential change $t = T_0 + 1$ will occur when the multivariate data $\mathbf{y}_t$ has been observed up to time $2T_0 + 1$. Each successive point t in the sequence will undergo investigation as a potential change when the right endpoint reaches $T_{\text{end}} = t + T_0$. In closed-ended detection (if the data has a defined endpoint T) this definition is altered to $T_{\text{end}} = \min\{t + T_0,\ T\}$.

The cosine similarities $s_{r,t}$ from Equation 2.3, for each ESN featurization r, form univariate sequences contained in the interval $[0, 1]$ (because each conceptor matrix $\mathbf{C}_r$ is positive semidefinite) that quantify the similarity between reservoir state behavior

and the baseline training window. To guard against the variability inherent to each random ESN featurization and extract a general trend, these individual sequences are aggregated as an ensemble of weak learners, like in bagging (Breiman, 1996). The method considers the average similarity sequence $S_t = \mathcal{R}^{-1} \sum_{r=1}^{\mathcal{R}} s_{r,t}$ as presented in Section 2.2.2.

The sequence S_t is divided into two sections that collectively span the region $[T_0 + 1, T_{\text{end}}]$. Equations 2.4 and 2.5 are modified to those shown below with the right endpoint adjusted to adhere to the specified window.

$$\hat{\mathcal{F}}_{(T_0+1):t}(s) = \frac{1}{t - T_0} \sum_{i=T_0+1}^{t} \mathbf{1}\{S_i \leq s\} \tag{3.4}$$

$$\hat{\mathcal{F}}_{(t+1):T_{\text{end}}}(s) = \frac{1}{T_{\text{end}} - t} \sum_{i=t+1}^{T_{\text{end}}} \mathbf{1}\{S_i \leq s\} \tag{3.5}$$

The sequence of scaled statistics in Equation 3.6 resembles the equivalent form in Equation 2.6 (Gombay and Horváth, 1995; O. H. M. Padilla, Y. Yu, et al., 2021). For time points in the domain of a potential change,

$$K_t = \frac{(t - T_0)(T_{\text{end}} - t)}{q(t)(T_{\text{end}} - T_0)^2} \sup_s \left| \hat{\mathcal{F}}_{(T_0+1):t}(s) - \hat{\mathcal{F}}_{(t+1):T_{\text{end}}}(s) \right|, \tag{3.6}$$

and $K_t = 0$ otherwise. The scaling function $q(t)$ defined in Equation 3.7 is equivalent to the Chapter 2 form in Equation 2.8.

$$q(t) = \max\left\{ \left(\frac{t - T_0}{T_{\text{end}} - T_0} \right)^{\nu} \left(1 - \frac{t - T_0}{T_{\text{end}} - T_0} \right)^{\nu}, \kappa \right\} \tag{3.7}$$

Like in Chapter 2 and Gade and Rodu (2023a), values $\nu = 1/2$ and $\kappa = 0.01$ normalize the behavior of the Kolmogorov-Smirnov statistic near the edges of the se-

quence (Csörgő and Szyszkowicz, 1994a; Csörgő and Szyszkowicz, 1994b; Csörgő and Horváth, 1997; Csörgő, Horváth, and Szyszkowicz, 1997; Kojadinovic and Verdier, 2021).

MBB Inferential Procedure

Local maxima in the sequence K_t indicate regions where the point t divides the time window into sections with greater distributional dissimilarity. Define $\hat{\boldsymbol{\tau}}^*$ in Equation 3.8 as the set of flagged potential change points that require an inferential testing procedure, and label the associated local maximum statistics K^j.

$$\hat{\boldsymbol{\tau}}^* = \left\{ t \;\middle|\; K_t \geq K_{t'} \; \forall \; |t - t'| \leq T_0 \right\} \tag{3.8}$$

The set $\hat{\boldsymbol{\tau}}^*$ contains all points $\hat{\tau}_j^*$ that correspond to local maxima in the sequence K_t with separation at least equal to the assumed minimum spacing $\gamma^* = T_0 + 1$.

If time point t results in a flagged local maximum for inclusion in $\hat{\boldsymbol{\tau}}^*$, a moving block bootstrap, similar to that performed in Chapter 2 and Gade and Rodu (2023a), samples and concatenates potentially overlapping blocks of data to form a bootstrapped time series with dimension equivalent to the original dataset (Kunsch, 1989; R. Y. Liu and Singh, 1992). Dependence within blocks of the time series remains intact, but longer-term dependence, like a shift in the data due to a change point, will be shuffled to generate an approximate null distribution of the scaled maximum Kolmogorov-Smirnov statistic. Further discussion of the applicability and implications of the MBB is presented in Section 2.2.3.

As in Section 2.2.3, the assumption of no change in the baseline training window leaves the conceptor matrices, and all reservoir states prior to T_0, unchanged. Each

time point in the interval $[T_0+1, T_{\text{end}}]$ is equally likely to be chosen as the beginning of a bootstrap interval of length L, where a point within L of T_{end} will cycle back to T_0+1 ensuring equal inclusion probability. With the process to generate bootstrapped data in Procedure B.2, similar to Procedure A.4, $b = 1, \ldots, B$ bootstrapped time series are generated. The associated maximum statistic K_b^j is computed for each bootstrapped time series as above simulating an approximate null distribution for the local maximum K^j. The bootstrap distributional quantile $p_j = B^{-1} \sum_{b=1}^{B} \mathbf{1}\left\{K_b^j < K^j\right\}$ provides an estimate for the strength of evidence for the proposed point t corresponding to a true change point, and the full approach is outlined in Procedure B.3.

For a cutoff value c_q, the elementary sequential method terminates after labelling the first flagged point $\hat{\tau}_1^*$ as a change point if $p_1 \leq c_q$, or proceeds to the next time point in the sequence if $p_1 > c_q$. In the case of the latter, the method will proceed until some $p_j \leq c_q$ for a potential change point $\hat{\tau}_j^*$, or for closed-ended online detection, result in no found change point after examining all points in the time series. The full process for the elementary sequential framework, that serves as a building block for SCCP, is given in Algorithm 3.1.

Online Sequential Conceptor Change Point Method

The SCCP method performs repeated iterations of the elementary sequential methodology, stepping forward in time until all estimated change points have been identified $\hat{\tau}_j \in \hat{\boldsymbol{\tau}}$. Designate the first change point estimate, returned by an initial run of Algorithm 3.1, as $\hat{\tau}_1$. After locating $\hat{\tau}_1$, the elementary sequential method is performed on the truncated time series $\mathbf{y}_t \in \mathbb{R}^d$, $t = \hat{\tau}_1 + 1, \ldots, T$. Designate time points $t \in [\hat{\tau}_1 + 1,\ \hat{\tau}_1 + T_{\text{wash}}]$ for reservoir washout, and define $t \in [\hat{\tau}_1 + T_{\text{wash}} + 1,\ \hat{\tau}_1 + T_0]$ as the new baseline training window for the next change point.

Algorithm 3.1 Elementary Sequential Method

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; washout length T_{wash}; testing cutoff threshold c_q

Outputs: estimated change point $\hat{\tau}$; statistic K; p-value; MBB null distribution K_b

perform Procedure B.1 to obtain ESN scaling, N, all $\mathbf{C}_r$, $\mathbf{W^i}_r$, $\mathbf{b}_r$, $\mathbf{W^h}_r$

for r in $1 : \mathcal{R}$ **do**

  $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$ for $t = 1, \ldots, T_{\text{wash}}$

  $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \tilde{\mathbf{h}}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$;

  $\tilde{\mathbf{h}}_{r,t} \leftarrow \mathbf{C}_r \mathbf{h}_{r,t}$ for $t = T_{\text{wash}} + 1, \ldots, 2T_0$

  $s_{r,t} \leftarrow \frac{\tilde{\mathbf{h}}_{r,t}^\top \mathbf{h}_{r,t}}{||\tilde{\mathbf{h}}_{r,t}||||\mathbf{h}_{r,t}||}$ for $t = T_0 + 1, \ldots, 2T_0$

end for

$S_t \leftarrow \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s_{r,t}$ for $t = T_0 + 1, \ldots, 2T_0$

for t in $(T_0 + 1) : (T - 1)$ **do**

  $T_{\text{end}} \leftarrow \min\{t + T_0,\ T\}$

  if $S_{T_{\text{end}}} = \text{NULL}$ **then**

    $s_{r,T_{\text{end}}} \leftarrow \frac{\tilde{\mathbf{h}}_{r,T_{\text{end}}}^\top \mathbf{h}_{r,T_{\text{end}}}}{||\tilde{\mathbf{h}}_{r,T_{\text{end}}}||||\mathbf{h}_{r,T_{\text{end}}}||}$ for $r = 1, \ldots, \mathcal{R}$; $S_{T_{\text{end}}} \leftarrow \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s_{r,T_{\text{end}}}$

  end if

  $\hat{\mathcal{F}}_{(T_0+1):t}(s) \leftarrow \frac{1}{t - T_0} \sum_{i=T_0+1}^{t} \mathbf{1}\{S_i \le s\}$

  $\hat{\mathcal{F}}_{(t+1):T_{\text{end}}}(s) \leftarrow \frac{1}{T_{\text{end}} - t} \sum_{i=t+1}^{T_{\text{end}}} \mathbf{1}\{S_i \le s\}$

  $K_t \leftarrow \frac{(t-T_0)(T_{\text{end}}-t)}{q(t)(T_{\text{end}}-T_0)^2} \sup_s \left|\hat{\mathcal{F}}_{(T_0+1):t}(s) - \hat{\mathcal{F}}_{(t+1):T_{\text{end}}}(s)\right|$

  perform Procedure B.4 to obtain $\hat{\tau}$, K, p, all K_b

  if $\hat{\tau} \neq \text{NULL}$ **then**

    break

  end if

end for

return estimated change point $\hat{\tau}$, K, p, all K_b

This framework introduces a multiple testing scenario, where up to $n(\hat{\boldsymbol{\tau}}^*) \in \mathbb{N}$ hypothesis tests are performed via moving block bootstraps on the local maxima included in the set $\hat{\boldsymbol{\tau}}^*$. The tests are conducted in succession; if test j rejects the null hypothesis, the elementary sequential method resets, starting at a new time point. If test j fails to reject the null, test $j+1$ is performed on the next local maximum in the sequence. For closed-ended online detection, the process repeats until a run of the elementary sequential method fails to identify a change and returns a null value, or until an estimated change point $\hat{\tau}_j$ is within the assumed minimum spacing to the endpoint, $\hat{\tau}_j \geq T - \gamma^* = T - T_0 - 1$. If observed simultaneously, a procedure to control the FWER at the Type 1 error threshold q could consider the $n(\hat{\boldsymbol{\tau}}^*)$ hypotheses collectively, and follow one of several well-known methods (Bonferroni, 1936; Šidák, 1967; Holm, 1979; Shaffer, 1986; Hochberg, 1988; Romano and Wolf, 2005). This, however, is not feasible; the sequential nature of the testing procedure requires a decision rule prior to observing $n(\hat{\boldsymbol{\tau}}^*)$.

Define $\max\{n(\hat{\boldsymbol{\tau}}^*)\}$ as the maximum number of hypothesis tests potentially conducted in the SCCP method. The sequential procedure adapted from the simultaneous process in Holm (1979) with cutoff value,

$$c_{q,j} = \frac{q}{\max\{n(\hat{\boldsymbol{\tau}}^*)\} - k_j}, \tag{3.9}$$

where k_j is the number of rejected null hypotheses prior to test $j = 1, \ldots, \max\{n(\hat{\boldsymbol{\tau}}^*)\}$, conservatively controls the FWER at the defined threshold q for any possible $0 \leq n(\hat{\boldsymbol{\tau}}^*) \leq \max\{n(\hat{\boldsymbol{\tau}}^*)\}$. Further details of the sequential testing procedure are presented in Section 3.3.3. For the first hypothesis test, $k = 0$ and the value in Equation 3.9 is equivalent to the Bonferroni correction (Bonferroni, 1936). After rejection of hypothesis j, $k_{j+1} = k_j + 1$ and $c_{q,j+1} > c_{q,j}$, and the adapted sequential procedure

is uniformly more powerful than Bonferroni (Bonferroni, 1936; Holm, 1979). Due to the assumed spacing requirements of consecutive change points and selection of local maxima, the method will flag at most

$$\max\left\{n(\hat{\boldsymbol{\tau}}^*)\right\} = \lfloor (T-1)/\gamma^* \rfloor = \lfloor (T-1)/(T_0+1) \rfloor \tag{3.10}$$

points for testing by means of a moving block bootstrap, and this value is used in Equation 3.9 for the cutoff threshold.

The procedure for level-q FWER control requires the number of bootstrapped time series $B \geq \lfloor (T-1)/(T_0+1) \rfloor / q$ to avoid a trivial scenario that the MBB is not sensitive enough to clear the initial rejection threshold $c_{q,1}$. This trivial case is also observed when $T \gg T_0$, and in open-ended detection when there is no defined endpoint of the time series ($\max\left\{n(\hat{\boldsymbol{\tau}}^*)\right\} \to \infty$), resulting in $c_{q,1} \to 0$. These settings require a switch to FDR Type 1 error control, where $c_{q,j} \approx q$ for all j, and the threshold represents the *approximate* expected proportion of erroneous change points identified (Benjamini and Hochberg, 1995). FDR error control may be also used in closed-ended detection when there is reasonable tolerance for false alarm (H. Li, Munk, and Sieling, 2016; Gösmann, Stoehr, et al., 2022). For more thorough procedures that control the FDR, refer to Storey (2002) and Foster and Stine (2008). Algorithm 3.2 presents the full process for SCCP.

3.2.2 Offline Multiple Change Point Methodology

Like SCCP, the MCCP methodology builds on repetitions of the elementary sequential method, but with the complete section of the time series available, the algorithm uses both forward and backward-looking processes to return $\hat{\boldsymbol{\tau}}$, the estimate of the change

Algorithm 3.2 Online Sequential Conceptor Change Point Method

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; washout length T_{wash}; Type 1 error threshold q; Error control {FWER, FDR}

Outputs: estimated change point set $\hat{\boldsymbol{\tau}}$

$\hat{\boldsymbol{\tau}} \leftarrow \emptyset$; $i \leftarrow 1$; $k \leftarrow 0$; EndLoop $\leftarrow$ FALSE
while EndLoop = FALSE **do**
    if FWER **then**
        $c_q \leftarrow \frac{q}{\lfloor (T-1)/(T_0+1) \rfloor - k}$
    else
        $c_q \leftarrow q$
    end if
    perform Algorithm 3.1 to obtain $\hat{\tau}_i$
    if $\hat{\tau}_i$ = NULL **then**
        EndLoop $\leftarrow$ TRUE
    else
        $\hat{\boldsymbol{\tau}} \leftarrow \hat{\boldsymbol{\tau}} \cup \{\hat{\tau}_i\}$
        if $\hat{\tau}_i \geq T - T_0 - 1$ **then**
            EndLoop $\leftarrow$ TRUE
        end if
        $i \leftarrow i + 1$; $k \leftarrow k + 1$
    end if
end while
return estimated change point set $\hat{\boldsymbol{\tau}}$

point set $\boldsymbol{\tau}$.

To test the hypothesis in Equation 3.3, MCCP begins with selection of an initial training window length T_{train}, like in Section 3.2.1. Because all relevant data has been observed, MCCP can aggregate information from both forward and backward algorithms. The forward and backward algorithms follow equivalent processes; the first considers multivariate time series data $\mathbf{y}_t \in \mathbb{R}^d$ for $t = 1, \ldots, T$ in order. The baseline training window is defined as $t \in [T_{\text{wash}} + 1,\ T_0]$ for $T_0 = T_{\text{wash}} + T_{\text{train}}$ and some integer length T_{wash} that pads the baseline state to wash out initial ESN reservoir conditions. The backward algorithm considers the reversed time series $\mathbf{y}'_t = \mathbf{y}_{T-t+1}$. This defines a baseline training window as $t \in [T - T_0 + 1,\ T - T_{\text{wash}}]$. As in Section 3.2.1, assume the initial distributions produce a time series that is at least wide-sense cyclostationary, no change points take place in the washout or training regions of data, and the minimum spacing between consecutive change points is at least $\gamma^* = T_0 + 1$. Combining the assumed regions of consistent behavior, this restricts changes to the domain $\tau_j \in [T_0 + 1,\ T - T_0]$.

Estimated change points are accumulated from the elementary sequential method in Algorithm 3.1, similar to the SCCP method of Algorithm 3.2. The forward-looking procedure returns the set of estimated change points $\hat{\boldsymbol{\tau}}^f$, and the backward-looking procedure produces the corresponding set $\hat{\boldsymbol{\tau}}^b$. Level-$q$ FWER control for two repetitions of SCCP modifies the cutoff for change point identification with a Bonferroni correction, and adjusts Equation 3.10 to $\max\left\{n\left(\hat{\boldsymbol{\tau}}^*\right)\right\} = \lfloor (T - T_0) \,/\, (T_0 + 1) \rfloor$.

$$c_{q,j} = \frac{q}{2\left[\max\left\{n(\hat{\boldsymbol{\tau}}^*)\right\} - k_j\right]} \tag{3.11}$$

The FWER of each individual instance of SCCP is at most $q/2$ in a strong sense, and

thus, the collective FWER for MCCP is at most q (Bonferroni, 1936; Holm, 1979). *Approximate* FDR control, like in Section 3.2.1, sets $c_{q,j} = q/2$.

MCCP requires reconciliation between the two estimated change point sets $\hat{\boldsymbol{\tau}}^f$ and $\hat{\boldsymbol{\tau}}^b$. Denote $\hat{\boldsymbol{\tau}}^f = \left\{\hat{\tau}_1^f, \ldots, \hat{\tau}_{n(\hat{\boldsymbol{\tau}}^f)}^f\right\}$ and $\hat{\boldsymbol{\tau}}^b = \left\{\hat{\tau}_1^b, \ldots, \hat{\tau}_{n(\hat{\boldsymbol{\tau}}^b)}^b\right\}$, and from the sequential nature of the procedure, $\hat{\tau}_1^f < \cdots < \hat{\tau}_{n(\hat{\boldsymbol{\tau}}^f)}^f$, and $\hat{\tau}_1^b > \cdots > \hat{\tau}_{n(\hat{\boldsymbol{\tau}}^b)}^b$. Define the minimum spacing between two consecutive change points $\gamma^* = T_0 + 1$, where $|\hat{\tau}_i^\delta - \hat{\tau}_j^\delta| \geq \gamma^*$ for $\delta = f, b$ and $i \neq j$. Flagged change points within the minimum spacing are aggregated to regions of dynamic behavior, or perhaps transition, that warrant further examination.

Let $\hat{\boldsymbol{\tau}}^{fb} = \hat{\boldsymbol{\tau}}^f \cap \hat{\boldsymbol{\tau}}^b$, and write $\hat{\tau}_{(1)}^{fb}, \ldots, \hat{\tau}_{(n_{fb})}^{fb}$ as the ordered sequence of $n_{fb} = n\left(\hat{\boldsymbol{\tau}}^{fb}\right)$ points in the intersection, where $n_{fb} \leq n(\hat{\boldsymbol{\tau}}^f) + n(\hat{\boldsymbol{\tau}}^b)$. Define a neighborhood around each $\hat{\tau}_{(i)}^{fb}$ as $\mathcal{N}\left(\hat{\tau}_{(i)}^{fb}; \gamma^*\right) = \left\{t \in \mathbb{N} : |t - \hat{\tau}_{(i)}^{fb}| < \gamma^*\right\}$. Consider the intersections

$$\mathcal{A}_i = \mathcal{N}\left(\hat{\tau}_{(i)}^{fb}; \gamma^*\right) \cap \hat{\boldsymbol{\tau}}^{fb} \tag{3.12}$$

for $i = 1, \ldots, n_{fb}$, where $1 \leq n(\mathcal{A}_i) \leq 3$, that create groups of adjacent points within the assumed minimum spacing. The goal is to identify sets $\mathcal{B}_k = \{b_{k,l}\}$ for $k = 1, \ldots, n_{\mathcal{B}}$ and $l = 1, \ldots, n(\mathcal{B}_k)$ as groupings of the sets $\mathcal{A}_i$, where each element $b_{k,l} \in \mathcal{N}(b_{k,l'}; \gamma^*)$ for $l \neq l'$, and $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ for $k \neq k'$. As shown in Equations 3.13 and 3.14, a set $\mathcal{B}_k$ contains a single estimated change point at the midpoint of the region if $\max\{\mathcal{B}_k\} - \min\{\mathcal{B}_k\} < \gamma^*$ (*i.e.*, $n(\mathcal{B}_k) \leq 2$), or two estimated change points on either side of a region of dynamic behavior ("transition region") if

$\max\{\mathcal{B}_k\} - \min\{\mathcal{B}_k\} \geq \gamma^*$ (*i.e.*, $n(\mathcal{B}_k) > 2$).

$$\hat{\tau}_{k_1} = \begin{cases} \lfloor n(\mathcal{B}_k)^{-1} \sum_{l=1}^{n(\mathcal{B}_k)} b_{k,l} \rfloor & \text{for } n(\mathcal{B}_k) \leq 2 \\ \min\{\mathcal{B}_k\} & \text{for } n(\mathcal{B}_k) > 2 \end{cases} \tag{3.13}$$

$$\hat{\tau}_{k_2} = \max\{\mathcal{B}_k\} \qquad \text{for } n(\mathcal{B}_k) > 2 \tag{3.14}$$

The complete process to reconcile between the estimated change point sets is given in Procedure B.5, and Algorithm 3.3 presents the full MCCP method for estimation of the change point set $\hat{\boldsymbol{\tau}}$.

Algorithm 3.3 Offline Multiple Conceptor Change Point Method

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; washout length T_{wash}; Type 1 error threshold q; Error control {FWER, FDR}

Outputs: estimated change point set $\hat{\boldsymbol{\tau}}$

1: $\hat{\boldsymbol{\tau}}^f \leftarrow$ result from Algorithm 3.2 with $\mathbf{y}_t$ and Type 1 error threshold $q/2$

2: $\hat{\boldsymbol{\tau}}^f \leftarrow \hat{\boldsymbol{\tau}}^f \cap [T_0 + 1,\ T - T_0]$

3: $\mathbf{y}'_t \leftarrow \mathbf{y}_{T-t+1}$ for $t = 1, \ldots, T$

4: $T - \hat{\boldsymbol{\tau}}^b \leftarrow$ result from Algorithm 3.2 with $\mathbf{y}'_t$ and Type 1 error threshold $q/2$

5: $\hat{\boldsymbol{\tau}}^b \leftarrow \hat{\boldsymbol{\tau}}^b \cap [T_0 + 1,\ T - T_0]$

6: perform Procedure B.5 to obtain $\hat{\boldsymbol{\tau}}$

return estimated change point set $\hat{\boldsymbol{\tau}}$

3.3 Theory

Asymptotic theory for SCCP and MCCP methodology builds on that found in Section 2.3. The hypotheses for online and offline change point detection are tested using the

processes outlined in Section 3.2. Like in Chapter 2, the cosine similarity values S_t are expected to exhibit relatively consistent behavior under the null. Under the alternative wit at least one change point, the reservoir state activity (summarized by S_t) will indicate one or more shifts and deviate from the behavior in the baseline window.

For a multivariate time series $\mathbf{y}_t \in \mathbb{R}^d$, $t = 1, \ldots, T$, let $\gamma^* = T_{\text{wash}} + T_{\text{train}} + 1$ be the minimum spacing between two consecutive change points determined from the specified washout and training lengths of data, placing a constraint on the alternatives in Equations 3.2 and 3.3. Define $\zeta^* = \gamma^*/T$ as the fractional minimum spacing relative to the length time series. The assumption that any change takes place after the washout and training windows restricts the domain to $\tau_j \in [\zeta^* T,\ T_{\max}]$, where $T_{\max} = T - 1$ for online detection and $T_{\max} = (1 - \zeta^*)\, T + 1$ for offline detection.

Define S_t as the sequence of cosine similarity values, and suppose $S_t \sim \mathcal{F}_t(s)$ for all $s \in [0, 1]$, with each $\mathcal{F}_t(s)$ a defined distribution function. Let $\boldsymbol{\tau} = \{\tau_j\}$ be the set of change points in data, producing $\mathcal{F}_{\tau_j}(s_0) \neq \mathcal{F}^i_{\tau_j+1}(s_0)$ for some $s_0 \in [0, 1]$, and further define the empirical distribution function $\hat{\mathcal{F}}(s; a, b) = (b - a)^{-1} \sum_{i=a}^{b} \mathbf{1}\{S_i \leq s\}$.

3.3.1 Null Hypothesis

Define the preceding change point in a sequence $\tau^+_{j-1} = \max\{0, \tau_{j-1}\}$, and examine time $t \in \mathcal{T}_j$ in the potential change point domain, where $\mathcal{T}_j = \left[\tau^+_{j-1} + \zeta^* T,\ (1 - \zeta^*)\, T + 1\right]$. Label $\boldsymbol{\zeta}_j = \left[\tau^+_{j-1} + \zeta^* T,\ t + \zeta^* T - 1\right)$ as the domain defined by selection of t.

Under the null hypothesis, the sequence S_z, $z \in \boldsymbol{\zeta}_j$, is stationary by construction, and denote the common distribution function $\mathcal{F}(s) \equiv \mathcal{F}_{\tau^+_{j-1}+\zeta^* T}(s) = \cdots =$

$\mathcal{F}_{t+\zeta^* T-1}(s)$. Define the function $q : [0,1] \to (0,1)$ for some small $\kappa > 0$ by

$$q(\delta) = \max\{\delta^{1/2}(1-\delta)^{1/2}, \kappa\} \tag{3.15}$$

$$\text{where } \delta = \frac{z - \tau_{j-1}^{+} - \zeta^* T + 1}{z - \tau_{j-1}^{+}}, \tag{3.16}$$

for $z \in \boldsymbol{\zeta}_j$ resembling Equation 3.7 with $\nu = 1/2$, and the statistic

$$K(z) = \frac{\delta\,(1-\delta)}{q\,(\delta)} \quad \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}(s; \tau_{j-1}^{+} + \zeta^* T, z) - \hat{\mathcal{F}}(s; z+1, z+\zeta^* T - 1) \right| \tag{3.17}$$

akin to Equation 3.6 in Section 3.2. Let $\mathcal{K}(s,\delta)$ be a Gaussian process with

$$\begin{aligned} &\mathbb{E}\left[\mathcal{K}(s,\delta)\right] = 0, \\ &\mathbb{E}\left[\mathcal{K}(s,\delta)\,\mathcal{K}(s',\delta')\right] = (\delta \wedge \delta')\,\Gamma(s,s'), \\ \text{and} \quad &\Gamma(s,s') = \sum_{-\infty<z<\infty} \mathbb{E}\left[S_{\tau_{j-1}^{+}+\zeta^* T}(s) S_z(s')\right]. \end{aligned} \tag{3.18}$$

Theorem 3.1. *For some $t \in \mathcal{T}_j$ and a fixed minimum fractional spacing $\zeta^* \in \left(0, \frac{1}{2}\right)$, suppose $\mathcal{F}(s)$ is Lipschitz continuous of order $C > 0$ and the sequence S_z, $z \in \boldsymbol{\zeta}_j$, is $\mathcal{S}$-mixing, where condition (1) of Definition 2.1 holds with $\gamma_m = m^{-AC}$, $\delta_m = m^{-A}$ for some $A > 4$. Under the null hypothesis, for any $\kappa \in \left(0, \frac{1}{2}\right)$ as $T \to \infty$,*

$$\sqrt{t - \tau_{j-1}^{+}} \max_{z\in\boldsymbol{\zeta}_j} K(z) \xrightarrow{D} \sup_{\delta\in[0,1]} \sup_{s\in[0,1]} \frac{|\mathcal{K}(s,\delta)|}{q(\delta)} \tag{3.19}$$

and $\max_{z\in\boldsymbol{\zeta}_j} K(z) = o(1)$.

Under mild assumptions, Berkes, Hörmann, and Schauer (2009) show a $\mathcal{S}$-mixing sequence can be represented as a shift process of i.i.d. random variables. Section

2.3.1 briefly discusses the intuition of $\mathcal{S}$-mixing and construction of the approximating sequence. Like in Chapter 2, these assumptions only apply to S_t and are not imposed on the original dataset $\mathbf{y}_t$. The proof of Theorem 3.1, shown in Appendix B follows the result of Csörgő and Horváth (1997), Theorem 2.6.1, and is similar to that of Theorem 2.2.

In offline detection, Theorem 3.1 holds for the full domain of potential change points $\tau_j \in [\zeta^* T,\, T_{\max}]$. As $t \to T$ in closed-ended online detection, the length of the final window of the sequence becomes finite, and convergence is not realized for $t > (1 - \zeta^*)\, T + 1$.

The minimum fractional spacing parameter ζ^* effectively controls the number of potential identified change points in a dataset. As the parameter approaches the upper limit $\zeta^* \to 1/2$, the domain of potential change points is compressed toward the center of the data. When $\zeta^* \to 0$ such that $\zeta^* T = \mathcal{O}(1)$ the number of potential change points becomes infinitely large and the span of the domain $\boldsymbol{\zeta}_j$ is finite, so convergence in distribution is not attained. Smaller $\zeta^* > 0$ allows for flexibility of the methodology (as long as it is sufficiently large to produce a well specified training window) and leads to stationarity of the resulting sequence.

3.3.2 Alternative Hypothesis

Consider the online multiple change point problem and the hypothesis in Equation 3.2 to test for change point τ_j. As in Section 3.3.1, define the preceding change point as $\tau_{j-1}^{+} = \max\{0, \tau_{j-1}\}$. Further define the next point $\tau_{j+1}^{T} = \min\{\tau_{j+1}, T\}$, and examine time $t \in \mathcal{T}_j$, where $\mathcal{T}_j = \left[\tau_{j-1}^{+} + \zeta^* T,\; \tau_{j+1}^{T} - \zeta^* T + 1\right]$. Note the spacing requirement of neighboring change points asserts $\tau_{j+1}^{T} - \tau_{j-1}^{+} \geq 2\zeta^* T$. Label $\boldsymbol{\zeta}_j =$

$\left[\tau_{j-1}^{+} + \zeta^{*}T,\ t + \zeta^{*}T - 1\right)$ as the domain defined by selection of t.

By construction of the alternative S_z, $z \in \boldsymbol{\zeta}_j$, divides into two stationary ergodic pieces on either side of τ_j for some $t \in \mathcal{T}_j$, with $\mathcal{F}_{\tau_j}(s_0) \neq \mathcal{F}_{\tau_j+1}(s_0)$. Let $\hat{\tau}_j^f$ be the forward SCCP estimate, where

$$\hat{\tau}_j^f = \arg\max_{z \in \boldsymbol{\zeta}_j^f} K(z). \tag{3.20}$$

Theorem 3.2. *For some $t \in \mathcal{T}_j$ and a fixed minimum fractional spacing $\zeta^{*} \in \left(0, \frac{1}{2}\right)$, suppose the sequence S_z, $z \in \boldsymbol{\zeta}_j$, divides into two stationary ergodic pieces on either side of a change point τ_j. Then, for any $\kappa \in \left(0, \frac{1}{2}\right)$ as $T \to \infty$, $\hat{\tau}_j^f \xrightarrow{P} \tau_j$ provided $\tau_j \in (\boldsymbol{\zeta}_j \cap \boldsymbol{\Delta}_j)$, where*

$$\boldsymbol{\Delta}_j - \zeta^{*}T = \left[\frac{t + \tau_{j-1}^{+}}{2} - \frac{t - \tau_{j-1}^{+}}{2}\sqrt{1 - 4\kappa^2},\ \frac{t + \tau_{j-1}^{+}}{2} + \frac{t - \tau_{j-1}^{+}}{2}\sqrt{1 - 4\kappa^2}\right]. \tag{3.21}$$

The statement is similar to Theorem 2.3 in Chapter 2 and the proof follows the outline of Theorem 2.1 from Newey and McFadden (1994). The domain restriction in Equation 3.21 does not narrow the window of allowable change points in practice if $\kappa < \sqrt{\frac{1}{4} - \frac{1}{4}\left(1 - \frac{2}{t - \tau_{j-1}^{+}}\right)^2}$.

In the offline multiple change point problem, rejecting the null hypothesis in Equation 3.3 produces the estimate

$$\hat{\tau}_j = \left\lfloor \left(\hat{\tau}_j^f + \hat{\tau}_j^b\right)/2 \right\rfloor, \tag{3.22}$$

where $\hat{\tau}_j^b$ is obtained from a backward-looking algorithm run. Define $\mathbf{y}_t' = \mathbf{y}_{T-t+1}$, the corresponding cosine similarity sequence S_z', $z \in \boldsymbol{\zeta}_j'$, for $\boldsymbol{\zeta}_j' = \left(t - \zeta^{*}T + 1,\ \tau_{j+1}^{T} - \zeta^{*}T\right]$,

and domain restriction $\tau_j \in \left(\boldsymbol{\zeta}'_j \cap \boldsymbol{\Delta}'_j\right)$, where

$$\boldsymbol{\Delta}'_j + \zeta^* T = \left[\frac{t+\tau_{j+1}^T}{2} - \frac{\tau_{j+1}^T - t}{2}\sqrt{1-4\kappa^2},\ \frac{t+\tau_{j+1}^T}{2} + \frac{\tau_{j+1}^T - t}{2}\sqrt{1-4\kappa^2}\right]. \quad (3.23)$$

The result of Theorem 3.2 implies that $\hat{\tau}_j^b \xrightarrow{P} \tau_j$. Thus, the estimate $\hat{\tau}_j$, defined in Equation 3.22, is consistent for the change point τ_j.

Corollary 3.3. *For some $t \in \mathcal{T}_j$ and a fixed minimum fractional spacing $\zeta^* \in \left(0, \frac{1}{2}\right)$, suppose the sequences S_z, $z \in \boldsymbol{\zeta}_j$, and S'_z, $z \in \boldsymbol{\zeta}'_j$, divide into two stationary ergodic pieces on either side of a change point τ_j. Then, for any $\kappa \in \left(0, \frac{1}{2}\right)$ as $T \to \infty$, $\hat{\tau}_j \xrightarrow{P} \tau_j$ provided $\tau_j \in \left[(\boldsymbol{\zeta}_j \cap \boldsymbol{\Delta}_j) \cap \left(\boldsymbol{\zeta}'_j \cap \boldsymbol{\Delta}'_j\right)\right]$.*

As above, the condition does not restrict the domain in practice if the chosen κ is small.

3.3.3 Type 1 Error Control

The sequential testing procedure adapted from Holm (1979), with cutoff values shown in Equation 3.9 for online detection and Equation 3.11 for offline detection, addresses the multiplicity problem and conservatively controls the FWER at a defined Type 1 error threshold.

Proposition 3.4. *Suppose H_i, $i = 1, \ldots, m$, are null hypotheses for testing in a sequential framework, where a decision must be rendered on H_{i-1} prior to testing H_i. Let p_i be the associated p-value for hypothesis H_i, and k_i the number of rejected hypotheses prior to H_i in the set H_1, ..., H_{i-1}. A sequential procedure that rejects H_i for $p_i \leq \alpha/(m-k_i)$ ensures the FWER is at most α in the strong sense.*

The threshold for the first hypothesis test is identical to the Bonferroni correction; after rejection of at least one hypothesis, the sequential procedure adapted from Holm (1979) is uniformly more powerful than Bonferroni.

3.4 Simulation Study

Performance of SCCP and MCCP is investigated via simulation of the multiple change point problem. For online detection, SCCP is compared to the Scan-B method of S. Li et al. (2019) (ScanB) and the kernel-based CUSUM method of Wei and Y. Xie (2023) (kCUSUM). Comparator methods for MCCP in the offline problem include the e-divisive method of Matteson and James (2014) (EDiv), the kernel change point algorithm of Arlot, Celisse, and Harchaoui (2019) (KCP), Type 1 and 2 sparsified binary segmentation methods of Cho and Fryzlewicz (2015) (SBS1/2), and the NP-MOJO method of McGonigle and Cho (2023).

Methods are evaluated by examination of the true positive rate (TPR) (also referred to as the "sensitivity"), the positive predictive value (PPV) (the "precision" or $1 - \text{FDR}$), and the Matthews correlation coefficient (MCC). Larger values of all metrics indicate better performance, with $0 \leq \text{TPR, PPV, MCC} \leq 1$. A method is determined to have correctly located a change point if placed within a radius of the reservoir washout length $r_{\gamma^*} = T_{\text{wash}}$, avoiding overlap between adjacent regions.

$$\text{TPR} = \frac{\sum_{j=1}^{n(\boldsymbol{\tau})} \mathbf{1}\left\{|\hat{\tau}_j - \tau_j| < r_{\gamma^*}\right\}}{n\left(\boldsymbol{\tau}\right)} \tag{3.24}$$

$$\text{PPV} = \frac{\sum_{j=1}^{n(\boldsymbol{\tau})} \mathbf{1}\left\{|\hat{\tau}_j - \tau_j| < r_{\gamma^*}\right\}}{n\left(\hat{\boldsymbol{\tau}}\right)} \tag{3.25}$$

The TPR represents the ability to correctly identify true change points, and the PPV

indicates propensity to limit the number of false discoveries. The MCC summarizes several binary classification metrics into a single value,

$$\begin{aligned} \text{MCC} = &\sqrt{\text{TPR} \times \text{PPV} \times (1 - \text{FPR}) \times (1 - \text{FOR})} \\ &- \sqrt{(1 - \text{TPR}) \times \text{FDR} \times \text{FPR} \times \text{FOR}}, \end{aligned} \tag{3.26}$$

where FPR is the false positive rate and FOR the false omission rate (Matthews, 1975).

3.4.1 Simulated Data

The goal of this study is to generate processes that retain consistent mean and variance structures after change points, testing the ability of the methods to capture information about the dependence structure. All simulated time series $\mathbf{y}_t \in \mathbb{R}^2$, $t = 1, \ldots, T$, have length $T = 800$. Define $\boldsymbol{\tau}$ the set of change points, where $n(\boldsymbol{\tau}) \in \{0, 1, 2\}$, and generate $\mathcal{S} = 100$ simulated data sets for each setting. Change points are randomly selected from the interval $\tau_j \in [161, 640]$ such that two consecutive have a minimum spacing of $\gamma^* = 161$.

Gaussian Process Model

The first section of data follows a Gaussian process with a periodic covariance function shown in Equation 3.27, based on the separation of two time points $t - t'$ and a parameter f that determines the periodicity.

$$\text{Cov}\,(t, t') = \exp\left\{-32 \sin^2\left(\frac{t - t'}{f}\right)\right\} \tag{3.27}$$

The covariates in the time series $\mathbf{y}_t$ are joint random realizations from the Gaussian process model with a randomly specified correlation $\rho \sim \mathcal{U}(-0.8, 0.8)$ and periodicity parameter $f \sim \mathcal{U}(10, 30)$. Each covariate in the data is scaled to mean zero and unit variance for consistency before and after change points. After encountering a change, the covariates will negate, altering the process dependence while retaining the first and second-order structure.

Threshold Autoregressive Model

The second half of the data follows a self-exciting TAR(2) model that allows for regime changes in the autoregressive parameters based on a threshold variable(Tong and Lim, 1980).

$$\mathbf{y}_t = \mathbf{A}_s^{(k)} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \end{bmatrix} + \boldsymbol{\varepsilon}_t \tag{3.28}$$

The coefficient matrix $\mathbf{A}_s^{(k)} \in \mathbb{R}^{2\times 4}$ depends on the state of the data s and the regime k. The data takes two states and shifts back and forth after a change point is encountered. Each state has two regimes, where $k = 1$ corresponds to the case where $\sum_{i=1}^{2}\sum_{j=1}^{4} y_{i,t-j} < 0$, and $k = 2$ the opposing inequality, adding dependence up to $t-4$. The error term $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I})$. Each coefficient matrix is selected (from several in a random generation) to ensure constant absolute eigenvalues (within a small tolerance) over all potential states and regimes. Basic structures of the coefficient matrices are given below, where each $a_{s(i,j)}^k \sim \mathcal{U}(-0.8, 0.8)$ is shrunk to zero if $\left|a_{s(i,j)}^k\right| < 0.2$. The absolute eigenvalues for each simulated dataset are randomly determined from $\mathbf{A}_1^{(1)}$

and restricted to nonzero values that produce stable time series realizations.

$$\mathbf{A}_1^{(1)} = \begin{bmatrix} 0 & a_{1(1,2)}^{(1)} & 0 & a_{1(1,4)}^{(1)} \\ a_{1(2,1)}^{(1)} & a_{1(2,2)}^{(1)} & a_{1(2,3)}^{(1)} & a_{1(2,4)}^{(1)} \end{bmatrix}$$

$$\mathbf{A}_1^{(2)} = \begin{bmatrix} a_{1(1,1)}^{(2)} & a_{1(1,2)}^{(2)} & a_{1(1,3)}^{(2)} & a_{1(1,4)}^{(2)} \\ a_{1(2,1)}^{(2)} & 0 & a_{1(2,3)}^{(2)} & 0 \end{bmatrix} \tag{3.29}$$

$$\mathbf{A}_2^{(1)} = \begin{bmatrix} a_{2(1,1)}^{(1)} & a_{2(1,2)}^{(1)} & a_{2(1,3)}^{(1)} & a_{2(1,4)}^{(1)} \\ 0 & a_{2(2,2)}^{(1)} & 0 & a_{2(2,4)}^{(1)} \end{bmatrix}$$

$$\mathbf{A}_2^{(2)} = \begin{bmatrix} a_{2(1,1)}^{(2)} & 0 & a_{2(1,3)}^{(2)} & 0 \\ a_{2(2,1)}^{(2)} & a_{2(2,2)}^{(2)} & a_{2(2,3)}^{(2)} & a_{2(2,4)}^{(2)} \end{bmatrix} \tag{3.30}$$

3.4.2 Simulation Settings

The SCCP and MCCP methods use a washout length of $T_{\text{wash}} = 40$ and a training length $T_{\text{train}} = 120$ such that potential change points are identified in the interval $\hat{\tau}_j \in [161, 640]$. These values set the minimum sufficient spacing for detection of two consecutive change points at $\gamma^* = 161$. For consistency, comparator methods are restricted to identifying changes in the same domain. Methods that do not accept a parameter for the minimum spacing reconcile the final set with the process outlined in Section 3.2.2 and Procedure B.5.

Ranges of parameters are chosen to investigate the balance between sensitivity and precision of each method. Many methods directly accept a Type 1 error parameter q; KCP takes an analogous penalty scaling parameter C and the online comparator

methods employ the average run length (ARL) to control the rate of false discovery, where larger values of each encourage more stringent error control. While SCCP and MCCP (along with other methods) establish a theoretical upper bound for the FWER, observed Type 1 error often deviates from this bound, and it can behave much like the detached "scaling" parameter in KCP (Arlot, Celisse, and Harchaoui, 2019).

The KCP method requires specification of the maximum number of allowable changes, and it is granted a slight oracle advantage by defining $\max\{n(\hat{\boldsymbol{\tau}})\} = 2$ after accounting for the reconciliation process. NP-MOJO is instructed to look for changes within a relevant window of lagged time points, $t, \ldots, t-2$. Table B.1 summarizes the settings for each method used in the simulation study.

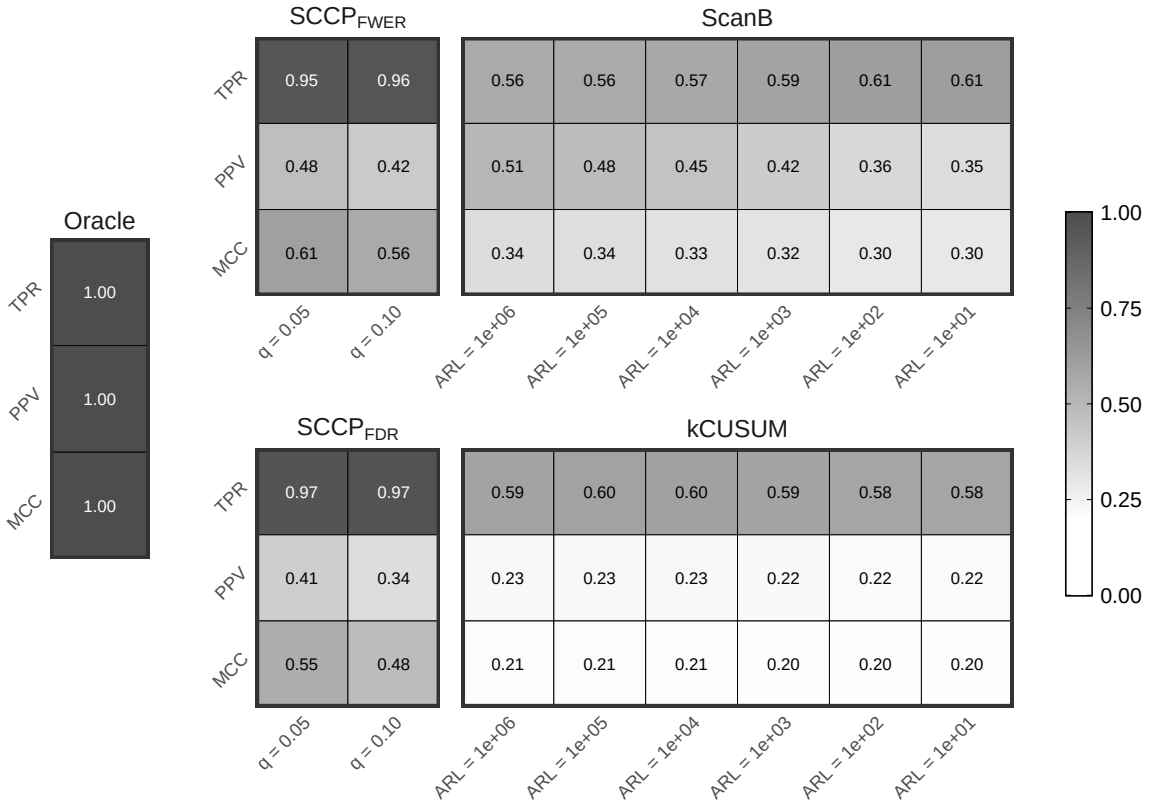
3.4.3 Simulation Results

Displays of performance for online change point detection in the simulated data are presented in Figures 3.1 and 3.2, and in Figures 3.3 and 3.4 for offline change point detection. Additional figures for more specific settings are relegated to Appendix B.

Online Simulation Results

SCCP outperforms ScanB and kCUSUM in both the Gaussian process and TAR simulations, but performance lags significantly behind an oracle method.

In the online Gaussian process simulations, SCCP boasts a high TPR, identifying a large fraction of the true change points present in the simulated data, but the theoretical Type 1 error threshold does not cover the true error for either FWER or FDR detection schemes. SCCP struggles with the TAR process, but still retains an



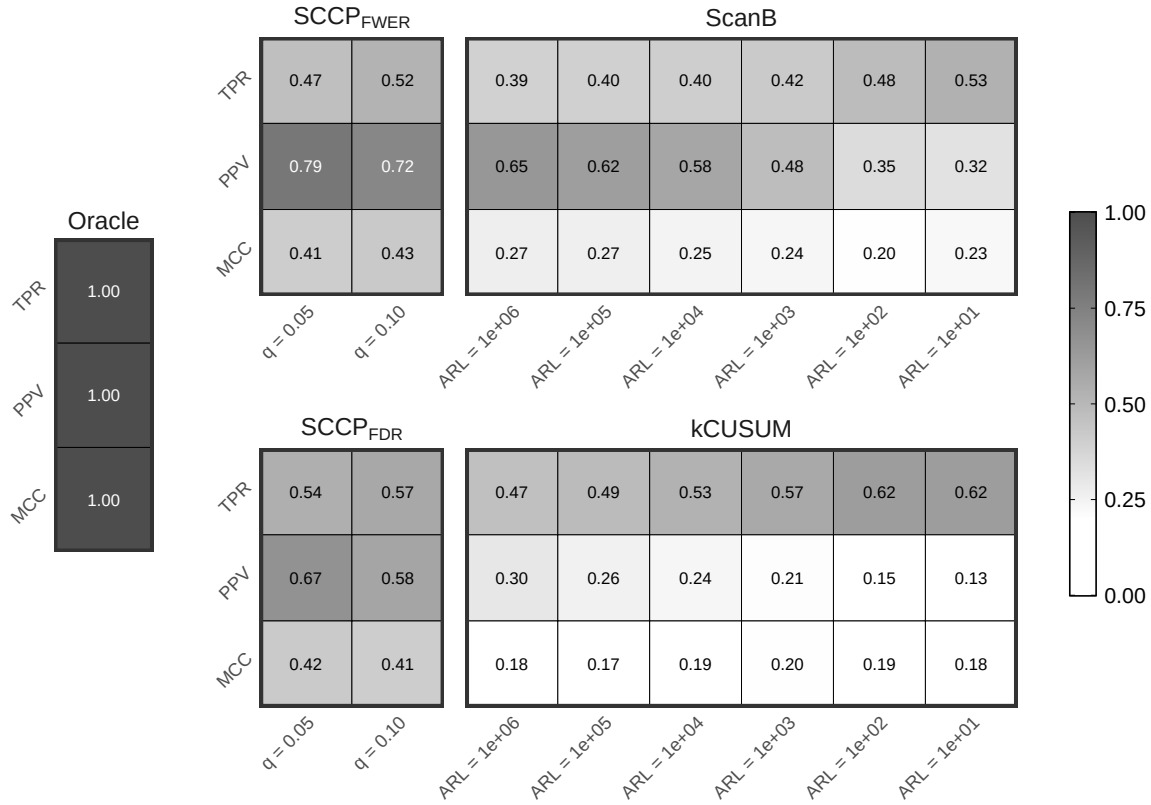
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure 3.1: Gaussian process online detection simulation results.

edge on the online comparators. The ScanB and kCUSUM methods show similar detection ability, with ScanB demonstrating a propensity for slightly better error control. Figures depicting results for each set of simulations $n(\boldsymbol{\tau}) = \{0, 1, 2\}$ are shown in Appendix B.

Offline Simulation Results

MCCP demonstrates relatively strong performance in the offline Gaussian process simulations. Both error control regimes still show sizeable undercoverage, but the methods correctly identify nearly all change points within the radius of tolerance,

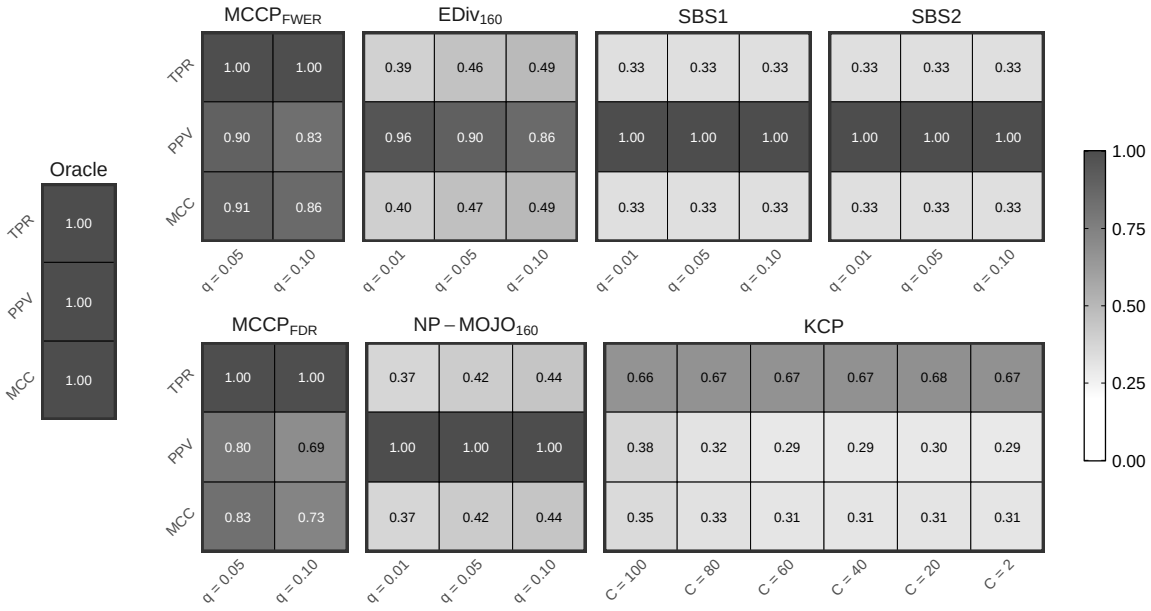


Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure 3.2: Threshold autoregressive process online detection simulation results.

r_{γ^*}. The SBS methods return empty change point sets for all simulations (with one-third of those correctly identified as $n(\boldsymbol{\tau}) = 0$). KCP is the closest competitor for the TPR metric, but error control lags well behind MCCP.

For offline detection in the simulated TAR data, MCCP struggles in locating the true change points. While marginally outperforming some comparator methods that assume independence (e.g., EDiv), MCCP methods are surpassed by NP-MOJO in terms of the TPR, and consequently the MCC. Error control for all methods that show decent performance exhibit slight undercoverage relative to the specified theoretical Type 1 error rate.



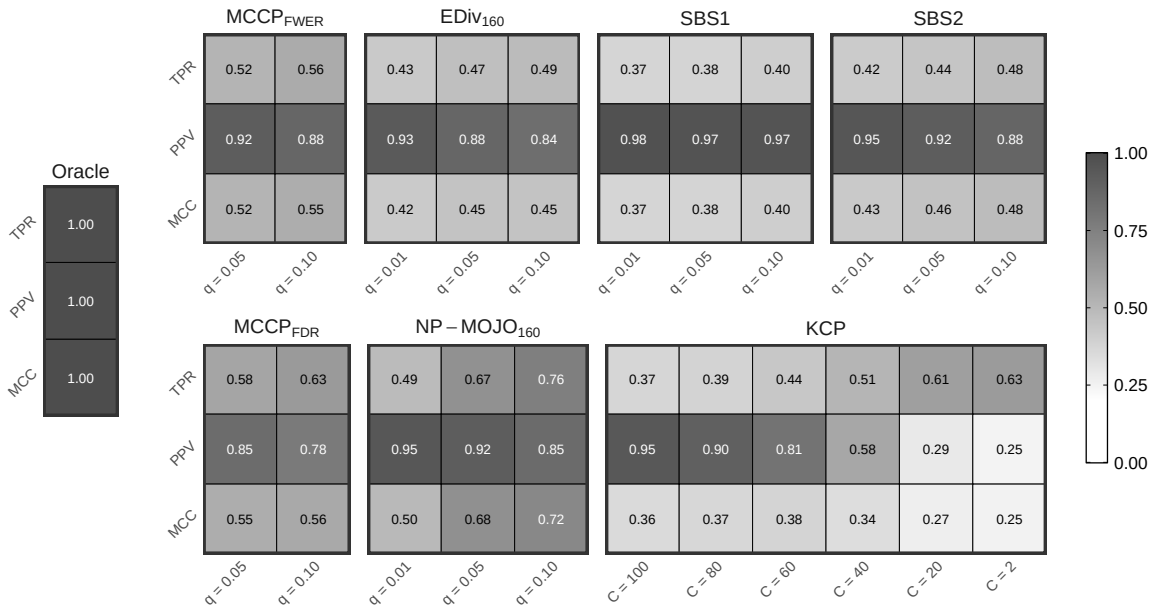
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure 3.3: Gaussian process offline detection simulation results.

The figures illustrate the power and Type 1 error trade-off for all change point methods, with a clear picture shown in the KCP results of Figure 3.4. Balance between these competing objectives inches a method closer to the non-oracle ideal method where TPR $\to 1$ and PPV $= 1 - q$. Figures depicting offline results for each $n(\boldsymbol{\tau}) = \{0, 1, 2\}$ are in Appendix B.

3.5 Discussion

For the online comparator methods, ScanB and kCUSUM are designed for change point identification problems where there is access to a long window of training data (reference section) prior to seeking changes in a shorter sequence of interest (S. Li et al., 2019; Wei and Y. Xie, 2023). In the simulation study, as in many applications



Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure 3.4: Threshold autoregressive process offline detection simulation results.

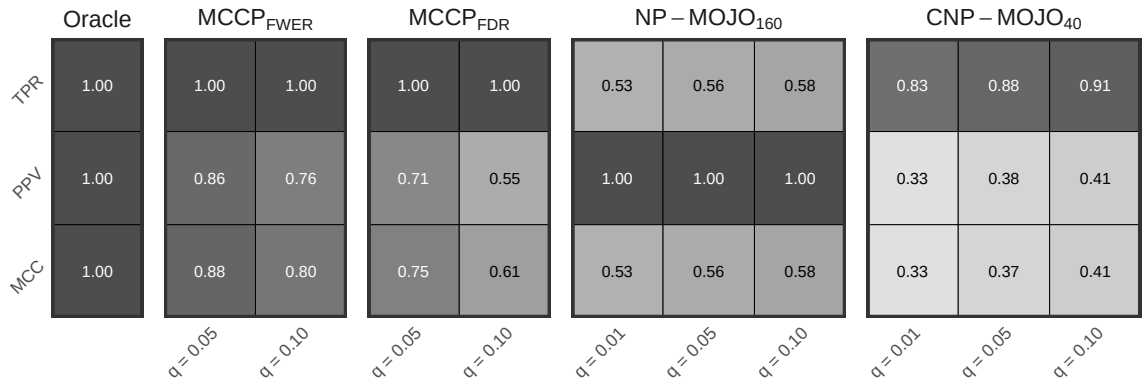
where spending this length of data on training is not feasible, methods were not granted this long running reference window and some lose strength as a result.

Offline comparator methods that assume independence of the observations are, as expected, less effective in the presence of temporal dependence. The contrast between the relative success in scope of application between the MCCP and NP-MOJO methods may lie in the elemental construction of the procedures. NP-MOJO seeks information directly from the lagged covarite structure, like VAR or TAR models, and should excel in these situations. In non-stationary (or cyclostationary data), the methodology may break down and fail (McGonigle and Cho, 2023). MCCP encodes information about interactions between network states with a conceptor matrix, and aggregates them over a given time window (the training data). This transforms and captures changes in cyclostationary data, but may miss key short-term information

lumped together in the aggregation process. The contrived TAR data in Section 3.4 retains a constant mean and variance structure with nearly identical absolute eigenvalues of the autoregressive coefficient matrices. Change points, then, must be sought from duration of time spent between regimes within a state, or from global information about loadings onto certain network states when a coefficient changes. This may have more to do with the specific featurization process, and perhaps is only sporadically captured by a random functional transformation and rotation.

To illustrate this point, the NP-MOJO method was performed on the simulated data in $n(\boldsymbol{\tau}) = \{0, 1\}$ (post transformation to a sequence of cosine similarities) by the CCP method of Chapter 2 (referred to by CNP-MOJO in Figures 3.5 and 3.6). Ignoring the propensity for false discovery and examining the TPR, the combined method outperforms NP-MOJO for Gaussian process data and lags behind in the TAR data suggesting information is lost in the featurization and aggregation process and not the back-end change point identification algorithm. In aggregate, the combined method performs worse than either of the two individual methods. Examination of $n(\boldsymbol{\tau}) = 2$ was not considered due to the methodological inconsistencies in the multiple change point problem. Individual figures for each class of simulated data and $n(\boldsymbol{\tau}) = \{0, 1\}$ are shown in Appendix B.

Comparing the online and offline simulation results from Section 3.4.3 demonstrates the advantage of access to the complete time series rather than just a brief window near a time point of interest.



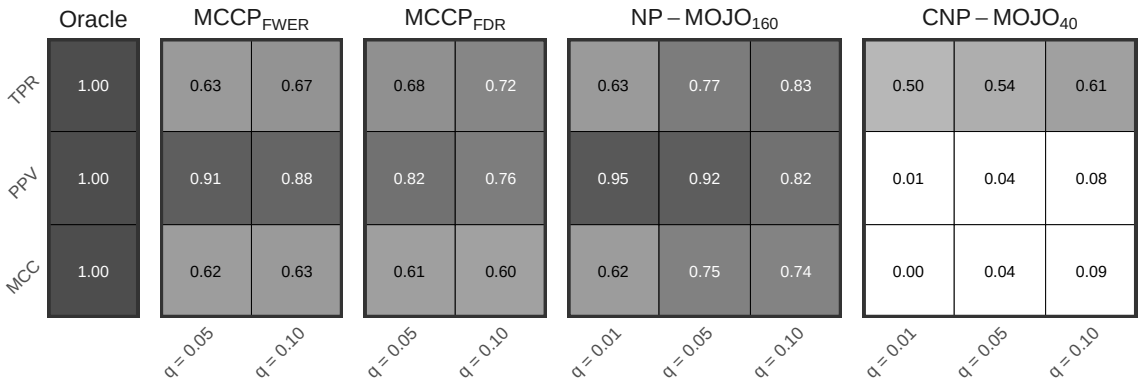
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

Figure 3.5: Gaussian process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2.

3.5.1 Naive Variance Change Simulations

MCCP assumes a minimum spacing γ^* between two consecutive change points. Similar assumptions are imposed by many methods in literature, like the binary segmentation of Cho and Fryzlewicz (2012) and the window selection of EDiv and NP-MOJO (Matteson and James, 2014; McGonigle and Cho, 2023). To investigate the ability of MCCP to detect change points near (or approaching) the minimum spacing, and the behavior of the method under violation of this assumption, naive variance change simulations were performed. Simulated white-noise datasets of length $T = 800$ encounter a change point and switch from $\sigma_t = 1$ to $\sigma_t = 0.2$. Five scenarios explore the performance under these settings, with the spacing varying from well under the minimum $\gamma = 80$, to stepping slowly away from the minimum $\gamma = \{161, 171, 181, 191\}$, where $\gamma^* = 161$.

MCCP becomes more powerful as the spacing between consecutive points gets



Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

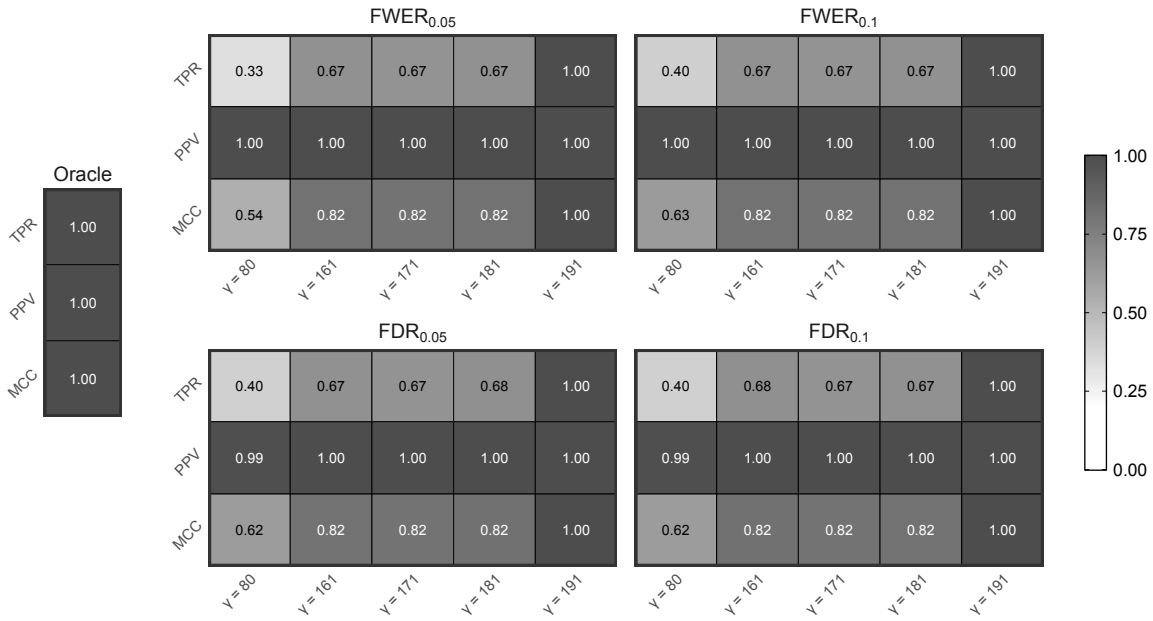
Figure 3.6: Threshold autoregressive process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2.

further away from the assumed minimum. For spacing in $\gamma = \{80, 161, 171, 181\}$, MCCP consistently identifies the first and last shifts in a time series, but the inner changes are lost as "regions of transition" due to their proximity.

3.6 Future Work

Inconsistent performance of SCCP and MCCP likely stems from the process of aggregating information over several time points and random featurizations. Major simplification of the methods may alleviate this challenge, allow for increased interpretability, and decrease the computational burden.

Suppose $\mathbf{y}_t \in \mathbb{R}^d$, with $t = 1, \ldots, T$, is a time series of interest in the offline multiple change point problem. With $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_T]^\top \in \mathbb{R}^{T \times d}$ the concatenated matrix of temporal observations, examine the singular value decomposition $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{T \times T}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$ are the orthonormal bases of the temporal and variable



Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Label refer to the error control procedure of the MCCP method.

Figure 3.7: MCCP performance for naive variance change simulations as the spacing between consecutive change points γ fluctuates.

axes, respectively. The matrix $\mathbf{V}$ captures information about the covariate directions of linear variability in the original dataset.

As in Equation 1.13, post transformation from an artificial network $\mathcal{N}$ of dimension N that generates a faithful representation of the data, analogously write the matrix of network state vectors $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is an orthonormal basis of the states. The conceptor-based methodology here and in Chapter 2 examines properties based on $\mathbf{V}$, with $\mathbf{H}_{\text{train}}^\top \mathbf{H}_{\text{train}} = \mathbf{V}_{\text{train}} \mathbf{\Sigma}_{\text{train}}^\top \mathbf{\Sigma}_{\text{train}} \mathbf{V}_{\text{train}}^\top$ and

$$\mathbf{C} = \mathbf{V}_{\text{train}} \mathbf{\Sigma}_{\text{train}}^\top \mathbf{\Sigma}_{\text{train}} \mathbf{V}_{\text{train}}^\top \left(\mathbf{V}_{\text{train}} \mathbf{\Sigma}_{\text{train}}^\top \mathbf{\Sigma}_{\text{train}} \mathbf{V}_{\text{train}}^\top + \frac{T_{\text{train}}}{\alpha^2} \mathbf{I} \right)^{-1}. \tag{3.31}$$

The conceptor matrix encodes how the interactions of a network state relate to those in the training window of data. Similar information may be captured through exam-

ination of the loadings on the orthonormal vectors composing $\mathbf{V}$,

$$\mathbf{\Lambda} = \mathbf{HV} = \mathbf{U\Sigma}. \tag{3.32}$$

A row $\boldsymbol{\lambda}_t$ in the matrix $\mathbf{\Lambda} \in \mathbb{R}^{T \times N}$ encodes a measure of the quantity of information contributed by a direction of the orthonormal basis of the network states. With N sequences of information, change points can be identified from shifts in the loadings on this "linearized" orthonormal basis (via the universal function approximator transformation). In practice, N should likely be restricted so that the relevant information is not spread thin over a large number of dimensions.

This simplified approach eliminates of the training window of data, provides flexibility for changes sought, shrinks the required window between consecutive changes, and permits qualitative statements about the nature of regimes in a dataset (like temporal clustering based on the loading vectors). The potential methodology will require careful choice of ANN (featurization process) such that dependence is captured without fabricating temporal relationships not present in the original data (through various MLP, RNN, and LSTM architectures). The approach would lose applicability to the online change point problem as the full vector sequence of data has not been realized for generation of the orthonormal basis matrix. The conceptor matrix captures periodic structure in the data from the time domain, and steps to retain this capability should be emphasized.

In the ANN featurization process, all covariate specific information is lost and changes are sought from the joint distribution of the covariate set. If a change takes place in sequence $y_{t,i}$, and not $y_{t,j}$, $i \neq j$, the conceptor-based methods here and in Chapter 2 can only mark a global change point; they provide zero intuition on

the covariate(s) involved or the nature of the change. Introducing architectures of featurization that retain covariate specific information, while allowing for the representative transformation harnessing the universal approximator ability of ANNs, will add noteworthy benefits in the interpretability of these methods.

Chapter 4

Nonlinear Permuted Granger Causality

This chapter is adapted from the article *Nonlinear Permuted Granger Causality* (Gade and Rodu, 2023b).

Granger causal inference investigates the ability of a time series $\mathbf{x}_t \in \mathbb{R}^p$, $t = 1, \ldots, T$, to predict future values of a response $\mathbf{y}_t \in \mathbb{R}^d$ (Wiener, 1956; Granger, 1969). The effect is traditionally measured through the variance of residuals in restricted and unrestricted models, as shown in Definition 4.1, where $\mathcal{P}$ represents the optimal prediction function, $\mathcal{I}_{<t}$ is all information prior to time t, and $\mathbf{X}_{<t}$ is a matrix of compiled values of $\mathbf{x}_t$ prior to time t.

Definition 4.1. Time series $\mathbf{x}_t \in \mathbb{R}^p$ is Granger causal for $\mathbf{y}_t \in \mathbb{R}^d$ if

$$\text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t | \mathcal{I}_{<t}\right)\right] < \text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t | \mathcal{I}_{<t} \backslash \mathbf{X}_{<t}\right)\right]. \tag{4.1}$$

Modern methods for adapting Granger causality to nonlinear functional relationships leverage deep learning and representation learning for capturing dependence between variables (Moodie and Stephens, 2022). These tools, when paired with other machine learning techniques, are not necessarily reliable or precise. Penalized vari-

able selection as a screening method potentially removes signal by not considering the collective covariate set of the system and tends to up-weight contributions of the chosen nonzero covariates. Erroneous conclusions can result from deep learning because variable specific inference is muddled. This work explicitly redefines Granger causality in terms of a permuted framework with out-of-sample testing (NPGC) that retains the flexibility of representation learning and has specific advantages when seeking nonlinear functional connections.

4.1 Granger Causality

The form of Granger causality presented in Definition 4.1 is inherently conditional on additional information included in the modeling process. The optimal prediction $\mathcal{P}\left(\mathbf{y}_t|\mathcal{I}_{<t}\right)$ is unattainable in practice, and the notion of Granger causality is a conditional model on some included explanatory covariate set $\mathbf{z}_t \in \mathbb{R}^q$ and the history of the response prior to time t, $\mathbf{Y}_{<t}$, as in Definition 4.2.

Definition 4.2. Time series $\mathbf{x}_t \in \mathbb{R}^p$ is conditionally Granger causal for $\mathbf{y}_t \in \mathbb{R}^d$ given $\mathbf{z}_t \in \mathbb{R}^q$ and the relevant history of the response if

$$\text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t|\mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \mathbf{X}_{<t}\right)\right] < \text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t|\mathbf{Y}_{<t}, \mathbf{Z}_{<t}\right)\right]. \tag{4.2}$$

Model selection is implicit in the test for presence of Granger causality, and the inferential conclusion is coupled with a written form (Friston, Moran, and Seth, 2013). Inclusion of additional variables strengthens the condition for establishing Granger causality; rejection of the null implies the covariate set $\mathbf{x}_t$ is found to provide unique

and useful information for prediction of $\mathbf{y}_t$ beyond that contained in both $\mathbf{z}_t$ and the lagged response (Granger, 1980). The basis for Granger causality requires fulfillment of several conditions including a sufficient length of continuous-valued stationary data, exact and complete specification of the model, error-free observation of the variables, and a sampling frequency on a regular discrete grid that contains the known lag relationship (Granger, 1969; Granger, 1980; Granger, 1988). In the form of Definition 4.2, the second condition can be relaxed provided that inference is accordingly narrowed to the conditional statement.

The framework can also be defined in terms of non-causality as a statement of conditional independence, where inclusion of additional variables $\mathbf{z}_t$ is a simple extension (Granger, 1980; Florens and Mouchart, 1982).

Definition 4.3. Time series $\mathbf{x}_t \in \mathbb{R}^p$ does not Granger cause $\mathbf{y}_t \in \mathbb{R}^d$ if and only if

$$\mathbf{Y}_{<t+1} \perp \mathbf{X}_{<t} \text{ given } \mathbf{Y}_{<t}. \tag{4.3}$$

This statement is perhaps more powerful because it is defined in terms of the distributions of the variables, allowing for extension to several other forms of statistical tests. It may be prone to misuse if interpreted to place the burden of proof on establishing independence. In a definition from Section 4 of Shojaie and Fox (2022), column j in time series $\mathbf{x}_t$ is Granger non-causal for time series $\mathbf{y}_t$ if and only if $\forall t$,

$$\mathcal{P}\left(\mathbf{y}_t|\mathbf{x}_{<t1}, \ldots, \mathbf{x}_{<tj}, \ldots, \mathbf{x}_{<tp}\right) = \mathcal{P}\left(\mathbf{y}_t|\mathbf{x}_{<t1}, \ldots, \mathbf{x}_{<t(j-1)}, \mathbf{x}_{<t(j+1)}, \ldots, \mathbf{x}_{<tp}\right) \tag{4.4}$$

that implies the equality of these predictions holds for *all* time points t. If misin-

terpreted to mean finding correlation at *one* time point in series $\mathbf{y}_t$ is enough to claim functional dependence of two time series, the presence of a causal connection effectively becomes the null hypothesis and Granger causality is reduced to an exceptionally weak statement.

Even when the question is framed with the onus on demonstrating a functional relationship, all Granger causal methods overreach their scope of reliable application when inference is performed on the individual variables included in the covariate set rather than on their collective behavior. Applications of the framework to interpreting individual model coefficients introduce a hidden multiplicity problem of repeated testing on subsets of $\mathbf{x}_t$, and the conclusion requires amendment to conditional non-causality of $\mathbf{y}_t$ given an exhaustive list of all other components $\mathbf{x}_{tj}$, $j = 1, \ldots, p$ in the model after adequate Type 1 error control. Methods that select causal covariate pairs through the use of penalized optimization problems do not always allow for easy extension to the multiple testing problem, and may require repetitive sub-sampling approaches such as stability selection (Meinshausen and Bühlmann, 2010). A thorough definition of Granger causality provides a clear representation of the collective conclusion to be drawn on the covariate set $\mathbf{x}_t$, clarifies the conditional nature of the result on the specified model and included variables, and stresses the philosophical ordering from the null hypothesis implying no causal structure to the alternative that demands evidence of the contrary, all while retaining any general functional form of $\mathcal{P}\left(\mathbf{y}_t | \mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \mathbf{X}_{<t}\right)$.

Holland (1986) relates the definition to that of Suppes (1970) and criticizes its fragile reliance on the specified pre-exposure variables that may completely change an inferential result. Maziarz (2015) writes that "Granger causality does not meet the requirements of an investigator who uses this method due to epistemic reasons" and

the methodology should be used "only if the theoretical background is insufficient," noting the common cause fallacy, indirect causality, and problems related to sampling frequency. In this tone, the predictive nature of these definitions can relate to a causal structure between two variable groups, but alone is not enough to establish *effective* connectivity, distinguishing a direct influence of one population on another (Bressler and Seth, 2011; Friston, 1994). Even in the presence of the optimal set $\mathcal{I}_{<t}$, association and precedence are not enough to distinguish true causality if slight redundancies are included or an effect does not remain constant in direction through time (Maziarz, 2015). Granger causality exists in the realm of *functional* connectivity that identifies correlation at one or more time lags (Friston, 1994). Appropriate use of Granger causality is contentious, but the method has been applied to a variety of fields like economics, environmental sciences, and neuroscience (Bernanke, 1990; F. Chen et al., 2021; Cox Jr. and Popken, 2015; Dey et al., 2020; Holland, 1986; Reid et al., 2019; Seth, Barrett, and Barnett, 2015; Sims, 1972). Cautious and targeted use of the Granger causal framework can elucidate predictive relationships between variables that warrant further study when prior knowledge of potential causal relationships is limited.

In the linear realm, $\mathcal{P}(\mathbf{y}_t|\mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \mathbf{X}_{<t})$ is often sought from a VAR(MA) model and evaluated with in-sample testing, and Himdi and Roy (1997) examines formulation via a non-causal hypothesis (Granger, 1969; Granger, 1980). Geweke (1982) proposed a spectral decomposition form of linear Granger causality for application to stationary Gaussian processes. Inference is performed with the estimated covariance of the restricted model (with $\mathbf{X}_{<t}$ excluded) and that of the unrestricted model, where under the null it is assumed the two are equal. Increasing dimension of the response variable often requires implementation of an approximate test or a switch

to permutation-like testing (Anderson and Robinson, 2001; Barnett and Seth, 2011). In-sample testing differs from the true notion of predictive ability, and out-of-sample methods align closer to the essence of Granger causality (Chao, Corradi, and Swanson, 2001; Inoue and Kilian, 2005; Peters, Bühlmann, and Meinshausen, 2016). This distinction is especially important when using deep learning techniques to model complex, nonlinear dynamics.

4.1.1 Nonlinear Adaptations

Nonparametric methods provide the basis for many nonlinear adaptations of Granger causality; specification of the exact functional form can be challenging. Parametric attempts, like the ordinary differential equations approaches of Henderson and Michailidis (2014) and H. Wu et al. (2014), allow for flexible definition of a series of functions to capture dependence, but are limited to modeling additive dynamics when the true mechanism of interaction may be more complicated. Some model-free information theoretic methods detect more elaborate forms of nonlinear dependence with minimal assumptions, but suffer from highly variable estimates and challenges of application to multivariate systems (Amblard and Michel, 2011; Runge et al., 2012; Vicente et al., 2011). Kernel Granger causality examines the linear form in a transformed feature space, but model comparison can become difficult (Marinazzo, Pellicoro, and Stramaglia, 2008; Marinazzo, Liao, et al., 2011). ANNs allow for general forms of nonlinear dependence in a similar feature space.

Tank et al. (2022) extend Granger causality to the nonlinear space using component-wise MLPs (cMLP), which model individual variables in the response $\mathbf{y}_{ti}$, $i = 1, \ldots, d$, with separate artificial networks. Parameters are sought via a penalized optimiza-

tion approach and proximal gradient descent, and no Granger causal connection is inferred for an individual covariate if the corresponding row in a component input parameter matrix ($\mathbf{W^{1i}}$ in Equation 1.4) is zero (Tank et al., 2022). The penalized approach encourages sparse solutions that block the inclusion of information from less predictive components in the hidden states $\mathbf{h}_t$, but selecting the regularization parameter is not an easy task and values may produce vastly different results. Tank et al. (2022) further introduce a component-wise LSTM (cLSTM) model that harnesses the recurrent structure to circumvent selection of the optimal lag for inclusion in the covariate set $\mathbf{X}_{<t}$. This formulation, while making model specification as it relates to the time lag components easier, has the consequence of mixing inferential results across several lags.

Khanna and Tan (2019) builds on the cMLP framework with statistical recurrent units (Oliva, Póczos, and Schneider, 2017), Biswas and Ombao (2022) discusses the application of the component network structure to frequency-specific relationships and non-stationary data, and Marcinkevičs and Vogt (2021) introduces generalized vector autoregressive (GVAR) methodology aimed at interpretability of potential functional relationships. The Jacobian Granger causality method of Suryadi, Chew, and Ong (2023) uses the Jacobian matrix, and Nauta, Bucur, and Seifert (2019) (TCDF) uses convolutional neural networks and attention scores to serve as measures of variable importance. Jointly estimating a large number of parameters is computationally expensive, and Duggento, Guerrisi, and Toschi (2021) instead use randomly initialized ESNs. Because computation is performed using linear techniques rather than a gradient descent algorithm, complexity decreases; however, inference can only be performed on the output coefficients if information mixing does not occur in the hidden states $\mathbf{h}_t$. They formulate $\mathbf{W^h}$ as a block diagonal matrix, which limits the

scope of application to a specific subset of additive nonlinear interactions (Duggento, Guerrisi, and Toschi, 2021).

Many of these methods ignore the collective inference principle of Granger causality and instead take the eager approach of performing individual covariate inference, sometimes with disregard for the multiplicity problem. Evaluation of these predictive relationships is often performed using in-sample tests. As a universal approximator, artificial neural networks of sufficient width or depth can approximate *any* functional relationship between two covariate sets, even if it is data-specific and the model is overfit, making them prone to link variables that do not have a predictive relationship as the dimension of the network increases. Sparsity inducing penalties may marginally improve reliability of in-sample tests, but out-of-sample testing helps control the overfitting problem to identify only *useful* functional relationships. In this vein, Horvath, Sultan, and Ombao (2022) develop the Learned Kernel VAR (LeKVAR) method that proposes use of a kernel parameterized by an artificial neural network they argue is less prone to overfitting from a decoupling importance measure of the individual series and the selected lags, and the TCDF method employs a permutation-like testing procedure after variable selection (Nauta, Bucur, and Seifert, 2019).

Rather than comparison of the inherently unequal restricted and unrestricted model errors, focus in this chapter is shifted to out-of-sample predictability by implementing a permutation structure. There is precedence for the use of permutation-type procedures on general linear models, and their asymptotics are well studied in literature (Anderson and Legendre, 1999; Anderson and Robinson, 2001; DiCiccio and Romano, 2017; Winkler et al., 2014). Nauta, Bucur, and Seifert (2019) employ this type of permutation procedure after their complicated convolutional neural network (CNN) screening procedure. This chapter explicitly defines the methodology for its

widespread use as a decision framework in Granger causal inference. Importance of a chronologically ordered variable can be interpreted as "causal" (predictive) effect, and the strategy builds on the concept of exchangeability (like the directed graph method of Caron and Fox (2017)), where if $\tilde{\mathbf{X}}$ is a random permutation of the rows of $\mathbf{X}$, $\mathbf{Y}_{<t+1} \perp \mathbf{X}$ given $\mathbf{Y}_{<t}$ implies $\mathbf{Y}_{<t+1} \perp \tilde{\mathbf{X}}$ given $\mathbf{Y}_{<t}$, but the converse is not always true (Van der Laan, 2006).

The following definition pair, adjusting Definitions 4.2 and 4.3 to a permutation structure for the covariate matrix, are proposed to investigate if $\mathbf{x}_t$ Granger causes $\mathbf{y}_t$. The unrestricted and restricted models are replaced by a null model and a permuted model, where $\tilde{\mathbf{X}}_{<t}$ is a copy of $\mathbf{X}_{<t}$ with the time axis (rows) randomly permuted.

Definition 4.4. Time series $\mathbf{x}_t \in \mathbb{R}^p$ is not conditionally Granger causal for $\mathbf{y}_t \in \mathbb{R}^d$ given $\mathbf{z}_t \in \mathbb{R}^q$ and the relevant history of the response $\mathbf{Y}_{<t}$ if and only if

$$\mathbf{Y}_{<t+1} | \mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \mathbf{X}_{<t} \stackrel{d}{=} \mathbf{Y}_{<t+1} | \mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \tilde{\mathbf{X}}_{<t}. \tag{4.5}$$

Definition 4.5. Time series $\mathbf{x}_t$ is conditionally Granger causal for $\mathbf{y}_t$ given $\mathbf{z}_t$ and the relevant history of the response $\mathbf{Y}_{<t}$ if

$$\text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t | \mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \mathbf{X}_{<t}\right)\right] < \text{Var}\left[\mathbf{y}_t - \mathcal{P}\left(\mathbf{y}_t | \mathbf{Y}_{<t}, \mathbf{Z}_{<t}, \tilde{\mathbf{X}}_{<t}\right)\right]. \tag{4.6}$$

Permutations of $\mathbf{X}_{<t}$ (augmented with lagged observations to maintain short-term dependence in $\mathbf{X}$ (Kunsch, 1989; R. Y. Liu and Singh, 1992)) break the dependence structure between $\mathbf{X}$ and the response while retaining the intradependence of the

covariates. Restructuring the Granger causal framework allows for use of out-of-sample estimated prediction errors, corrects the imbalance of comparison between restricted and unrestricted models, and presents a clear path to account for multiple testing, aligning the methodology closer to its inferential utility.

4.2 Methodology

Suppose observed realizations of the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$ arise from the set of all potential realizations $\omega \in \Omega$. Define Ω_{obs} as the size φ set of observations, $\Omega_{\text{obs}} = \{1, \ldots, \varphi\} \subseteq \Omega$, and note that usually $\varphi = 1$. Instances for $\varphi > 1$ may occur with repeated trials of a controlled experiment. For simplicity of the original presentation, the subscript ω notation specifying an observed realization is omitted until the end of this subsection.

4.2.1 Structure

Nonlinear functional dependence in the data is captured with FNNs of Equations 1.2 and 1.3, where $\mathbf{W^i}$ and $\mathbf{b}$ are randomly generated. The exact formulation of this transformation to a representative space (akin to Ψ in Equation 1.13) is not the main focus of this chapter and the dimension of the feature space N is fixed for direct comparison to other methods. For demonstration of NPGC, a simple model agnostic structure was selected. Other, more targeted featurization strategies will likely more effectively describe data specific dependence. For the simple FNN, adaptation to an automated selection of the feature dimension N can be found in Appendix C.

With familiarity of a dataset, a researcher selects γ lagged values of $\mathbf{y}_t$ that serve as a representative history, $\mathbf{Y}_{\text{lag}} \in \mathbb{R}^{(T+\gamma)\times\gamma d}$ (corresponding to $\mathbf{Y}_{<t}$ in Definitions

4.4 and 4.5), and assume that an appropriate number of lags is selected such that autocorrelation in $\mathbf{y}_t$ is fully explained across all potential realizations. Selection of the truncation lag γ is outside the scope of this work; Ng and Perron (2001), Ivanov and Kilian (2005), Shojaie and Michailidis (2010), and Nicholson, Matteson, and Bien (2017) provide detailed discussions. With a finite data length $T + \gamma$, this process restricts the usable portion of $\mathbf{X}$, $\mathbf{Y}$, and any additional covariates $\mathbf{Z}$ to the last T rows. All columns (individual variables) are standardized for consistent behavior in a random FNN featurization process with an activation function (Goodfellow, Bengio, and Courville, 2016).

Dependence across rows of $\mathbf{X}$, for example a covariate lag structure, is captured by augmenting the matrix with additional columns. The rows of the (augmented) covariate matrix $\mathbf{X}$ are randomly reorganized via $\mathbf{\Pi}_m$ to generate several permutations $\tilde{\mathbf{X}}_m = \mathbf{\Pi}_m \mathbf{X}$ for $m = 1, \ldots, M$. The designated first permutation, $m = 1$, corresponds to the original ordering of the data where $\mathbf{\Pi}_1 = \mathbf{I}$.

After permutation, the predictor matrices $\left[\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \tilde{\mathbf{X}}_m\right] \in \mathbb{R}^{T \times (1+\gamma d+q+p)}$ are compiled, and the FNN is rewritten to the structure in Equation 4.7 (with activation function $g = \tanh$). $\mathbf{W^i}$ and $\mathbf{b}$ are combined into a single parameter matrix $\mathbf{W} \in \mathbb{R}^{(1+\gamma d+q+p) \times N}$ after inclusion of the intercept term in the predictor matrices, and each matrix entry is an independent Gaussian realization $w_{ij} \sim \mathcal{N}(0, 1)$.

$$\mathbf{H}_m = g\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \tilde{\mathbf{X}}_m\right]\mathbf{W}\right) = \tanh\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \tilde{\mathbf{X}}_m\right]\mathbf{W}\right) \tag{4.7}$$

Added uncertainty arising from the random generation is mitigated through several featurizations, $\mathbf{W}_r$ for $r = 1, \ldots, \mathcal{R}$, and extracting the aggregate behavior. The models can be written in terms of the original ($m = 1$) and permuted ($m = 2, \ldots, M$)

feature spaces, where $\mathbf{U}_{m,r}$ is the variation in $\mathbf{Y}$ not captured by the functional relationship with the feature space $\mathbf{H}_{m,r}$.

$$\mathbf{Y} = \mathbf{H}_{m,r}\mathbf{W^o}_{m,r} + \mathbf{U}_{m,r} \tag{4.8}$$

Define $\mathbf{\Theta}_m$ as the underlying covariance matrix of the prediction for $\mathbf{Y}$ given the relevant history of the response $\mathbf{Y}_{\text{lag}}$, the additional variables $\mathbf{Z}$, and the permuted covariate set $\tilde{\mathbf{X}}_m$. Denote $\vartheta_m = \text{tr}\left(\mathbf{\Theta}_m\right)$ as the corresponding parameter over all potential realizations $\omega \in \Omega$. Variation in the estimate arises from potential realizations of the data $\omega \in \Omega$ and via randomly generated FNNs approximating the nonlinear functional form. For a given data realization $\left(\mathbf{X}, \mathbf{Y}, \mathbf{Z}\right)_\omega$, and under the true functional form f, define the specific covariance matrix of the prediction $\mathbf{\Sigma}_{m,\omega}$. For random featurization r, define the covariance matrix of prediction $\mathbf{S}_{m,\omega,r}$ as an estimate of $\mathbf{\Sigma}_{m,\omega}$. The out-of-sample variation parameter ϑ_m is estimated for each permutation via a cross-validation approach.

4.2.2 Estimating Granger Causal Influence

A sufficiently large featurization dimension N is chosen to "linearize" any existing functional relationship, but not so large that the network is able to memorize inputs or fabricate dependence between the permuted data and a response. This implicitly assumes the existence of some nonlinear functional relationship between a covariate set and a response will be "easier" for an ANN to learn than random matching of inputs to outputs in the permuted data, and an exact form of this condition is proposed in Section 4.3. The data is split into K sets for model computation and testing, and define the number of observations in each set $k = 1, \ldots, K$ as $T_k =$

$\lfloor T/K \rfloor + \mathbf{1}\{(T \bmod K) \geq k\}$.

Under the form of Equation 4.7, several random FNNs are generated $r = 1, \ldots, \mathcal{R}$. The model matrices $\mathbf{W}_r$ are held fixed over all permutations m and observations $\omega \in \Omega_{\text{obs}}$ for a consistent error estimation framework. For each permutation, with $m = 1$ corresponding to the original data, the predictor matrices are projected into the respective feature spaces, and the residuals for test set k can be written as in Equation 4.9, where the where the training data (subscript $-k$) excludes set k.

$$\mathbf{R}_{m,\omega,r,k} = \mathbf{H}_{m,\omega,r,k} \left(\mathbf{H}_{m,\omega,r,-k}^{\top} \mathbf{H}_{m,\omega,r,-k}\right)^{-1} \mathbf{H}_{m,\omega,r,-k}^{\top} \mathbf{Y}_{\omega,-k}, -\mathbf{Y}_{\omega,k} \tag{4.9}$$

The out-of-sample prediction residuals $\mathbf{R}$ from the test set align closer to the original definition of predictive ability in Granger causal inference than the in-sample model variation. This distinction is especially important with the use of ANNs and the ability to learn any arbitrary, data-specific dependence structure. The estimate for the out-of-sample variation in prediction residuals $\hat{\vartheta}_m$ is shown in Equation 4.10.

$$\hat{\vartheta}_m = \frac{1}{\varphi \mathcal{R} K} \sum_{\omega=1}^{\varphi} \sum_{r=1}^{\mathcal{R}} \sum_{k=1}^{K} \frac{1}{T_k} \text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top} \mathbf{R}_{m,\omega,r,k}\right) \tag{4.10}$$

A null distribution of variation from each model $m = 1, \ldots, M$ is approximated from the permutation structure. The random permutations of the covariate set $\mathbf{X}$ break the potential dependence structure present in the form of some predictive relationship with the response $\mathbf{Y}$. Under the null hypothesis, the original data is viewed as one of M random permutations, and the original "permutation" $\tilde{\mathbf{X}}_1 = \mathbf{X}$ should exhibit similar properties to $\tilde{\mathbf{X}}_m$ for $m = 2, \ldots, M$. Analogously, if the time observations $\mathbf{X}$ are exchangeable for prediction of $\mathbf{Y}$, the conditional distribution of the

prediction will not change.

The variation estimates are drawn from the distribution of all possible permutations in Equation 4.11, and $\hat{\vartheta}_1$ is expected to fall above some lower tail portion.

$$\hat{\vartheta}_m \sim \hat{\mathcal{H}}(s) = (T!)^{-1} \sum_{i=1}^{T!} \mathbf{1}\{\hat{\vartheta}_i \leq s\}, \tag{4.11}$$

This leads to an approximate null distribution where the estimate $\hat{\vartheta}_1$ is at quantile $\hat{Q}_M$ of the empirical distribution $\hat{\mathcal{H}}_M(s)$, defined in Equation 4.12, formed from a subsample of size $M \leq T!$. A decision rule is formulated from comparison to a chosen level of test α.

$$\hat{\mathcal{H}}_M(s) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\hat{\vartheta}_m \leq s\} \tag{4.12}$$

$$\hat{Q}_M = \hat{\mathcal{H}}_M(\hat{\vartheta}_1) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{\hat{\vartheta}_m \leq \hat{\vartheta}_1\right\} \tag{4.13}$$

Rejection of the null hypothesis in this framework, $\hat{Q}_M \leq \alpha$, presents evidence for $\mathbf{X}$ as Granger causal of $\mathbf{Y}$ *conditional* on the additional variables $\mathbf{Z}$ and the relevant history of the response $\mathbf{Y}_{\text{lag}}$. For the case when $\varphi = 1$, this conclusion is conditional on *error free observation* of the dataset. Inferential results must either include this assumption, or the scope narrowed to the specific observation ω. The full algorithmic process is shown in Algorithm 4.1, and an explicit outline of the theoretical behavior of these estimates and development of the underlying framework is given in Section 4.3.

Algorithm 4.1 Nonlinear Permuted Granger Causality

Inputs: $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$ for all φ realizations $\omega \in \Omega_{\text{obs}}$; lag selection γ; # permutations M; # random featurizations $\mathcal{R}$; feature space dimension N; # cross-validation folds K

Outputs: $\hat{Q}_M$; $\hat{\vartheta}_m$ for each permutation $m = 1, \ldots, M$

Generate permutations $\tilde{\mathbf{X}}_m = \mathbf{\Pi}_m \mathbf{X}$ for $m = 1, \ldots, M$ with $\mathbf{\Pi}_1 = \mathbf{I}$
Initialize $\mathbf{W}_r \in \mathbb{R}^{(1+\gamma d+q+p)\times N}$ where each element $w_{r,ij} \sim \mathcal{N}(0,1)$ for all $\mathcal{R}$
for m in $1 : M$ **do**
    for ω in $1 : \varphi$ **do**
        for r in $1 : \mathcal{R}$ **do**
            $\mathbf{H}_{m,\omega,r} \leftarrow \tanh\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag},\omega}\ \mathbf{Z}_\omega\ \tilde{\mathbf{X}}_{m,\omega}\right]\mathbf{W}_r\right)$
            for k in $1 : K$ **do**
                $\mathbf{R}_{m,\omega,r,k} \leftarrow \mathbf{H}_{m,\omega,r,k}\left(\mathbf{H}_{m,\omega,r,-k}^\top \mathbf{H}_{m,\omega,r,-k}\right)^{-1}\mathbf{H}_{m,\omega,r,-k}^\top \mathbf{Y}_{\omega,-k} - \mathbf{Y}_{\omega,k}$
            end for
        end for
    end for
    $\hat{\vartheta}_m \leftarrow (\varphi\mathcal{R}K)^{-1}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K} T_k^{-1}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^\top \mathbf{R}_{m,\omega,r,k}\right)$
end for
$\hat{Q}_M \leftarrow M^{-1}\sum_{m=1}^{M}\mathbf{1}\left\{\hat{\vartheta}_m \leq \hat{\vartheta}_1\right\}$
return $\hat{Q}_M$; $\hat{\vartheta}_m$ for $m = 1, \ldots, M$

4.3 Theory

Define $\mathbf{\Theta}_m$ as the underlying covariance matrix of the predictive ability of permutation m, and the quantity $\vartheta_m = \text{tr}(\mathbf{\Theta}_m)$. For each potential realization of the data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$, $\omega \in \Omega$, define the realization-specific covariance matrix $\mathbf{\Sigma}_{m,\omega}$ drawn from some distribution with expectation $\mathbb{E}[\text{tr}(\mathbf{\Sigma}_{m,\omega})] = \vartheta_m$ and variance $\tau_\omega^2 < \infty$ that is constant over all permutations. Each random generated FNN for the featurization process $r = 1, \ldots, \mathcal{R}$ produces $\mathbf{S}_{m,\omega,r}$ as an estimate of $\mathbf{\Sigma}_{m,\omega}$, where $\mathbb{E}[\text{tr}(\mathbf{S}_{m,\omega,r})|\mathbf{\Sigma}_{m,\omega}] = \text{tr}(\mathbf{\Sigma}_{m,\omega})$ and $\text{Var}[\text{tr}(\mathbf{S}_{m,\omega,r})|\mathbf{\Sigma}_{m,\omega}] = \tau_r^2 < \infty$.

The null and permuted models are evaluated with the out-of-sample prediction residuals, shown in Equation 4.9, and define the estimate for total variation as in Equation 4.10. Under the null hypothesis when the conditional distribution of the prediction is invariant to permutation of the covariate set $\mathbf{X}$, $\vartheta_1 = \vartheta_2 = \cdots = \vartheta_{T!} = \vartheta$, leading to the null and alternative hypotheses in Equation 4.14.

$$\begin{aligned} H_0 &: \vartheta_1 = \vartheta_2 = \cdots = \vartheta_{T!-1} = \vartheta_{T!} \\ H_A &: \vartheta_1 < \vartheta_i \text{ for all } i = 2, \ldots, T! \end{aligned} \tag{4.14}$$

The null hypothesis is tested using the sample quantile $\hat{Q}_M$ from the empirical distribution $\hat{\mathcal{H}}_M(s)$ defined in Equations 4.12 and 4.13.

4.3.1 Conditions for Theoretical Results

Theoretical results in this section rely on a set of three mild conditions comparable to those found in relevant literature. Four additional conditions provide regularity to the featurization process. Theoretical results are derived for a generic activation

function g in Equation 4.7 with the constraints of Condition 4.7.

Condition 4.6. The data is continuous and stationary, and the discrete, regular sampling grid $t = 1, \ldots, T$ is sufficiently fine to capture any potential functional dependence in the variable matrices $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$ for any realization $\omega \in \Omega$.

Condition 4.7. The nonlinear function activation function g in Equation 4.7 is bounded, $g : \mathbb{R} \to [a, b]$ for some $a < b$, such that for any $x \in \mathbb{R}$, $|g(x)| \leq G = \max\{|a|, |b|\}$ and $G < \infty$.

Condition 4.8. Let $\mathcal{W}$ be the space of all element-wise randomly generated FNNs, $w_{ij} \sim \mathcal{N}(0, 1)$, such that for all $\mathbf{W}_r \in \mathcal{W}$, the matrix $\mathbf{W}_r$ is full rank and generates a feature matrix that is full rank with a finite condition number. For permutation m, realization ω, and random featurization r,

$$\text{rank}\left(\mathbf{H}_{m,\omega,r}\right) = N \tag{4.15}$$

$$\text{and} \quad \kappa(\mathbf{H}_{m,\omega,r}) = \sigma_1(\mathbf{H}_{m,\omega,r})/\sigma_N(\mathbf{H}_{m,\omega,r}) \leq \kappa_{\max} < \infty. \tag{4.16}$$

All initialized model matrices $\mathbf{W}_r \in \mathcal{W}$.

For the matrix $\mathbf{H}_{m,\omega,r} = (h_{m,\omega,r,ij})$, define G as the maximal element from Condition 4.7 and the average squared entry as $\bar{h}^2_{m,\omega,r}$.

Condition 4.9. The following expected values exist and are finite:

$$\mathbb{E}\left[G^{-2}\bar{h}^2_{m,\omega,r}|\mathbf{S}_{m,\omega,r}\right] = \nu^2 < \infty, \tag{4.17}$$

$$\mathbb{E}\left[G^2\left(\bar{h}^2_{m,\omega,r}\right)^{-1}|\mathbf{S}_{m,\omega,r}\right] = \xi^2 < \infty, \tag{4.18}$$

$$\text{and} \quad \mathbb{E}\left[G^4\left(\bar{h}^2_{m,\omega,r}\right)^{-2}|\mathbf{S}_{m,\omega,r}\right] = \varrho^4 < \infty. \tag{4.19}$$

Combining Conditions 4.7, 4.8, and 4.9, $0 < \nu^2 \leq 1$, $1 \leq \xi^2 < \infty$, and $1 \leq \varrho^4 < \infty$. Define f as the true functional relationship between the response and the unpermuted predictor matrix for all $\omega \in \Omega$,

$$\mathbf{Y}_\omega = f\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag},\omega}\ \mathbf{Z}_\omega\ \tilde{\mathbf{X}}_{1,\omega}\right]\right) + \boldsymbol{\mathcal{U}}_\omega, \tag{4.20}$$

and the approximating model form for permutations $m = 1, \ldots, M$ and featurizations $r = 1, \ldots, \mathcal{R}$.

$$\mathbf{Y}_\omega = \mathbf{H}_{m,\omega,r}\mathbf{W^o}_{m,\omega,r} + \mathbf{U}_{m,\omega,r} \tag{4.21}$$

Condition 4.10. For every $\eta > 0$, there exists a fixed N, where $1 + \gamma d + q + p \leq N < \infty$, such that as the number of random featurizations $\mathcal{R} \to \infty$,

$$\sup_{\omega\in\Omega}\left\|\frac{1}{\mathcal{R}}\sum_{r=1}^{\mathcal{R}}\mathbf{H}_{1,\omega,r}\mathbf{W^o}_{1,\omega,r} - f\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag},\omega}\ \mathbf{Z}_\omega\ \tilde{\mathbf{X}}_{1,\omega}\right]\right)\right\| < \eta, \tag{4.22}$$

where $\mathbf{W^o}_{1,\omega,r}$ is the true coefficient matrix of feature space r for the unpermuted covariate set, $\mathbf{H}_{1,\omega,r} = g\left(\left[\mathbf{1}\ \mathbf{Y}_{\text{lag},\omega}\ \mathbf{Z}_\omega\ \tilde{\mathbf{X}}_{1,\omega}\right]\mathbf{W}_r\right)$.

Note that Condition 4.10 specifically pertains to the f piece of the true functional relationship; no assumption is made of the closeness of a transformed response $\mathbf{H}_{1,\omega,r}\mathbf{W^o}_{1,\omega,r}$ and the true values $\mathbf{Y}_\omega$ if the predictors themselves are not reliable and the entries of $\boldsymbol{\mathcal{U}}_\omega$ in Equation 4.20 are large.

Condition 4.11. For all realizations $\omega \in \Omega$, and random generated FNNs $\mathbf{W}_r \in \mathcal{W}$, $r = 1, \ldots, \mathcal{R}$, the quantities $\text{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}$ are independently drawn from continuous distributions with defined expectation in Equation 4.23 and constant, finite variance

across all permutations and potential realizations of the data.

$$\mathbb{E}\left[\text{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right] = \text{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right) \tag{4.23}$$

$$\text{Var}\left(\text{tr}\left[\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right) = \tau_r^2 < \infty \tag{4.24}$$

Similarly, $\boldsymbol{\Sigma}_{m,\omega}$ are independently drawn from continuous distributions with expectation in Equation 4.25 and constant, finite variance across permutations.

$$\mathbb{E}\left[\text{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right)\right] = \vartheta_m \tag{4.25}$$

$$\text{Var}\left(\text{tr}\left[\boldsymbol{\Sigma}_{m,\omega}\right]\right) = \tau_\omega^2 < \infty \tag{4.26}$$

Condition 4.12. The relevant history of the response, $\mathbf{Y}_{<t,\omega} = \mathbf{Y}_{\text{lag},\omega} \in \mathbb{R}^{T\times\gamma d}$ is appropriately chosen such that the model errors $\mathbf{U}_{m,\omega,r}$ are independent, or $\mathbf{u}_{m,\omega,r,t} \perp \mathbf{u}_{m,\omega,r,t'}$ for any realization $\omega \in \Omega$, permutation $m = 1,\ldots,M$, featurization $r = 1,\ldots,\mathcal{R}$, and time point $t = 1,\ldots,T$ where $t \neq t'$. Further, the model errors follow multivariate normal distributions with mean zero and constant variation, leading to the result

$$\mathbf{u}_{m,\omega,r,t} \overset{i.i.d.}{\sim} \mathcal{N}_d\left(\mathbf{0}, \mathbf{S}_{m,\omega,r}\right). \tag{4.27}$$

4.3.2 Asymptotic Properties

The asymptotic behavior of the estimates and established testing framework is examined under the listed conditions in Section 4.3.1. Define the underlying variation parameter for permutation m as ϑ_m, and the estimate $\hat{\vartheta}_m$ as in Equation 4.10. Assume a fixed test set size T_k and allow the training set $T_{-k} = T - T_k$ and number of

folds K to grow as $T \to \infty$.

Theorem 4.13. *Under the conditions listed in Section 4.3.1, with Condition 4.11 modified such that $\tau_\omega^2 = 0$ (i.e., error free observation of the data), for all $\varepsilon > 0$,*

$$\lim_{\mathcal{R}\to\infty} \lim_{T\to\infty} \mathbb{P}\left(|\hat{\vartheta}_m - \vartheta_m| \leq \varepsilon\right) = 1. \tag{4.28}$$

Alternatively, under the conditions listed in Section 4.3.1, for all $\varepsilon > 0$,

$$\lim_{\varphi\to\infty} \lim_{T\to\infty} \mathbb{P}\left(|\hat{\vartheta}_m - \vartheta_m| \leq \varepsilon\right) = 1. \tag{4.29}$$

Without error free observation of the data, the underlying variation τ_ω^2 remains present in the estimate, but shrinks as the number of observations gets large.

Theorem 4.14. *Under the conditions listed in Section 4.3.1, the estimate for the variation parameter $\hat{\vartheta}_m$ admits a Central Limit Theorem with respect to the number of observations φ in Ω_{obs}.*

$$\lim_{\mathcal{R}\to\infty} \lim_{T\to\infty} \sqrt{\varphi}\left(\hat{\vartheta}_m - \vartheta_m\right) \xrightarrow{D} \mathcal{N}\left(0, \tau_\omega^2\right) \tag{4.30}$$

As a direct result of Theorem 4.14, $\hat{\vartheta}_m$ is a consistent estimate for ϑ_m when $\tau_\omega^2 > 0$ as φ tends to infinity along with the size of the training and test sets.

Under the Null Hypothesis

Under the null hypothesis, the underlying variation parameter is constant for every possible permutation, $\vartheta_1 = \cdots = \vartheta_T$!. Define the quantile estimate $\hat{Q}_M$ as in Equation 4.13, and establish the limiting uniform distribution from the result of Theorem 4.13.

Theorem 4.15. *Under the null hypothesis and the conditions listed in Section 4.3.1,*

$$\lim_{T\to\infty}\lim_{M\to T!}\hat{Q}_M \xrightarrow{D} \mathit{Uniform}(0,1). \tag{4.31}$$

Under the Alternative Hypothesis

Define Q as the true quantile for parameter ϑ_1 over all $T!$ possible permutations. In the limit, the quantile estimate defined in Equation 4.13 converges in probability to Q.

Theorem 4.16. *Under the alternative hypothesis and the conditions listed in Section 4.3.1, with Condition 4.11 modified such that $\tau_\omega^2 = 0$ (i.e., error free observation of the data), for all $\varepsilon > 0$,*

$$\lim_{M\to T!}\lim_{\mathcal{R}\to\infty}\lim_{T\to\infty}\mathbb{P}\left(|\hat{Q}_M - Q| \leq \varepsilon\right) = 1. \tag{4.32}$$

Similarly, under the alternative hypothesis and the conditions listed in Section 4.3.1, for all $\varepsilon > 0$,

$$\lim_{M\to T!}\lim_{\varphi\to\infty}\lim_{T\to\infty}\mathbb{P}\left(|\hat{Q}_M - Q| \leq \varepsilon\right) = 1. \tag{4.33}$$

4.3.3 Finite Sample Distribution

For a finite sample, the distribution of the estimated quantity $\hat{\vartheta}_m$ is derived as a sum of linear combinations of chi-square random variables. Define the following chi-square

random variables

$$X_{m,\omega,r,k,i}, Y_{m,\omega,r,k,ij}, Z_{m,\omega,r,k,ij} \sim \chi^2_1 \tag{4.34}$$

for all $\omega \in \Omega_{\text{obs}}$, $r = 1, \ldots, \mathcal{R}$, $k = 1, \ldots, K$, $i = 1, \ldots, T_k d$ and $j < i$. Denote

$$\mathbf{H}_{m,\omega,r,k} \left[\mathbf{H}^{\top}_{m,\omega,r,-k} \mathbf{H}_{m,\omega,r,-k} \right]^{-1} \mathbf{H}^{\top}_{m,\omega,r,k} = (\phi_{m,\omega,r,k,ij}) \tag{4.35}$$

and $\mathbf{S}_{m,\omega,r} = (s_{m,\omega,r,ij})$, with $i' = \lceil i/d \rceil$, $j' = \lceil j/d \rceil$, $i^* = i \bmod d$, and $j^* = j \bmod d$.

Theorem 4.17. *Under the conditions listed in Section 4.3.1, a finite sample containing T observations, $\mathcal{R}$ random generated FNNs, and φ realizations in the set Ω_{obs}, the estimate for the variation parameter $\hat{\vartheta}_m$ defined in Equation 4.10 follows a generalized chi-square distribution.*

$$\begin{aligned}
\hat{\vartheta}_m \sim \frac{1}{\varphi \mathcal{R} K} \sum_{\omega=1}^{\varphi} \sum_{r=1}^{\mathcal{R}} \sum_{k=1}^{K} \frac{1}{T_k} \sum_{i=1}^{T_k d} \Bigg[& \left(\phi_{m,\omega,r,k,i'i'} + 1\right) s_{m,\omega,r,k,i^*i^*} X_{m,\omega,r,k,i} \\
& + \sum_{j=1}^{d\lfloor (i-1)/d \rfloor} \phi_{m,\omega,r,k,i'j'} s_{m,\omega,r,k,i^*j^*} \left(Y_{m,\omega,r,k,ij} - Z_{m,\omega,r,k,ij}\right) \\
& + \sum_{j=d\lfloor (i-1)/d \rfloor + 1}^{i-1} \left(\phi_{m,\omega,r,k,i'j'} + 1\right) s_{m,\omega,r,k,i^*j^*} \left(Y_{m,\omega,r,j,ij} - Z_{m,\omega,r,k,ij}\right) \Bigg]
\end{aligned} \tag{4.36}$$

4.4 Simulation Study

The ability of the permutation-based methodology of NPGC (similar to the decision rule formulation of TCDF) to detect the presence of functional connectivity and control for false positive results is evaluated with a consistent FNN framework (like the cMLP nonlinear transformation). All methods featurize the data to dimension

$N = 100$. Three of the five comparison methods use in-sample methodology and variants of the lasso penalty within cMLP models (Tank et al., 2022; Tibshirani, 1996). The other two employ random FNN generation like NPGC, and in-sample testing from comparison of restricted and unrestricted models or the substitution of random Gaussian noise.

Multiple testing on bivariate pairs is not examined; a group decision process is considered where the covariate set $\mathbf{X}$ is (or is not) collectively Granger causal for the response $\mathbf{Y}$. Ability to detect a Granger causal result will almost certainly improve with a more sophisticated nonlinear structure; relative differences between methods are demonstrated at a baseline level. A full list of included methods is given in Table 4.1.

Table 4.1: Testing frameworks for Granger causal inference simulations.

(i)	Nonlinear Permuted Granger Causality (NPGC)
(ii)	cMLP, Group Lasso Penalty (cMLP-GL)
(iii)	cMLP, Group Sparse Group Lasso Penalty (cMLP-GSGL)
(iv)	cMLP, Hierarchical Lasso Penalty (cMLP-H)
(v)	Restricted vs. Unrestricted Models (R/U)
(vi)	Gaussian Noise Substitution (GNS)

The lasso type objectives of cMLP methods (ii) - (iv) penalize nonzero rows of the matrix $\mathbf{W^{1i}}$ in the formulation of Equation 1.4 (or $\mathbf{W}$ in Equation 4.7). The corresponding null and alternative hypotheses are adjusted to those shown in Equation 4.37, and a Granger causal connection is present for a covariate set if any correspond-

ing rows of $\mathbf{W^{1i}}$ contain nonzero components (Tank et al., 2022).

$$\begin{aligned} H_0 &: \mathbf{W^{1i}}_j = \mathbf{0} \text{ for all } j \text{ corresponding to the covariate set } \mathbf{X} \\ H_A &: \mathbf{W^{1i}}_j \neq \mathbf{0} \text{ for at least one } j \text{ corresponding to the covariate set } \mathbf{X} \end{aligned} \tag{4.37}$$

A group lasso penalty is applied to each variable j in the model matrix $\mathbf{W^{1i}}$ corresponding to the covariate set $\mathbf{X}$ over all time lags $t, \ldots, t-\gamma$ (Yuan and Y. Lin, 2006). As in Tank et al. (2022), define $\mathbf{W^{1i}}_j = \left[\mathbf{W^{1i}}_{j,t} \cdots \mathbf{W^{1i}}_{j,t-\gamma}\right]$, with each entry as the weights of the model matrix row for variable j and lagged data up to $t-\gamma$. Penalties for the cMLP methods take the general form $\lambda \sum_{j=1}^{p} \beta\left(\mathbf{W^{1i}}_j\right)$, with specific β for each defined below.

$$\beta_{GL}\left(\mathbf{W^{1i}}_j\right) = \left\|\mathbf{W^{1i}}_j\right\|_F \tag{4.38}$$

$$\beta_{GSGL}\left(\mathbf{W^{1i}}_j\right) = \left\|\mathbf{W^{1i}}_j\right\|_F + \sum_{k=0}^{\gamma} \left\|\mathbf{W^{1i}}_{j,t-k}\right\|_2 \tag{4.39}$$

$$\beta_{H}\left(\mathbf{W^{1i}}_j\right) = \sum_{k=0}^{\gamma} \left\|\left[\mathbf{W^{1i}}_{j,t-k} \cdots \mathbf{W^{1i}}_{j,t-\gamma}\right]\right\|_F \tag{4.40}$$

The group sparse group lasso penalty combines sparsity of included variables and their lagged values like in Simon et al. (2013), and the novel hierarchical penalty of Tank et al. (2022) retains information about the natural ordering of the variables, encouraging solutions where for some lag k^*, $k > k^*$ implies $\mathbf{W^{1i}}_{j,t-k} = \mathbf{0}$. A regularization parameter of $\lambda = 0.5$ is chosen based on a cross-validation like trial and error approach. Conclusions drawn from these methods can vary greatly depending on the chosen λ; smaller values retain all matrix entries and larger values penalize the matrix to zero. Effects of varying λ are shown for the application in Section 4.5.

Two additional naive methods are included for comparison to NPGC. The in-

sample restricted and unrestricted method examines the ratio of model residuals $\hat{\sigma}^2_{\text{Res}}/\hat{\sigma}^2_{\text{Unres}}$ in a randomly generated FNN. The in-sample Gaussian noise substitution is methodologically similar, but instead substitutes Gaussian white noise in place of the covariate set $\mathbf{X}$ to examine $\hat{\sigma}^2_{\text{Noise}}/\hat{\sigma}^2_{\text{Unres}}$. The three model errors are shown in Equations 4.41 to 4.43, where $\mathbf{E} \sim \mathcal{MN}_{T\times p}(\mathbf{0}, \mathbf{I}, \mathbf{I})$, and the corresponding $\hat{\sigma}^2$ shown in Equation 4.44.

$$\mathbf{U}_{\text{Unres},r} = \mathbf{Y} - \tanh\left([\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \mathbf{X}]\,\mathbf{W}_r\right)\mathbf{W^o}_{\text{Unres},r} \tag{4.41}$$

$$\mathbf{U}_{\text{Res},r} = \mathbf{Y} - \tanh\left([\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \mathbf{0}]\,\mathbf{W}_r\right)\mathbf{W^o}_{\text{Res},r} \tag{4.42}$$

$$\mathbf{U}_{\text{Noise},r} = \mathbf{Y} - \tanh\left([\mathbf{1}\ \mathbf{Y}_{\text{lag}}\ \mathbf{Z}\ \mathbf{E}]\,\mathbf{W}_r\right)\mathbf{W^o}_{\text{Noise},r} \tag{4.43}$$

$$\hat{\sigma}^2_r = T^{-1}\mathbf{U}_r^\top\mathbf{U}_r \tag{4.44}$$

The individual terms in Equation 4.44 follow chi-square distributions with degrees of freedom $T - N$, and each ratio for an individual generated FNN is distributed $\mathcal{F}_{T-N,T-N}$. The sum of a large number of these independent (via Condition 4.12) statistics, all random FNNs $r = 1, \ldots, \mathcal{R}$, is approximately normal. As $\mathcal{R} \to \infty$ under the null hypothesis,

$$\mathcal{R}^{-1}\sum_{r=1}^{\mathcal{R}}\frac{\hat{\sigma}^2_{0,r}}{\hat{\sigma}^2_{1,r}} \xrightarrow{D} \mathcal{N}\left[\frac{T-N}{T-N-2}, \frac{4(T-N)^2(T-N-1)}{\mathcal{R}(T-N)(T-N-2)^2(T-N-4)}\right]. \tag{4.45}$$

This is used to formulate a naive decision rule with a large number ($\mathcal{R} = 1000$) of randomly generated FNNs.

The NPGC methodology is performed with $M = 400$ permutations, $K = 5$ cross validation folds, and $\mathcal{R} = 50$ randomly generated FNNs. The methods are successful when correctly flagging a Granger causal result when direct functional dependence

is present, or correctly labelling a non-causal result when it is absent, leading to the potential outcomes in Table 4.2. For the NPGC and naive ratio methods, a result is flagged if the quantile estimate of the associated statistic under the null hypotheses is under a specified level α. The lasso-type methods lack the direct translation to a traditional hypothesis testing framework.

Table 4.2: Potential simulation outcomes.

	Truth	
Result	Granger causal (GC = 1)	Not causal (GC = 0)
Granger causal (GC = 1)	$\rho_1 = \sum_{\mathcal{S}} \mathbf{1}\{\mathcal{D} = 1\} / N_1$	$\rho_{10} = \sum_{\mathcal{S}} \mathbf{1}\{\mathcal{D} = 1\} / N_0$
Not causal (GC = 0)	$\rho_{01} = \sum_{\mathcal{S}} \mathbf{1}\{\mathcal{D} = 0\} / N_1$	$\rho_0 = \sum_{\mathcal{S}} \mathbf{1}\{\mathcal{D} = 0\} / N_0$

Proportion of potential outcomes for the decision $\mathcal{D}$ of each of the methods shown in Table 4.1. The set $\mathcal{S}$ represents the space of all simulations (both GC = 1 and GC = 0) within a given setting, $N_1 = \sum_{\mathcal{S}} \mathbf{1}\{\text{GC} = 1\}$, and $N_0 = \sum_{\mathcal{S}} \mathbf{1}\{\text{GC} = 0\}$.

4.4.1 Simulation Settings

NPGC and the comparator methods in Table 4.1 are tested on two nonlinear processes: Lorenz-96 models of Karimi and Paul (2010), and TAR models introduced by Tong and Lim (1980). The p-dimensional Lorenz-96 model ($p \geq 4$) is governed by the continuous differential equation

$$\frac{dx_{i,t}}{dt} = (x_{i+1,t} - x_{i-2,t})\, x_{i-1,t} - x_{i,t} + F, \tag{4.46}$$

with $i = 1, \ldots, p$ and boundary series $x_{-1,t} = x_{p-1,t}$, $x_{0,t} = x_{p,t}$, and $x_{p+1,t} = x_{1,t}$. F is a forcing constant generated as $F \sim \text{Uniform}(5, 20)$ with higher values introducing a larger degree of nonlinear, chaotic behavior. For data generation, a sampling rate

of $\Delta t = 0.05$ and a burn-in period of 500 time steps are used.

The TAR(2) model is governed by a similar skeleton to a VAR, but allows for changes in parameters based on the value of a threshold variable (Tong and Lim, 1980).

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^{(k)} & \mathbf{A}_2^{(k)} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-1} \\ \mathbf{x}_{t-2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_t^{(k)} \\ \mathbf{0} \end{bmatrix}. \tag{4.47}$$

The threshold is defined on the values in $\mathbf{x}_{t-2}$ with regime $k = 1$ corresponding to the case when $\sum_{i=1}^{p} x_{i,t-2} \leq 0$ and $k = 2$ to $\sum_{i=1}^{p} x_{i,t-2} > 0$. For the TAR(2) process, $\boldsymbol{\varepsilon}_t^{(k)} \sim \mathcal{N}_p(\mathbf{0}, \sigma_k^2 \mathbf{I})$ where $\sigma_1 = 0.5$ and $\sigma_2 = 0.2$. Each $\mathbf{A}_1^{(k)}, \mathbf{A}_2^{(k)} \in \mathbb{R}^{p \times p}$ has elements $\mathbf{A}^{(k)} = (a_{ij}^{(k)}) \sim \text{Uniform}(-0.5, 0.5)$ that are thresholded to zero if $|a_{ij}^{(k)}| < 0.1$, and their spectral radii are at most 0.8 to ensure stationary data generation. Like the Lorenz-96 data generation procedure, there is a burn-in period of 500 time steps.

Two groups of dependent data observations are generated with $p = 6$ (one from each process), containing $\mathbf{x}_i$ for $i = 1, \ldots, 12$. The series $i = 1, \ldots, 6$ are labelled as the Lorenz-96 process and $i = 7, \ldots, 12$ as the TAR(2) process. The samples are of length $T = 250$, 500, and 1000 after burn-in and truncation for lag selection to include in the model, $\gamma = 3$. One index is selected to serve as the response, three to serve as the set of additional variables $\mathbf{Z}$ and three for the covariate set $\mathbf{X}$ depending on a Granger causal designation. A dataset is designated "Granger causal" if at least one series in $\mathbf{X}$ has a direct influence on the selected response (*i.e.*, one series in $\mathbf{X}$ is chosen from the same generating process). For each of the settings in Table 4.3, 200 datasets are generated for a total of 4800 trials.

Table 4.3: NPGC simulation settings.

Response Variable	T	# Causal in $\mathbf{X}$	# Causal in $\mathbf{Z}$
Lorenz-96	$\{250, 500, 1000\}$	$\{0, 2\}$	$\{0, 2\}$
TAR(2)	$\{250, 500, 1000\}$	$\{0, 2\}$	$\{0, 2\}$

A variable is labelled *causal* if from the same generating process as the response $\mathbf{Y}$. Variables are randomly selected from within their *causal* or *non-causal* groups.

4.4.2 Simulation Results

Simulation results are split by Granger causal designation and process of the chosen response variable. The control for false positive results and the ability to detect a Granger causal connection is examined in each setting. AUROC is not considered as it does not implicate a decision rule a priori. Tables 4.4 and 4.5 list the proportion of correctly identified causal relationships ρ_1 for TAR(2) and Lorenz-96 response variables, respectively. Results for the finer grid of designations from Table 4.3 are given in Appendix C.

Table 4.4: TAR(2) response compiled simulation results.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	0.915	0.870	0.155	0.555	0.752	0.989
$T = 500$	0.965	0.852	0.132	0.558	0.842	0.998
$T = 1000$	0.985	0.850	0.122	0.500	0.850	1.000

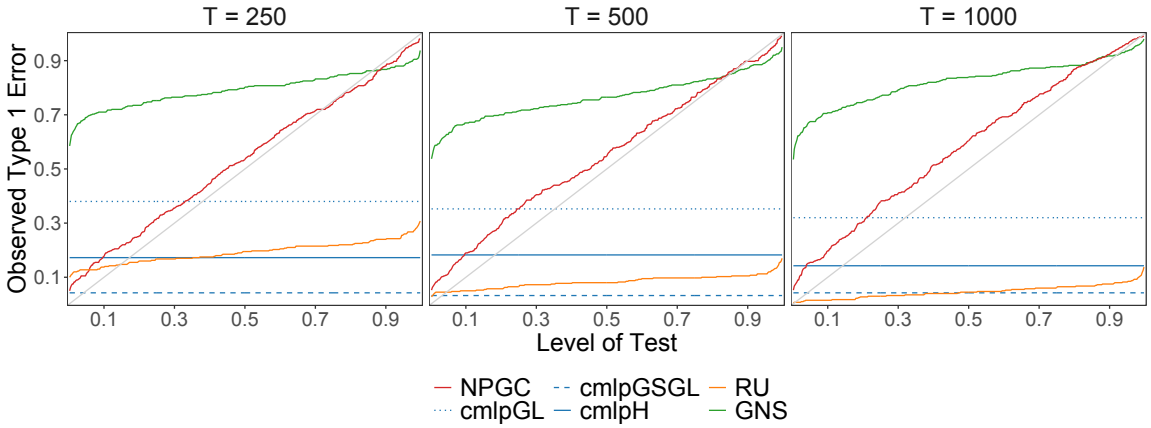
Proportion of correctly labelled Granger causal outcomes (GC $= 1$, ρ_1 in Table 4.2).

Figures 4.1 and 4.2 plot the observed Type 1 error by specified level of test, with a target 45 degree line included for reference. The NPGC method succeeds in identifying many cases of Granger causal influence in both processes, and as expected,

Table 4.5: Lorenz-96 response compiled simulation results.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	0.978	0.765	0.232	0.482	0.025	1.000
$T = 500$	0.988	0.728	0.160	0.442	0.001	1.000
$T = 1000$	0.992	0.715	0.128	0.412	0.020	1.000

Proportion of correctly labelled Granger causal outcomes (GC = 1, ρ_1 in Table 4.2).

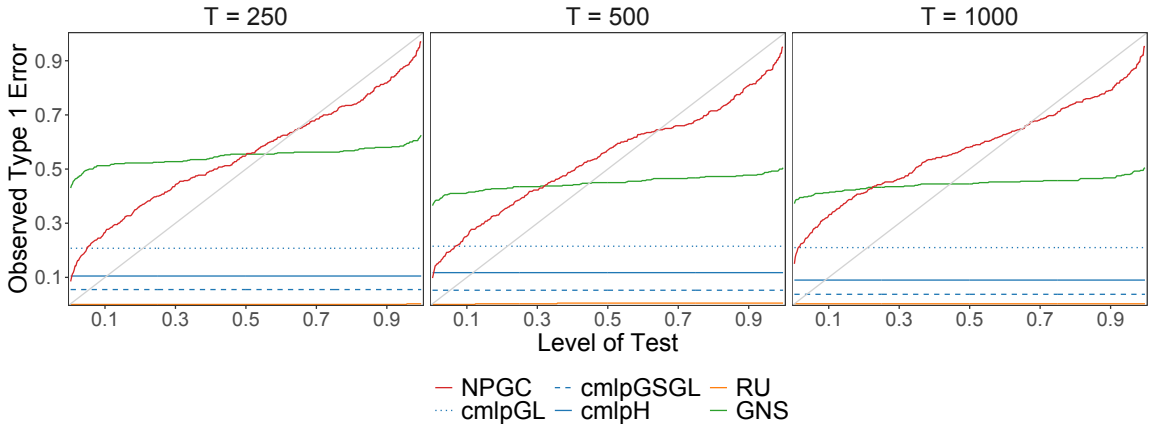


Methods that adhere to a chosen level of test will lie closer to the uniform CDF (gray) included for clarity that indicates the reference quantile α.

Figure 4.1: Type 1 error control for TAR(2) simulations.

this ability approaches the upper limit as the number of time points increases. The cMLP methods provide spotty results, and these are greatly influenced by the selected parameter λ. Without guidance for a specific penalty selection, the global decision of "Granger causal" or "non-causal" is uncertain. The naive restricted and unrestricted model method fails to identify many Granger causal pairings in the Lorenz-96 trials. The Gaussian noise substitution method correctly identifies nearly all Granger causal pairs, but this comes at the cost of uncontrolled Type 1 error (see Figures 4.1 and 4.2).

Under the null hypothesis, moderate adherence to the asymptotic properties shown



Methods that adhere to a chosen level of test will lie closer to the uniform CDF (gray) included for clarity that indicates the reference quantile α.

Figure 4.2: Type 1 error control for Lorenz-96 simulations.

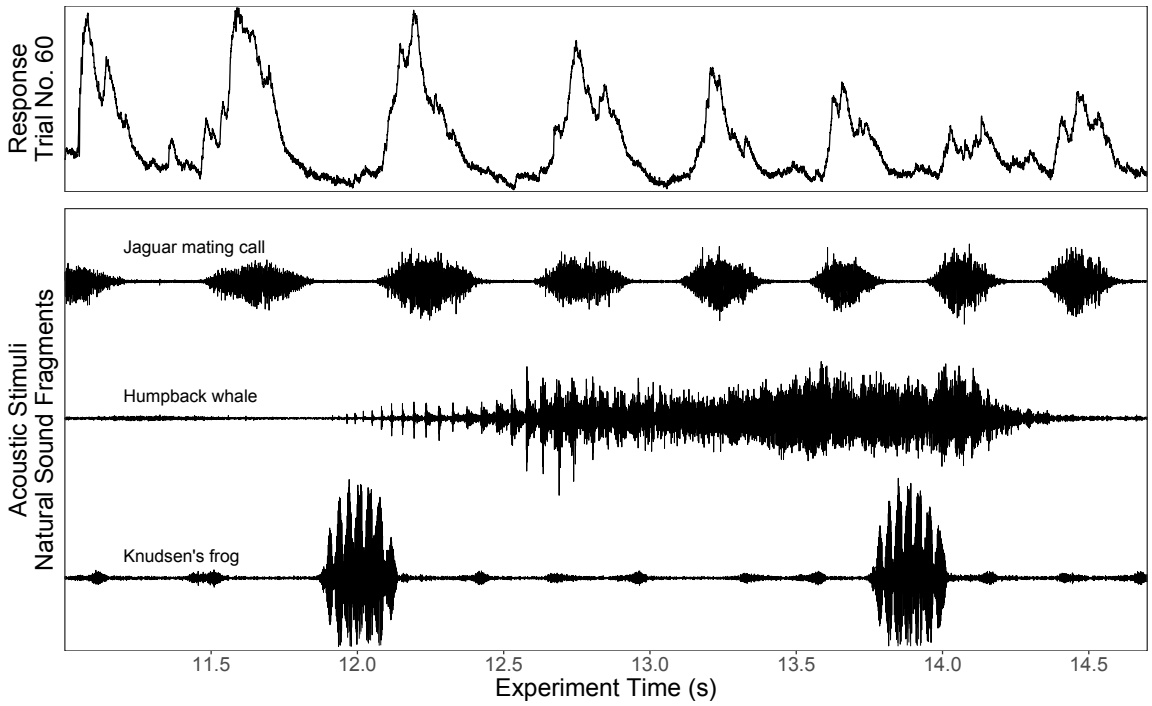
in Theorem 4.15 is demonstrated. The NPGC permutation methodology tracks relatively close to the included reference CDFs in Figure 4.1 for TAR(2) simulation settings, and exhibits mild to moderate deviations for Lorenz-96 settings in Figure 4.2. These deviations can likely be attributed to the complicated nature of the chaotic Lorenz system; the same pattern is not observed in other simulations when both groups are generated by a TAR(2) process (see Figure C.3 in the Appendix C). The cMLP lasso methods are included as horizontal lines in these plots; they do not have a straightforward extension for a specific level of test, other than computationally expensive iteration over several penalties λ. The naive methods do not adhere to their theoretical Type 1 error control, but approaches in the noise substitution realm appear promising if this issue can be resolved.

Granger causal identification methods that use a penalized objective can be useful, but do not allow for the global decision to label a group of variables "Granger causal" or "non-causal". These methods should be applied to narrow the scope of research to individual variables of the covariate set after a global method is applied.

4.5 Application Study

The out-of-sample permutation framework is applied to neuronal responses to acoustic stimuli in the primary auditory cortex of an anesthetized (ketamine–medetomidine) rat. The data, taken from the Collaborative Research in Computational Neuroscience data sharing website, consists of in vivo whole-cell recordings (mV) sampled at 4kHz in response to natural sound fragments (Machens, Wehr, and Zador, 2004; Asari et al., 2009). A time region of interest between 11 and 15 seconds is isolated from experimental trial 60 of the 050802mw03 data (partially shown in Figures 2D and 2E of Machens, Wehr, and Zador (2004)). The subset of data corresponds to the recordings in response to a jaguar (*Panthera onca*) mating call sound fragment presented at 97.656 kHz; this is resampled to a frequency of 4kHz matching that of the response. Additional sound fragments of a Humpback whale (*Megaptera novaeangliae*) and Knudsen's frog (*Leptodactylus knudseni*), played in other trials throughout the experiment, are included to construct a simplistic, non-causal scenario. Figure 4.3 displays the whole-cell recordings to the jaguar mating call and all sound fragments examined. Machens, Wehr, and Zador (2004) contains additional detail on data collection and experimental methods.

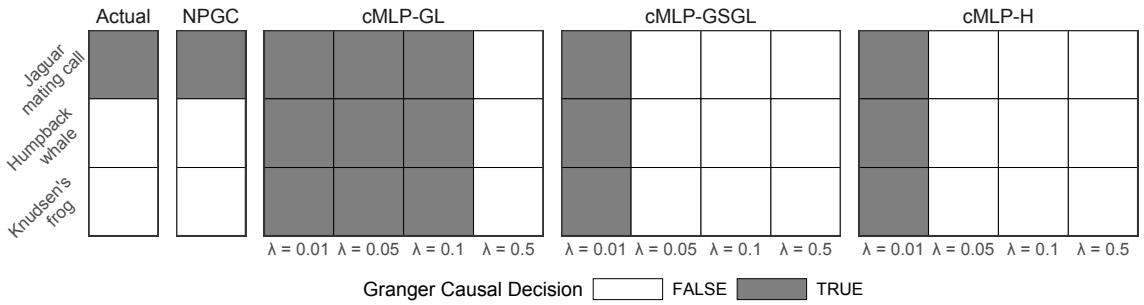
Performance of the Granger causal detection methodology is evaluated by formulating a farcical example in which any suitable method should be able to recover the correct causal structure. The methods are expected to flag the jaguar sound fragment as causal, while correctly labelling the other two included fragments non-causal. Lagged information from 30ms to 40ms is isolated for inclusion, consistent with the measured the half-maximal synaptic conductance in Wehr and Zador (2003). One acoustic stimulus fragment is chosen as the covariate set $\mathbf{X}$ and the others are ad-



Top: In vivo whole-cell recordings (mV) from the contralateral (left) primary auditory cortex of an anesthetized rat. *Bottom*: Jaguar mating call acoustic stimulus fragment resampled at 4kHz. Humpback whale and Knudsen's frog acoustic stimuli are included as non-causal examples for the methodology.

Figure 4.3: Responses to acoustic stimuli in the primary auditory cortex of an anesthetized rat.

ditional variables $\mathbf{Z}$ in the model. Lagged values of the response are included up to 10ms. The NPGC method is implemented with $M = 400$ permutations, $K = 5$ cross validation folds, $\mathcal{R} = 50$ randomly generated FNNs, and a feature dimension of $N = 250$ due to the large number of covariates. The comparator penalized optimization methods are examined at the same dimension and a variety of penalty values. The selection of the penalty λ presents a major challenge, and this value must be chosen prior to analysis; it should not be fine-tuned after the fact to produce a desired (already known) result. Results for the fabricated, naive scenario are compiled in Figure 4.4. Permutation based methodology extracts the correct relationship between the jaguar mating call sound fragment and the whole-cell recording response,



NPGC quantiles $\hat{Q}_M = 0.0025, 1.000, 0.4425$, from top to bottom. The NPGC method recovers the correct causal structure, while the lasso-based cMLP methods fail to isolate the jaguar stimulus.

Figure 4.4: Estimated Granger causal relationship between each acoustic stimulus and primary auditory cortex response.

and the penalized variable selection approaches do not.

4.6 Discussion

The NPGC methodology illustrates the use of the permutation-based shift from in-sample to out-of-sample testing. Permutation tests are used widely in literature, and Nauta, Bucur, and Seifert (2019) implement a decision rule similar to NPGC. This chapter explicitly defines Granger causality in this framework, and it provides a theoretical analysis of the corresponding variance estimates and decision rule. The shift allows for control in identifying useful functional relationships when overfitting from an artificial network becomes a concern, and removes the burden of specifying regularization parameters that have a major impact on the observed outcome. Once a Granger causal outcome has been determined, penalized variable selection techniques provide a practical screening method for identifying the prominent relationships, but they do not effectively estimate the global presence or absence of functional connectivity.

The permutation framework is able to detect the presence of functional connectivity, and provide a safeguard against misidentification of data-specific noise as functional dependence when the artificial network is overfit. Alternative formulations of the featurizing function Ψ in Equation 4.7 can utilize the same methodological structure and retain the theoretical guarantees provided the transformation meets the required conditions listed at the beginning of Section 4.3.1 and the construction allows for row-wise permutation without breaking the dependence structure across variables in $\mathbf{X}$. Misspecification of the dimension of the feature space should only produce additional false negative results as long as N is less than the number of data points used in model estimation. Using a large enough dimension to capture any potential nonlinear dependence structure that may exist in the dataset is suggested. The NPGC method may exhibit slight undercoverage for a chosen level of test, but *minor* deviations in this realm are not of great concern if the result is correctly interpreted as a potential functional connection and not an outright causal effect. The permutation method circumvents the need for a penalty parameter selection, and provides ease of extension to multiple testing problems and control of family-wise error rate.

Potential misuses of Granger causality include, but are not limited to, repeated application to subsets of a selected covariate group without adequate Type 1 error correction, disregard for the conditional nature of the inferential conclusion, attempts at individual rather than collective covariate inference (unless model specification is complete and exact), inclusion of a covariate without careful scientific or logical reasoning, and use as an outright mechanism for identifying causal relationships rather than predictive links for future study. Prudent selection of the length of lag response included in the covariate matrix is required, and two values may produce different

results. In nonlinear processes, selection of the relevant history of the response for inclusion in the model remains an area of future study.

Chapter 5

Conclusion

Representation learning harnesses the ability to uncover information about a dataset $\mathbf{Y}$ via a functional transformation Ψ, as defined in Equation 1.13. With careful choice of Ψ, nonlinear features of the data become linearly separable, and methods can examine the transformed data $\mathbf{H}$ to glean information about the behavior of the original process. This idea is broadly applicable to many areas of research including including classification, natural language processing, signal processing, transfer learning, finance, economics, and biology applications (Bishop, 2006; Bengio, Courville, and Vincent, 2013; LeCun, Bengio, and Hinton, 2015; Goodfellow, Bengio, and Courville, 2016).

This dissertation focuses on using ANNs as universal function approximators to transform data of interest $\mathbf{Y}$ into a representative set of network states $\mathbf{H}$ (Cybenko, 1989; Hornik, 1991). Featurization by an ANN $\mathcal{N}$ escapes the burden of specifying a "close enough" functional form for nonlinear temporal dependence in multivariate time series, and linear techniques are applicable to the transformed network states. Chapters 2 and 3 employ ESNs to identify change points in arbitrarily dependent data by examining the relationships between the hidden state vectors in regions of time, and Chapter 4 uses a simple MLP to extract information about functional relationships in multivariate time series as it relates to Granger causality.

Future work can extend methodology to classification of temporal regimes in a

multivariate time series, allow for inspection of functional relationships in the presence of generic (non-Gaussian) and autocorrelated errors, retain covariate specific information streams for interpretability and individual inference, bolster the theoretical underpinnings of the artificial network transformations, and explain the representative ability of the network states relative to chosen network architectures and specifications. In examination of arbitrarily dependent and multivariate time series data, representation learning is a promising strategy for prospective research when model recovery is not the main goal.

Bibliography

Amblard, Pierre Olivier and Olivier J. J. Michel (2011). "On directed information theory and Granger causality graphs". In: *Journal of Computational Neuroscience* 30.1, pp. 7–16. DOI: 10.1007/s10827-010-0231-x.

Anderson, Marti J. and Pierre Legendre (1999). "An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model". In: *Journal of statistical computation and simulation* 62.3, pp. 271–303. DOI: 10.1080/00949659908811936.

Anderson, Marti J. and John Robinson (2001). "Permutation tests for linear models". In: *Australian & New Zealand Journal of Statistics* 43.1, pp. 75–88. DOI: 10.1111/1467-842X.00156.

Arlot, Sylvain, Alain Celisse, and Zaid Harchaoui (2019). "A kernel multiple change-point algorithm via model selection". In: *Journal of Machine Learning Research* 20.162, pp. 1–56.

Asari, Hiroki et al. (2009). *Auditory cortex and thalamic neuronal responses to various natural and synthetic sounds*. Collaborative Research in Computational Neuroscience. DOI: 10.6080/K0KW5CXR.

Bai, Jushan (1997a). "Estimating Multiple Breaks One at a Time". In: *Econometric Theory* 13.3, pp. 315–352. DOI: 10.1017/S0266466600005831.

— (Nov. 1997b). "Estimation of a Change Point in Multiple Regression Models". In: *The Review of Economics and Statistics* 79.4, pp. 551–563. DOI: 10.1162/003465397557132.

Bao, Jiao et al. (2016). "Action recognition based on conceptors of skeleton joint trajectories". In: *Revista Facultad de Ingenieria* 31.4, pp. 11–22. DOI: 10.21311/002.31.4.02.

Barnett, Lionel and Anil K. Seth (2011). "Behaviour of Granger causality under filtering: Theoretical invariance and practical application". In: *Journal of Neuroscience Methods* 201.2, pp. 404–419. DOI: 10.1016/j.jneumeth.2011.08.010.

Barron, A.R. (1993). "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information Theory* 39.3, pp. 930–945. DOI: 10.1109/18.256500.

Baum, Leonard E. and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6, pp. 1554–1563. DOI: 10.1214/aoms/1177699147.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2, pp. 157–166. DOI: 10.1109/72.279181.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B* 57.1, pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Berkes, István, Siegfried Hörmann, and Johannes Schauer (2009). "Asymptotic results for the empirical process of stationary sequences". In: *Stochastic Processes and their Applications* 119.4, pp. 1298–1324. DOI: 10.1016/j.spa.2008.06.010.

Bernanke, Ben S. (Oct. 1990). *The federal funds rate and the channels of monetary transnission*. Working Paper 3487. National Bureau of Economic Research. DOI: 10.3386/w3487.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Biswas, Archishman and Hernando Ombao (2022). "Frequency-Specific Non-Linear Granger Causality in a Network of Brain Signals". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1401–1405. DOI: 10.1109/ICASSP43922.2022.9746794.

Black, Fischer and Myron Scholes (1973). "The Pricing of Options and Corporate Liabilities". In: *Journal of Political Economy* 81.3, pp. 637–654. DOI: 10.1086/260062.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1.

Bonferroni, Carlo (1936). "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, pp. 3–62.

Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24, pp. 123–140.

Bressler, Steven L. and Anil K. Seth (2011). "Wiener–Granger Causality: A well established methodology". In: *NeuroImage* 58.2, pp. 323–329. DOI: 10.1016/j.neuroimage.2010.02.059.

Bühlmann, Peter and Hans R. Künsch (1999). "Block length selection in the bootstrap for time series". In: *Computational Statistics & Data Analysis* 31.3, pp. 295–310. DOI: 10.1016/S0167-9473(99)00014-6.

Cappello, Lorenzo, Oscar Hernan Madrid Padilla, and Julia A. Palacios (2023). "Bayesian change point detection with spike and slab priors". In: *Journal of Computational and Graphical Statistics*. DOI: 10.1080/10618600.2023.2182312.

Caron, François and Emily B. Fox (Sept. 2017). "Sparse Graphs Using Exchangeable Random Measures". In: *Journal of the Royal Statistical Society: Series B* 79.5, pp. 1295–1366. DOI: 10.1111/rssb.12233.

Chan, Kung Sik and Howell Tong (1986). "On estimating thresholds in autoregressive models". In: *Journal of time series analysis* 7.3, pp. 179–190. DOI: 10.1111/j.1467-9892.1986.tb00501.x.

Chao, John, Valentina Corradi, and Norman R. Swanson (2001). "Out-of-sample tests for Granger causality". In: *Macroeconomic Dynamics* 5.4, pp. 598–620. DOI: 10.1017/S1365100501023070.

Chen, Fang et al. (2021). "Machine learning in/for blockchain: Future and challenges". In: *Canadian Journal of Statistics* 49.4, pp. 1364–1382. DOI: 10.1002/cjs.11623.

Cho, Haeran and Piotr Fryzlewicz (2012). "Multiscale and multilevel technique for consistent segmentation of nonstationary time series". In: *Statistica Sinica* 22.1, pp. 207–229. DOI: 10.5705/ss.2009.280.

— (2015). "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation". In: *Journal of the Royal Statistical Society: Series B* 77.2, pp. 475–507. DOI: 10.1111/rssb.12079.

Cox Jr., Louis Anthony Tony and Douglas A. Popken (2015). "Has reducing fine particulate matter and ozone caused reduced mortality rates in the United States?"

In: *Annals of epidemiology* 25.3, pp. 162–173. DOI: 10.1016/j.annepidem.2014.11.006.

Csörgő, Miklós and Lajos Horváth (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Csörgő, Miklós, Lajos Horváth, and Barbara Szyszkowicz (1997). "Integral tests for suprema of Kiefer processes with application". In: *Statistics & Risk Modeling* 15.4, pp. 365–378. DOI: 10.1524/strm.1997.15.4.365.

Csörgő, Miklós and Barbara Szyszkowicz (1994a). "Applications of multi-time parameter processes to change-point analysis". In: *Probability Theory and Mathematical Statistics: Proceedings of the Sixth Vilnius Conference*. VSP/TEV, pp. 159–222.

— (1994b). "Weighted multivariate empirical processes and contiguous change-point analysis". In: *IMS Lecture Notes Monograph Series* 23.1, pp. 93–98. DOI: 10.1214/lnms/1215463116.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4, pp. 303–314. DOI: 10.1007/BF02551274.

Dehling, Herold, Roland Fried, et al. (2015). "Change-Point Detection Under Dependence Based on Two-Sample U-Statistics". In: *Asymptotic Laws and Methods in Stochastics: A Volume in Honour of Miklós Csörgő*. Ed. by Donald Dawson et al. New York, NY: Springer New York, pp. 195–220. DOI: 10.1007/978-1-4939-3076-0_12.

Dehling, Herold and Walter Philipp (2002). "Empirical process techniques for dependent data". In: *Empirical process techniques for dependent data*. Birkhäuser, pp. 3–113. DOI: 10.1007/978-1-4612-0099-4.

Dehling, Herold, Kata Vuk, and Martin Wendler (2022). "Change-point detection based on weighted two-sample U-statistics". In: *Electronic Journal of Statistics* 16.1, pp. 862–891. DOI: 10.1214/21-EJS1964.

Del Moral, Pierre (1997). "Nonlinear filtering: Interacting particle resolution". In: *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* 325.6, pp. 653–658. DOI: 10.1016/S0764-4442(97)84778-7.

Dette, Holger, Theresa Eckle, and Mathias Vetter (2020). "Multiscale change point detection for dependent data". In: *Scandinavian Journal of Statistics* 47.4, pp. 1243–1274. DOI: 10.1111/sjos.12465.

Dette, Holger and Josua Gösmann (2020). "A Likelihood Ratio Approach to Sequential Change Point Detection for a General Class of Parameters". In: *Journal of the American Statistical Association* 115.531, pp. 1361–1377. DOI: 10.1080/01621459.2019.1630562.

Dette, Holger and Dominik Wied (May 2015). "Detecting Relevant Changes in Time Series Models". In: *Journal of the Royal Statistical Society: Series B* 78.2, pp. 371–394. DOI: 10.1111/rssb.12121.

Dette, Holger, Weichi Wu, and Zhou Zhou (2019). "Change point analysis of correlation in non-stationary time series". In: *Statistica Sinica* 29.2, pp. 611–643. DOI: 10.5705/ss.202016.0493.

Dey, Asim K. et al. (2020). "On the role of local blockchain network features in cryptocurrency price formation". In: *Canadian Journal of Statistics* 48.3, pp. 561–581. DOI: 10.1002/cjs.11547.

DiCiccio, Cyrus J. and Joseph P. Romano (2017). "Robust permutation tests for correlation and regression coefficients". In: *Journal of the American Statistical Association* 112.519, pp. 1211–1220. DOI: 10.1080/01621459.2016.1202117.

Duggento, Andrea, Maria Guerrisi, and Nicola Toschi (2021). "Echo state network models for nonlinear granger causality". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2212, p. 20200256. DOI: 10.1098/rsta.2020.0256.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552.

Einstein, Albert et al. (1905). "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat". In: *Annals of Physics* 322.8, pp. 549–560. DOI: 10.1002/andp.19053220806.

El Hihi, Salah and Yoshua Bengio (1995). "Hierarchical recurrent neural networks for long-term dependencies". In: *NIPS'95: Proceedings of the 8th International Conference on Neural Information Processing Systems*, pp. 493–499.

Engle, Robert F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica*, pp. 987–1007. DOI: 10.2307/1912773.

Florens, Jean Pierre and Michel Mouchart (1982). "A note on noncausality". In: *Econometrica* 50.3, pp. 583–591. DOI: 10.2307/1912602.

Fokker, Adriaan Daniël (1914). "Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld". In: *Annals of Physics* 348.5, pp. 810–820. DOI: 10.1002/andp.19143480507.

Foster, Dean P and Robert A Stine (2008). "α-investing: a procedure for sequential control of expected false discoveries". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.2, pp. 429–444.

Frick, Klaus, Axel Munk, and Hannes Sieling (May 2014). "Multiscale Change Point Inference". In: *Journal of the Royal Statistical Society: Series B* 76.3, pp. 495–580. DOI: 10.1111/rssb.12047.

Friston, Karl J. (1994). "Functional and effective connectivity in neuroimaging: a synthesis". In: *Human brain mapping* 2.1-2, pp. 56–78. DOI: 10.1002/hbm.460020107.

Friston, Karl J., Rosalyn Moran, and Anil K. Seth (2013). "Analysing connectivity with Granger causality and dynamic causal modelling". In: *Current Opinion in Neurobiology* 23.2, pp. 172–178. DOI: 10.1016/j.conb.2012.11.010.

Fryzlewicz, Piotr (2014). "Wild binary segmentation for multiple change-point detection". In: *The Annals of Statistics* 42.6, pp. 2243–2281. DOI: 10.1214/14-AOS1245.

— (2018). "Tail-greedy bottom-up data decompositions and fast multiple change-point detection". In: *The Annals of Statistics* 46.6B, pp. 3390–3421. DOI: 10.1214/17-AOS1662.

Gade, Noah D. and Jordan Rodu (2023a). "Change Point Detection With Conceptors". In: *arXiv preprint arXiv:2308.06213*. DOI: 10.48550/arXiv.2308.06213.

— (2023b). "Nonlinear Permuted Granger Causality". In: *arXiv preprint arXiv:2308.06220*. DOI: 10.48550/arXiv.2308.06220.

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins (2000). "Learning to forget: Continual prediction with LSTM". In: *Neural computation* 12.10, pp. 2451–2471. DOI: 10.1162/089976600300015015.

Gerstenberger, Carina (2018). "Robust Wilcoxon-Type Estimation of Change-Point Location Under Short-Range Dependence". In: *Journal of Time Series Analysis* 39, pp. 90–104. DOI: 10.1111/jtsa.12268.

Geweke, John (1982). "Measurement of linear dependence and feedback between multiple time series". In: *Journal of the American Statistical Association* 77.378, pp. 304–313. DOI: 10.1080/01621459.1982.10477803.

Gombay, Edit and Lajos Horváth (1995). "An application of U-statistics to change-point analysis". In: *Acta Scientiarum Mathematicarum* 60.1, pp. 345–358.

Gombay, Edit and Lajos Horváth (1999). "Change-points and bootstrap". In: *Environmetrics* 10.6, pp. 725–736. DOI: 10.1002/(SICI)1099-095X(199911/12)10:6<725::AID-ENV387>3.0.CO;2-K.

Gonon, Lukas and Juan Pablo Ortega (2021). "Fading memory echo state networks are universal". In: *Neural Networks* 138, pp. 10–13. DOI: 10.1016/j.neunet.2021.01.025.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press.

Gösmann, Josua, Tobias Kley, and Holger Dette (2021). "A new approach for open-end sequential change point monitoring". In: *Journal of Time Series Analysis* 42.1, pp. 63–84. DOI: 10.1111/jtsa.12555.

Gösmann, Josua, Christina Stoehr, et al. (2022). "Sequential change point detection in high dimensional time series". In: *Electronic Journal of Statistics* 16.1, pp. 3608–3671. DOI: 10.1214/22-EJS2027.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438. DOI: 10.2307/1912791.

— (1980). "Testing for causality: A personal viewpoint". In: *Journal of Economic Dynamics and Control* 2, pp. 329–352. DOI: 10.1016/0165-1889(80)90069-X.

— (1988). "Some recent development in a concept of causality". In: *Journal of Econometrics* 39.1, pp. 199–211. DOI: 10.1016/0304-4076(88)90045-0.

Grigoryeva, Lyudmila and Juan Pablo Ortega (2018). "Echo state networks are universal". In: *Neural Networks* 108, pp. 495–508. DOI: 10.1016/j.neunet.2018.08.025.

Guliyev, Namig J. and Vugar E. Ismailov (2018a). "Approximation capability of two hidden layer feedforward neural networks with fixed weights". In: *Neurocomputing* 316, pp. 262–269. DOI: 10.1016/j.neucom.2018.07.075.

— (2018b). "On the approximation by single hidden layer feedforward neural networks with fixed weights". In: *Neural Networks* 98, pp. 296–304. DOI: 10.1016/j.neunet.2017.12.007.

Gupta, Arjun K. and Daya K. Nagar (2018). *Matrix variate distributions*. Chapman and Hall/CRC. DOI: 10.1201/9780203749289.

Hall, Peter, Joel L. Horowitz, and Bing Yi Jing (1995). "On blocking rules for the bootstrap with dependent data". In: *Biometrika* 82.3, pp. 561–574. DOI: 10.1093/biomet/82.3.561.

Hart, Allen, James Hook, and Jonathan Dawes (2020). "Embedding and approximation theorems for echo state networks". In: *Neural Networks* 128, pp. 234–247. DOI: 10.1016/j.neunet.2020.05.013.

Haynes, Kaylea, Idris A. Eckley, and Paul Fearnhead (2017). "Computationally Efficient Changepoint Detection for a Range of Penalties". In: *Journal of Computational and Graphical Statistics* 26.1, pp. 134–143. DOI: 10.1080/10618600.2015.1116445.

Haynes, Kaylea, Paul Fearnhead, and Idris A. Eckley (2017). "A computationally efficient nonparametric approach for changepoint detection". In: *Statistics and Computing* 27.5, pp. 1293–1305. DOI: 10.1007/s11222-016-9687-5.

Henderson, James and George Michailidis (Apr. 2014). "Network Reconstruction Using Nonparametric Additive ODE Models". In: *PLOS ONE* 9.4, pp. 1–15. DOI: 10.1371/journal.pone.0094003.

Himdi, Khalid El and Roch Roy (1997). "Tests for noncorrelation of two multivariate ARMA time series". In: *Canadian Journal of Statistics* 25.2, pp. 233–256. DOI: 10.2307/3315734.

Hochberg, Yosef (1988). "A sharper Bonferroni procedure for multiple tests of significance". In: *Biometrika* 75.4, pp. 800–802.

Hochreiter, Sepp (1991). "Untersuchungen zu dynamischen neuronalen Netzen". In: *Diploma, Technische Universität München* 91.1, p. 31.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735..

Holland, Paul W. (1986). "Statistics and causal inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960. DOI: 10.1080/01621459.1986.10478354.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics*, pp. 65–70.

Holmes, Mark and Ivan Kojadinovic (2021). "Open-end nonparametric sequential change-point detection based on the retrospective CUSUM statistic". In: *Electronic Journal of Statistics* 15.1, pp. 2288–2335. DOI: 10.1214/21-EJS1840.

Holmes, Mark, Ivan Kojadinovic, and Jean François Quessy (2013). "Nonparametric tests for change-point detection à la Gombay and Horváth". In: *Journal of Multivariate Analysis* 115, pp. 16–32. DOI: 10.1016/j.jmva.2012.10.004.

Hooker, Reginald H. (1901). "Correlation of the marriage-rate with trade". In: *Journal of the Royal Statistical Society* 64.3, pp. 485–492. DOI: 10.1111/j.2397-2335.1901.tb03810.x.

Hoon Song, Chang et al. (2023). "Minimal Width for Universal Property of Deep RNN". In: *Journal of Machine Learning Research* 24.121, pp. 1–41.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2, pp. 251–257. DOI: 10.1016/0893-6080(91)90009-T.

Horvath, Samuel, Malik Shahid Sultan, and Hernando Ombao (2022). "Granger causality using neural networks". In: *arXiv preprint arXiv:2208.03703*. DOI: 10.48550/arXiv.2208.03703.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2.1, pp. 193–218. DOI: 10.1007/BF01908075.

Hušková, Marie (2004). "Permutation principle and bootstrap in change point analysis". In: *Fields Institute Communications* 44, pp. 273–291. DOI: 10.1090/fic/044/15.

Inoue, Atsushi and Lutz Kilian (2005). "In-sample or out-of-sample tests of predictability: Which one should we use?" In: *Econometric Reviews* 23.4, pp. 371–402. DOI: 10.1081/ETC-200040785.

Ivanov, Ventzislav and Lutz Kilian (2005). "A practitioner's guide to lag order selection for VAR impulse response analysis". In: *Studies in Nonlinear Dynamics & Econometrics* 9.1, p. 2. DOI: 10.2202/1558-3708.1219.

Jaeger, Herbert (2001). *The "echo state" approach to analysing and training recurrent neural networks-with an erratum note*. GMD Technical Report 148. German National Research Center for Information Technology, p. 13.

— (2002). "Adaptive nonlinear system identification with echo state networks". In: *Advances in Neural Information Processing Systems*. Vol. 15, pp. 609–616.

— (2007). "Echo State Network". In: *Scholarpedia* 2.9, p. 2330. DOI: 10.4249/scholarpedia.2330.

— (Mar. 2014). *Controlling Recurrent Neural Networks by Conceptors*. Tech. rep. 31. Jacobs University Bremen.

Jaeger, Herbert (2017). "Using Conceptors to Manage Neural Long-Term Memories for Temporal Patterns". In: *Journal of Machine Learning Research* 18, pp. 1–43.

James, Nicholas A. and David S. Matteson (2015). "ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data". In: *Journal of Statistical Software* 62.7, pp. 1–25. DOI: 10.18637/jss.v062.i07.

Kalman, Rudolph Emil (1960). "A new approach to linear filtering and prediction problems". In: *Journal of Fluids Engineering* 82.1, pp. 35–45. DOI: 10.1115/1.3662552.

Karimi, Alireza and Mark R. Paul (2010). "Extensive chaos in the Lorenz-96 model". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.4, p. 043105. DOI: 10.1063/1.3496397.

Khanna, Saurabh and Vincent Y. F. Tan (2019). "Economy statistical recurrent units for inferring nonlinear granger causality". In: *arXiv preprint arXiv:1911.09879*. DOI: 10.48550/arXiv.1911.09879.

Kiefer, Chris (2019). "Sample-level sound synthesis with recurrent neural networks and conceptors". In: *PeerJ Computer Science* 5, e205. DOI: 10.7717/peerj-cs.205.

Kiefer, Jack (1972). "Skorohod embedding of multivariate RV's, and the sample DF". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 24.1, pp. 1–35. DOI: 10.1007/BF00532460.

Killick, R., Paul Fearnhead, and Idris A. Eckley (2012). "Optimal Detection of Changepoints With a Linear Computational Cost". In: *Journal of the American Statistical Association* 107.500, pp. 1590–1598. DOI: 10.1080/01621459.2012.737745.

Kirch, Claudia, Birte Muhsal, and Hernando Ombao (2015). "Detection of changes in multivariate time series with application to EEG data". In: *Journal of the*

American Statistical Association 110.511, pp. 1197–1216. DOI: 10.1080/01621459.2014.957545.

Kojadinovic, Ivan and Ghislain Verdier (2021). "Nonparametric sequential change-point detection for multivariate time series based on empirical distribution functions". In: *Electronic Journal of Statistics* 15.1, pp. 773–829. DOI: 10.1214/21-EJS1798.

Kolmogorov, Andrei (1931). "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung". In: *Mathematische Annalen* 104, pp. 415–458. DOI: 10.1007/BF01457949.

Korkas, Karolos K. and Piotr Fryzlewicz (2017). "Multiple change-point detection for non-stationary time series using wild binary segmentation". In: *Statistica Sinica* 27.1, pp. 287–311. DOI: 10.5705/ss.202015.0262.

Kunsch, Hans R. (1989). "The jackknife and the bootstrap for general stationary observations". In: *The Annals of Statistics* 17.3, pp. 1217–1241. DOI: 10.1214/aos/1176347265.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Series in Statistics. Springer Science & Business Media. DOI: 10.1007/978-1-4757-3803-2.

Lahiri, S. N., K. Furukawa, and Y. D. Lee (2007). "A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods". In: *Statistical methodology* 4.3, pp. 292–321. DOI: 10.1016/j.stamet.2006.08.002.

Lai, Tze Leung (1995). "Sequential changepoint detection in quality control and dynamical systems". In: *Journal of the Royal Statistical Society: Series B* 57.4, pp. 613–644. DOI: 10.1111/j.2517-6161.1995.tb02052.x.

Langevin, Paul (1908). "Sur la théorie du mouvement brownien". In: *Comptes-rendus de l'Académie des sciences* 146, pp. 530–533.

Lau, Tze Siong, Wee Peng Tay, and Venugopal V. Veeravalli (2019). "A Binning Approach to Quickest Change Detection With Unknown Post-Change Distribution". In: *IEEE Transactions on Signal Processing* 67.3, pp. 609–621. DOI: 10.1109/TSP.2018.2881666.

Lavielle, Marc and Gilles Teyssiere (2006). "Detection of multiple change-points in multivariate time series". In: *Lithuanian Mathematical Journal* 46.3, pp. 287–306. DOI: 10.1007/s10986-006-0028-9.

LeCun, Yann, Yoshua Bengio, and Geoffrey E. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539.

Li, Housen, Qinghai Guo, and Axel Munk (2019). "Multiscale change-point segmentation: beyond step functions". In: *Electronic Journal of Statistics* 13.2, pp. 3254–3296. DOI: 10.1214/19-EJS1608.

Li, Housen, Axel Munk, and Hannes Sieling (2016). "FDR-control in multiscale change-point segmentation". In: *Electronic Journal of Statistics* 10.1, pp. 918–959. DOI: 10.1214/16-EJS1131.

Li, Shuang et al. (2019). "Scan B-statistic for kernel change-point detection". In: *Sequential Analysis* 38.4, pp. 503–544. DOI: 10.1080/07474946.2019.1686886.

Lin, Tsungnan et al. (1996). "Learning long-term dependencies in NARX recurrent neural networks". In: *IEEE transactions on neural networks* 7.6, pp. 1329–1338. DOI: 10.1109/72.548162.

Liu, Jun S. and Rong Chen (1998). "Sequential Monte Carlo methods for dynamic systems". In: *Journal of the American statistical association* 93.443, pp. 1032–1044. DOI: 10.1080/01621459.1998.10473765.

Liu, Regina Y. and Kesar Singh (1992). "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence". In: *Exploring the Limits of the Bootstrap*. Ed. by R. Lepage and L. Billard. Wiley, New York.

Lorden, G. (1971). "Procedures for Reacting to a Change in Distribution". In: *The Annals of Mathematical Statistics* 42.6, pp. 1897–1908. DOI: 10.1214/aoms/1177693055.

Lu, Zhou et al. (2017). "The expressive power of neural networks: A view from the width". In: *Advances in Neural Information Processing Systems*. Vol. 30.

Lukoševičius, Mantas (2012). "Neural Networks: Tricks of the Trade". In: ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus Robert Müller. 2nd ed. Vol. 7700. Lecture Notes in Computer Science. Springer. Chap. A Practical Guide to Applying Echo State Networks, pp. 659–686. DOI: 10.1007/978-3-642-35289-8.

Lukoševičius, Mantas and Herbert Jaeger (2009). "Reservoir computing approaches to recurrent neural network training". In: *Computer science review* 3.3, pp. 127–149. DOI: 10.1016/j.cosrev.2009.03.005.

Maass, Wolfgang, Thomas Natschläger, and Henry Markram (2002). "Real-time computing without stable states: A new framework for neural computation based on perturbations". In: *Neural Computation* 14.11, pp. 2531–2560. DOI: 10.1162/089976602760407955..

Machens, Christian K., Michael Wehr, and Anthony M. Zador (2004). "Linearity of cortical receptive fields measured with natural sounds". In: *Journal of Neuroscience* 24.5, pp. 1089–1100. DOI: 10.1523/JNEUROSCI.4445-03.2004.

Maiorov, Vitaly and Allan Pinkus (1999). "Lower bounds for approximation by MLP neural networks". In: *Neurocomputing* 25.1-3, pp. 81–91. DOI: 10.1016/S0925-2312(98)00111-8.

Marcinkevičs, Ričards and Julia E. Vogt (2021). "Interpretable Models for Granger Causality Using Self-explaining Neural Networks". In: *arXiv preprint arXiv:2101.07600*. DOI: 10.48550/arXiv.2101.07600.

Marinazzo, Daniele, Wei Liao, et al. (2011). "Nonlinear connectivity by Granger causality". In: *NeuroImage* 58.2, pp. 330–338. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2010.01.099.

Marinazzo, Daniele, Mario Pellicoro, and Sebastiano Stramaglia (2008). "Kernel method for nonlinear Granger causality". In: *Physical review letters* 100.14, p. 144103. DOI: 10.1103/PhysRevLett.100.144103.

Markov, Andrey Andreyevich (1906). "Extension of the law of large numbers to dependent quantities". In: *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)* 15.1, pp. 135–156.

Matteson, David S. and Nicholas A. James (2014). "A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data". In: *Journal of the American Statistical Association* 109.505, pp. 334–345. DOI: 10.1080/01621459.2013.849605.

Matthews, B.W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2, pp. 442–451. DOI: 10.1016/0005-2795(75)90109-9.

Maziarz, Mariusz (2015). "A review of the Granger-causality fallacy". In: *The journal of philosophical economics: Reflections on economic and social issues* 8.2, pp. 86–105.

McCulloch, Warren S. and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5, pp. 115–133. DOI: 10.1007/BF02478259.

McGonigle, Euan T. and Haeran Cho (2023). "Nonparametric data segmentation in multivariate time series via joint characteristic functions". In: *arXiv preprint arXiv:2305.07581*. DOI: 10.48550/arXiv.2305.07581.

Meinshausen, Nicolai and Peter Bühlmann (Aug. 2010). "Stability Selection". In: *Journal of the Royal Statistical Society: Series B* 72.4, pp. 417–473. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2010.00740.x.

Messer, Michael (2022). "Bivariate change point detection: Joint detection of changes in expectation and variance". In: *Scandinavian Journal of Statistics* 49.2, pp. 886–916. DOI: 10.1111/sjos.12547.

Moodie, Erica E. M. and David A. Stephens (2022). "Causal inference: Critical developments, past and future". In: *Canadian Journal of Statistics* 50.4, pp. 1299–1320. DOI: 10.1002/cjs.11718.

Moustakides, George V. (1986). "Optimal Stopping Times for Detecting Changes in Distributions". In: *The Annals of Statistics* 14.4, pp. 1379–1387. DOI: 10.1214/aos/1176350164.

Mozer, Michael C. (1991). "Induction of multiscale temporal structure". In: *Advances in Neural Information Processing Systems*. Vol. 4, pp. 275–282.

Nauta, Meike, Doina Bucur, and Christin Seifert (2019). "Causal discovery with attention-based convolutional neural networks". In: *Machine Learning and Knowledge Extraction* 1.1, pp. 312–340. DOI: 10.3390/make1010019.

Newey, Whitney K. (1991). "Uniform convergence in probability and stochastic equicontinuity". In: *Econometrica* 59.4, pp. 1161–1167. DOI: 10.2307/2938179.

Newey, Whitney K. and Daniel McFadden (1994). "Large sample estimation and hypothesis testing". In: *Handbook of econometrics* 4, pp. 2111–2245. DOI: 10.1016/S1573-4412(05)80005-4.

Ng, Serena and Pierre Perron (2001). "Lag length selection and the construction of unit root tests with good size and power". In: *Econometrica* 69.6, pp. 1519–1554. DOI: 10.1111/1468-0262.00256.

Nicholson, William B., David S. Matteson, and Jacob Bien (2017). "VARX-L: Structured regularization for large vector autoregressions with exogenous variables". In: *International Journal of Forecasting* 33.3, pp. 627–651. DOI: 10.1016/j.ijforecast.2017.01.003.

Niu, Yue S., Ning Hao, and Heping Zhang (2016). "Multiple Change-Point Detection: A Selective Overview". In: *Statistical Science* 31.4, pp. 611–623. DOI: 10.1214/16-STS587.

Oliva, Junier B., Barnabás Póczos, and Jeff Schneider (2017). "The Statistical Recurrent Unit". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2671–2680.

Ombao, Hernando, Rainer von Sachs, and Wensheng Guo (2005). "SLEX Analysis of Multivariate Nonstationary Time Series". In: *Journal of the American Statistical Association* 100.470, pp. 519–531. DOI: 10.1198/016214504000001448.

Padilla, Carlos Misael Madrid et al. (2023). "Change point detection and inference in multivariable nonparametric models under mixing conditions". In: *arXiv preprint arXiv:2301.11491*. DOI: 10.48550/arXiv.2301.11491.

Padilla, Oscar Hernan Madrid, Alex Athey, et al. (2019). "Sequential Nonparametric Tests for a Change in Distribution: An Application to Detecting Radiological Anomalies". In: *Journal of the American Statistical Association* 114.526, pp. 514–528. DOI: 10.1080/01621459.2018.1476245.

Padilla, Oscar Hernan Madrid, Yi Yu, et al. (2021). "Optimal nonparametric change point analysis". In: *Electronic Journal of Statistics* 15.1, pp. 1154–1201. DOI: 10.1214/21-EJS1809.

Page, E. S. (1954). "Continuous Inspection Schemes". In: *Biometrika* 41.1/2, pp. 100–115. DOI: 10.2307/2333009.

— (1955). “A Test for a Change in a Parameter Occurring at an Unknown Point”. In: *Biometrika* 42.3/4, pp. 523–527. DOI: 10.2307/2333401.

Patton, Andrew, Dimitris N. Politis, and Halbert White (2009). “Correction to “Automatic block-length selection for the dependent bootstrap” by D. Politis and H. White”. In: *Econometric Reviews* 28.4, pp. 372–375. DOI: 10.1080/07474930802459016.

Pein, Florian, Hannes Sieling, and Axel Munk (Aug. 2016). “Heterogeneous Change Point Inference”. In: *Journal of the Royal Statistical Society: Series B* 79.4, pp. 1207–1227. DOI: 10.1111/rssb.12202.

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (Oct. 2016). “Causal Inference by using Invariant Prediction: Identification and Confidence Intervals”. In: *Journal of the Royal Statistical Society: Series B* 78.5, pp. 947–1012. ISSN: 1369-7412. DOI: 10.1111/rssb.12167.

Picard, Dominique (1985). “Testing and estimating change-points in time series”. In: *Advances in Applied Probability* 17.4, pp. 841–867. DOI: 10.2307/1427090.

Planck, VM (1917). “Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie”. In: *Sitzungberichte der*.

Polikar, Robi (2012). “Ensemble Machine Learning: Methods and Applications”. In: ed. by Yunqian Ma Cha Zhang. Springer. Chap. Ensemble learning, pp. 1–34. DOI: 10.1007/978-1-4419-9326-7.

Politis, Dimitris N. and Halbert White (2004). “Automatic block-length selection for the dependent bootstrap”. In: *Econometric Reviews* 23.1, pp. 53–70. DOI: 10.1081/ETC-120028836.

Pollak, Moshe (1985). “Optimal Detection of a Change in Distribution”. In: *The Annals of Statistics* 13.1, pp. 206–227. DOI: 10.1214/aos/1176346587.

Reid, Andrew T et al. (2019). "Advancing functional connectivity research from association to causation". In: *Nature Neuroscience* 22.11, pp. 1751–1760. DOI: 10.1038/s41593-019-0510-4.

Ritov, Y. (1990). "Decision Theoretic Optimality of the Cusum Procedure". In: *The Annals of Statistics* 18.3, pp. 1464–1469. DOI: 10.1214/aos/1176347761.

Romano, Joseph P. and Michael Wolf (2005). "Exact and approximate stepdown methods for multiple hypothesis testing". In: *Journal of the American Statistical Association* 100.469, pp. 94–108.

Rosenblatt, Frank (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Report 1196-G-8. Cornell Aerunautical Laboratory.

Rovatsos, Georgios et al. (2017). "Statistical Power System Line Outage Detection Under Transient Dynamics". In: *IEEE Transactions on Signal Processing* 65.11, pp. 2787–2797. DOI: 10.1109/TSP.2017.2673802.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0.

Runge, Jakob et al. (2012). "Escaping the curse of dimensionality in estimating multivariate transfer entropy". In: *Physical review letters* 108.25, p. 258701. DOI: 10.1103/PhysRevLett.108.258701.

Schäfer, Anton Maximillian and Hans Georg Zimmerman (2007). "Recurrent Neural Networks are Universal Approximators". In: *International Journal of Neural Systems* 17.04, pp. 253–263. DOI: 10.1142/S0129065707001111.

Schrödinger, Erwin (1926). "An undulatory theory of the mechanics of atoms and molecules". In: *Physical review* 28.6, p. 1049. DOI: 10.1103/PhysRev.28.1049.

Schuster, Arthur (1898). "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena". In: *Terrestrial Magnetism* 3.1, pp. 13–41. DOI: 10.1029/TM003i001p00013.

— (1906a). "II. On the periodicities of sunspots". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 206.402-412, pp. 69–100. DOI: 10.1098/rsta.1906.0016.

— (1906b). "The periodogram and its optical analogy". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 77.515, pp. 136–140. DOI: 10.1098/rspa.1906.0011.

Seth, Anil K., Adam B. Barrett, and Lionel Barnett (2015). "Granger causality analysis in neuroscience and neuroimaging". In: *Journal of Neuroscience* 35.8, pp. 3293–3297. DOI: 10.1523/JNEUROSCI.4399-14.2015.

Shaffer, Juliet Popper (1986). "Modified sequentially rejective multiple test procedures". In: *Journal of the American Statistical Association* 81.395, pp. 826–831.

Shao, Xiaofeng and Xianyang Zhang (2010). "Testing for Change Points in Time Series". In: *Journal of the American Statistical Association* 105.491, pp. 1228–1240. DOI: 10.1198/jasa.2010.tm10103.

Shewhart, W. A. (1925). "The Application of Statistics as an Aid in Maintaining Quality of a Manufactured Product". In: *Journal of the American Statistical Association* 20.152, pp. 546–548. DOI: 10.1080/01621459.1925.10502930.

Shiryaev, A. N. (1963). "On Optimum Methods in Quickest Detection Problems". In: *Theory of Probability & Its Applications* 8.1, pp. 22–46. DOI: 10.1137/1108002.

Shojaie, Ali and Emily B. Fox (2022). "Granger Causality: A Review and Recent Advances". In: *Annual Review of Statistics and Its Application* 9.1, pp. 289–319. DOI: 10.1146/annurev-statistics-040120-010930.

Shojaie, Ali and George Michailidis (2010). "Discovering graphical Granger causality using the truncating lasso penalty". In: *Bioinformatics* 26.18, pp. i517–i523. DOI: 10.1093/bioinformatics/btq377.

Shumway, Robert H. and David S. Stoffer (2017). *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer International Publishing. DOI: 10.1007/978-3-319-52452-8.

Šidák, Zbyněk (1967). "Rectangular confidence regions for the means of multivariate normal distributions". In: *Journal of the American Statistical Association* 62.318, pp. 626–633.

Simon, Noah et al. (2013). "A sparse-group lasso". In: *Journal of Computational and Graphical Statistics* 22.2, pp. 231–245. DOI: 10.1080/10618600.2012.681250.

Sims, Christopher A. (1972). "Money, Income, and Causality". In: *The American Economic Review* 62.4, pp. 540–552.

— (1980). "Macroeconomics and Reality". In: *Econometrica*, pp. 1–48. DOI: 10.2307/1912017.

Smirnov, Nikolai V. (1933). "Estimate of deviation between empirical distribution functions in two independent samples". In: *Bulletin Moscow University* 2.2, pp. 3–16.

Storey, John D (2002). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 479–498.

Suppes, Patrick (1970). *A Probabilistic Theory of Causality*. Vol. 24. 4. Amsterdam: North Holland Publishing Co., pp. 409–410. DOI: 10.1086/288485.

Suryadi, Lock Yue Chew, and Yew Soon Ong (2023). "Granger causality using Jacobian in neural networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 33.2, p. 023126. DOI: 10.1063/5.0106666.

Synowiecki, Rafal (2007). "Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure". In: *Bernoulli* 13.4, pp. 1151–1178. DOI: 10.3150/07-BEJ102.

Tank, Alex et al. (2022). "Neural Granger Causality". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8, pp. 4267–4279. DOI: 10.1109/TPAMI.2021.3065601.

Tartakovsky, Alexander, Igor Nikiforov, and Michele Basseville (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Tickle, S. O. et al. (2020). "Parallelization of a Common Changepoint Detection Method". In: *Journal of Computational and Graphical Statistics* 29.1, pp. 149–161. DOI: 10.1080/10618600.2019.1647216.

Tong, Howell and Keng S. Lim (1980). "Threshold autoregression, limit cycles and cyclical data". In: *Journal of the Royal Statistical Society: Series B* 42.3, pp. 245–268. DOI: 10.1111/j.2517-6161.1980.tb01126.x.

Truong, Charles, Laurent Oudre, and Nicolas Vayatis (2020). "Selective review of offline change point detection methods". In: *Signal Processing* 167, p. 107299. DOI: 10.1016/j.sigpro.2019.107299.

Tucker, Howard G. (1959). "A generalization of the Glivenko-Cantelli theorem". In: *The Annals of Mathematical Statistics* 30.3, pp. 828–830. DOI: 10.1214/aoms/1177706212.

Van der Laan, Mark J. (2006). "Statistical inference for variable importance". In: *The International Journal of Biostatistics* 2.1, p. 2. DOI: 10.2202/1557-4679.1008.

Vanegas, Laura Jula, Merle Behr, and Axel Munk (2022). "Multiscale Quantile Segmentation". In: *Journal of the American Statistical Association* 117.539, pp. 1384–1397. DOI: 10.1080/01621459.2020.1859380.

Varela, Carmen and Matthew A. Wilson (2019). *Simultaneous extracellular recordings from midline thamalic neclei, median prefrontal cortex, and CA1 from rats cycling through bouts of sleep and wakefulness*. Collaborative Research in Computational Neuroscience. DOI: 10.6080/K0K35RVG.

— (2020). "mPFC spindle cycles organize sparse thalamic activation and recently active CA1 cells during non-REM sleep". In: *eLife* 9, e48881. DOI: 10.7554/eLife.48881.

Vicente, Raul et al. (2011). "Transfer entropy—a model-free measure of effective connectivity for the neurosciences". In: *Journal of Computational Neuroscience* 30.1, pp. 45–67. DOI: 10.1007/s10827-010-0262-3.

Vostrikova, L. Yu. (1981). "Detecting "disorder" in multidimensional random processes". In: *Doklady Akademii Nauk SSSR* 259.2, pp. 270–274.

Wang, Tengyao and Richard J. Samworth (2018). "High dimensional change point estimation via sparse projection". In: *Journal of the Royal Statistical Society: Series B* 80.1, pp. 57–83. DOI: 10.1111/rssb.12243.

Wehr, Michael and Anthony M. Zador (2003). "Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex". In: *Nature* 426.6965, pp. 442–446. DOI: 10.1038/nature02116.

Wei, Song and Yao Xie (2023). "Online Kernel CUSUM for Change-Point Detection". In: *arXiv preprint arXiv:2211.15070*. DOI: 10.48550/arXiv.2211.15070.

Widrow, Bernard, Marcian E Hoff, et al. (1960). "Adaptive switching circuits". In: *IRE WESCON Convention Record*. Vol. 4. 1. New York, pp. 96–104. DOI: 10.7551/mitpress/4943.003.0012.

Wiener, Norbert (1956). "The theory of prediction". In: *Modern Mathematics for Engineers*. Ed. by E. F. Beckenback. 1. McGraw-Hill, New York. Chap. 8.

Winkler, Anderson M. et al. (2014). "Permutation inference for the general linear model". In: *Neuroimage* 92, pp. 381–397. DOI: 10.1016/j.neuroimage.2014.01.060.

Wold, Herman (1938). "A study in the analysis of stationary time series". PhD thesis. Almqvist & Wiksell.

Wu, Hulin et al. (2014). "Sparse additive ordinary differential equations for dynamic gene regulatory network modeling". In: *Journal of the American Statistical Association* 109.506, pp. 700–716. DOI: 10.1080/01621459.2013.859617.

Xie, Liyan et al. (2021). "Sequential (Quickest) Change Detection: Classical Results and New Directions". In: *IEEE Journal on Selected Areas in Information Theory* 2.2, pp. 494–514. DOI: 10.1109/JSAIT.2021.3072962.

Yau, Chun Yip and Zifeng Zhao (Nov. 2015). "Inference for Multiple Change Points in Time Series via Likelihood Ratio Scan Statistics". In: *Journal of the Royal Statistical Society: Series B* 78.4, pp. 895–916. DOI: 10.1111/rssb.12139.

Yi Yu Oscar Hernan Madrid Padilla, Daren Wang and Alessandro Rinaldo (2023). "A note on online change point detection". In: *Sequential Analysis* 42.4, pp. 438–471. DOI: 10.1080/07474946.2023.2276170.

Yildiz, Izzet B., Herbert Jaeger, and Stefan J. Kiebel (2012). "Re-visiting the echo state property". In: *Neural Networks* 35, pp. 1–9. DOI: 10.1016/j.neunet.2012.07.005.

Yu, Hao (1993). "A Glivenko-Cantelli lemma and weak convergence for empirical processes of associated sequences". In: *Probability theory and related fields* 95.3, pp. 357–370. DOI: 10.1007/BF01192169.

Yu, Mengjia and Xiaohui Chen (Dec. 2020). "Finite Sample Change Point Inference and Identification for High-Dimensional Mean Vectors". In: *Journal of the Royal Statistical Society: Series B* 83.2, pp. 247–270. DOI: 10.1111/rssb.12406.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B* 68.1, pp. 49–67. DOI: 10.1111/j.1467-9868.2005.00532.x.

Yule, George U (1909). "The applications of the method of correlation to social and economic statistics". In: *Journal of the Royal Statistical Society* 72.4, pp. 721–730. DOI: 10.2307/2340140.

— (1921). "On the time-correlation problem, with especial reference to the variate-difference correlation method". In: *Journal of the Royal Statistical Society* 84.4, pp. 497–537. DOI: 10.2307/2341101.

— (1927). "VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 226.636-646, pp. 267–298. DOI: 10.1098/rsta.1927.0007.

Zhang, Anguo and Zheng Xu (2020). "Chaotic time series prediction using phase space reconstruction based conceptor network". In: *Cognitive Neurodynamics* 14.6, pp. 849–857. DOI: 10.1007/s11571-020-09612-7.

Zou, Changliang et al. (2014). "Nonparametric maximum likelihood approach to multiple change-point problems". In: *The Annals of Statistics* 42.3, pp. 970–1002. DOI: 10.1214/14-AOS1210.

Zou, Shaofeng, Georgios Fellouris, and Venugopal V. Veeravalli (2017). "Asymptotic optimality of D-CuSum for quickest change detection under transient dynamics". In: *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2263–2267. DOI: 10.1109/ISIT.2017.8006932.

Appendices

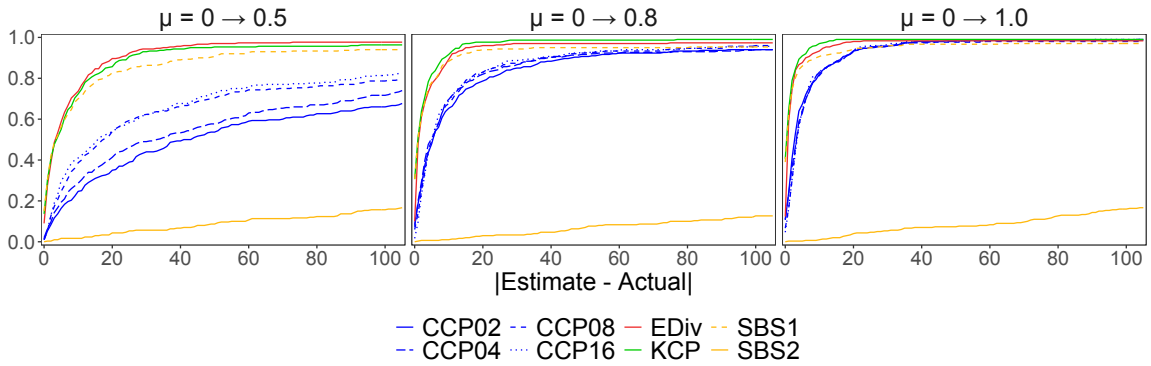
Appendix A

Additional Material for Chapter 2

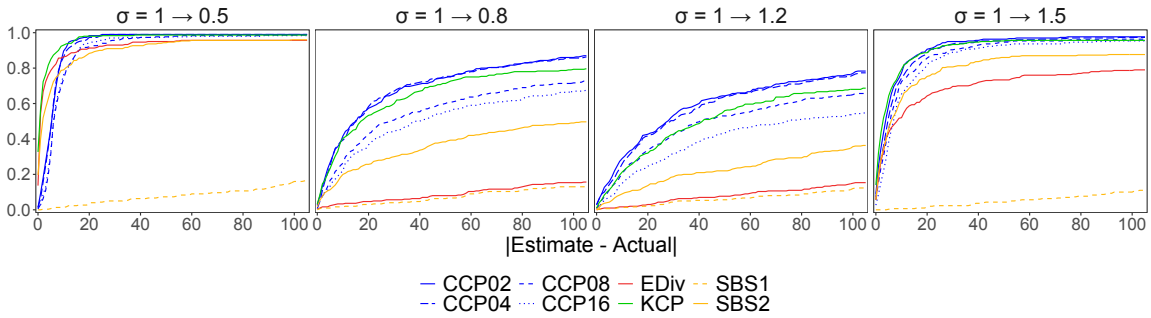
Code and supporting material: Files (.RData) used to assess performance of change point methods, code (.R) used to generate results and figures, and data (.mat and .csv) and code (.R) used to generate output in Section 2.5 can be found at

`github.com/noahgade/ChangePointDetectionWithConceptors`.

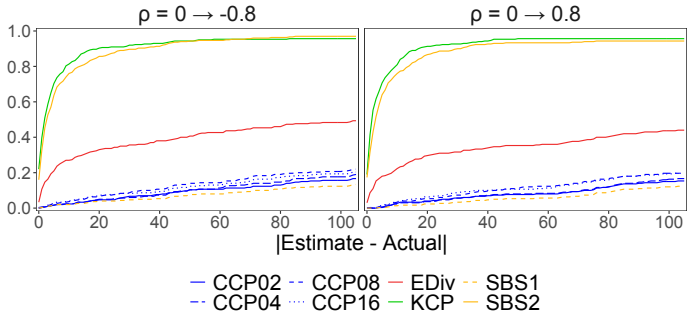
A.1 Additional Figures



(a) Fraction of identified points within error, white noise simulations with mean change μ, IDs (5a-c).

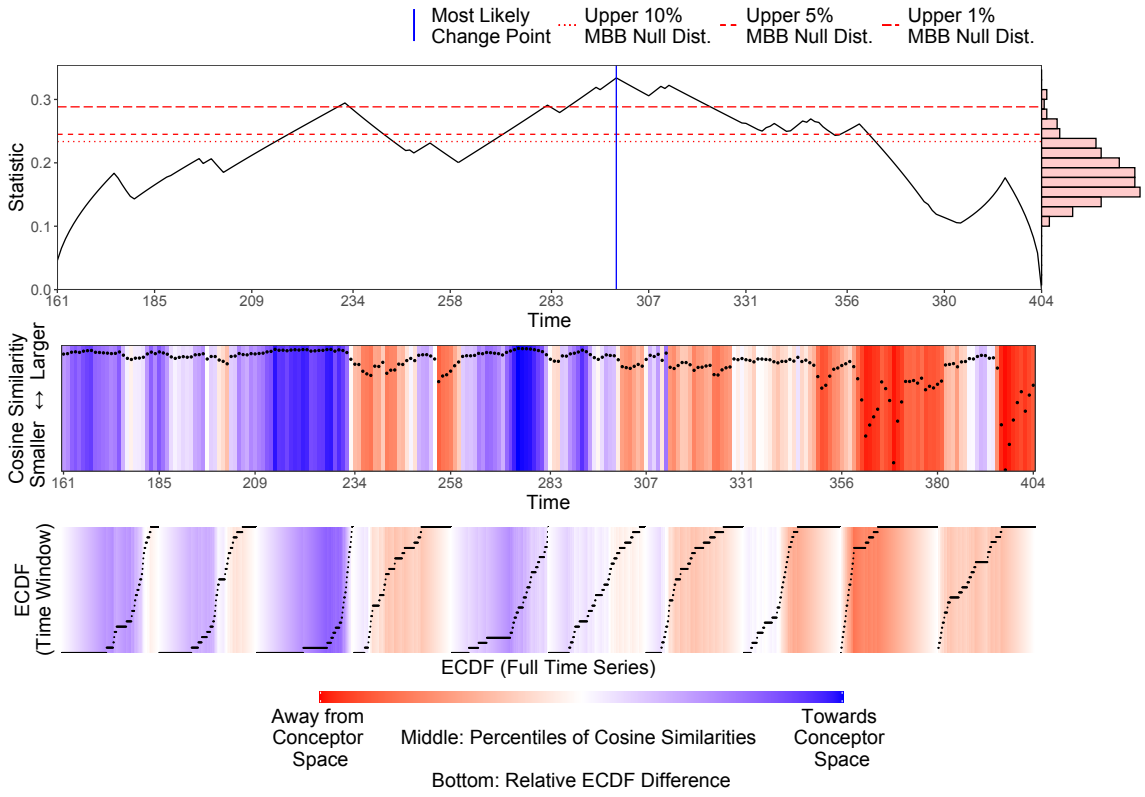


(b) Fraction of identified points within error, white noise simulations with variance change σ, IDs (5d-g).



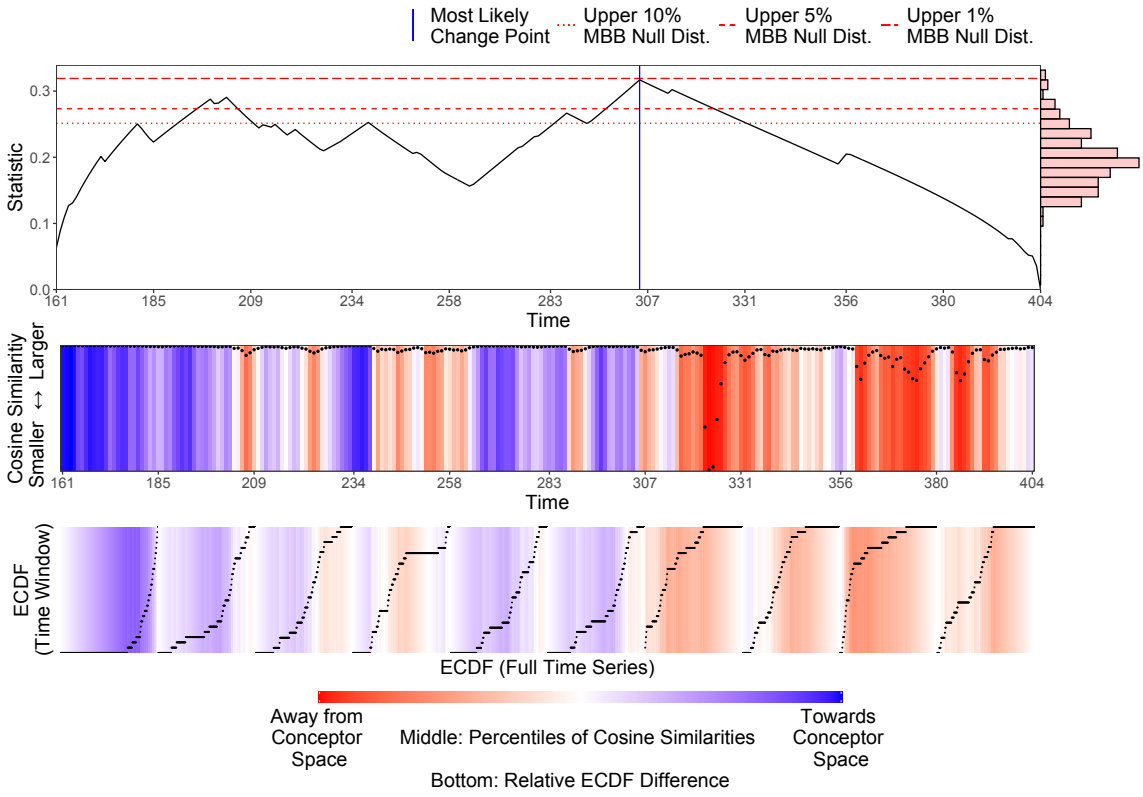
(c) Fraction of identified points within error, white noise simulations with covariance change ρ, ID (5h).

Figure A.1: Gaussian white noise simulation results.



Proposed Change Point: 299 (813.8s), Statistic = 0.334, MBB quantile = 0. *Top*: Identification of most likely change point from CUSUM-like statistics. Null bootstrap distribution included on the right vertical axis with estimated quantiles. *Middle*: Cosine similarities between conceptor and reservoir space over the time series. Shading represents percentiles of cosine similarities over the full time series. *Bottom*: Compares segment specific cosine similarity empirical CDFs to the full time series. Shading represents a relative difference of empirical CDFs.

Figure A.2: CCP method visualization of Figure 2.5 (*middle*).



Proposed Change Point: 305 (1155.4s), Statistic = 0.317, MBB quantile = 0.013. *Top*: Identification of most likely change point from CUSUM-like statistics. Null bootstrap distribution included on the right vertical axis with estimated quantiles. *Middle*: Cosine similarities between conceptor and reservoir space over the time series. Shading represents percentiles of cosine similarities over the full time series. *Bottom*: Compares segment specific cosine similarity empirical CDFs to the full time series. Shading represents a relative difference of empirical CDFs.

Figure A.3: CCP method visualization of Figure 2.5 (*bottom*).

A.2 Additional Tables

Table A.1: VAR(1) simulation results.

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\rho = 0.5 \to 0.5$	0.576	0.620	**0.716**	0.700	0.192	0.556	0.011	0.584
$\rho = 0.5 \to 0.8$	0.755	0.751	0.784	**0.795**	0.385	0.600	0.029	0.606
$\rho = 0.8 \to 0.5$	0.549	0.609	**0.698**	0.682	0.393	0.668	0.044	0.679
$\rho = 0.8 \to 0.8$	0.685	0.719	0.766	**0.775**	0.288	0.523	0.032	0.656

Mean ARI of VAR(1) spectral radius ρ changes, IDs (1a-d).

Table A.2: VAR(2) simulation results.

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\rho = 0.5 \to 0.5$	0.595	0.694	0.774	**0.769**	0.239	0.607	0.012	0.575
$\rho = 0.5 \to 0.8$	0.837	0.848	0.873	**0.883**	0.293	0.677	0.012	0.654
$\rho = 0.8 \to 0.5$	0.610	0.697	0.769	**0.816**	0.302	0.655	0.016	0.622
$\rho = 0.8 \to 0.8$	0.776	0.822	0.839	**0.880**	0.310	0.606	0.027	0.621

Mean ARI of VAR(2) spectral radius ρ changes, IDs (2a-d).

Table A.3: Periodic simulation results.

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\omega = 1 \to 0.5$	0.943	0.936	0.953	**0.961**	0.000	0.057	0.000	0.057
$\omega = 1 \to 0.8$	**0.649**	0.597	0.559	0.586	0.000	0.056	0.000	0.000
$\omega = 1 \to 1.2$	0.587	**0.605**	0.538	0.532	0.000	0.056	0.000	0.000
$\omega = 1 \to 1.5$	0.932	0.935	0.940	**0.941**	0.000	0.055	0.000	0.000

Mean ARI of periodic frequency ω changes, IDs (3a-d).

Table A.4: Ornstein-Uhlenbeck simulation results.

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\theta = 0.5 \to 0$	0.709	**0.718**	0.716	0.701	0.516	0.633	0.419	0.199
$\theta = 0.5 \to 1$	0.056	0.124	0.479	**0.538**	0.241	0.359	0.018	0.368
$\theta = 1 \to 0$	0.738	**0.743**	0.738	0.736	0.523	0.633	0.406	0.720
$\theta = 1 \to 0.5$	0.328	**0.367**	0.365	0.311	0.304	0.348	0.020	0.313

(a) Mean ARI of mean reverting θ changes, IDs (4a-d).

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\lambda = 0.5 \to 0.2$	0.922	0.924	0.924	0.926	0.874	**0.949**	0.053	0.927
$\lambda = 0.5 \to 0.8$	**0.909**	0.896	0.882	0.870	0.558	0.861	0.005	0.836
$\lambda = 0.5 \to 1$	0.936	0.935	0.934	0.931	0.767	**0.942**	0.016	0.886

(b) Mean ARI of volatility λ changes, IDs (4e-g).

Table A.5: Gaussian white noise simulation results.

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\mu = 0 \to 0.5$	0.478	0.532	0.620	0.638	0.868	**0.904**	0.802	0.000
$\mu = 0 \to 0.8$	0.840	0.846	0.850	0.845	0.922	**0.962**	0.899	0.002
$\mu = 0 \to 1$	0.899	0.894	0.897	0.898	0.940	**0.974**	0.898	0.000

(a) Mean ARI of mean μ changes, IDs (5a-c).

	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
$\sigma = 1 \to 0.5$	0.921	0.916	0.905	0.905	0.883	**0.964**	0.011	0.866
$\sigma = 1 \to 0.8$	**0.720**	0.718	0.578	0.527	0.039	0.716	0.007	0.345
$\sigma = 1 \to 1.2$	**0.631**	0.604	0.503	0.409	0.032	0.599	0.001	0.214
$\sigma = 1 \to 1.5$	0.917	0.903	0.891	0.872	0.690	**0.920**	0.005	0.796
$\rho = 0 \to 0.8$	0.034	0.040	0.082	0.087	0.299	**0.916**	0.000	0.857

(b) Mean ARI of variance σ and covariance ρ changes, IDs (5d-h).

Table A.6: No change point simulation results.

ID	Class	CCP02	CCP04	CCP08	CCP16	EDiv	KCP	SBS1	SBS2
(1e)	VAR(1)	0.07	0.07	0.04	0.06	0.41	1.00	0.01	0.01
(1f)	VAR(1)	0.06	0.07	0.05	0.08	0.54	1.00	0.01	0.00
(2e)	VAR(2)	0.06	0.06	0.07	0.06	0.44	1.00	0.01	0.02
(2f)	VAR(2)	0.09	0.08	0.08	0.08	0.35	1.00	0.01	0.00
(3e)	Periodic	0.07	0.09	0.06	0.09	0.00	1.00	0.00	0.00
(4h)	Orn.-Uhl.	0.04	0.04	0.03	0.04	0.99	1.00	0.02	0.00
(4i)	Orn.-Uhl.	0.05	0.05	0.04	0.05	0.03	1.00	0.00	0.01
(5i)	Wh. Noise	0.06	0.08	0.05	0.05	0.05	1.00	0.00	0.01

Observed Type 1 error for $q = 0.05$ with no change point.

A.3 Additional Algorithms

This section presents additional procedures composing the main algorithms in Chapter 2.

Procedure A.1 ESN Featurization: I. Scaling

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}

Outputs: ESN scaling parameters

Default Parameters: ESN spectral radius $\rho \leftarrow 0.8$; grid of possible $\mathbf{W^i}_r$, $\mathbf{b}_r$ scalings $G \leftarrow \{c_{\text{input}} \leftarrow \{0.2, 0.6, 1.0, 1.4\}, c_{\text{bias}} \leftarrow \{0.1, 0.3, 0.5\}\}$; test reservoir size $N \leftarrow 10d$; number of initializations $\mathcal{R} \leftarrow 10$; approximate washout length $T^*_{\text{wash}} \leftarrow 50$; output regularization parameter $\lambda \leftarrow 10^{-4}$

for each grid scaling combination of c_{input} and c_{bias} in G **do**
  for r in $1 : \mathcal{R}$ **do**
    initialize $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$ where each element $\mathcal{N}(0, 1)$, and $\mathbf{W^h}_r$ is sparse
    $\mathbf{W^i}_r \leftarrow c_{\text{input}} \mathbf{W^i}_r$; $\mathbf{b}_r \leftarrow c_{\text{bias}} \mathbf{b}_r$
    $\mathbf{W^h}_r \leftarrow \rho \mathbf{W^h}_r / \max \left\{ \mathbf{v}^\top \mathbf{W^h}_r \mathbf{v} : ||\mathbf{v}|| = 1 \right\}$
    $\mathbf{h}_{r,t} \leftarrow \tanh \left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r \right)$ for $t = 1, \ldots, T^*_{\text{wash}} + T_{\text{train}}$
    $\mathbf{W^o}_r \leftarrow \left(\mathbf{H}_r^\top \mathbf{H}_r + \lambda \mathbf{I} \right)^{-1} \mathbf{H}_r^\top \mathbf{Y}$ where $\mathbf{H}_r = \left[\mathbf{h}_{r, T^*_{\text{wash}}+1} \ \cdots \ \mathbf{h}_{r, T^*_{\text{wash}}+T_{\text{train}}} \right]^\top$
  end for
  $\text{NRMSE} \leftarrow \dfrac{1}{\mathcal{R}} \displaystyle\sum_{j=1}^{\mathcal{R}} \sqrt{\dfrac{(\mathbf{Y} - \mathbf{H}_r \mathbf{W^o}_r)^2}{\frac{1}{2}\text{Var}(\mathbf{Y}) + \frac{1}{2}\text{Var}(\mathbf{H}_r \mathbf{W^o}_r)}}$
end for
return ESN scaling : $\{c_{\text{input}}, c_{\text{bias}}, \rho\} \leftarrow \arg\min_G \{\text{NRMSE}\}$

Procedure A.2 ESN Featurization: II. Washout Length

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; ESN scaling : $\{c_{\text{input}}, c_{\text{bias}}, \rho\}$ from Procedure A.1; reservoir size N

Outputs: washout length T_{wash}

Default Parameters: initial reservoir states $\mathbf{h}_{r,0,0} \leftarrow 0$, $\mathbf{h}_{r,0,1} \leftarrow 1$; washout tolerance $\varepsilon_{\text{wash}} \leftarrow 10^{-6}$; initial time state $t \leftarrow 0$; number of initializations $\mathcal{R} \leftarrow 10$

$\delta_{01} \leftarrow |\mathbf{h}_{r,0,0} - \mathbf{h}_{r,0,1}|$
while $\delta_{01} > \varepsilon_{\text{wash}}$ **do**
    for r in $1 : \mathcal{R}$ **do**
        initialize $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$ where each element $\mathcal{N}(0,1)$, and $\mathbf{W^h}_r$ is sparse
        $\mathbf{W^i}_r \leftarrow c_{\text{input}} \mathbf{W^i}_j$; $\mathbf{b}_r \leftarrow c_{\text{bias}} \mathbf{b}_r$
        $\mathbf{W^h}_r \leftarrow \rho \mathbf{W^h}_r / \max \left\{ \mathbf{v}^\top \mathbf{W^h}_r \mathbf{v} : ||\mathbf{v}|| = 1 \right\}$
        $\mathbf{h}_{r,t,0} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{r,t-1,0} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r \right)$
        $\mathbf{h}_{r,t,1} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{r,t-1,1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r \right)$
    end for
    $\delta_{01} \leftarrow \max_j |\mathbf{h}_{r,t,0} - \mathbf{h}_{r,t,1}|$
    if $\delta_{01} > \varepsilon_{\text{wash}}$ **then**
        $t \leftarrow t + 1$
    else
        $T_{\text{wash}} \leftarrow t$
    end if
end while
    return T_{wash}, $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$

Procedure A.3 ESN Featurization: III. Parameter Computation

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; ESN scaling (c_{input}, c_{bias}, and $\rho \leftarrow 0.8$) from Procedure A.1; T_{wash} from specified value or Procedure A.2

Outputs: ESN reservoir size N; aperture α

Default Parameters: training error tolerance $\varepsilon_{\text{train}} \leftarrow 0.04$; initial reservoir size $N \leftarrow 10d$; initial aperture value $\alpha \leftarrow N$; number of initializations $\mathcal{R} \leftarrow 10$; output regularization parameter $\lambda \leftarrow 10^{-4}$

while NRMSE $> \varepsilon_{\text{train}}$ **do**
  perform Procedure A.2 to obtain T_{wash}, $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$
  for r in $1 : \mathcal{R}$ **do**
    $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$ for $t = 1, \ldots, T_{\text{wash}} + T_{\text{train}}$
    $\mathbf{C}_r \leftarrow \frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r \left(\frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r + \alpha^{-2}\mathbf{I}\right)^{-1}$
      where $\mathbf{H}_r = \left[\mathbf{h}_{r,T_{\text{wash}}+1} \ \cdots \ \mathbf{h}_{r,T_{\text{wash}}+T_{\text{train}}}\right]^\top$
    $\mathbf{h}_{r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \tilde{\mathbf{h}}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r\right)$
    $\tilde{\mathbf{h}}_{r,t} \leftarrow \mathbf{C}_r \mathbf{h}_{r,t}$ for $t = T_{\text{wash}} + 1, \ldots, T_{\text{wash}} + T_{\text{train}}$
    $\mathbf{W^o}_r \leftarrow \left(\tilde{\mathbf{H}}_r^\top \tilde{\mathbf{H}}_r + \lambda \mathbf{I}\right)^{-1} \tilde{\mathbf{H}}_r^\top \mathbf{Y}$
      where $\tilde{\mathbf{H}}_r = \left[\tilde{\mathbf{h}}_{r,T_{\text{wash}}+1} \ \cdots \ \tilde{\mathbf{h}}_{r,T_{\text{wash}}+T_{\text{train}}}\right]^\top$
  end for
  NRMSE $\leftarrow \frac{1}{\mathcal{R}} \sum_{j=1}^{\mathcal{R}} \sqrt{\frac{\left(\mathbf{Y} - \tilde{\mathbf{H}}_r \mathbf{W^o}_r\right)^2}{\frac{1}{2}\text{Var}(\mathbf{Y}) + \frac{1}{2}\text{Var}\left(\tilde{\mathbf{H}}_r \mathbf{W^o}_r\right)}}$
  if NRMSE $> \varepsilon_{\text{train}}$ and $\alpha \leq 100N$ **then**
    $\alpha \leftarrow \sqrt{10}\alpha$
  else
    $N \leftarrow dN$; $\alpha \leftarrow N$
  end if
end while
  return N, α, T_{wash}

Procedure A.4 Generating Bootstrapped Time Series

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training length T_{train}; T_{wash} from Algorithm 2.1

Outputs: bootstrapped time series $\mathbf{y}_{b,t}$

Default Parameters: number of bootstraps $B \leftarrow 240$; pilot block length $\ell \leftarrow T_{\text{wash}}$

1: Perform Hall, Horowitz, and Jing (1995) algorithm with ℓ, five proposed block lengths equally spaced between $\lceil T^{1/5} \rceil$ and $\lceil T^{1/2} \rceil$, and 40 bootstrapped series each to obtain MBB block length parameter L.

2: **for** b in $1 : B$ **do**

3: **for** i in $1 : \lceil (T - T_{\text{wash}} - T_{\text{train}})/L \rceil$ **do**

4: $\beta_{b,i} \leftarrow \beta \sim \text{Uniform}\{T_{\text{wash}} + T_{\text{train}} + 1,\ T\}$

5: $\mathbf{b}_i \leftarrow \mathbf{y}_{\beta_{b,i}:(\beta_{b,i}+L-1)}$

6: **end for**

7: $\mathbf{y}_t^b \leftarrow \left[\mathbf{y}_{1:(T_{\text{wash}}+T_{\text{train}})}^\top \ \mathbf{b}_1^\top \ \cdots \ \mathbf{b}_{\lceil (T-T_{\text{wash}}-T_{\text{train}})/L \rceil}^\top\right]_{1:T}^\top$

8: **end for**

return all $\mathbf{y}_{b,t}$

A.4 Proofs

Proof of Theorem 2.2 follows Csörgő and Horváth (1997), Theorem 2.6.1 with the relaxation of the i.i.d. sequence to a stationary, $\mathcal{S}$-mixing sequence. The proof consists of two major steps. First, we show the statistic converges to a sequence of Gaussian processes (Lemma A.1). Because of the relaxation from an i.i.d. sequence to a $\mathcal{S}$-mixing sequence, the rates of convergence shown in Csörgő and Horváth (1997), Lemma 2.6.1 are adjusted. Next we show that the sequence of Gaussian processes, in turn, converges in distribution to the desired limiting process (Lemma A.2).

Define the quantity $K_T(s,t)$ in Equation A.1 on the domain $1 \leq t \leq T-1$ as a scaled difference between the two empirical CDFs. This is a common form akin to that used in Csörgő and Horváth (1997), Chapter 2 with a modification of the denominator.

$$K_T(s,t) = \left[\frac{t(T-t)}{T^2}\right]\left(\hat{\mathcal{F}}_{1:t}(s) - \hat{\mathcal{F}}_{(t+1):T}(s)\right) \tag{A.1}$$

We scale the time domain to $\delta \in [0,1]$ such that $\delta = t/T$, and define $\mathcal{F}_{1:\delta T}(s) = \mathcal{F}_1(s)$ and $\mathcal{F}_{\delta T:T}(s) = \mathcal{F}_T(s)$ as the distribution functions on the intervals $(0,\delta]$ and $[\delta,1)$, respectively, with $\hat{\mathcal{F}}_{1:\delta T}(s)$ and $\hat{\mathcal{F}}_{\delta T:T}(s)$ as their corresponding empirical estimates. Equation A.1 can be adjusted to the analogous form,

$$K_T(s,\delta) = \left[\frac{\delta T(T-\delta T)}{T^2}\right]\left(\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right). \tag{A.2}$$

We now state and prove our Lemma A.1. The proof requires Theorem 2 from Berkes, Hörmann, and Schauer (2009), which we restate at the end of this section, without proof, for convenience.

Lemma A.1. *Let S_t be a stationary sequence such that $\mathcal{F}(s) = P(S_1 \leq s)$ is Lipschitz continuous of order $C > 0$. Assume S_t is $\mathcal{S}$-mixing and that condition (1) of Definition 2.1 holds with $\gamma_m = m^{-AC}$, $\delta_m = m^{-A}$ for some $A > 4$. Then under the null hypothesis with $q(\delta)$ a positive function on $(0,1)$ that increases in a neighborhood of zero and decreases in a neighborhood of one,*

$$\sup_{1/T \leq \delta \leq (T-1)/T} \sup_{s \in [0,1]} \left| \sqrt{T}\, K_T(s,\delta) - \mathcal{K}_T(s,\delta) \right| / q(\delta) = o(1), \tag{A.3}$$

where $K_T(s,\delta)$ is defined in Equation A.2, $\{\mathcal{K}_T(s,\delta),\ 0 \leq \delta \leq 1\}$ is a sequence of Gaussian processes with

$$\begin{aligned} &\mathbb{E}\left[\mathcal{K}_T(s,\delta)\right] = 0, \\ &\mathbb{E}\left[\mathcal{K}_T(s,\delta)\, \mathcal{K}_T(s',\delta')\right] = (\delta \wedge \delta')\, \Gamma(s,s'), \\ \textit{and} \quad &\Gamma(s,s') = \sum_{-\infty < t < \infty} \mathbb{E}\left[S_1(s) S_t(s')\right], \end{aligned} \tag{A.4}$$

provided $I_{0,1}(q,c) < \infty$ for all $c > 0$, where

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{\delta(1-\delta)} \exp\left\{ -\frac{c q^2(\delta)}{\delta(1-\delta)} \right\} d\delta. \tag{A.5}$$

Proof of Lemma A.1. With $\mathcal{F}(s)$ denoting the true distribution function of all S_t

under the null, we can write $K_T(s,t)$ from Equation A.1 as

$$\sqrt{T}\ K_T(s,t) = \begin{cases} \frac{1}{\sqrt{T}}\left(\sum_{i=1}^{t}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right) - \frac{t}{T}\sum_{i=1}^{T}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right)\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } 1 \le t \le T/2, \\ \frac{1}{\sqrt{T}}\left(\frac{T-t}{T}\sum_{i=1}^{T}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right) - \sum_{i=t+1}^{T}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right)\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } T/2 \le t \le T-1. \end{cases} \tag{A.6}$$

Replacing the strong approximation of empirical processes used in Csörgő and Horváth (1997), Lemma 2.6.1 we use the $\mathcal{S}$-mixing conditions and Theorem 2 from Berkes, Hörmann, and Schauer (2009). Define two Gaussian processes $\{\mathcal{K}_1(s,t),\ 1 \le t \le T/2\}$ and $\{\mathcal{K}_2(s,t),\ T/2 \le t \le T\}$ such that

$$\sup_{1\le t\le T/2}\sup_{s\in[0,1]}\left|\sum_{i=1}^{t}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right) - \mathcal{K}_1(s,t)\right| = o\left(\left(\frac{T}{2}\right)^{1/2}\left(\log\frac{T}{2}\right)^{-\alpha}\right), \tag{A.7}$$

and

$$\sup_{T/2\le t\le T}\sup_{s\in[0,1]}\left|\sum_{i=t+1}^{T}\left(\mathbf{1}\{S_i \le s\} - \mathcal{F}(s)\right) - \mathcal{K}_2(s,t)\right| = o\left(\left(\frac{T}{2}\right)^{1/2}\left(\log\frac{T}{2}\right)^{-\alpha}\right), \tag{A.8}$$

for some $\alpha > 0$ where the two processes have identical expected value and covariance

functions.

$$\mathbb{E}\left[\mathcal{K}_1(s,t)\right] = \mathbb{E}\left[\mathcal{K}_2(s,t)\right] = 0$$

$$\mathbb{E}\left[\mathcal{K}_1(s,t)\ \mathcal{K}_1(s',t')\right] = \mathbb{E}\left[\mathcal{K}_2(s,t)\ \mathcal{K}_2(s',t')\right] = (t \wedge t')\ \Gamma\left(s,s'\right)$$

$$\Gamma(s,s') = \sum_{-\infty<t<\infty} \mathbb{E}\left[S_1(s)S_t(s')\right] \tag{A.9}$$

From the strong approximation in Equations A.7 and A.8, we define a sequence of Gaussian processes $\mathcal{K}_T(s,\delta)$ based on the partial sums in Equation A.6 using the scaled time domain $\delta \in [0,1]$.

$$\mathcal{K}_T(s,\delta) = \begin{cases} \frac{1}{\sqrt{T}}\left(\mathcal{K}_1(s,\delta T) - \delta\left[\mathcal{K}_1(s,T/2) + \mathcal{K}_2(s,T/2)\right]\right) & \text{for } 0 \le \delta \le 1/2, \\ \frac{1}{\sqrt{T}}\left(-\mathcal{K}_2(s,\delta T) + (1-\delta)\left[\mathcal{K}_1(s,T/2) + \mathcal{K}_2(s,T/2)\right]\right) & \text{for } 1/2 \le \delta \le 1 \end{cases} \tag{A.10}$$

Assembling Equations A.6, A.7, and A.8, we immediately obtain the result stated in Equation A.3. □

Lemma A.2. *Let* $\{\mathcal{K}_T(s,\delta),\ 0 \le \delta \le 1\}$ *be a sequence of Gaussian processes defined in Equation A.4 and* $\{\mathcal{K}(s,\delta),\ 0 \le \delta \le 1\}$ *be a Gaussian process defined in Equation 2.11. Under the null hypothesis with* $q(\delta)$ *a positive function on* $(0,1)$ *that increases in a neighborhood of zero and decreases in a neighborhood of one,*

$$\sup_{\delta\in[0,1]} \sup_{s\in[0,1]} \left|\mathcal{K}_T(s,\delta)\right| / q(\delta) \xrightarrow{D} \sup_{\delta\in[0,1]} \sup_{s\in[0,1]} \left|\mathcal{K}(s,\delta)\right| / q(\delta), \tag{A.11}$$

provided the result of Lemma A.1 holds for $K_T(s,\delta)$ *defined in Equation A.2, and* $I_{0,1}(q,c) < \infty$ *for all* $c > 0$*, where* $I_{0,1}(q,c)$ *is defined in Equation A.5.*

Proof of Lemma A.2. We continue as in Theorem 2.6.1 from Csörgő and Horváth (1997). From the definition of $K_T(s,\delta)$ in Equation A.2 at the extreme ends of the domain,

$$\sup_{0<\delta<1/T}\sup_{s\in[0,1]}\left|\sqrt{T}\,K_T(s,\delta)\right|/q(\delta)=0, \tag{A.12}$$

and

$$\sup_{(T-1)/T<\delta<1}\sup_{s\in[0,1]}\left|\sqrt{T}\,K_T(s,\delta)\right|/q(\delta)=0. \tag{A.13}$$

From the result of Lemma A.1 in Equation A.3 with the condition $I_{0,1}(q,c)<\infty$ for all $c>0$, we can obtain Equations A.14 and A.15.

$$\sup_{0<\delta<1/T}\sup_{s\in[0,1]}\left|\mathcal{K}_T(s,\delta)\right|/q(\delta)=o(1) \tag{A.14}$$

$$\sup_{(T-1)/T<\delta<1}\sup_{s\in[0,1]}\left|\mathcal{K}_T(s,\delta)\right|/q(\delta)=o(1) \tag{A.15}$$

Examining the covariance structure of $\mathcal{K}_T(s,\delta)$ will verify that

$$\{\mathcal{K}_T(s,\delta),\ 0\le\delta\le1\}\xrightarrow{D}\{\mathcal{K}(s,\delta),\ 0\le\delta\le1\}, \tag{A.16}$$

and via Kolmogorov's zero-one law and Theorem A.7.3 from Csörgő and Horváth (1997),

$$\lim_{\delta\downarrow0}\sup_{s\in[0,1]}\left|\mathcal{K}(s,\delta)\right|/q(\delta)=0\ \text{ a.s.} \tag{A.17}$$

$$\text{and}\quad\lim_{\delta\uparrow1}\sup_{s\in[0,1]}\left|\mathcal{K}(s,\delta)\right|/q(\delta)=0\ \text{ a.s.}, \tag{A.18}$$

if and only if $I_{0,1}(q,c) < \infty$ for all $c > 0$. Thus, we obtain the Lemma A.2 result stated in Equation A.11. □

We now use Lemmas A.1 and A.2 to prove Theorem 2.2.

Proof of Theorem 2.2. We combine the result from Lemma A.1 in Equation A.3 with that of Lemma A.2 in Equation A.11 and can write

$$\max_{1\le t<T} \sup_{s\in[0,1]} \sqrt{T}\,\frac{K_T(s,t)}{q\left(\frac{t}{T}\right)} \xrightarrow{D} \sup_{\delta\in[0,1]} \sup_{s\in[0,1]} |\mathcal{K}(s,\delta)|\,/q(\delta), \tag{A.19}$$

as shown in in Equation 2.10 provided the necessary condition, $I_{0,1}(q,c) < \infty$ for all $c > 0$, is met with $I_{0,1}(q,c)$ defined in Equation A.5. We expand $q(\delta)$, defined in in Theorem 2.2, as a piecewise function on $\delta \in [0,1]$.

$$q(\delta) = \begin{cases} \kappa & \text{for } 0 \le \delta < \frac{1}{2} - \frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}, \\ \delta^{\nu}(1-\delta)^{\nu} & \text{for } \frac{1}{2} - \frac{1}{2}\sqrt{1-4\kappa^{1/\nu}} \le \delta \le \frac{1}{2} + \frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}, \\ \kappa & \text{for } \frac{1}{2} + \frac{1}{2}\sqrt{1-4\kappa^{1/\nu}} < \delta \le 1 \end{cases} \tag{A.20}$$

The integral from Equation A.5 becomes

$$\begin{aligned} I_{0,1}(q,c) = &\int_0^{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta \\ &+ \int_{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\delta^{2\nu-1}(1-\delta)^{2\nu-1}\right\} d\delta \\ &+ \int_{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^{1} \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta, \end{aligned} \tag{A.21}$$

and for any $c > 0$ the boundary terms are finite with $\kappa > 0$.

$$\int_0^{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta < \infty \tag{A.22}$$

$$\int_{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^1 \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta < \infty \tag{A.23}$$

For the middle term,

$$\int_{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\delta^{2\nu-1}(1-\delta)^{2\nu-1}\right\} d\delta < \infty \tag{A.24}$$

provided $\nu < 1/2$ if $\kappa \to 0$. For $\kappa > 0$, the range of values satisfying Equation A.24 includes $\nu = 1/2$. Thus, from the specification of $q(\delta)$ in Theorem 2.2 with $\nu = 1/2$ and $\kappa > 0$, $I_{0,1}(q, c) < \infty$ for all $c > 0$. □

Proof of Theorem 2.3 follows the outline of Theorem 2.1 from Newey and McFadden (1994) for consistency of extremum estimators. To meet the first three of the four conditions, we show the statistic for selection of a change point is uniquely maximized at the true change point τ, the set used for estimation is compact and bounded away from the endpoints, and the statistic is continuous. To show the estimate of the statistic converges uniformly in probability to the true values, we employ the almost sure convergence of the empirical CDF of a stationary, ergodic sequence from the Glivenko-Cantelli Theorem and the definition of stochastic equicontinuity and Theorem 1 from Newey (1991), along with Lemmas A.3 and A.4. We include the definition of stochastic equicontinuity from Newey (1991) and restate Theorem 2.1, without proof, from Newey and McFadden (1994) at the end of the section. We first proceed with our proof of Lemma A.3.

Lemma A.3. *A sequence of functions $\hat{Q}_T(\delta)$ is stochastically equicontinuous if there*

exists $\alpha > 0$, $\hat{A}_T = o(1)$, *and* $\hat{B}_T = \mathcal{O}(1)$ *such that for all* $\tilde{\delta}, \delta \in \Delta$, $|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)| \leq \hat{A}_T + \hat{B}_T \|\tilde{\delta} - \delta\|^\alpha$.

Proof of Lemma A.3. We follow the strategy of the proof for Lemma 2.9 in Newey and McFadden (1994). Pick $\varepsilon, \eta > 0$. By $\hat{A}_T = o(1)$, $\mathbb{P}(|\hat{A}_T| > \frac{\varepsilon}{2}) < \frac{\eta}{2}$ for T large enough. Likewise, by $\hat{B}_T = \mathcal{O}(1)$, there is M such that $\mathbb{P}(|\hat{B}_T| > M) < \frac{\eta}{2}$ for all T large enough. Let $\Gamma_T(\varepsilon, \eta) = \hat{A}_T + \hat{B}_T \frac{\varepsilon}{2M}$ and $\mathcal{N}(\delta, \varepsilon, \eta) = \{\tilde{\delta} : \|\tilde{\delta} - \delta\|^\alpha \leq \frac{\varepsilon}{2M}\}$. Then, $\mathbb{P}(|\Gamma_T(\varepsilon, \eta)| > \varepsilon) = \mathbb{P}(|\hat{A}_T + \hat{B}_T \frac{\varepsilon}{2M}| > \varepsilon) \leq \mathbb{P}(|\hat{A}_T| + |\hat{B}_T \frac{\varepsilon}{2M}| > \varepsilon) \leq \mathbb{P}(|\hat{A}_T| > \frac{\varepsilon}{2}) + \mathbb{P}(|\hat{B}_T| > M) < \eta$ and for all $\tilde{\delta}, \delta \in \mathcal{N}(\delta, \varepsilon, \eta)$, $|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)| \leq \hat{A}_T + \hat{B}_T \|\tilde{\delta} - \delta\|^\alpha \leq \Gamma_T(\varepsilon, \eta)$. □

For the statement and proof of Lemma A.4, we scale the time domain to $\delta \in [0, 1]$ such that $\delta = t/T$, with the true change point at $\delta_0 = \tau/T$, and define $\mathcal{F}_{1:\delta T}(s)$ and $\mathcal{F}_{\delta T:T}(s)$ as the distribution functions on the intervals $(0, \delta]$ and $[\delta, 1)$, respectively, with $\hat{\mathcal{F}}_{1:\delta T}(s)$ and $\hat{\mathcal{F}}_{\delta T:T}(s)$ as their corresponding empirical estimates. Suppose the true change point occurs at $\delta_0 \in \Delta$, and divides the sequence into two distinct pieces with distribution functions $\mathcal{F}_1(s) = \mathcal{F}_{1:\delta_0 T}(s)$ and $\mathcal{F}_T(s) = \mathcal{F}_{\delta_0 T:T}(s)$, where $\mathcal{F}_1(s_0) \neq \mathcal{F}_T(s_0)$ for some $s_0 \in [0, 1]$. Define the supremum of the difference between the two distributions $\theta > 0$, where the quantity is maximized at s_0.

$$\theta = \sup_{s \in [0,1]} |\mathcal{F}_1(s) - \mathcal{F}_T(s)| = |\mathcal{F}_1(s_0) - \mathcal{F}_T(s_0)| \tag{A.25}$$

Write $Q_0(\delta)$ as below for $\delta \in \Delta = \left[\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\kappa^2}, \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\kappa^2}\right]$, $q(\delta) = \max\{\delta^{1/2}(1 - \delta)^{1/2}, \kappa\}$, and $\kappa > 0$ a small constant.

$$Q_0(\delta) = \frac{1}{q(\delta)} \left[\frac{\delta T\,(T - \delta T)}{T^2}\right] \sup_{s \in [0,1]} |\mathcal{F}_{1:\delta T}(s) - \mathcal{F}_{\delta T:T}(s)| \tag{A.26}$$

We construct $\hat{Q}_T(\delta)$ in Equation A.27 comparably to $Q_0(\delta)$ while including the em-

pirical CDFs $\hat{\mathcal{F}}_{1:\delta T}(s)$ and $\hat{\mathcal{F}}_{\delta T:T}(s)$.

$$\hat{Q}_T(\delta) = \frac{1}{q(\delta)} \left[\frac{\delta T(T - \delta T)}{T^2} \right] \sup_{s \in [0,1]} \left| \hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s) \right| \tag{A.27}$$

Lemma A.4. *Suppose the sequence S_t, $1, \ldots, T$ divides into two stationary ergodic pieces on either side of the change point τ, where $1 \leq \tau < T$, and $\mathcal{F}_{1:\tau}(s_0) \neq \mathcal{F}_{(\tau+1):T}(s_0)$ for some $s_0 \in [0, 1]$. Then, $\hat{Q}_T(\delta)$ from Equation A.27 converges uniformly in probability to $Q_0(\delta)$ from Equation A.26 on the interval*

$$\delta, \delta_0 \in \Delta = \left[\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\kappa^2}, \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\kappa^2} \right] \tag{A.28}$$

with $q(\delta) = \max\{\delta^{1/2}(1 - \delta)^{1/2}, \kappa\}$ for and some small $\kappa > 0$.

Proof of Lemma A.4. We use Theorem 1 from Newey (1991) and show uniform convergence in probability with pointwise convergence and stochastic equicontinuity. For any $\delta, \delta_0 \in \Delta$, we can write the distributions $\mathcal{F}_{1:\delta T}(s)$ and $\mathcal{F}_{\delta T:T}(s)$ as a mixture between the distributions $\mathcal{F}_1(s)$ and $\mathcal{F}_T(s)$.

$$\mathcal{F}_{1:\delta T}(s) = \mathbf{1}\{\delta \leq \delta_0\}\mathcal{F}_1(s) + \mathbf{1}\{\delta > \delta_0\} \left[\frac{\delta_0}{\delta}\mathcal{F}_1(s) + \frac{\delta - \delta_0}{\delta}\mathcal{F}_T(s) \right] \tag{A.29}$$

$$\mathcal{F}_{\delta T:T}(s) = \mathbf{1}\{\delta \leq \delta_0\} \left[\frac{\delta_0 - \delta}{1 - \delta}\mathcal{F}_1(s) + \frac{1 - \delta_0}{1 - \delta}\mathcal{F}_T(s) \right] + \mathbf{1}\{\delta > \delta_0\}\mathcal{F}_T(s) \tag{A.30}$$

Taking the supremum of the difference between the distributions in Equations A.29

and A.30 as it appears in $Q_0(\delta)$,

$$\begin{aligned}
&\sup_{s\in[0,1]} |\mathcal{F}_{1:\delta T}(s) - \mathcal{F}_{\delta T:T}(s)| \\
&\quad= \sup_{s\in[0,1]} \Bigg| \mathbf{1}\{\delta \le \delta_0\}\mathcal{F}_1(s) + \mathbf{1}\{\delta > \delta_0\}\left[\frac{\delta_0}{\delta}\mathcal{F}_1(s) + \frac{\delta-\delta_0}{\delta}\mathcal{F}_T(s)\right] \\
&\qquad - \mathbf{1}\{\delta \le \delta_0\}\left[\frac{\delta_0-\delta}{1-\delta}\mathcal{F}_1(s) + \frac{1-\delta_0}{1-\delta}\mathcal{F}_T(s)\right] + \mathbf{1}\{\delta > \delta_0\}\mathcal{F}_T(s)\Bigg| \qquad \text{(A.31)} \\
&\quad= \sup_{s\in[0,1]} \Bigg| \mathbf{1}\{\delta \le \delta_0\}\left[\frac{1-\delta_0}{1-\delta}\mathcal{F}_1(s) - \frac{1-\delta_0}{1-\delta}\mathcal{F}_T(s)\right] \\
&\qquad - \mathbf{1}\{\delta > \delta_0\}\left[\frac{\delta_0}{\delta}\mathcal{F}_1(s) - \frac{\delta_0}{\delta}\mathcal{F}_T(s)\right]\Bigg| \qquad \text{(A.32)} \\
&\quad= \sup_{s\in[0,1]} \left| \left[\mathbf{1}\{\delta \le \delta_0\}\frac{1-\delta_0}{1-\delta} - \mathbf{1}\{\delta > \delta_0\}\frac{\delta_0}{\delta}\right]\big(\mathcal{F}_1(s) - \mathcal{F}_T(s)\big)\right| \qquad \text{(A.33)} \\
&\quad= \left|\mathbf{1}\{\delta \le \delta_0\}\frac{1-\delta_0}{1-\delta} - \mathbf{1}\{\delta > \delta_0\}\frac{\delta_0}{\delta}\right| \sup_{s\in[0,1]} |\mathcal{F}_1(s) - \mathcal{F}_T(s)| \qquad \text{(A.34)} \\
&\quad= \left|\mathbf{1}\{\delta \le \delta_0\}\frac{1-\delta_0}{1-\delta} - \mathbf{1}\{\delta > \delta_0\}\frac{\delta_0}{\delta}\right| \theta. \qquad \text{(A.35)}
\end{aligned}$$

Considering the leading term in Equation A.26,

$$Q_0(\delta) = \begin{cases} \theta\dfrac{\delta^{1/2}(1-\delta_0)}{(1-\delta)^{1/2}} & \text{for } \delta, \delta_0 \in \Delta \text{ and } \delta \le \delta_0, \\ \theta\dfrac{\delta_0(1-\delta)^{1/2}}{\delta^{1/2}} & \text{for } \delta, \delta_0 \in \Delta \text{ and } \delta > \delta_0. \end{cases} \qquad \text{(A.36)}$$

We expand the supremum term of $\hat{Q}_T(\delta)$ in a similar fashion depending on the rela-

tionship between δ and δ_0. For $\delta \leq \delta_0$,

$$\begin{aligned}
\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| &= \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \frac{\delta_0-\delta}{1-\delta}\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \frac{1-\delta_0}{1-\delta}\hat{\mathcal{F}}_{\delta_0 T:T}(s)\right| \\
&= \sup_{s\in[0,1]} \left| \left(\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right) + \frac{\delta_0-\delta}{1-\delta}\left(\mathcal{F}_1(s) - \hat{\mathcal{F}}_{\delta T:\delta_0 T}(s)\right)\right. \\
&\quad \left. + \frac{1-\delta_0}{1-\delta}\left(\mathcal{F}_T(s) - \hat{\mathcal{F}}_{\delta_0 T:T}(s)\right)\right| \qquad \text{(A.37)} \\
&\leq \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right| + \frac{\delta_0-\delta}{1-\delta}\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\
&\quad + \frac{1-\delta_0}{1-\delta}\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \\
&\quad + \frac{1-\delta_0}{1-\delta}\sup_{s\in[0,1]} \left|\mathcal{F}_1(s) - \mathcal{F}_T(s)\right|. \qquad \text{(A.38)}
\end{aligned}$$

And for $\delta > \delta_0$,

$$\begin{aligned}
\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| &\leq \frac{\delta_0}{\delta}\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \frac{\delta-\delta_0}{\delta}\sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\delta T}(s) - \mathcal{F}_T(s)\right| \\
&\quad + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:T}(s) - \mathcal{F}_T(s)\right| + \frac{\delta_0}{\delta}\sup_{s\in[0,1]} \left|\mathcal{F}_1(s) - \mathcal{F}_T(s)\right|. \qquad \text{(A.39)}
\end{aligned}$$

Using an extension of the Glivenko-Cantelli theorem to stationary and ergodic sequences, the supremum term almost surely converges to a scaled difference in the true distribution functions as the number of time points in each section gets large

(Tucker, 1959; H. Yu, 1993; Dehling and Philipp, 2002).

$$\mathbb{P}\left[\left(\sup_{s\in[0,1]}\left|\hat{\mathcal{F}}_{1:\delta T}(s)-\hat{\mathcal{F}}_{\delta T:T}(s)\right|-\frac{1-\delta_0}{1-\delta}\theta\right)\to 0\right]=1 \quad \text{for } \delta\leq\delta_0, \tag{A.40}$$

$$\text{and} \quad \mathbb{P}\left[\left(\sup_{s\in[0,1]}\left|\hat{\mathcal{F}}_{1:\delta T}(s)-\hat{\mathcal{F}}_{\delta T:T}(s)\right|-\frac{\delta_0}{\delta}\theta\right)\to 0\right]=1 \quad \text{for } \delta>\delta_0. \tag{A.41}$$

Including the leading coefficient,

$$\hat{Q}_T(\delta)\xrightarrow{a.s.}\begin{cases}\theta\dfrac{\delta^{1/2}(1-\delta_0)}{(1-\delta)^{1/2}} & \text{for } \delta\leq\delta_0,\\[2ex] \theta\dfrac{\delta_0(1-\delta)^{1/2}}{\delta^{1/2}} & \text{for } \delta>\delta_0,\end{cases} \tag{A.42}$$

and we obtain $\hat{Q}_T(\delta)\to Q_0(\delta)$ with probability 1 for all $\delta,\delta_0\in\Delta$.

To show $\hat{Q}_T(\delta)$ is stochastically equicontinuous, we use the structure of Lemma A.3. Define $\alpha=\kappa$, $\hat{B}_T=C\theta$, where $C=(2/\kappa)\sqrt{1-\kappa^2}+1$, and $\hat{A}_T$ as below in Equations A.43 through A.48 depending on the relationship between $\tilde{\delta}$, δ, and δ_0 where each is contained in Δ. For each case $i=1,\ldots,6$, $\hat{A}_T^i=o(1)$ is a weaker condition than the extended Glivenko-Cantelli result, and $\hat{B}_T=\mathcal{O}(1)$ is verified with $C<\infty$, $0<\theta\leq 1$

(Tucker, 1959; H. Yu, 1993).

$$\hat{A}_T^1 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \tilde{\delta} < \delta \leq \delta_0, \tag{A.43}$$

$$\hat{A}_T^2 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \delta < \tilde{\delta} \leq \delta_0, \tag{A.44}$$

$$\hat{A}_T^3 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\delta T}(s) - \mathcal{F}_T(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \tilde{\delta} \leq \delta_0 < \delta, \tag{A.45}$$

$$\hat{A}_T^4 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\tilde{\delta}T}(s) - \mathcal{F}_T(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \delta \leq \delta_0 < \tilde{\delta}, \tag{A.46}$$

$$\hat{A}_T^5 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\tilde{\delta} T}(s) - \mathcal{F}_T(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta} T:\delta T}(s) - \mathcal{F}_T(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \delta_0 < \tilde{\delta} < \delta, \tag{A.47}$$

$$\hat{A}_T^6 = \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta_0 T}(s) - \mathcal{F}_1(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\delta T}(s) - \mathcal{F}_T(s)\right| + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\tilde{\delta} T}(s) - \mathcal{F}_T(s)\right| + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta} T:T}(s) - \mathcal{F}_T(s)\right| \quad \text{for } \delta_0 < \delta < \tilde{\delta} \tag{A.48}$$

We simplify the form of $\hat{Q}_T$ in Equation A.27 and write $q(\delta) = \delta^{1/2}(1-\delta)^{1/2}$ for the domain restriction $\tilde{\delta}, \delta \in \Delta$ such that

$$\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| = \left| q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta} T}(s) - \hat{\mathcal{F}}_{\tilde{\delta} T:T}(s)\right| - q(\delta) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \right|, \tag{A.49}$$

and then alter the expression to a convenient final form shown in Equation A.52,

where all absolute value terms are separated for easy manipulation.

$$\begin{aligned}
&\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \\
&\quad = \Bigg| q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \hat{\mathcal{F}}_{1:\delta T}(s) + \hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) + \hat{\mathcal{F}}_{\delta T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \\
&\qquad - q(\delta) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \Bigg| \qquad \text{(A.50)} \\
&\quad \leq \Bigg| q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \hat{\mathcal{F}}_{1:\delta T}(s)\right| + q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \\
&\qquad + \Big(q(\tilde{\delta}) - q(\delta)\Big) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \Bigg| \qquad \text{(A.51)} \\
&\quad \leq q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \hat{\mathcal{F}}_{1:\delta T}(s)\right| + q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \\
&\qquad + \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \qquad \text{(A.52)}
\end{aligned}$$

For each piece in Equation A.52, we examine the result based on the relationship between $\tilde{\delta}$, δ, and δ_0. The full exposition is given below for the case where $\tilde{\delta} < \delta \leq \delta_0$ to obtain $\hat{A}_T^1$, and abbreviated versions are included for the remaining pieces $\hat{A}_T^2$ through $\hat{A}_T^6$ that follow a similar structure.

We proceed under condition one of six, where $\tilde{\delta} < \delta \leq \delta_0$. For the terms in Equation A.52, we separate the form into distinct pieces of the empirical CDFs and introduce

the true distribution functions $\mathcal{F}_1(s)$ and $\mathcal{F}_T(s)$.

$$
\begin{aligned}
& q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \hat{\mathcal{F}}_{1:\delta T}(s) \right| \\
& \quad = q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \frac{\tilde{\delta}}{\delta}\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \frac{\delta-\tilde{\delta}}{\delta}\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) \right| && \text{(A.53)} \\
& \quad = q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \frac{\delta-\tilde{\delta}}{\delta}\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \frac{\delta-\tilde{\delta}}{\delta}\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) \right| && \text{(A.54)} \\
& \quad = q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \frac{\delta-\tilde{\delta}}{\delta}\left(\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right) - \frac{\delta-\tilde{\delta}}{\delta}\left(\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s)\right) \right| && \text{(A.55)} \\
& \quad \leq q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s) \right| + q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s) \right| && \text{(A.56)}
\end{aligned}
$$

For the second term,

$$
\begin{aligned}
& q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s) \right| \\
& \quad = q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \frac{\delta-\tilde{\delta}}{1-\tilde{\delta}}\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) + \frac{1-\delta}{1-\tilde{\delta}}\hat{\mathcal{F}}_{\delta T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s) \right| && \text{(A.57)} \\
& \quad = q(\tilde{\delta}) \sup_{s\in[0,1]} \left| \frac{\delta-\tilde{\delta}}{1-\tilde{\delta}}\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \frac{\delta-\tilde{\delta}}{1-\tilde{\delta}}\hat{\mathcal{F}}_{\delta T:T}(s) \right| && \text{(A.58)} \\
& \quad = q(\tilde{\delta})\frac{\delta-\tilde{\delta}}{1-\tilde{\delta}} \sup_{s\in[0,1]} \left| \hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \frac{\delta_0-\delta}{1-\delta}\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \frac{1-\delta_0}{1-\delta}\hat{\mathcal{F}}_{\delta_0 T:T}(s) \right|, && \text{(A.59)}
\end{aligned}
$$

and following the same outline as above,

$$
\begin{aligned}
& q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \\
& \quad \leq q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s)\right| + q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\
& \quad\quad + q(\tilde{\delta}) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| + q(\tilde{\delta}) \frac{\delta - \tilde{\delta}}{1-\tilde{\delta}}\theta. \qquad \text{(A.60)}
\end{aligned}
$$

For the third term,

$$
\begin{aligned}
& \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \hat{\mathcal{F}}_{\delta T:T}(s)\right| \\
& \quad \leq \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| + \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T} T(s) - \mathcal{F}_1(s)\right| \\
& \quad\quad + \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T} T(s) - \mathcal{F}_1(s)\right| \\
& \quad\quad + \left|q(\tilde{\delta}) - q(\delta)\right| \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T} T(s) - \mathcal{F}_T(s)\right| + \left|q(\tilde{\delta}) - q(\delta)\right| \theta. \qquad \text{(A.61)}
\end{aligned}
$$

Combining the terms from Equations A.56, A.60, and A.61, we obtain

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{1-\tilde{\delta}} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta. \quad \text{(Case 1: } \tilde{\delta} < \delta \leq \delta_0\text{)}\end{aligned} \tag{A.62}$$

Each q term in the first four pieces of Equation A.62 is bounded above: $q(\tilde{\delta}) \leq 1/2$, $q(\delta) \leq 1/2$, and their difference $|q(\tilde{\delta}) - q(\delta)| < 1/2$. We can substitute the coefficients 1, 3/2, 1, and 1 that appear below in Equation A.68 to resemble $\hat{A}_T^1$ in Equation A.43. All that remains is to manipulate the final term of Equation A.62 to look like the form in Lemma A.3.

In the initial scenario $\tilde{\delta} < \delta \leq \delta_0$ where $\tilde{\delta}, \delta, \delta_0 \in \left[\frac{1}{2} - \frac{1}{2}\sqrt{1-4\kappa^2}, \frac{1}{2} + \frac{1}{2}\sqrt{1-4\kappa^2}\right]$, we examine the final coefficient term of Equation A.62. For any $q(\tilde{\delta})$ and $q(\delta)$, we can write $|q(\tilde{\delta}) - q(\delta)| \leq \sqrt{|\tilde{\delta}(1-\tilde{\delta}) - \delta(1-\delta)|} \leq \sqrt{|\tilde{\delta} - \delta|}$, and the expression becomes

$$q(\tilde{\delta})\frac{(\delta - \tilde{\delta})}{(1-\tilde{\delta})} + \left|q(\tilde{\delta}) - q(\delta)\right| \leq q(\tilde{\delta})\frac{(\delta - \tilde{\delta})}{(1-\tilde{\delta})} + (\delta - \tilde{\delta})^{1/2}. \tag{A.63}$$

For some constant $0 \leq \alpha \leq 1$, we note that $(\delta - \tilde{\delta}) \leq (\delta - \tilde{\delta})^{\alpha}$, and similarly for

$\kappa < 1/2$, $(\delta - \tilde{\delta})^{1/2} < (\delta - \tilde{\delta})^{\kappa}$.

$$q(\tilde{\delta})\frac{(\delta - \tilde{\delta})}{(1 - \tilde{\delta})} + (\delta - \tilde{\delta})^{1/2} \leq \frac{q(\tilde{\delta})}{(1 - \tilde{\delta})}(\delta - \tilde{\delta})^{1/2} + (\delta - \tilde{\delta})^{1/2} \tag{A.64}$$

$$\leq \left(\frac{\tilde{\delta}^{1/2}}{(1 - \tilde{\delta})^{1/2}} + 1\right)(\delta - \tilde{\delta})^{1/2} \tag{A.65}$$

$$< \left(\frac{\tilde{\delta}^{1/2}}{(1 - \tilde{\delta})^{1/2}} + 1\right)(\delta - \tilde{\delta})^{\kappa} \tag{A.66}$$

Equation A.66 takes the form

$$\left(\frac{\tilde{\delta}^{1/2}}{(1 - \tilde{\delta})^{1/2}} + 1\right)(\delta - \tilde{\delta})^{\kappa} \leq C_1 \left|\tilde{\delta} - \delta\right|^{\kappa}, \tag{A.67}$$

where the maximum value of C_1 will occur at large $\tilde{\delta}$. We use a slightly cleaner form of the domain restriction to write $\kappa^2 \leq \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\kappa^2} \leq \tilde{\delta}, \delta \leq \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\kappa^2} \leq 1 - \kappa^2$, and set $C_1 = (1/\kappa)\sqrt{1 - \kappa^2} + 1$. With $C = (2/\kappa)\sqrt{1 - \kappa^2} + 1 < \infty$ defined above and $C_1 \leq C$, we adjust the final term to look like that in Lemma A.3.

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| \\ & + \frac{3}{2} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta T}(s) - \mathcal{F}_1(s)\right| \\ & + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \\ & + C\theta \left|\tilde{\delta} - \delta\right|^{\kappa} \end{aligned} \tag{A.68}$$

$$\leq \hat{A}_T^1 + \hat{B}_T \left\|\tilde{\delta} - \delta\right\|^{\kappa}, \tag{A.69}$$

where $\kappa > 0$ is the small constant from $q(\delta)$ that appears in the domain restriction. The last term in Equation A.62 is bounded above by the equivalent in Equation A.68 for all $\tilde{\delta} < \delta \leq \delta_0$ where $\tilde{\delta}, \delta, \delta_0 \in \Delta$ and when the trivial boundary condition $\kappa < 1/2$ is satisfied.

In case two of six where $\delta < \tilde{\delta} \leq \delta_0$, we can write

$$\begin{aligned}
\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right| \\
& + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\tilde{\delta} T}(s) - \mathcal{F}_1(s)\right| \\
& + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta} T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\
& + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:T}(s) - \mathcal{F}_T(s)\right| \\
& + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{1 - \tilde{\delta}} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta. \quad \text{(Case 2: } \delta < \tilde{\delta} \leq \delta_0\text{)}
\end{aligned} \tag{A.70}$$

We obtain the coefficients for $\hat{A}_T^2$ in Equation A.44 of 1, 3/2, 1, and 1 from the bounds $q(\tilde{\delta}) \leq 1/2$, $q(\delta) \leq 1/2$, and $|q(\tilde{\delta}) - q(\delta)| < 1/2$. In the same process as above, we can show the final piece is bounded by $C\theta|\tilde{\delta} - \delta|^\kappa$ where $C = (2/\kappa)\sqrt{1 - \kappa^2} + 1$.

For case three,

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\tilde{\delta}T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\delta T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{\delta} + q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{1-\tilde{\delta}} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta, \quad \text{(Case 3: } \tilde{\delta} \leq \delta_0 < \delta\text{)}\end{aligned} \tag{A.71}$$

we obtain the coefficients for $\hat{A}_T^3$ in Equation A.45 of 1, 3/2, 3/2, and 1, and $C = (2/\kappa)\sqrt{1-\kappa^2} + 1$ as above when $\tilde{\delta}$ is large and δ small.

For case four,

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\tilde{\delta} T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta}T:T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{\tilde{\delta}} + q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{1-\delta} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta, \quad \text{(Case 4: } \delta \leq \delta_0 < \tilde{\delta}\text{)}\end{aligned} \tag{A.72}$$

we obtain the coefficients for $\hat{A}_T^4$ in Equation A.46 of 1, 3/2, 3/2, and 1, and $C = (2/\kappa)\sqrt{1-\kappa^2}+1$ as above when $\tilde{\delta}$ is small and δ large.

For case five,

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\tilde{\delta} T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta} T:\delta T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{\tilde{\delta}} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta, \quad \text{(Case 5: } \delta_0 < \tilde{\delta} < \delta\text{)}\end{aligned} \tag{A.73}$$

we obtain the coefficients for $\hat{A}_T^5$ in Equation A.47 of 1, 1, 3/2, and 1, and $C = (2/\kappa)\sqrt{1-\kappa^2}+1$ as above when $\tilde{\delta}$ is small.

For the sixth and final case,

$$\begin{aligned}\left|\hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta)\right| \leq & \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{1:\delta_0 T}(s) - \mathcal{F}_1(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta_0 T:\delta T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(2q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\delta T:\tilde{\delta} T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta}) + \left|q(\tilde{\delta}) - q(\delta)\right|\right) \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{\tilde{\delta} T:T}(s) - \mathcal{F}_T(s)\right| \\ & + \left(q(\tilde{\delta})\frac{\delta - \tilde{\delta}}{\tilde{\delta}} + \left|q(\tilde{\delta}) - q(\delta)\right|\right)\theta, \quad \text{(Case 6: } \delta_0 < \delta < \tilde{\delta}\text{)}\end{aligned} \tag{A.74}$$

we obtain the coefficients for $\hat{A}_T^6$ in Equation A.48 of 1, 1, 3/2, and 1, and $C = (2/\kappa)\sqrt{1-\kappa^2} + 1$ as above when $\tilde{\delta}$ is small.

For all cases $\tilde{\delta}, \delta, \delta_0 \in \Delta$, the result of Lemma A.3 holds with the corresponding $\hat{A}_T^i$, $\hat{B}_T = \theta\left[(2/\kappa)\sqrt{1-\kappa^2} + 1\right]$, and $\alpha = \kappa$. Thus, $\hat{Q}_T(\delta)$ is stochastically equicontinuous and uniformly converges in probability to $Q_0(\delta)$ on the interval Δ. □

We now employ Theorem 2.1 from Newey and McFadden (1994) and Lemma A.4 to complete the proof of Theorem 2.3.

Proof of Theorem 2.3. We begin with condition (1). It can easily be seen that the quantity in Equation A.35 is uniquely maximized to a value of θ at $\delta = \delta_0$, thus $Q_0(\delta)$ in Equation A.36 has a unique maximum at $\delta = \delta_0$. If we relax the boundary restriction $\delta \in \Delta$ and let $\delta \in (0, 1)$, the unique maximum will still hold provided $\kappa \leq \delta_0^{1/2}(1-\delta_0)^{1/2}$, or $\delta_0 \in \Delta$.

Condition (2) is satisfied as Δ is a closed and bounded set on $\mathbb{R}$.

For condition (3), the individual pieces of Equation A.36 are continuous, and we examine the extreme points of each subdomain. Because the set Δ is closed, continuity is trivially satisfied at the outward extremes. For the point δ_0,

$$\lim_{\delta \to \delta_0^+} Q_0(\delta_0) = Q_0(\delta_0) = \theta \delta_0^{1/2} (1 - \delta_0)^{1/2} \tag{A.75}$$

$$\text{and} \quad \lim_{\delta \to \delta_0^-} Q_0(\delta_0) = \lim_{\delta \to \delta_0^-} \theta \frac{\delta_0 (1 - \delta)^{1/2}}{\delta^{1/2}} = \theta \delta_0^{1/2} (1 - \delta_0)^{1/2}. \tag{A.76}$$

Condition (4) is satisfied via Lemma A.4.

The four conditions of Theorem 2.1 from Newey and McFadden (1994) are satisfied and $\hat{\delta} \xrightarrow{P} \delta_0$ implies $\hat{\tau} \xrightarrow{P} \tau$ for $t, \tau \in \left[\frac{T}{2} - \frac{T}{2}\sqrt{1 - 4\kappa^2}, \frac{T}{2} + \frac{T}{2}\sqrt{1 - 4\kappa^2}\right]$. □

Berkes, Hörmann, and Schauer (2009), Theorem 2. *Let $\{S_t,\ t \in \mathbb{Z}\}$ be a stationary sequence such that $\mathcal{F}(s) = P(S_1 \leq s)$ is Lipschitz continuous of order $C > 0$. Assume that $\{S_t,\ t \in \mathbb{Z}\}$ is $\mathcal{S}$-mixing and condition (1) of Definition 2.1 holds with $\gamma_m = m^{-AC}$, $\delta_m = m^{-A}$ for some $A > 4$. Then the series*

$$\Gamma(s, s') = \sum_{-\infty < t < \infty} \mathbb{E}\left[S_1(s) S_t(s')\right] \tag{A.77}$$

converges absolutely for every choice of parameters $(s, s') \in \mathbb{R}^2$. Moreover, there exists a two-parameter Gaussian process $\mathcal{K}(s, t)$ such that $\mathbb{E}\left[\mathcal{K}(s, t)\right] = 0$, $\mathbb{E}\left[\mathcal{K}(s, t)\ \mathcal{K}(s', t')\right] = (t \wedge t')\ \Gamma(s, s')$, and for some $\alpha > 0$,

$$\sup_{1 \leq t \leq T} \sup_{s \in [0,1]} \left| \sum_{i=1}^{t} \left(\mathbf{1}\{S_i \leq s\} - \mathcal{F}(s)\right) - \mathcal{K}(s, t) \right| = o\left(T^{1/2} (\log T)^{-\alpha}\right) \quad a.s. \tag{A.78}$$

Definition A.5. A sequence of functions $\hat{Q}_T(\delta)$ is stochastically equicontinuous if for every $\varepsilon, \eta > 0$ there exists a random quantity $\Gamma_T(\varepsilon, \eta)$ and a constant $T_0(\varepsilon, \eta)$ such that for $T \geq T_0(\varepsilon, \eta)$, $\mathbb{P}\left(|\Gamma_T(\varepsilon, \eta)| > \varepsilon\right) < \eta$ and for each δ there is an open set $\mathcal{N}(\delta, \varepsilon, \eta)$ containing δ with

$$\sup_{\tilde{\delta} \in \mathcal{N}(\delta, \varepsilon, \eta)} \left| \hat{Q}_T(\tilde{\delta}) - \hat{Q}_T(\delta) \right| \leq \Gamma_T(\varepsilon, \eta), \quad \text{for } T > T_0(\varepsilon, \eta). \tag{A.79}$$

Newey and McFadden (1994), Theorem 2.1. *If there is a function $Q_0(\delta)$ such that*

(1) $Q_0(\delta)$ is uniquely maximized at δ_0, $\delta_0 = \arg\max_{\delta \in \Delta} Q_0(\delta)$;

(2) Δ is compact;

(3) $Q_0(\delta)$ is continuous;

(4) $\hat{Q}_T(\delta)$ converges uniformly in probability to $Q_0(\delta)$, $\sup_{\delta \in \Delta} \left| \hat{Q}_T(\delta) - Q_0(\delta) \right| \xrightarrow{P} 0$;

then $\hat{\delta} = \arg\max_{\delta \in \Delta} \hat{Q}_T(\delta) \xrightarrow{P} \delta_0$.

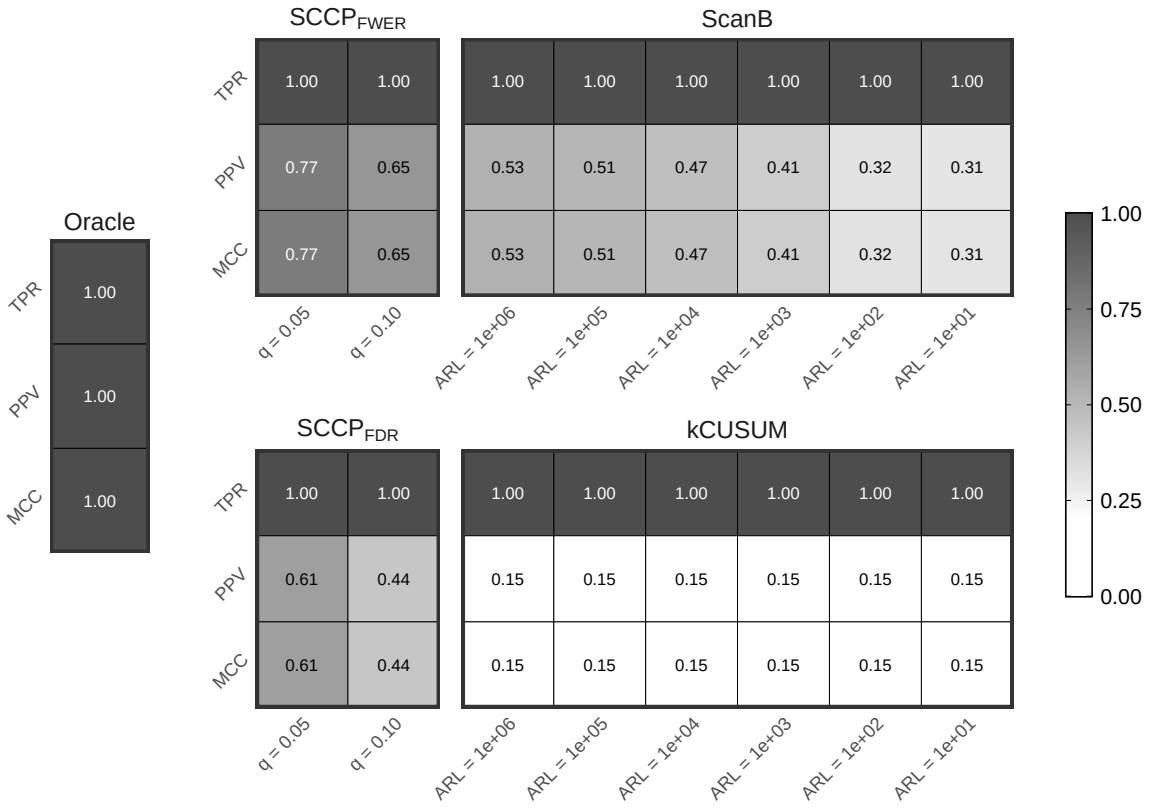
Appendix B

Additional Material for Chapter 3

Code and supporting material: Files (.RData, .csv, and .mat) used to assess performance of change point methods, code (.R, .cpp, and .m) used to generate results and figures, and data for generating output in Sections 3.4 and 3.5 can be found at

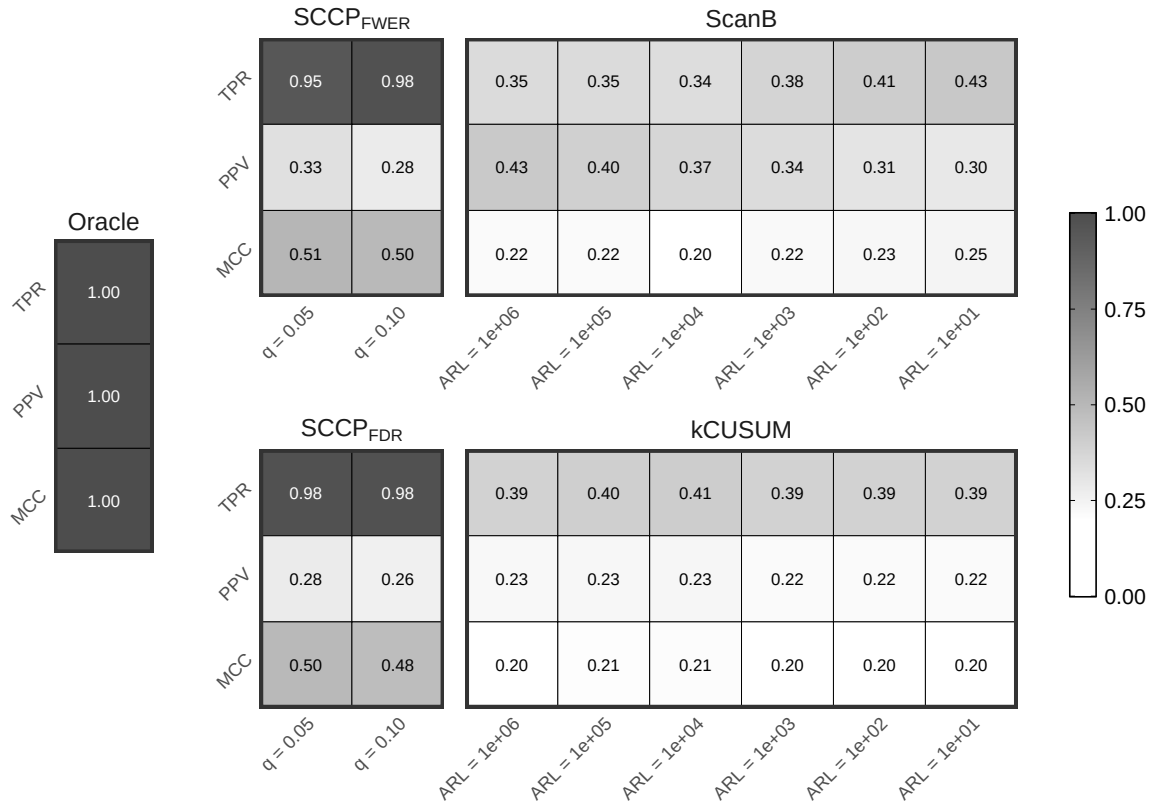
`github.com/noahgade/MCCP`.

B.1 Additional Figures



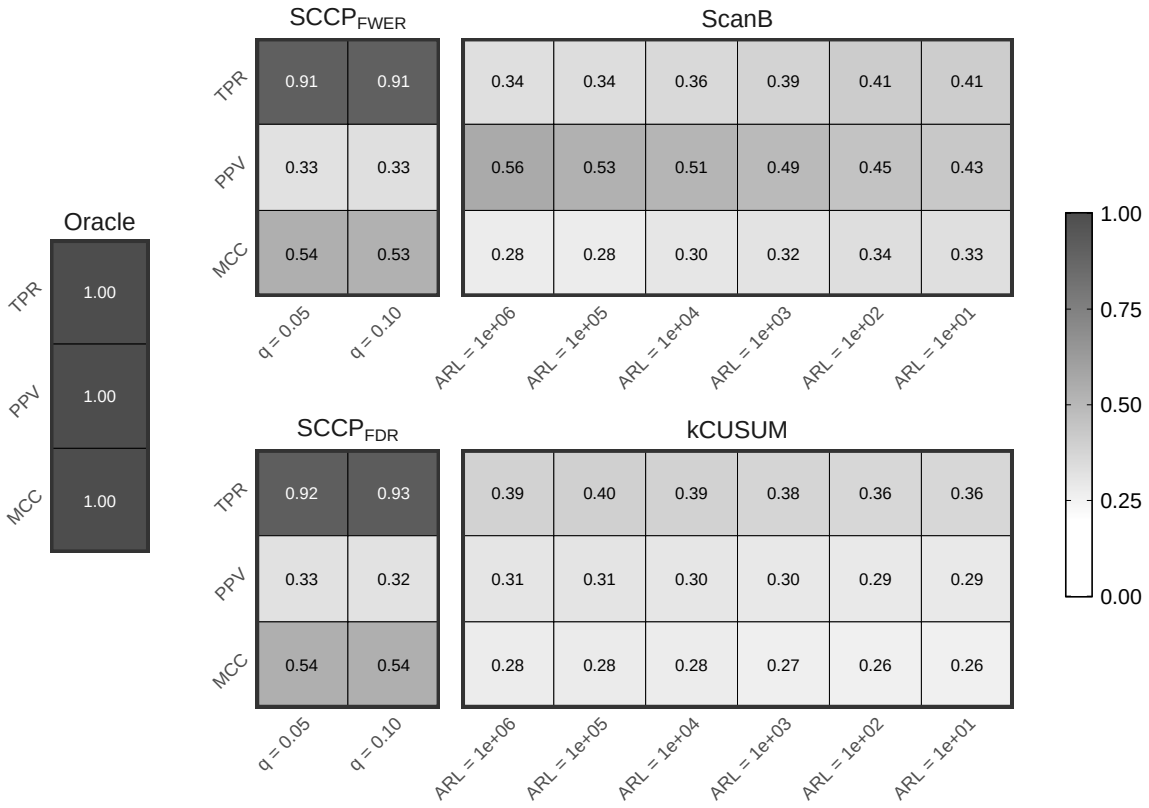
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.1: Gaussian process online detection simulation results for $n(\boldsymbol{\tau}) = 0$.



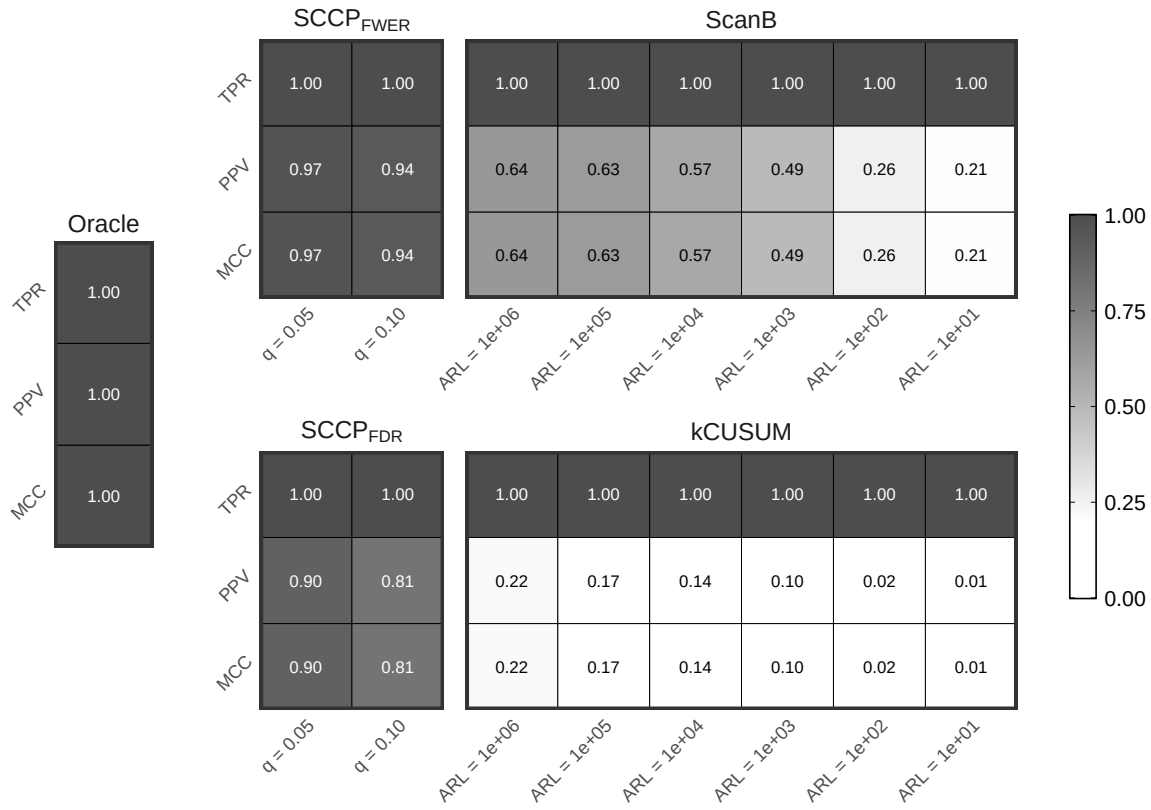
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.2: Gaussian process online detection simulation results for $n(\boldsymbol{\tau}) = 1$.



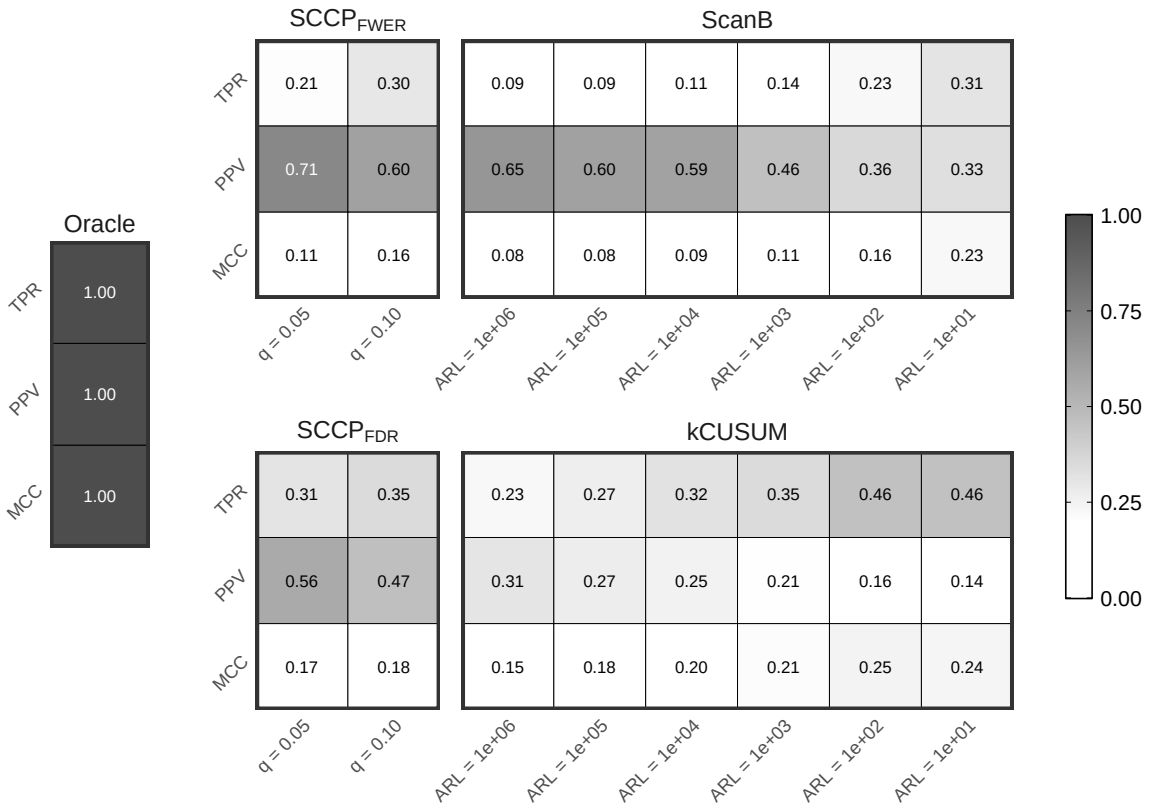
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.3: Gaussian process online detection simulation results for $n(\boldsymbol{\tau}) = 2$.



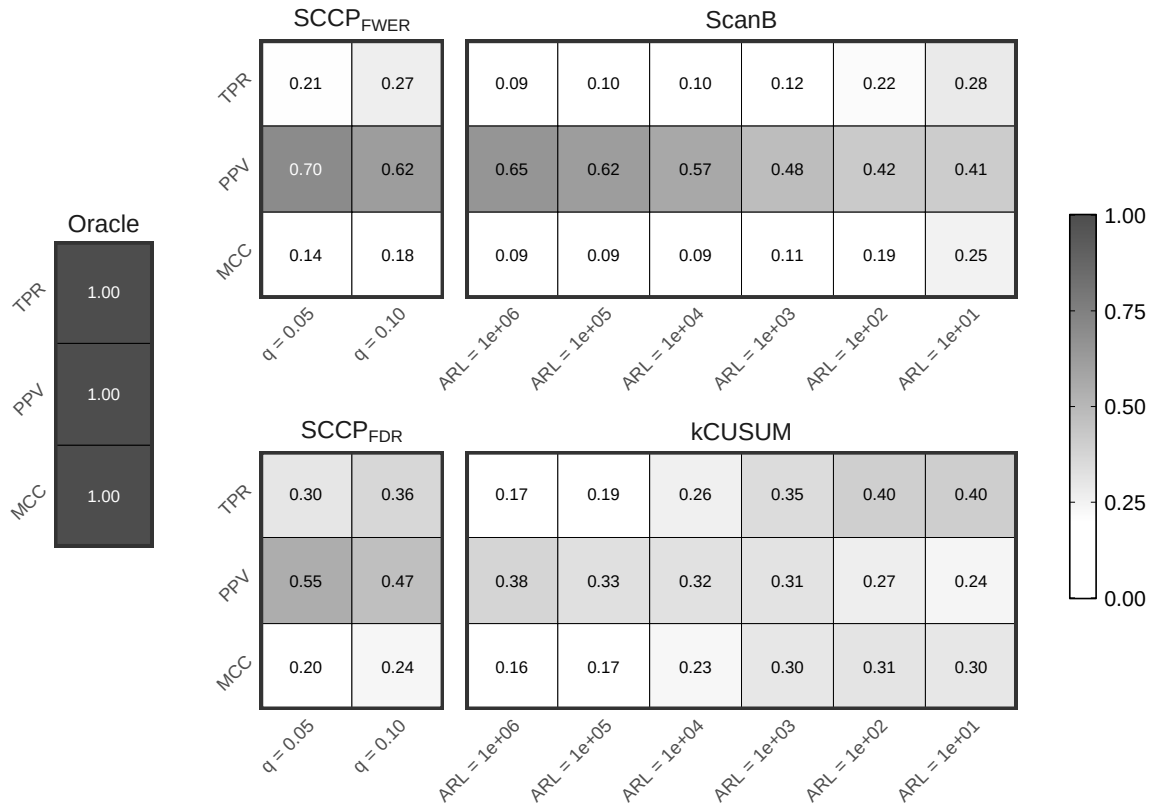
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.4: Threshold autoregressive process online detection simulation results for $n(\boldsymbol{\tau}) = 0$.



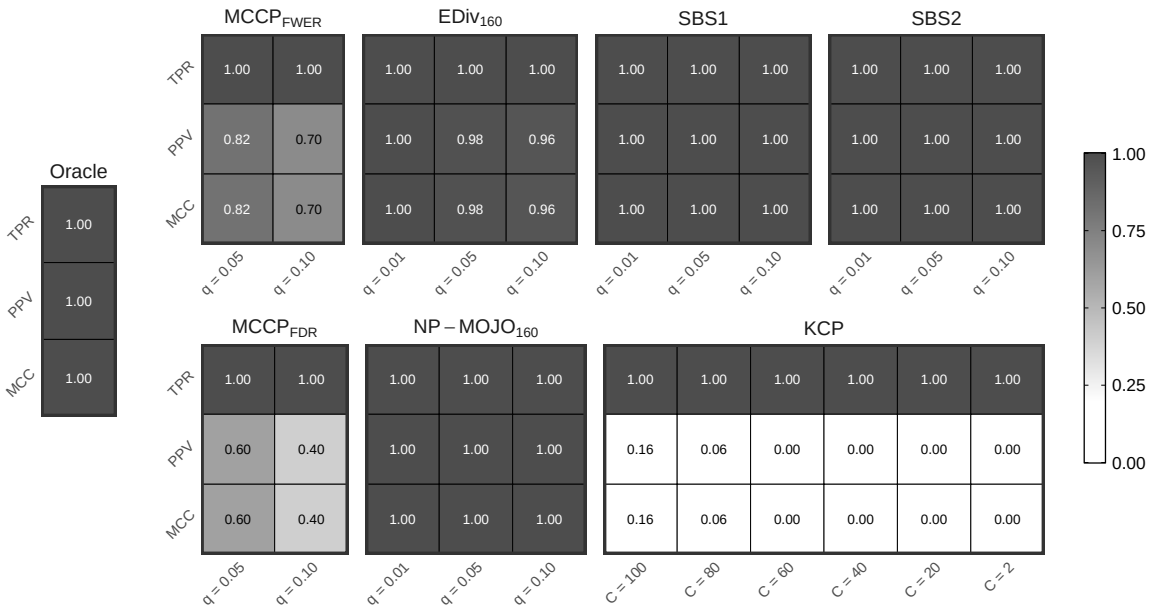
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.5: Threshold autoregressive process online detection simulation results for $n(\boldsymbol{\tau}) = 1$.



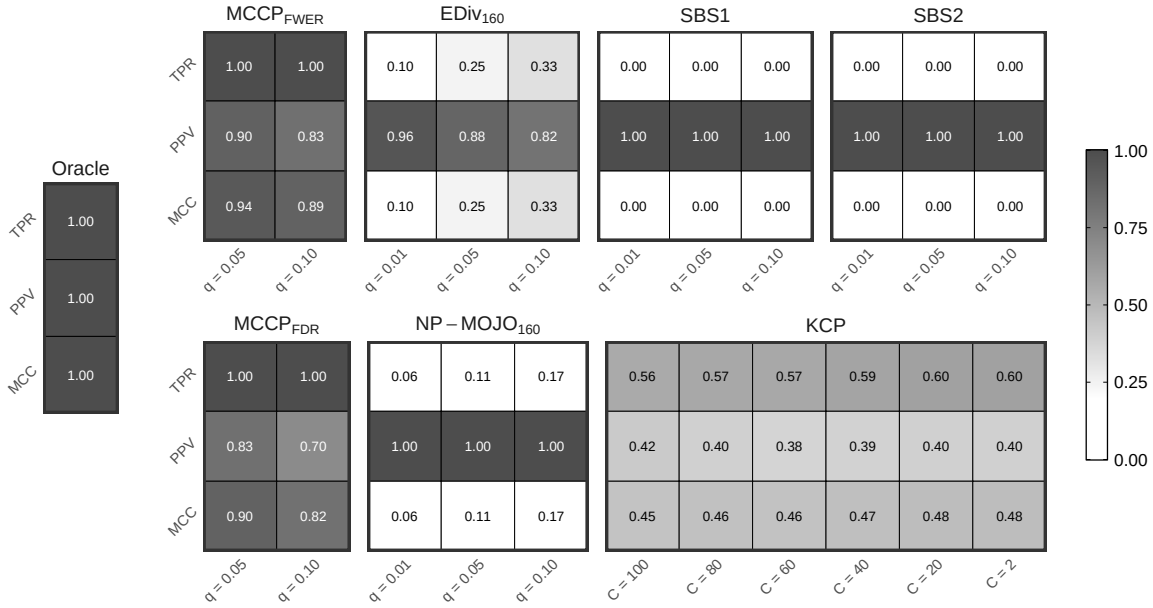
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on SCCP methods refer to the error control procedure.

Figure B.6: Threshold autoregressive process online detection simulation results for $n(\boldsymbol{\tau}) = 2$.



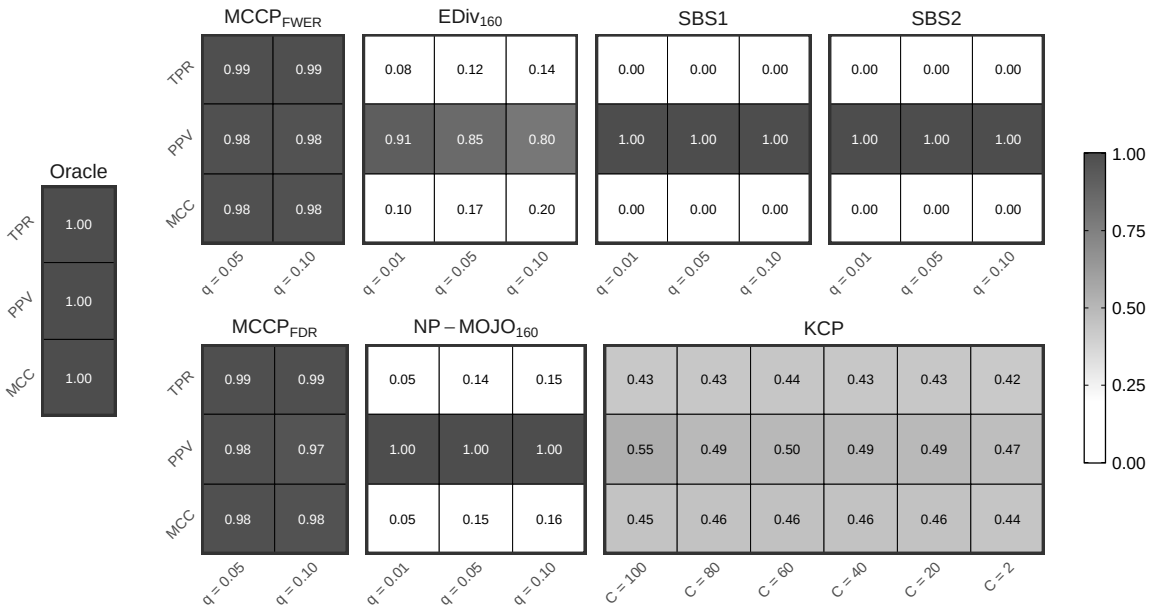
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.7: Gaussian process offline detection simulation results for $n(\boldsymbol{\tau}) = 0$.



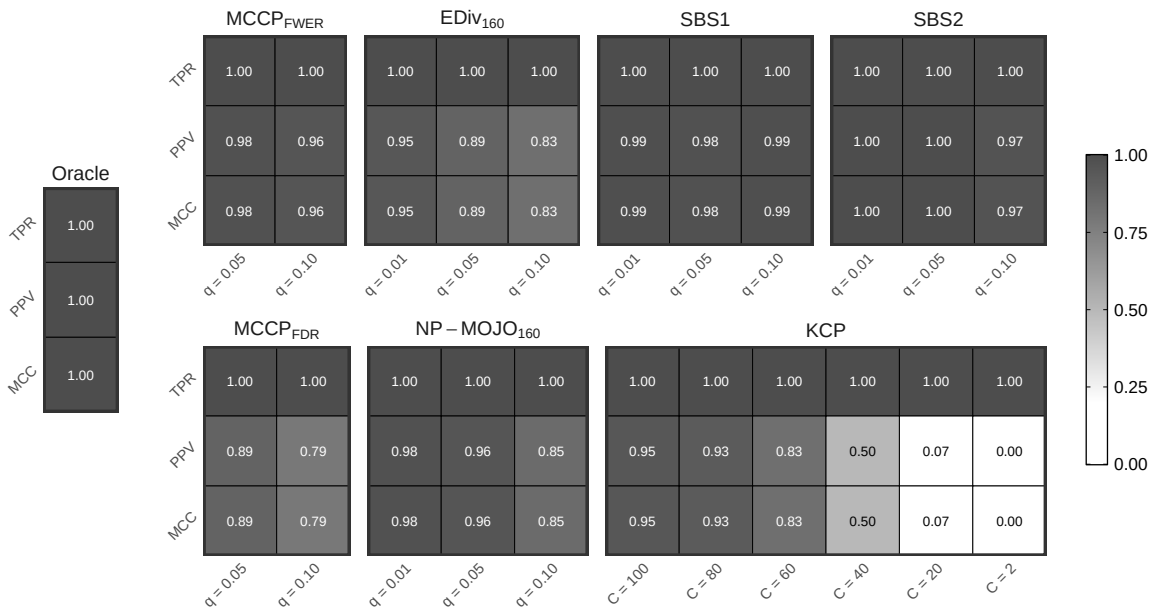
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.8: Gaussian process offline detection simulation results for $n(\boldsymbol{\tau}) = 1$.



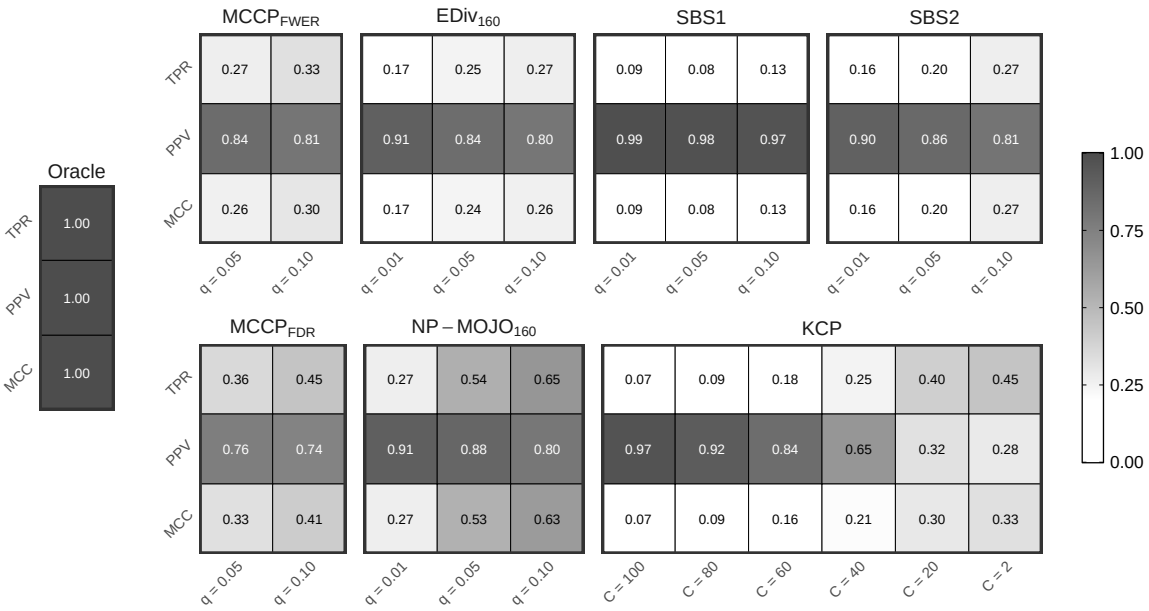
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.9: Gaussian process offline detection simulation results for $n(\boldsymbol{\tau}) = 2$.



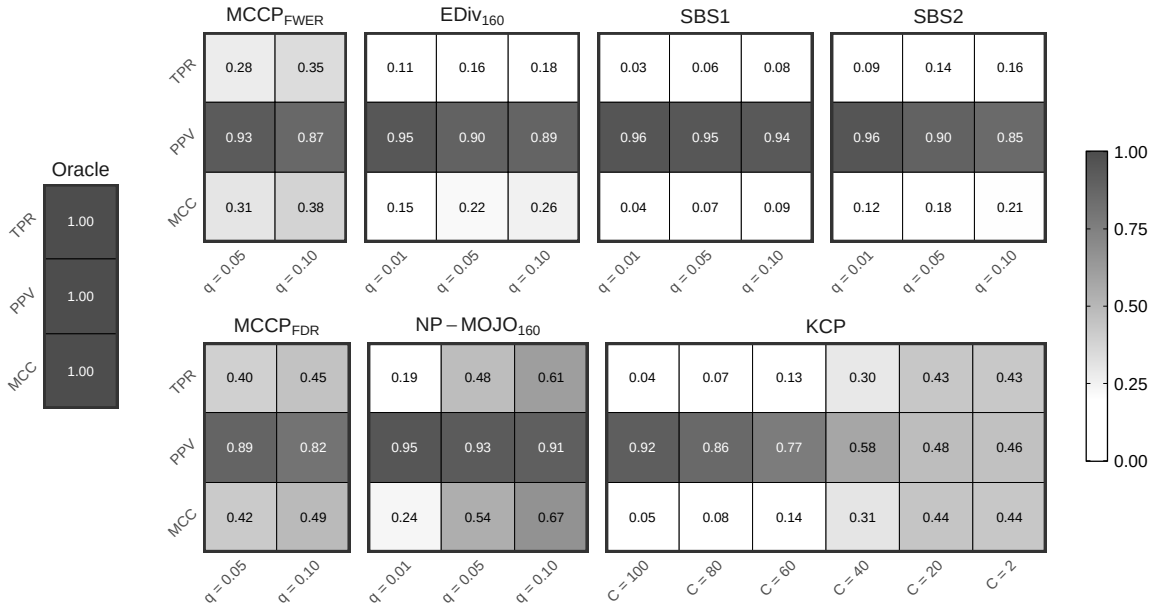
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.10: Threshold autoregressive process offline detection simulation results for $n(\boldsymbol{\tau}) = 0$.



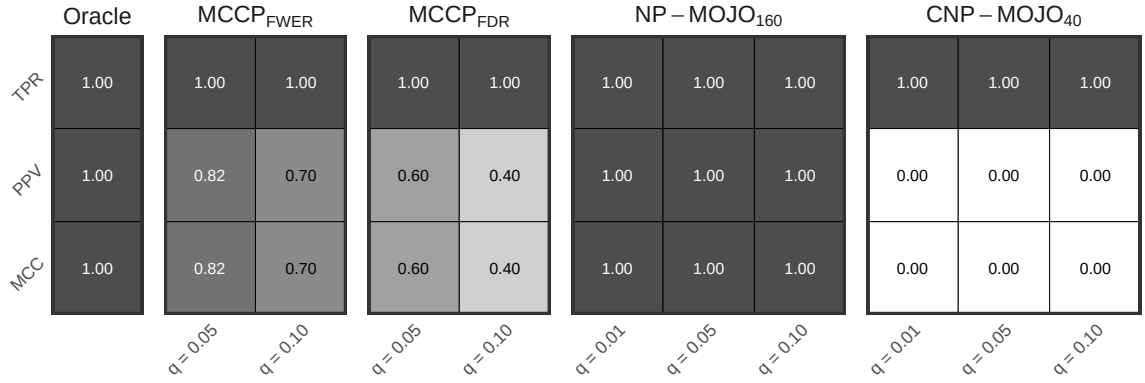
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.11: Threshold autoregressive process offline detection simulation results for $n(\boldsymbol{\tau}) = 1$.



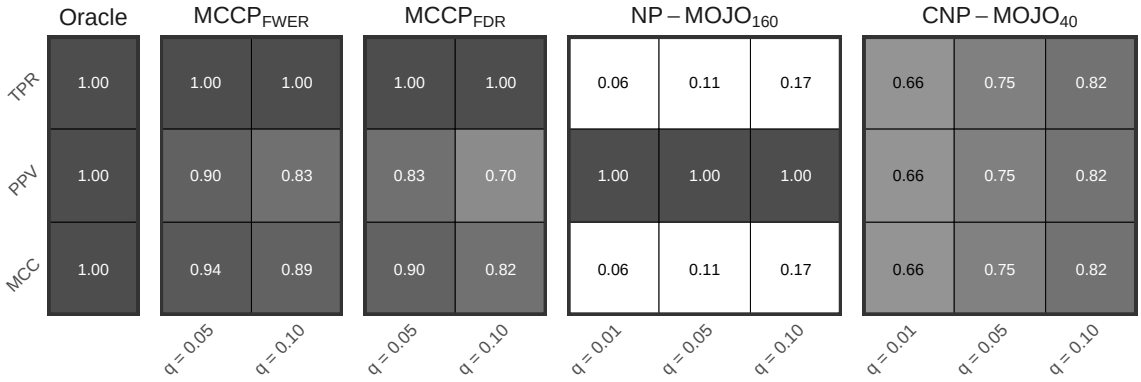
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on EDiv and NP-MOJO refer to the defined minimum separation (window size).

Figure B.12: Threshold autoregressive process offline detection simulation results for $n(\boldsymbol{\tau}) = 2$.



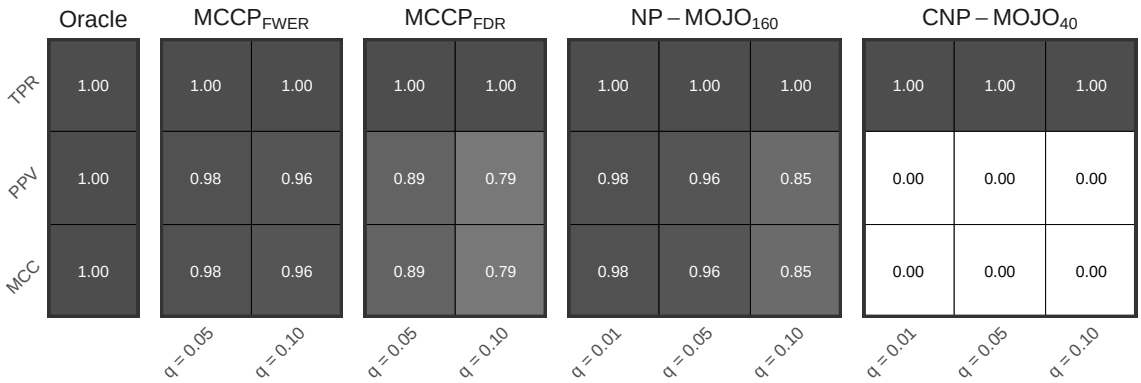
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

Figure B.13: Gaussian process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2 with $n(\boldsymbol{\tau}) = 0$.



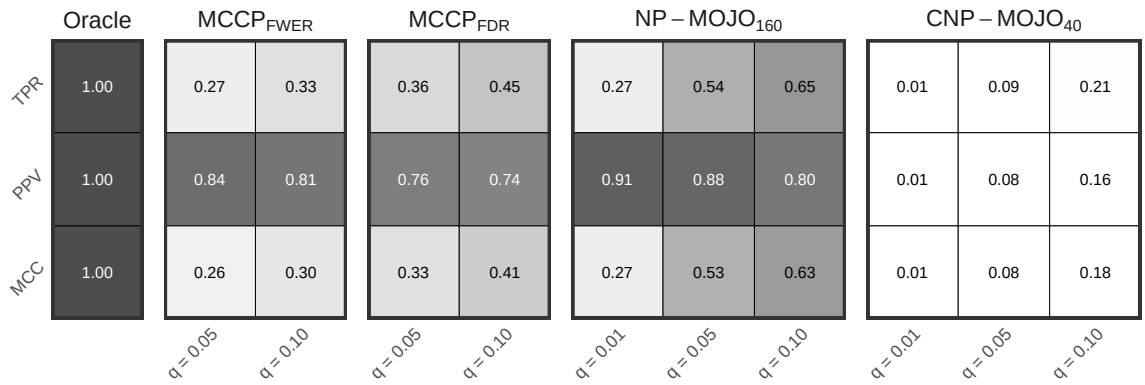
Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

Figure B.14: Gaussian process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2 with $n(\boldsymbol{\tau}) = 1$.



Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

Figure B.15: Threshold autoregressive process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2 with $n(\boldsymbol{\tau}) = 0$.



Change points correctly identified if within radius $r_{\gamma^*} = T_{\text{wash}} = 40$. Subscripts on MCCP methods refer to the error control procedure, and on (C)NP-MOJO refer to the defined minimum separation (window size).

Figure B.16: Threshold autoregressive process comparison between MCCP and NP-MOJO methods on simulated data and the transformed sequence of cosine similarities (CNP-MOJO) from the CCP method of Chapter 2 with $n(\boldsymbol{\tau}) = 1$.

B.2 Additional Tables

Table B.1: Parameter settings for methods in multiple change point simulation study.

Method		Settings
Sequential Conceptor Change Point	(SCCP)	$T_{\text{wash}} = 40$, $T_{\text{train}} = 120$, $\{\text{FWER}, \text{FDR}\}$, $q = \{0.05, 0.1\}$
Scan-B	(ScanB)	block length $= 31$, # blocks $= 5$, ARL $= \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$
Kernel CUSUM	(kCUSUM)	block length $= 31$, # blocks $= 5$, ARL $= \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$

(a) Online Change Point Detection Methods.

Method		Settings
Multiple Conceptor Change Point	(MCCP)	$T_{\text{wash}} = 60$, $T_{\text{train}} = 120$, $\{\text{FWER}, \text{FDR}\}$, $q = \{0.05, 0.1\}$
E-Divisive	(EDiv)	$\gamma^* = 161$, $q = \{0.01, 0.05, 0.1\}$
NP-MOJO	(NP-MOJO)	$\gamma^* = 181$, lags $= \{0, 1, 2\}$, $q = \{0.01, 0.05, 0.1\}$
Kernel Change Point	(KCP)	$\max\{n(\hat{\boldsymbol{\tau}})\} = 2$, $C = \{2, 20, 40, 60, 80, 100\}$
Sparsified Binary Segmentation	(SBS1/2)	Type $= \{1, 2\}$, $q = \{0.01, 0.05, 0.1\}$

(b) Offline Change Point Detection Methods.

B.3 Additional Algorithms

This section presents additional procedures composing the main algorithms in Chapter 3.

Procedure B.1 ESN Featurization & Conceptor Computation

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$, $t = 1, \ldots T_{\text{wash}} + T_{\text{train}}$; training window length T_{train}; washout length T_{wash}

Outputs: ESN parameters, reservoir size N, all $\mathbf{C}_r, \mathbf{W^i}_r$, $\mathbf{b}_r$, $\mathbf{W^h}_r$

Default Parameters: ESN spectral radius $\rho \leftarrow 0.8$; grid of possible $\mathbf{W^i}_r$, $\mathbf{b}_r$ scalings $G \leftarrow \{c_{\text{input}} \leftarrow \{0.6, 1.0, 1.4\}, c_{\text{bias}} \leftarrow \{0.1, 0.3, 0.5\}\}$; reservoir scaling $c_{\text{res}} \leftarrow 0.9$; number of test initializations $\mathcal{R}_{\text{test}} \leftarrow 5$; output regularization parameter $\lambda \leftarrow 10^{-6}$; number of featurizations $\mathcal{R}_{\text{test}} \leftarrow 100$; aperture $\alpha \leftarrow 100$

$N \leftarrow \lfloor c_{\text{res}} T_{\text{train}} \rfloor$; $T_0 \leftarrow T_{\text{wash}} + T_{\text{train}}$

for each grid scaling combination of c_{input} and c_{bias} in G **do**

  for r in $1 : \mathcal{R}_{\text{test}}$ **do**

    initialize $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$ where each element $\mathcal{N}(0, 1)$, and $\mathbf{W^h}_r$ is sparse

    $\mathbf{W^i}_r \leftarrow c_{\text{input}} \mathbf{W^i}_r$; $\mathbf{b}_r \leftarrow c_{\text{bias}} \mathbf{b}_r$

    $\mathbf{W^h}_r \leftarrow \rho \mathbf{W^h}_r / \max \left\{ \mathbf{v}^\top \mathbf{W^h}_r \mathbf{v} : ||\mathbf{v}|| = 1 \right\}$

    $\mathbf{h}_{r,t} \leftarrow \tanh \left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r \right)$ for $t = 1, \ldots, T_0$

    $\mathbf{W^o}_r \leftarrow \left(\mathbf{H}_r^\top \mathbf{H}_r + \lambda \mathbf{I} \right)^{-1} \mathbf{H}_r^\top \mathbf{Y}$ where $\mathbf{H}_r = \left[\mathbf{h}_{r, T_{\text{wash}}+1} \ \cdots \ \mathbf{h}_{r,T_0} \right]^\top$

  end for

  NRMSE $\leftarrow \frac{1}{\mathcal{R}} \sum_{j=1}^{\mathcal{R}} \sqrt{\frac{(\mathbf{Y} - \mathbf{H}_r \mathbf{W^o}_r)^2}{\frac{1}{2}\text{Var}(\mathbf{Y}) + \frac{1}{2}\text{Var}(\mathbf{H}_r \mathbf{W^o}_r)}}$

end for

ESN scaling : $\{c_{\text{input}}, c_{\text{bias}}, \rho\} \leftarrow \arg\min_G \{\text{NRMSE}\}$

for r in $1 : \mathcal{R}$ **do**

  initialize $\mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r$ where each element $\mathcal{N}(0, 1)$, and $\mathbf{W^h}_r$ is sparse

  $\mathbf{W^i}_r \leftarrow c_{\text{input}} \mathbf{W^i}_r$; $\mathbf{b}_r \leftarrow c_{\text{bias}} \mathbf{b}_r$

  $\mathbf{W^h}_r \leftarrow \rho \mathbf{W^h}_r / \max \left\{ \mathbf{v}^\top \mathbf{W^h}_r \mathbf{v} : ||\mathbf{v}|| = 1 \right\}$

  $\mathbf{h}_{r,t} \leftarrow \tanh \left(\mathbf{W^h}_r \mathbf{h}_{r,t-1} + \mathbf{W^i}_r \mathbf{y}_t + \mathbf{b}_r \right)$ for $t = 1, \ldots, T_0$

  $\mathbf{C}_r \leftarrow \frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r \left(\frac{1}{T_{\text{train}}} \mathbf{H}_r^\top \mathbf{H}_r + \alpha^{-2} \mathbf{I} \right)^{-1}$ where $\mathbf{H}_r = \left[\mathbf{h}_{r, T_{\text{wash}}+1} \ \cdots \ \mathbf{h}_{r,T_0} \right]^\top$

end for

return ESN parameters, reservoir size N, all $\mathbf{C}_r, \mathbf{W^i}_r$, $\mathbf{b}_r$, $\mathbf{W^h}_r$

Procedure B.2 Generating Bootstrapped Time Series

Inputs: potential change point $\hat{\tau}_j^* \in \hat{\boldsymbol{\tau}}^*$; time series data $\mathbf{y}_t \in \mathbb{R}^d$; training length T_{train}; washout length T_{wash}

Outputs: bootstrapped time series $\mathbf{y}_{b,t}$

Default Parameters: number of bootstraps $B \leftarrow 99$; block length $L \leftarrow \lceil T^{1/3} \rceil$

for b in $1 : B$ **do**
  $T_0 \leftarrow T_{\text{wash}} + T_{\text{train}}$
  $T_{\text{end}} \leftarrow \min\left\{\hat{\tau}_j^* + T_0,\ T\right\}$
  for i in $1 : \lceil (T_{\text{end}} - T_0)/L \rceil$ **do**
    $\beta_{b,i} \leftarrow \beta \sim \text{Uniform}\left\{T_0 + 1,\ T_{\text{end}}\right\}$
    $\mathbf{b}_i \leftarrow \mathbf{y}_{\beta_{b,i}:\left(\beta_{b,i}+L-1\right)}$
  end for
  $\mathbf{y}_t^b \leftarrow \left[\mathbf{y}_{1:T_0}^\top\ \mathbf{b}_1^\top\ \cdots\ \mathbf{b}_{\lceil (T_{\text{end}}-T_0)/L \rceil}^\top\right]_{1:T_{\text{end}}}^\top$
end for
  return all $\mathbf{y}_{b,t}$

Procedure B.3 Generate MBB Null Distribution for a Potential Change Point

Inputs: potential change point $\hat{\tau}_j^* \in \hat{\boldsymbol{\tau}}^*$; corresponding maximum statistic K^j; time series data $\mathbf{y}_t \in \mathbb{R}^d$; training window length T_{train}; washout length T_{wash}; all $\mathbf{C}_r$, $\mathbf{W^i}_r$, $\mathbf{b}_r$, $\mathbf{W^h}_r$, $\mathcal{R}$ from Procedure B.1

Outputs: MBB null distribution estimate K_b^j; quantile estimate p_j

perform Procedure B.2 to obtain bootstrapped data $\mathbf{y}_{b,t}$ and B
for b in $1 : B$ **do**
  for r in $1 : \mathcal{R}$ **do**
    $\mathbf{h}_{b,r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \mathbf{h}_{b,r,t-1} + \mathbf{W^i}_r \mathbf{y}_{b,t} + \mathbf{b}_r\right)$ for $t = 1, \ldots, T_{\text{wash}}$
    $\mathbf{h}_{b,r,t} \leftarrow \tanh\left(\mathbf{W^h}_r \tilde{\mathbf{h}}_{b,r,t-1} + \mathbf{W^i}_r \mathbf{y}_{b,t} + \mathbf{b}_r\right)$;
    $\tilde{\mathbf{h}}_{b,r,t} \leftarrow \mathbf{C}_r \mathbf{h}_{b,r,t}$ for $t = T_{\text{wash}} + 1, \ldots, T_{\text{end}}$
    $s_{b,r,t} \leftarrow \frac{\tilde{\mathbf{h}}_{b,r,t}^\top \mathbf{h}_{b,r,t}}{||\tilde{\mathbf{h}}_{b,r,t}|| \, ||\mathbf{h}_{b,r,t}||}$ for $t = T_0 + 1, \ldots, T_{\text{end}}$
  end for
  $S_{b,t} \leftarrow \frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} s_{b,r,t}$ for $t = T_0 + 1, \ldots, T_{\text{end}}$
  for t in $(T_0 + 1) : (T_{\text{end}} - 1)$ **do**
    $\hat{\mathcal{F}}^b_{(T_0+1):t}(s) \leftarrow \frac{1}{t - T_0} \sum_{i=T_0+1}^{t} \mathbf{1}\{S_{b,i} \leq s\}$
    $\hat{\mathcal{F}}^b_{(t+1):T_{\text{end}}}(s) \leftarrow \frac{1}{T_{\text{end}} - t} \sum_{i=t+1}^{T_{\text{end}}} \mathbf{1}\{S_{b,i} \leq s\}$
    $K_{b,t} \leftarrow \frac{(t-T_0)(T_{\text{end}}-t)}{q(t)(T_{\text{end}}-T_0)^{3/2}} \sup_s \left| \hat{\mathcal{F}}^b_{(T_0+1):t}(s) - \hat{\mathcal{F}}^b_{(t+1):T_{\text{end}}}(s) \right|$
  end for
  $K_b^j \leftarrow \max_t K_{b,t}$
end for
$p_j \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{K_b^j > K^j\right\}$
return p_j, all K_b^j

Procedure B.4 Inference for Potential Change Points

Inputs: time series data $\mathbf{y}_t \in \mathbb{R}^d$; training length T_{train}; washout length T_{wash}; sequence of statistics K_t; all $\mathbf{C}_r, \mathbf{W^i}_r, \mathbf{b}_r, \mathbf{W^h}_r, \mathcal{R}$ from Algorithm B.1; testing cutoff threshold c_q

Outputs: estimated change point $\hat{\tau}$; corresponding maximum statistic K; corresponding p-value p, moving block bootstrap estimated null distribution K_b

$\hat{\tau} \leftarrow$ NULL
if $K_t \geq K_{t'} \; \forall \; |t - t'| \leq T_0$ **then**
    perform Procedure B.3 to obtain p_j, all K_b^j
    if $p_j \leq c_q$ **then**
        $\hat{\tau} \leftarrow t$
        $K \leftarrow K_t$; $p \leftarrow p_j$; $K_b \leftarrow K_b^j$
    end if
end if
return $\hat{\tau}$, K, p, K_b

Procedure B.5 Reconciliation of Estimated Change Point Sets

Inputs: estimated sets from forward and backward-looking procedures, $\hat{\boldsymbol{\tau}}^f$ and $\hat{\boldsymbol{\tau}}^b$

Outputs: estimated change point set $\hat{\boldsymbol{\tau}}$

$\hat{\boldsymbol{\tau}} \leftarrow \emptyset$; $\hat{\boldsymbol{\tau}}^{fb} \leftarrow \hat{\boldsymbol{\tau}}^f \cap \hat{\boldsymbol{\tau}}^b$
$\mathcal{B}_1 \leftarrow \mathcal{N}\left(\hat{\tau}^{fb}_{(1)}; \gamma^*\right) \cap \hat{\boldsymbol{\tau}}^{fb}$
$j \leftarrow 1$; $i \leftarrow 1$
while $i < n\left(\hat{\boldsymbol{\tau}}^{fb}\right)$ **do**
    $i \leftarrow i + 1$
    if $\mathcal{B}_j \cap \left[\mathcal{N}\left(\hat{\tau}^{fb}_{(i)}; \gamma^*\right) \cap \hat{\boldsymbol{\tau}}^{fb}\right] \neq \emptyset$ **then**
        $\mathcal{B}_j \leftarrow \mathcal{B}_j \cup \left[\mathcal{N}\left(\hat{\tau}^{fb}_{(i)}; \gamma^*\right) \cap \hat{\boldsymbol{\tau}}^{fb}\right]$
    else
        $j \leftarrow j + 1$
        $\mathcal{B}_j \leftarrow \mathcal{N}\left(\hat{\tau}^{fb}_{(i)}; \gamma^*\right) \cap \hat{\boldsymbol{\tau}}^{fb}$
    end if
end while
$n_{\mathcal{B}} \leftarrow j$
for k in $1 : n_{\mathcal{B}}$ **do**
    if $\max\{\mathcal{B}_k\} - \min\{\mathcal{B}_k\} < \gamma^*$ **then**
        $\hat{\boldsymbol{\tau}} \leftarrow \left\{\hat{\boldsymbol{\tau}},\ \lfloor n\left(\mathcal{B}_k\right)^{-1} \sum_{l=1}^{n(\mathcal{B}_k)} b_{k,l} \rfloor\right\}$
    else
        $\hat{\boldsymbol{\tau}} \leftarrow \{\hat{\boldsymbol{\tau}},\ \min\left(\mathcal{B}_k\right),\ \max\left(\mathcal{B}_k\right)\}$
    end if
end for
    return estimated change point set $\hat{\boldsymbol{\tau}}$

B.4 Proofs

Proof of Theorem 3.1 follows the proof of Theorem 2.2 in Section A.4, and Theorem 2.6.1 from Csörgő, Horváth, and Szyszkowicz (1997) with the relaxation of the i.i.d. sequence to a stationary, $\mathcal{S}$-mixing sequence.

Proof of Theorem 3.1. Let $\mathcal{F}(s)$ denote the true distribution function of all points in the sequence S_z, $z \in \boldsymbol{\zeta}_j$, where $\boldsymbol{\zeta}_j = \left[\tau^+_{j-1} + \zeta^* T,\ t + \zeta^* T - 1\right)$ as in Section 3.3.1. Define $K_t(s,\delta)$ as in Equation B.1 to directly resemble that in Equation A.2.

$$K_t(s,\delta) = \delta\,(1-\delta)\left[\hat{\mathcal{F}}\left(s; \tau^+_{j-1} + \zeta^* T, \delta\left(t - \tau^+_{j-1}\right) + \tau^+_{j-1} + \zeta^* T + 1\right)\right.$$
$$\left. - \hat{\mathcal{F}}\left(s; \delta\left(t - \tau^+_{j-1}\right) + \tau^+_{j-1} + \zeta^* T + 2, t + \zeta^* T - 1\right)\right] \tag{B.1}$$

In the limit as $T \to \infty$ for a fixed $\zeta^* \in \left(0, \frac{1}{2}\right)$, $\zeta^* T \to \infty$, and $n\left(\boldsymbol{\zeta}_j\right) \to \infty$. By Lemma A.1,

$$\sup_{\delta \in \boldsymbol{\Delta}} \sup_{s \in [0,1]} \left| \sqrt{t - \tau^+_{j-1}}\, K_t(s,\delta) - \mathcal{K}_t(s,\delta) \right| / q(\delta) = o(1), \tag{B.2}$$

where $\boldsymbol{\Delta} = \left[\frac{1}{t - \tau^+_{j-1}},\ \frac{t - \tau^+_{j-1} - 1}{t - \tau^+_{j-1}}\right]$, and $\{\mathcal{K}_t(s,\delta),\ 0 \le \delta \le 1\}$ is a sequence of Gaussian processes with

$$\mathbb{E}\left[\mathcal{K}_t(s,\delta)\right] = 0,$$
$$\mathbb{E}\left[\mathcal{K}_t(s,\delta)\ \mathcal{K}_t(s',\delta')\right] = (\delta \wedge \delta')\ \Gamma(s,s'),$$
$$\text{and} \quad \Gamma(s,s') = \mathbb{E}\left[S_{\tau^+_{j-1} + \zeta^* T}(s) S_z(s')\right], \tag{B.3}$$

provided $I_{0,1}(q,c) < \infty$ for all $c > 0$, where

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{\delta(1-\delta)} \exp\left\{-\frac{cq^2(\delta)}{\delta(1-\delta)}\right\} d\delta. \tag{B.4}$$

By Lemma A.2, given the condition in Equation B.4 and the result in Equation B.1,

$$\sup_{\delta\in[0,1]} \sup_{s\in[0,1]} |\mathcal{K}_t(s,\delta)| / q(\delta) \xrightarrow{D} \sup_{\delta\in[0,1]} \sup_{s\in[0,1]} |\mathcal{K}(s,\delta)| / q(\delta), \tag{B.5}$$

where $\mathcal{K}(s,\delta)$ is a Gaussian process defined in Equation 3.18. All that remains is to show $I_{0,1}(q,c) < \infty$ for all $c > 0$. Like in the proof for Theorem 2.2, write $q(\delta)$ as a piecewise function on the domain $\delta \in [0,1]$, and the integral from Equation B.4 becomes

$$\begin{aligned} I_{0,1}(q,c) = &\int_0^{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta \\ &+ \int_{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\delta^{2\nu-1}(1-\delta)^{2\nu-1}\right\} d\delta \\ &+ \int_{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^1 \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta. \end{aligned} \tag{B.6}$$

For any $c > 0$ the boundary terms are finite with $\kappa > 0$,

$$\int_0^{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta < \infty \tag{B.7}$$

$$\int_{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^1 \frac{1}{\delta(1-\delta)} \exp\left\{-c\kappa^2\delta^{-1}(1-\delta)^{-1}\right\} d\delta < \infty, \tag{B.8}$$

and the middle term,

$$\int_{\frac{1}{2}-\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}}^{\frac{1}{2}+\frac{1}{2}\sqrt{1-4\kappa^{1/\nu}}} \frac{1}{\delta(1-\delta)} \exp\left\{-c\delta^{2\nu-1}(1-\delta)^{2\nu-1}\right\} d\delta < \infty \tag{B.9}$$

provided $\nu < 1/2$ if $\kappa \to 0$. For $\kappa > 0$, the range of values satisfying Equation B.9 includes $\nu = 1/2$. Thus, from the specification of $q(\delta)$ in Theorem 3.1 with $\nu = 1/2$ and $\kappa > 0$, $I_{0,1}(q, c) < \infty$ for all $c > 0$. □

Proof of Theorem 3.2 follows the structure of Theorem 2.3 in Section A.4 and the outline of Theorem 2.1 from Newey and McFadden (1994).

Proof of Theorem 3.2. Scale the time domain as in Section 3.3, where $\delta \in [0, 1]$ is defined in Equation 3.16, and let

$$\delta_0 = \frac{\tau_j - \tau_{j-1}^+ - \zeta^* T + 1}{t - \tau_{j-1}^+} \tag{B.10}$$

be the true change point. For the sequence S_z, $z \in \boldsymbol{\zeta}$, write

$$\mathcal{F}_{(0,\delta]}(s) = \mathcal{F}\left[s; \tau_{j-1}^+ + \zeta^* T, \delta\left(t - \tau_{j-1}^+\right) + \tau_{j-1}^+ + \zeta^* T + 1\right] \tag{B.11}$$

and

$$\mathcal{F}_{(\delta,1)}(s) = \mathcal{F}\left[s; \delta\left(t - \tau_{j-1}^+\right) + \tau_{j-1}^+ + \zeta^* T + 2, t + \zeta^* T - 1\right] \tag{B.12}$$

as the distribution functions on the intervals $(0, \delta]$ and $(\delta, 1)$, respectively, and label $\hat{\mathcal{F}}_{(0,\delta]}(s)$ and $\hat{\mathcal{F}}_{(\delta,1)}(s)$ the corresponding empirical estimates. The true change point δ_0 divides the sequence S_z into two distinct pieces with distribution functions $\mathcal{F}_{\tau_j}(s)$ and $\mathcal{F}_{\tau_j+1}(s)$, where $\mathcal{F}_{\tau_j}(s_0) \neq \mathcal{F}_{\tau_j+1}(s_0)$ for some $s_0 \in [0, 1]$. Like in Section A.4,

define the supremum of the difference between the two distributions $\theta > 0$, where the quantity is maximized at s_0.

$$\theta = \sup_{s\in[0,1]} \left|\mathcal{F}_{\tau_j}(s) - \mathcal{F}_{\tau_j+1}(s)\right| = \left|\mathcal{F}_{\tau_j}(s_0) - \mathcal{F}_{\tau_j+1}(s_0)\right| \tag{B.13}$$

Define $Q_0(\delta)$ as in Equation B.14 for $\delta \in \boldsymbol{\Delta} = \left[\frac{1}{2} - \frac{1}{2}\sqrt{1-4\kappa^2}, \frac{1}{2} + \frac{1}{2}\sqrt{1-4\kappa^2}\right]$, $q(\delta)$ as in Equation 3.15, and $\kappa > 0$ a small constant, and construct the empirical estimate $\hat{Q}_t(\delta)$ in Equation B.15.

$$Q_0(\delta) = \frac{\delta(1-\delta)}{q(\delta)} \sup_{s\in[0,1]} \left|\mathcal{F}_{(0,\delta]}(s) - \mathcal{F}_{(\delta,1)}(s)\right| \tag{B.14}$$

$$\hat{Q}_t(\delta) = \frac{\delta(1-\delta)}{q(\delta)} \sup_{s\in[0,1]} \left|\hat{\mathcal{F}}_{(0,\delta]}(s) - \hat{\mathcal{F}}_{(\delta,1)}(s)\right| \tag{B.15}$$

Condition (1) of Newey and McFadden (1994), Theorem 2.1 is verified by writing $Q_0(\delta)$ as in Equation A.35 of Section A.4 to show the quantity is uniquely maximized at $\delta = \delta_0$. Condition (2) is verified as $\boldsymbol{\Delta}_j$ is a closed and bounded set on $\mathbb{R}$. Continuity in condition (3) is verified from Equations A.75 and A.76. Like in the Theorem 3.1 proof, in the limit as $T \to \infty$ for a fixed $\zeta^* \in \left(0, \frac{1}{2}\right)$, $\zeta^* T \to \infty$, and $n\left(\boldsymbol{\zeta}_j\right) \to \infty$. Condition (4) is verified by Lemma A.4 that shows $\hat{Q}_t(\delta)$ converges uniformly in probability to $Q_0(\delta)$ on the interval $\delta, \delta_0 \in \boldsymbol{\Delta}$ (equivalent to $z, \tau_j \in \boldsymbol{\Delta}_j$).

Therefore, $\hat{\tau}_j^f \xrightarrow{P} \tau_j$ provided $\tau_j \in \boldsymbol{\Delta}_j$, with $\boldsymbol{\Delta}_j$ defined in Equation 3.21, and τ_j divides the sequence S_z, $z \in \boldsymbol{\zeta}_j$ (i.e., $\tau_j \in \boldsymbol{\zeta}_j$). □

The result of Corollary 3.3 is obtained directly from the definition of $\hat{\tau}_j$ in Equation 3.22, and the result of Theorem 3.2.

Proof of Corollary 3.3. By the result of Theorem 3.2, $\hat{\tau}_j^f \xrightarrow{P} \tau_j$ provided $\tau_j \in \boldsymbol{\Delta}_j$,

where $\boldsymbol{\Delta}_j$ is defined in Equation 3.21. Theorem 3.2 directly implies that the backward-looking algorithm yields the same guarantee for the corresponding sequence S'_z, $z \in \boldsymbol{\zeta}'_j$: $\hat{\tau}_j^b \xrightarrow{P} \tau_j$ provided $\tau_j \in \left(\boldsymbol{\zeta}'_j \cap \boldsymbol{\Delta}'_j\right)$, where $\boldsymbol{\zeta}'_j = \left(t - \zeta^* T + 1, \tau_{j+1}^T - \zeta^* T\right]$ and $\boldsymbol{\Delta}'_j$ is defined in Equation 3.23.

With $\hat{\tau}_j$ defined in Equation 3.22, $\hat{\tau}_j^f \xrightarrow{P} \tau_j$ and $\hat{\tau}_j^b \xrightarrow{P} \tau_j$ implies $\hat{\tau}_j \xrightarrow{P} \tau_j$ provided $\tau_j \in \left[(\boldsymbol{\zeta}_j \cap \boldsymbol{\Delta}_j) \cap \left(\boldsymbol{\zeta}'_j \cap \boldsymbol{\Delta}'_j\right)\right]$, where

$$\boldsymbol{\zeta}_j \cap \boldsymbol{\zeta}'_j = (t - \zeta^* T + 1,\ t + \zeta^* T - 1) \tag{B.16}$$

and

$$\boldsymbol{\Delta}_j \cap \boldsymbol{\Delta}'_j = \left[\frac{t + \tau_{j+1}^T}{2} - \frac{\tau_{j+1}^t - t}{2}\sqrt{1 - 4\kappa^2} - \zeta^* T, \frac{t + \tau_{j-1}^+}{2} + \frac{t - \tau_{j-1}^+}{2}\sqrt{1 - 4\kappa^2} + \zeta^* T\right]. \tag{B.17}$$

□

The proof of Proposition 3.4 examines the multiple testing problem in succession and shows conservative FWER control at a given Type 1 error threshold.

Proof of Proposition 3.4. In the sequential hypothesis testing framework where null hypotheses H_i, $i = 1, \ldots, m$ are tested in order, define p_i as the associated p-value for hypothesis H_i, and k_i the number of rejected hypotheses prior to H_i in the set H_1, …, H_{i-1}. Define the rejection rule for hypothesis H_i as in Proposition 3.4 to be $p_i \leq \alpha/(m - k_i)$, and k_j^* as the number of rejected null hypotheses in the set H_1, …, H_{j-1} from simultaneous observation as in the Holm (1979) procedure.

Examine the hypothesis H_j, and suppose $k_j = 0$. Rejection of the hypothesis $p_j \leq \alpha/m$ ensures that $p_{(1)} = \min_{i=1,\ldots,j}\{p_i\} \leq \alpha/m$. The remaining p-values $p_i > \alpha/m$ for all $i = 1, \ldots, j-1$ and the sequential procedure will reject at most the number of hypotheses as the simultaneous procedure $k_{j+1} \leq k^*_{j+1}$.

Next, suppose a generic number of rejected null hypotheses, k_j. Rejection of the hypothesis $p_j \leq \alpha/(m - k_j)$ implies that $p_{(1)} \leq \alpha/m$, $p_{(2)} \leq \alpha/(m-1)$, ..., $p_{(k_j+1)} \leq \alpha/(m - k_j)$, and the remaining p-values $p_{(k_j+2)}, \ldots, p_{(j)} > \alpha/m$. Thus, the sequential procedure will result in $k_{j+1} \leq k^*_{j+1}$, and the decision rule controls the FWER to at most the Holm (1979) procedure FWER. □

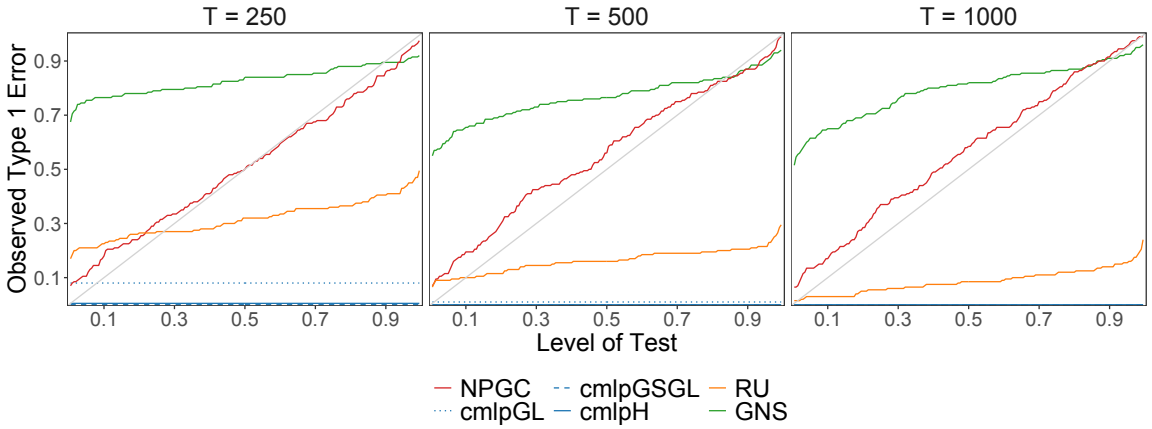
Appendix C

Additional Material for Chapter 4

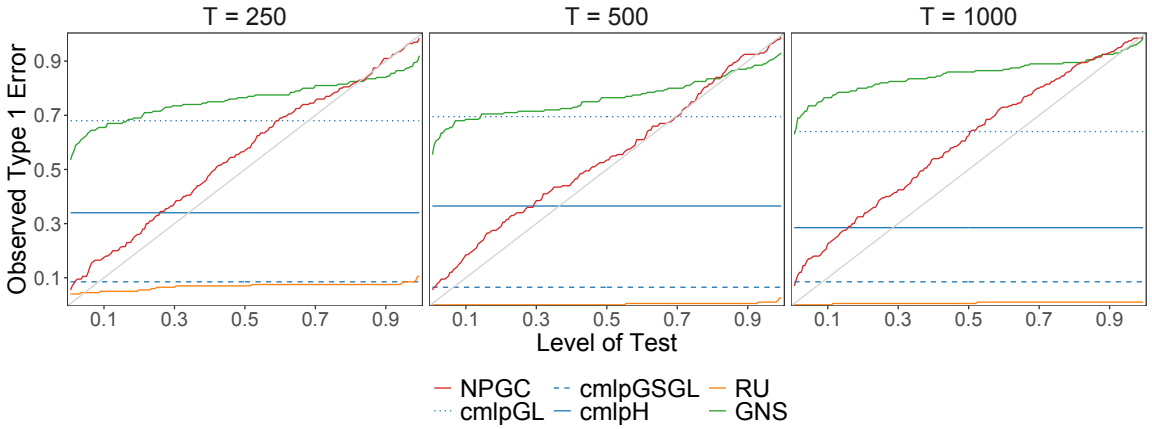
Code and supporting material: Files (.RData, .npy, and .csv) used to assess performance of change point methods, code (.R, .cpp, and .py) used to generate results and figures, and data (.mat) and code (.R, .py, and .m) used to generate output in Section 4.5 can be found at

`github.com/noahgade/NonlinearPermutedGrangerCausality`.

C.1 Additional Figures

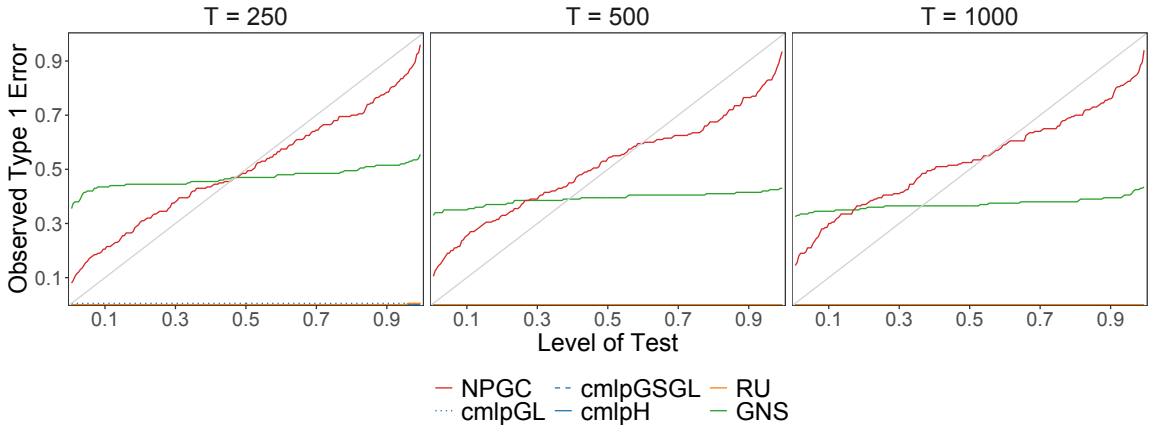


(a) Control by reference quantile α with zero Granger causal variables included in additional set $\mathbf{Z}$.

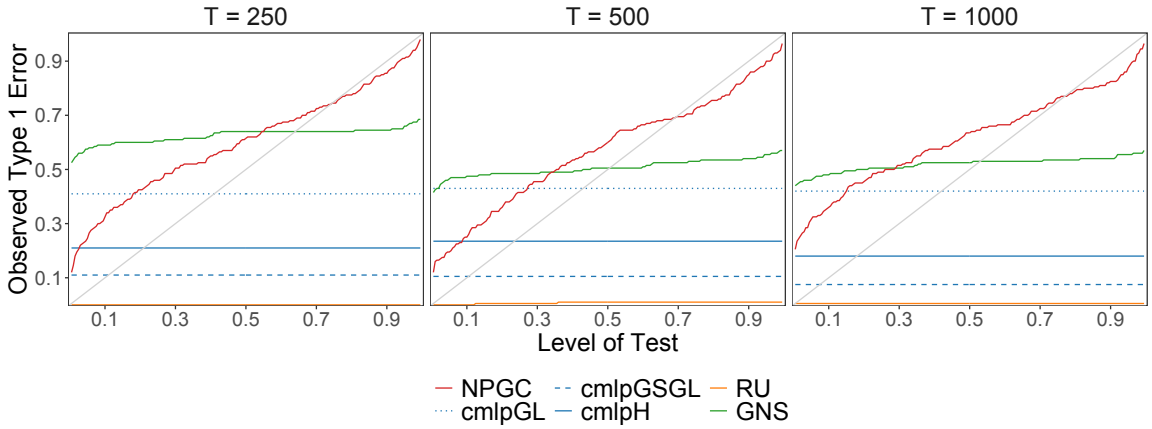


(b) Control by reference quantile α with two Granger causal variables included in additional set $\mathbf{Z}$.

Figure C.1: Type 1 error control for TAR(2) simulations.

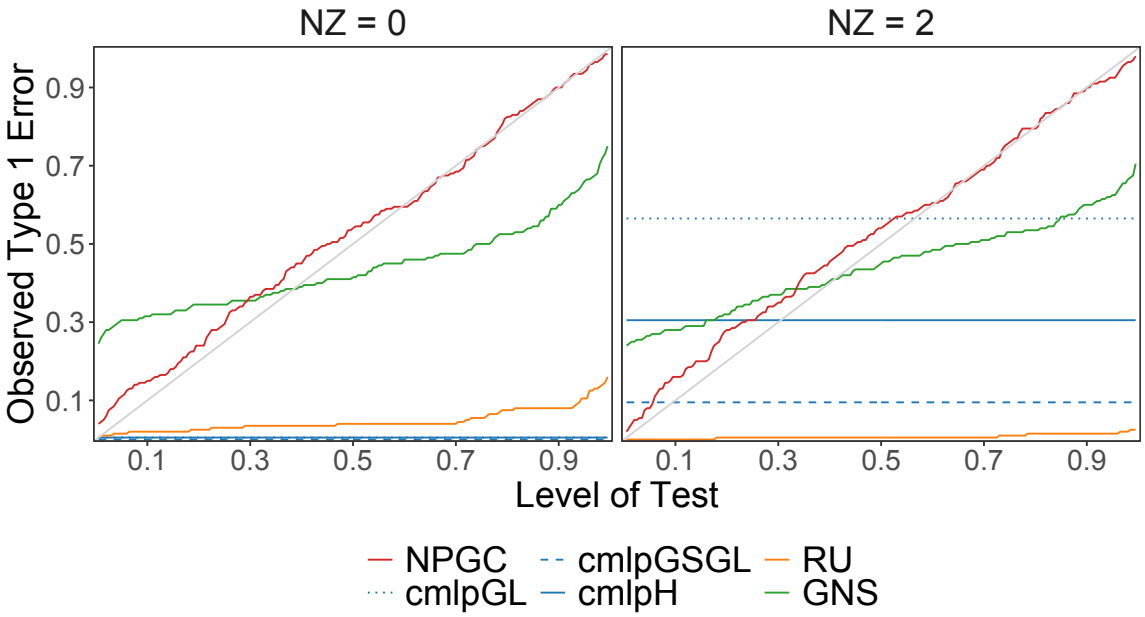


(a) Control by reference quantile α with zero Granger causal variables included in additional set $\mathbf{Z}$.



(b) Control by reference quantile α with two Granger causal variables included in additional set $\mathbf{Z}$.

Figure C.2: Type 1 error control for Lorenz-96 simulations.



Control by reference quantile α simulations without Lorenz-96 group included. *Left*: Zero Granger causal variables. *Right*: Two Granger causal variables included in the additional set $\mathbf{Z}$. All datasets are $T = 1000$.

Figure C.3: Type 1 error control for two-group TAR(2) simulations.

C.2 Additional Tables

Table C.1: TAR(2) simulation results.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	0.895	0.795	0.110	0.435	0.865	0.985
$T = 500$	0.975	0.805	0.130	0.505	0.910	1.000
$T = 1000$	0.990	0.805	0.090	0.420	0.915	1.000

(a) Proportion of correctly labelled Granger causal outcomes (GC $= 1$, ρ_1 in Table 4.2) for zero Granger causal variables included in additional set $\mathbf{Z}$.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	0.935	0.945	0.200	0.675	0.640	0.990
$T = 500$	0.955	0.900	0.135	0.610	0.775	0.995
$T = 1000$	0.980	0.895	0.155	0.580	0.785	1.000

(b) Proportion of correctly labelled Granger causal outcomes (GC $= 1$, ρ_1 in Table 4.2) for two Granger causal variables included in additional set $\mathbf{Z}$.

Table C.2: Lorenz-96 simulation results.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	0.955	0.705	0.220	0.435	0.050	1.000
$T = 500$	0.980	0.650	0.115	0.380	0.015	1.000
$T = 1000$	0.985	0.670	0.090	0.360	0.040	1.000

(a) Proportion of correctly labelled Granger causal outcomes (GC $= 1$, ρ_1 in Table 4.2) for zero Granger causal variables included in additional set $\mathbf{Z}$.

	NPGC	cMLP-GL	cMLP-GSGL	cMLP-H	R/U	GNS
$T = 250$	1.000	0.825	0.245	0.530	0.000	1.000
$T = 500$	0.995	0.805	0.205	0.505	0.000	1.000
$T = 1000$	1.000	0.760	0.165	0.465	0.000	1.000

(b) Proportion of correctly labelled Granger causal outcomes (GC $= 1$, ρ_1 in Table 4.2) for two Granger causal variables included in additional set $\mathbf{Z}$.

C.3 Additional Algorithms

Automated selection of the feature space dimension in Algorithm C.1 uses cross validation sets $1, \ldots, k^*$ of the K total to obtain N. Implementation of Algorithm C.1 requires alteration of Algorithm 4.1 to only use test sets $k = k^* + 1, \ldots, K$.

Algorithm C.1 Automated Feature Space Dimension Selection

Inputs: $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$ for all φ realizations $\omega \in \Omega_{\text{obs}}$; lag selection γ; # random FNN generations $\mathcal{R}$; # cross-validation folds K

Outputs: feature space dimension N

Remove set $k^* + 1, \ldots, K$ from $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})_\omega$ leaving only folds $1, \ldots, k^*$
$N_{\max} \leftarrow \sum_{k=1}^{k^*} T_k - 1$
for n in $10 : 10 : N_{\max}$ **do**
    Initialize $\mathbf{W}_r \in \mathbb{R}^{(1+\gamma d+q+p)\times n}$ where each element $w_{r,ij} \sim \mathcal{N}(0,1)$ for all $\mathcal{R}$
    for ω in $1 : \varphi$ **do**
        for r in $1 : \mathcal{R}$ **do**
            $\mathbf{H}_{n,\omega,r} \leftarrow \tanh\left([\mathbf{1}\ \mathbf{Y}_{\text{lag},\omega}\ \mathbf{Z}_\omega\ \mathbf{X}_\omega]\,\mathbf{W}_r\right)$
            for k in $1 : k^*$ **do**
                $\mathbf{R}_{n,\omega,r,k} \leftarrow \mathbf{H}_{n,\omega,r,k}\left(\mathbf{H}_{n,\omega,r,-k}^\top \mathbf{H}_{n,\omega,r,-k}\right)^{-1}\mathbf{H}_{n,\omega,r,-k}^\top \mathbf{Y}_{\omega,-k} - \mathbf{Y}_{\omega,k}$
            end for
        end for
    end for
end for
$N \leftarrow \arg\min_n \left\{(\varphi\mathcal{R}k^*)^{-1}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{k^*} T_k^{-1}\text{tr}\left(\mathbf{R}_{n,\omega,r,k}^\top \mathbf{R}_{n,\omega,r,k}\right)\right\}$
return N

C.4 Proofs

Lemma C.1. *Under the conditions listed in Section 4.3.1,* $\lim_{T\to\infty} \mathbb{E}\left[\hat{\vartheta}_m\right] = \vartheta_m$.

The proof of Lemma C.1 begins with the estimate for one test set with a single individual randomly generated FNN given the variation measure $\mathbf{S}_{m,\omega,r}$. We aggregate these expectations over the sets $k = 1, \ldots, K$, generated FNNs $r = 1, \ldots, \mathcal{R}$ and the $\varphi \geq 1$ observed datasets in Ω_{obs}.

Proof of Lemma C.1. We seek the conditional expectation $\mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]$, and can write the unconditional expectation of the underlying parameter based on the definition in Equation 4.10.

$$\mathbb{E}\left[\hat{\vartheta}_m\right] = \frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\mathbb{E}\left[\mathbb{E}\left(\mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\mathbf{\Sigma}_{m,\omega}\right)\right] \tag{C.1}$$

From Condition 4.12, we can write the distribution of the residual term $\mathbf{R}_{m,\omega,r,k}$ from that of $\mathbf{U}_{m,\omega,r}$.

$$\mathbf{U}_{m,\omega,r} \sim \mathcal{MN}_{T\times d}\left(\mathbf{0}, \mathbf{I}_T, \mathbf{S}_{m,\omega,r}\right) \tag{C.2}$$

$$\mathbf{R}_{m,\omega,r,k} = \mathbf{H}_{m,\omega,r,k}\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{Y}_{\omega,-k} - \mathbf{Y}_{\omega,k} \tag{C.3}$$

$$\mathbf{R}_{m,\omega,r,k} \sim \mathcal{MN}_{T_k\times d}\left(\mathbf{0}, \mathbf{H}_{m,\omega,r,k}\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top} + \mathbf{I}_{T_k}, \mathbf{S}_{m,\omega,r}\right) \tag{C.4}$$

Define $\mathbf{\Phi}_{m,\omega,r,k} = \mathbf{H}_{m,\omega,r,k}\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top}$, and further define $\mathbf{V}_{m,\omega,r,k} \sim$

$\mathcal{MN}_{T_k\times d}(\mathbf{0}, \mathbf{I}_{T_k}, \mathbf{I}_d)$ such that we can write $\mathbf{R}_{m,\omega,r,k} = \left(\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right)^{1/2} \mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}^{1/2}$.

$$\begin{aligned}
&\mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \\
&\qquad = \mathbb{E}\left[\text{tr}\left(\mathbf{S}_{m,\omega,r}^{1/2}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}^{1/2}\right)|\mathbf{S}_{m,\omega,r}\right] && \text{(C.5)} \\
&\qquad = \mathbb{E}\left[\text{tr}\left(\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] && \text{(C.6)} \\
&\qquad = \text{tr}\left(\mathbb{E}\left[\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right]\right) && \text{(C.7)} \\
&\qquad = \text{tr}\left(\mathbf{S}_{m,\omega,r}\mathbb{E}\left[\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right]\right) && \text{(C.8)}
\end{aligned}$$

We can write

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}\left(\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{\Phi}_{m,\omega,r,k}, \mathbf{S}_{m,\omega,r}\right)\Big|\mathbf{S}_{m,\omega,r}\right], && \text{(C.9)}
\end{aligned}$$

and from Gupta and Nagar (2018),

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{\Phi}_{m,\omega,r,k}, \mathbf{S}_{m,\omega,r}\right)] \\
&\qquad = \text{tr}\left(\mathbf{I}_{T_k}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\right)\mathbf{I}_d && \text{(C.10)} \\
&\qquad = \text{tr}\left(\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right)\mathbf{I}_d. && \text{(C.11)}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right] &= \mathbb{E}\left[\text{tr}\left(\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right)\mathbf{I}_d|\mathbf{S}_{m,\omega,r}\right] && \text{(C.12)} \\
&= \mathbb{E}\left[\text{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]\mathbf{I}_d + T_k\mathbf{I}_d. && \text{(C.13)}
\end{aligned}$$

We modify Equation C.13 by exchanging $\mathbf{\Phi}_{m,\omega,r,k}$ for its definition as written above.

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] = \mathbb{E}\left[\operatorname{tr}\left(\mathbf{H}_{m,\omega,r,k}\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top}\right)|\mathbf{S}_{m,\omega,r}\right] \tag{C.14}$$

$$= \mathbb{E}\left[\operatorname{tr}\left(\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top}\mathbf{H}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \tag{C.15}$$

We examine the matrices $\mathbf{H}_{m,\omega,r,-k}$ and $\mathbf{H}_{m,\omega,r,k}$ under Conditions 4.7, 4.8, 4.9, and 4.10. Recall T_k is the number of rows in the test set, and T_{-k} the number in the training set. We obtain a fixed and finite N from Condition 4.10, and $\operatorname{rank}(\mathbf{H}_{m,\omega,r,-k}) = N$ under Equation 4.15 from Condition 4.8. We label the singular values of the matrices $\sigma_i(\mathbf{H}_{m,\omega,r,-k}) > 0$ and $\sigma_i(\mathbf{H}_{m,\omega,r,k}) \geq 0$ for $i = 1, \ldots, N$. For any matrix $\mathbf{P}$, $\|\mathbf{P}\|_F = \sqrt{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|p_{ij}|^2} = \sqrt{\sum_{i=1}^{\operatorname{rank}(\mathbf{P})}\sigma_i(\mathbf{P})^2}$. From Condition 4.7, each entry in $\mathbf{H}_{m,\omega,r,-k}$ and $\mathbf{H}_{m,\omega,r,k}$ lies in a bounded interval, $h_{m,\omega,r,-k,ij}, h_{m,\omega,r,k,ij} \in [a,b]$ such that $|h_{m,\omega,r,-k,ij}|, |h_{m,\omega,r,k,ij}| \leq G$, and we use this to establish the upper bounds for the maximal singular values shown in Equations C.16 and C.17. To write the minimum values, we define the average squared matrix entry as in Condition 4.9, and note the combination of Conditions 4.7, 4.8, and 4.9 yields $0 < \nu^2 \leq 1$, $1 \leq \xi^2 < \infty$, and $1 \leq \varrho^4 < \infty$.

$$\sqrt{T_{-k}\bar{h}_{m,\omega,r,-k}^2} \leq \sigma_1\left(\mathbf{H}_{m,\omega,r,-k}\right) \leq \|\mathbf{H}_{m,\omega,r,-k}\|_F \leq \sqrt{T_{-k}NG^2} \tag{C.16}$$

$$\sqrt{T_k\bar{h}_{m,\omega,r,k}^2} \leq \sigma_1\left(\mathbf{H}_{m,\omega,r,k}\right) \leq \|\mathbf{H}_{m,\omega,r,k}\|_F \leq \sqrt{T_kNG^2} \tag{C.17}$$

Under Condition 4.8, $\mathbf{H}_{m,\omega,r,-k}$ has a finite condition number $\kappa\left(\mathbf{H}_{m,\omega,r,-k}\right) \leq \kappa_{\max}$, and we can use the bounds in Equations C.16 and C.17 to obtain bounds for the minimum singular values. The rank of $\mathbf{H}_{m,\omega,r,k}$ is not necessarily N, as T_k can be less

than N, but it is at least $\min\{T_k, N\}$.

$$\frac{\sqrt{T_{-k}\bar{h}^2_{m,\omega,r,-k}}}{\kappa_{\max}} \leq \sigma_N\left(\mathbf{H}_{m,\omega,r,-k}\right) \leq \sqrt{T_{-k}\bar{h}^2_{m,\omega,r,-k}} \tag{C.18}$$

$$0 \leq \sigma_N\left(\mathbf{H}_{m,\omega,r,k}\right) \leq \sqrt{T_k\bar{h}^2_{m,\omega,r,k}} \tag{C.19}$$

We return to the $\text{tr}\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right)$ piece of Equation C.15. Applying the above results, we can write $\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1} \succ 0$, $\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k} \succeq 0$, and the following bounds on the extreme eigenvalues.

$$\frac{1}{T_{-k}\bar{h}^2_{m,\omega,r,-k}} \leq \lambda_1\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right) \leq \frac{\kappa^2_{\max}}{T_{-k}\bar{h}^2_{m,\omega,r,-k}} \tag{C.20}$$

$$\frac{1}{T_{-k}NG^2} \leq \lambda_N\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right) \leq \frac{1}{T_{-k}\bar{h}^2_{m,\omega,r,-k}} \tag{C.21}$$

$$T_k\bar{h}^2_{m,\omega,r,k} \leq \lambda_1\left(\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right) \leq T_kNG^2 \tag{C.22}$$

$$0 \leq \lambda_N\left(\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right) \leq T_k\bar{h}^2_{m,\omega,r,k} \tag{C.23}$$

We can bound the trace of the product with the eigenvalues using Von Neumann's trace inequality.

$$\begin{aligned}
&\text{tr}\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right) \\
&\qquad \leq \sum_{i=1}^{N}\lambda_i\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right)\lambda_i\left(\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right) &\text{(C.24)}\\
&\qquad \leq N\lambda_1\left(\left[\mathbf{H}^\top_{m,\omega,r,-k}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right)\lambda_1\left(\mathbf{H}^\top_{m,\omega,r,k}\mathbf{H}_{m,\omega,r,k}\right) &\text{(C.25)}\\
&\qquad \leq \frac{T_kN^2G^2\kappa^2_{\max}}{T_{-k}\bar{h}^2_{m,\omega,r,-k}} &\text{(C.26)}
\end{aligned}$$

The upper bound in Equation C.26 is a positive, finite constant. Returning to Equa-

tion C.15,

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \leq \mathbb{E}\left[\frac{T_k N^2 G^2 \kappa_{\max}^2}{T_{-k}\bar{h}_{m,\omega,r,-k}^2}\middle|\mathbf{S}_{m,\omega,r}\right] \tag{C.27}$$

$$\leq \frac{T_k}{T_{-k}}\left(NG\kappa_{\max}\right)^2 \mathbb{E}\left[\left(\bar{h}_{m,\omega,r,-k}^2\right)^{-1}|\mathbf{S}_{m,\omega,r}\right] \tag{C.28}$$

$$\leq \frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2. \tag{C.29}$$

Using this result in Equation C.13,

$$\mathbb{E}\left[\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right] \leq \frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2\mathbf{I}_d + T_k\mathbf{I}_d, \tag{C.30}$$

and in Equation C.8,

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \leq \operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\left[\frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2\mathbf{I}_d + T_k\mathbf{I}_d\right]\right) \tag{C.31}$$

$$\leq \left[\frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2 + T_k\right]\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right). \tag{C.32}$$

We now require a lower bound for the expectation. We examine the trace inequality

$$\operatorname{tr}\left(\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top}\mathbf{H}_{m,\omega,r,k}\right)$$

$$\geq \sum_{i=1}^{N}\lambda_i\left(\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right)\lambda_{N-i+1}\left(\mathbf{H}_{m,\omega,r,k}^{\top}\mathbf{H}_{m,\omega,r,k}\right) \tag{C.33}$$

$$\geq \lambda_N\left(\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\right)\lambda_1\left(\mathbf{H}_{m,\omega,r,k}^{\top}\mathbf{H}_{m,\omega,r,k}\right) \tag{C.34}$$

$$\geq \frac{T_k\bar{h}_{m,\omega,r,k}^2}{T_{-k}NG^2}. \tag{C.35}$$

Again returning to Equation C.15,

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \geq \mathbb{E}\left[\frac{T_k \bar{h}^2_{m,\omega,r,k}}{T_{-k}NG^2}\Big|\mathbf{S}_{m,\omega,r}\right] \tag{C.36}$$

$$\geq \frac{T_k}{T_{-k}NG^2}\mathbb{E}\left[\bar{h}^2_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right] \tag{C.37}$$

$$\geq \frac{T_k\nu^2}{T_{-k}N}, \tag{C.38}$$

and using the result in Equation C.13,

$$\mathbb{E}\left[\mathbf{V}^{\top}_{m,\omega,r,k}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\right] \geq \frac{T_k\nu^2}{T_{-k}N}\mathbf{I}_d + T_k\mathbf{I}_d. \tag{C.39}$$

This yields the lower bound,

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{R}^{\top}_{m,\omega,r,k}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \geq \operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\left[\frac{T_k\nu^2}{T_{-k}N}\mathbf{I}_d + T_k\mathbf{I}_d\right]\right) \tag{C.40}$$

$$\geq \left[\frac{T_k\nu^2}{T_{-k}N} + T_k\right]\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right). \tag{C.41}$$

We return to the expectation equation given in Equation C.1 with the two bounds

from Equations C.32 and C.41.

$$\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\mathbb{E}\left[\mathbb{E}\left(\left[\frac{T_k\nu^2}{T_{-k}N}+T_k\right]\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right)\right]\leq\mathbb{E}\left[\hat{\vartheta}_m\right]$$
$$\leq\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\mathbb{E}\left[\mathbb{E}\left(\left[\frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2+T_k\right]\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \tag{C.42}$$

$$\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\mathbb{E}\left[\left(\frac{T_k\nu^2}{T_{-k}N}+T_k\right)\operatorname{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right)\right]\leq\mathbb{E}\left[\hat{\vartheta}_m\right]$$
$$\leq\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\mathbb{E}\left[\left(\frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2+T_k\right)\operatorname{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right)\right] \tag{C.43}$$

$$\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\left[\frac{\nu^2}{T_{-k}N}+1\right]\vartheta_m\leq\mathbb{E}\left[\hat{\vartheta}_m\right]$$
$$\leq\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\left[\frac{1}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2+1\right]\vartheta_m \tag{C.44}$$

In the limit as $T\rightarrow\infty$, with a fixed test set size $T_k<\infty$, $T_{-k}=T-T_k\rightarrow\infty$.

$$\lim_{T\rightarrow\infty}\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\left[\frac{\nu^2}{T_{-k}N}+1\right]\vartheta_m$$
$$=\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\lim_{T_{-k}\rightarrow\infty}\left[\frac{\nu^2}{T_{-k}N}+1\right]\vartheta_m=\vartheta_m \tag{C.45}$$

$$\lim_{T\rightarrow\infty}\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\left[\frac{1}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2+1\right]\vartheta_m$$
$$=\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\lim_{T_{-k}\rightarrow\infty}\left[\frac{1}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2+1\right]\vartheta_m=\vartheta_m \tag{C.46}$$

Thus, $\lim_{T\rightarrow\infty}\mathbb{E}\left[\hat{\vartheta}_m\right]=\vartheta_m$. □

Lemma C.2. *Under the conditions listed in Section 4.3.1,* $\lim_{T\rightarrow\infty}Var\left(\hat{\vartheta}_m\right)=\varphi^{-1}\tau_\omega^2+\left(\varphi\mathcal{R}\right)^{-1}\tau_r^2$.

For use in the proof of Lemma C.2, we first state and prove Remark C.3, and then proceed like in the proof of Lemma C.1.

Remark C.3. For any square, positive semidefinite matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}^2(\mathbf{P}) \leq (n^2 - n + 1)\,\mathrm{tr}(\mathbf{P}^2)$.

Proof of Remark C.3. Let $\lambda_1(\mathbf{P}), \lambda_2(\mathbf{P}), \ldots, \lambda_n(\mathbf{P}) \geq 0$ be the eigenvalues of $\mathbf{P} \in \mathbb{R}^{n\times n}$. For any $n \geq 1$ we can write,

$$\mathrm{tr}^2(\mathbf{P}) = \left(\sum_{i=1}^{n} \lambda_i(\mathbf{P})\right)^2 \tag{C.47}$$

$$= \sum_{i=1}^{n} \lambda_i(\mathbf{P})^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} \lambda_i(\mathbf{P})\,\lambda_j(\mathbf{P}) \tag{C.48}$$

$$\leq \sum_{i=1}^{n} \lambda_i(\mathbf{P})^2 + n(n-1)\sum_{i=1}^{n} \lambda_i(\mathbf{P})^2 \tag{C.49}$$

$$\leq \left(n^2 - n + 1\right)\sum_{i=1}^{n} \lambda_i(\mathbf{P})^2 \tag{C.50}$$

with the last line equal to $(n^2-n+1)\,\mathrm{tr}(\mathbf{P}^2)$. Thus, $\mathrm{tr}^2(\mathbf{P}) \leq (n^2-n+1)\,\mathrm{tr}(\mathbf{P}^2)$. □

Proof of Lemma C.2. We isolate the quantity $\mathrm{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)$ like in the proof of Lemma C.1 and examine the conditional variance $\mathrm{Var}\left(\mathrm{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)$. The unconditional variance follows from the law of total variance and from the defi-

nition in Equation 4.10.

$$\begin{aligned}\text{Var}\left(\hat{\vartheta}_m\right) = \mathbb{E}\left[\mathbb{E}\left(\text{Var}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \\ + \mathbb{E}\left[\text{Var}\left(\mathbb{E}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \\ + \text{Var}\left(\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right)\end{aligned} \tag{C.51}$$

We first examine $\text{Var}\left(\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)$.

$$\begin{aligned}\text{Var}\left(\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right) = \mathbb{E}\left[\text{tr}^2\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \\ - \mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]^2\end{aligned} \tag{C.52}$$

To establish an bounds on the variance, we seek both lower and upper bounds for the terms in Equation C.52. The bounds for the second term can be directly obtained from the derivation in the Proof of Lemma C.1.

$$\begin{aligned}\left[\frac{T_k\nu^2}{T_{-k}N} + T_k\right]^2\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right) &\leq \mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]^2 \\ &\leq \left[\frac{T_k}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2 + T_k\right]^2\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right) \qquad \text{(C.53)} \\ \left[\frac{T_k^2\nu^4}{T_{-k}^2N^2} + 2\frac{T_k^2\nu^2}{T_{-k}N} + T_k^2\right]\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right) &\leq \mathbb{E}\left[\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]^2 \\ &\leq \left[\frac{T_k^2}{T_{-k}^2}\left(N\xi\kappa_{\max}\right)^4 + 2\frac{T_k^2}{T_{-k}}\left(N\xi\kappa_{\max}\right)^2 + T_k^2\right]\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right)\end{aligned} \tag{C.54}$$

For the first term in Equation C.52, we follow the strategy of the proof of Lemma

C.1. The matrix in the trace of Equation C.55 is square and positive semidefinite, and we apply the result of Remark C.3.

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}^2\left(\mathbf{R}_{m,\omega,r,k}^\top \mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \\
&\quad= \mathbb{E}\left[\mathrm{tr}^2\left(\mathbf{S}_{m,\omega,r}^{1/2}\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}^{1/2}\right)|\mathbf{S}_{m,\omega,r}\right] && \text{(C.55)} \\
&\quad\leq (d^2-d+1)\,\mathbb{E}\Bigg[\mathrm{tr}\Bigg(\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r} \\
&\qquad\qquad \mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\Bigg)|\mathbf{S}_{m,\omega,r}\Bigg] && \text{(C.56)} \\
&\quad\leq (d^2-d+1)\,\mathrm{tr}\Bigg(\mathbf{S}_{m,\omega,r}\mathbb{E}\Bigg[\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r} \\
&\qquad\qquad \mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\Bigg]\Bigg) && \text{(C.57)}
\end{aligned}$$

We can write the expectation as

$$\begin{aligned}
&\mathbb{E}\Bigg[\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\Bigg] \\
&\quad= \mathbb{E}\Bigg[\mathbb{E}\Bigg(\mathbf{V}_{m,\omega,r,k}^\top\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^\top \\
&\qquad\qquad \left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{\Phi}_{m,\omega,r,k},\mathbf{S}_{m,\omega,r}\Bigg)|\mathbf{S}_{m,\omega,r}\Bigg] && \text{(C.58)}
\end{aligned}$$

and obtain the inner piece from Theorem 2.3.8 (v) of Gupta and Nagar (2018).

$$\begin{aligned}
&\mathbb{E}\Bigg(\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{\Phi}_{m,\omega,r,k},\mathbf{S}_{m,\omega,r}\Bigg)\\
&\quad=\operatorname{tr}\left(\mathbf{I}_{T_k}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{I}_{T_k}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\right)\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\mathbf{I}_d\right)\mathbf{I}_d\\
&\qquad+\operatorname{tr}\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{I}_{T_k}\right)\operatorname{tr}\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{I}_{T_k}\right)\mathbf{I}_d\mathbf{S}_{m,\omega,r}\mathbf{I}_d\\
&\qquad+\operatorname{tr}\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{I}_{T_k}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{I}_{T_k}\right)\mathbf{I}_d\mathbf{S}_{m,\omega,r}\mathbf{I}_d \qquad \text{(C.59)}\\
&\quad=\operatorname{tr}\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]^2\right)\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d\\
&\qquad+\operatorname{tr}^2\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+\operatorname{tr}\left(\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]^2\right)\mathbf{S}_{m,\omega,r} \qquad \text{(C.60)}\\
&\quad=\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}^2\right)\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}^2\right)\mathbf{S}_{m,\omega,r}+\operatorname{tr}^2\left(\mathbf{\Phi}_{m,\omega,r,k}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+2\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+2\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)\mathbf{S}_{m,\omega,r}+2T_k\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+T_k\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+T_k\mathbf{S}_{m,\omega,r}+T_k^2\mathbf{S}_{m,\omega,r} \qquad \text{(C.61)}
\end{aligned}$$

Using Von Neumann's trace inequality and the upper bound on the largest eigenvalue from Equation C.26,

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{\Phi}_{m,\omega,r,k}^2\right)&\leq\sum_{i=1}^{N}\lambda_i\left(\mathbf{\Phi}_{m,\omega,r,k}\right)^2 &\text{(C.62)}\\
&\leq N\lambda_1\left(\mathbf{\Phi}_{m,\omega,r,k}\right)^2 &\text{(C.63)}\\
&\leq\frac{T_k^2N^3G^4\kappa_{\max}^4}{T_{-k}^2\left(\bar{h}_{m,\omega,r,-k}^2\right)^2}, &\text{(C.64)}
\end{aligned}$$

and we input the result to Equation C.61.

$$\begin{aligned}
&\mathbb{E}\Bigg(\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{\Phi}_{m,\omega,r,k},\mathbf{S}_{m,\omega,r}\Bigg)\\
&\quad\leq\left(\frac{T_k^2N^3G^4\kappa_{\max}^4}{T_{-k}^2\left(\bar{h}_{m,\omega,r,-k}^2\right)^2}\right)\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+\left(\frac{T_k^2N^3G^4\kappa_{\max}^4}{T_{-k}^2\left(\bar{h}_{m,\omega,r,-k}^2\right)^2}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+\left(\frac{T_k^2N^4G^4\kappa_{\max}^4}{T_{-k}^2\left(\bar{h}_{m,\omega,r,-k}^2\right)^2}\right)\mathbf{S}_{m,\omega,r}+2\left(\frac{T_kN^2G^2\kappa_{\max}^2}{T_{-k}\bar{h}_{m,\omega,r,-k}^2}\right)\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d\\
&\qquad+2\left(\frac{T_kN^2G^2\kappa_{\max}^2}{T_{-k}\bar{h}_{m,\omega,r,-k}^2}\right)\mathbf{S}_{m,\omega,r}+2\left(\frac{T_k^2N^2G^2\kappa_{\max}^2}{T_{-k}\bar{h}_{m,\omega,r,-k}^2}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+T_k\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+T_k\mathbf{S}_{m,\omega,r}+T_k^2\mathbf{S}_{m,\omega,r}
\end{aligned}\tag{C.65}$$

With the expected values from Condition 4.9, we can simplify the form to that shown in Equation C.66.

$$\begin{aligned}
&\mathbb{E}\Bigg[\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}\mathbf{S}_{m,\omega,r}\mathbf{V}_{m,\omega,r,k}^{\top}\left[\mathbf{\Phi}_{m,\omega,r,k}+\mathbf{I}_{T_k}\right]\mathbf{V}_{m,\omega,r,k}|\mathbf{S}_{m,\omega,r}\Bigg]\\
&\quad\leq\left(\frac{T_k^2N^3\varrho^4\kappa_{\max}^4}{T_{-k}^2}\right)\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+\left(\frac{T_k^2N^3\varrho^4\kappa_{\max}^4}{T_{-k}^2}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+\left(\frac{T_k^2N^4\varrho^4\kappa_{\max}^4}{T_{-k}^2}\right)\mathbf{S}_{m,\omega,r}+2\left(\frac{T_kN^2\xi^2\kappa_{\max}^2}{T_{-k}}\right)\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d\\
&\qquad+2\left(\frac{T_kN^2\xi^2\kappa_{\max}^2}{T_{-k}}\right)\mathbf{S}_{m,\omega,r}+2\left(\frac{T_k^2N^2\xi^2\kappa_{\max}^2}{T_{-k}}\right)\mathbf{S}_{m,\omega,r}\\
&\qquad+T_k\mathrm{tr}\left(\mathbf{S}_{m,\omega,r}\right)\mathbf{I}_d+T_k\mathbf{S}_{m,\omega,r}+T_k^2\mathbf{S}_{m,\omega,r}
\end{aligned}\tag{C.66}$$

We return to Equation C.57, and note that for any square, positive semidefinite

matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$, $\text{tr}\left(\mathbf{P}^2\right) \leq \text{tr}^2\left(\mathbf{P}\right)$.

$$\begin{aligned}
&\mathbb{E}\left[\text{tr}^2\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right] \\
&\quad\leq (d^2-d+1)\,\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right)\left[3\left(\frac{T_k^2N^4\varrho^4\kappa_{\max}^4}{T_{-k}^2}\right)+4\left(\frac{T_kN^2\xi^2\kappa_{\max}^2}{T_{-k}}\right)\right. \\
&\qquad\left.+\,2\left(\frac{T_k^2N^2\xi^2\kappa_{\max}^2}{T_{-k}}\right)+2T_k+T_k^2\right] \qquad \text{(C.67)}
\end{aligned}$$

We can write the following inequality for $\text{Var}\left(T_k^{-1}\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)$.

$$\begin{aligned}
&\text{Var}\left(T_k^{-1}\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right) \\
&\quad\leq \text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right)\left(\frac{1}{T_{-k}^2}\left[\frac{3(d^2-d+1)N^6\varrho^4\kappa_{\max}^4-\nu^4}{N^2}\right]\right. \\
&\qquad+\frac{1}{T_kT_{-k}}\left[4(d^2-d+1)N^2\xi^2\kappa_{\max}^2\right] \\
&\qquad+\frac{1}{T_{-k}}\left[\frac{2(d^2-d+1)N^3\xi^2\kappa_{\max}^2-2\nu^2}{N}\right] \\
&\qquad\left.+(d^2-d)+\frac{2}{T_K}(d^2-d+1)\right) \qquad \text{(C.68)}
\end{aligned}$$

Taking the limit as $T \to \infty$, we obtain an upper bound for the variance of the estimate given all individual covariance matrices for one test set with each featurization and potential realization of the data. As the number of time points gets large with a fixed $T_k < \infty$, the size of each training set T_{-k} tends to infinity.

$$\lim_{T\to\infty}\text{Var}\left(T_k^{-1}\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right) \leq \left[(d^2-d)+\frac{2}{T_k}(d^2-d+1)\right]\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right) \qquad \text{(C.69)}$$

Returning to the total variance in Equation C.51, with the limit in Equation C.69, we can apply the Dominated Convergence Theorem to establish the limiting value

for $\text{Var}\left(\hat{\vartheta}_m\right)$. We also use the results in Equations C.32 and C.41 to establish the limiting expectation $\lim_{T\to\infty}\mathbb{E}\left[T_k^{-1}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]=\text{tr}\left(\mathbf{S}_{m,\omega,r}\right)$.

The first term in Equation C.51 has an inner term of the variance of the variation parameter given the featurization specific matrix $\mathbf{S}_{m,\omega,r}$ and the realization specific matrix $\mathbf{\Sigma}_{m,\omega}$. The only remaining source of variation arises from labelling the training and test sets $k=1,\ldots,K$. With random assignment, and fully explained temporal dependence as in Condition 4.12, we treat these splits as uncorrelated draws, and the variance can move inside the summation.

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left(\text{Var}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\mathbf{\Sigma}_{m,\omega}\right)\right]\\ =\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\text{Var}\left[\frac{1}{T_k}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\mathbf{\Sigma}_{m,\omega}\right)\right]\end{aligned} \tag{C.70}$$

The inner variance converges pointwise and is dominated by the integrable quantity

$$\text{tr}^2\left(\mathbf{S}_{m,\omega,r}\right)\left[3(d^2-d+1)N^6\varrho^4\kappa_{\max}^4+6(d^2-d+1)N^3\xi^2\kappa_{\max}^2+3d^2+1\right], \tag{C.71}$$

as a simplification of the form shown in Equation C.68. The expectation of the inner term converges pointwise, where

$$\begin{aligned}\lim_{T\to\infty}\mathbb{E}\left[\text{Var}\left(T_k^{-1}\text{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\mathbf{\Sigma}_{m,\omega}\right]\\ \leq\left[(d^2-d)+\frac{2}{T_k}(d^2-d+1)\right]\text{tr}^2\left(\mathbf{\Sigma}_{m,\omega}\right),\end{aligned} \tag{C.72}$$

and is dominated by the similar integrable quantity

$$\operatorname{tr}^2\left(\boldsymbol{\Sigma}_{m,\omega}\right)\left[3(d^2-d+1)N^6\varrho^4\kappa_{\max}^4+6(d^2-d+1)N^3\xi^2\kappa_{\max}^2+3d^2+1\right]. \tag{C.73}$$

We again assume a fixed $T_k < \infty$, and $T \to \infty$ implies $T_{-k} \to \infty$. The limit also implies $K \to \infty$ for a fixed T_k. When applying the Dominated Convergence Theorem twice,

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\operatorname{Var}\left[\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \\
&=\mathbb{E}\left[\lim_{T\to\infty}\mathbb{E}\left(\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\operatorname{Var}\left[\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.74)} \\
&=\mathbb{E}\left[\mathbb{E}\left(\lim_{T_{-k},K\to\infty}\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\operatorname{Var}\left[\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.75)} \\
&\leq\mathbb{E}\left[\mathbb{E}\left(\lim_{K\to\infty}\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\left[(d^2-d)+\frac{2}{T_k}(d^2-d+1)\right]\operatorname{tr}^2\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right)\right]. \qquad \text{(C.76)}
\end{aligned}$$

The inner piece tends to zero as K gets large, and

$$\lim_{T\to\infty}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi^2\mathcal{R}^2K^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\operatorname{Var}\left[\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right]=0. \tag{C.77}$$

For the second term in Equation C.51, the inner expectation converges pointwise to $\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)$ and is dominated by the integrable term $\left[\left(N\xi\kappa_{\max}\right)^2+1\right]\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)$. The middle expectation converges pointwise to $\operatorname{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right)$ and is dominated by the similar

term $\left[(N\xi\kappa_{\max})^2+1\right]\operatorname{tr}(\boldsymbol{\Sigma}_{m,\omega})$. We can take the limit inside the function $h(x)=x^2$ as it is continuous on the full domain $x\in\mathbb{R}$ (*i.e.*, inside the variance term).

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{E}\left[\operatorname{Var}\left(\mathbb{E}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right]\\
&=\mathbb{E}\left[\lim_{T\to\infty}\operatorname{Var}\left(\mathbb{E}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.78)}\\
&=\mathbb{E}\left[\operatorname{Var}\left(\lim_{T\to\infty}\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\mathbb{E}\left[\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.79)}\\
&=\mathbb{E}\left[\operatorname{Var}\left(\frac{1}{\varphi\mathcal{R}}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.80)}
\end{aligned}$$

Given the data realized covariance matrix $\boldsymbol{\Sigma}_{m,\omega}$, each featurization $\mathbf{S}_{m,\omega,r}$ is an uncorrelated draw, and we can distribute the variance term like above. From Condition 4.11, we extract the limiting value.

$$\begin{aligned}
&\lim_{T\to\infty}\mathbb{E}\left[\operatorname{Var}\left(\mathbb{E}\left[\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left(\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right)|\mathbf{S}_{m,\omega,r}\right]|\boldsymbol{\Sigma}_{m,\omega}\right)\right]\\
&=\mathbb{E}\left[\frac{1}{\varphi^2\mathcal{R}^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\operatorname{Var}\left(\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right)\right] \qquad \text{(C.81)}\\
&=\mathbb{E}\left[\frac{1}{\varphi^2\mathcal{R}^2}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\tau_r^2\right]=\frac{\tau_r^2}{\varphi\mathcal{R}} \qquad \text{(C.82)}
\end{aligned}$$

For the third term in Equation C.51, we note the same conditions as before, and

proceed like above.

$$\lim_{T\to\infty} \operatorname{Var}\left(\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right)$$
$$= \operatorname{Var}\left(\lim_{T\to\infty}\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right) \tag{C.83}$$
$$= \operatorname{Var}\left(\mathbb{E}\left[\lim_{T\to\infty}\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\mathbb{E}\left(\frac{1}{T_k}\operatorname{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right) \tag{C.84}$$
$$= \operatorname{Var}\left(\mathbb{E}\left[\frac{1}{\varphi\mathcal{R}}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\operatorname{tr}\left(\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right) \tag{C.85}$$
$$= \operatorname{Var}\left(\frac{1}{\varphi}\sum_{\omega=1}^{\varphi}\operatorname{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right)\right) \tag{C.86}$$

Each realization of the data is independent, and we can take the variance inside the summation.

$$\lim_{T\to\infty} \operatorname{Var}\left(\mathbb{E}\left[\mathbb{E}\left(\frac{1}{\varphi\mathcal{R}K}\sum_{\omega=1}^{\varphi}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K}\frac{1}{T_k}\operatorname{tr}\left[\mathbf{R}_{m,\omega,r,k}^{\top}\mathbf{R}_{m,\omega,r,k}\right]|\mathbf{S}_{m,\omega,r}\right)|\boldsymbol{\Sigma}_{m,\omega}\right]\right)$$
$$= \frac{1}{\varphi^2}\sum_{\omega=1}^{\varphi}\operatorname{Var}\left(\operatorname{tr}\left[\boldsymbol{\Sigma}_{m,\omega}\right]\right) = \frac{\tau_{\omega}^2}{\varphi} \tag{C.87}$$

Thus, we can write the limiting variance as,

$$\lim_{T\to\infty}\operatorname{Var}\left(\hat{\vartheta}_m\right) = \frac{\tau_{\omega}^2}{\varphi} + \frac{\tau_r^2}{\varphi\mathcal{R}}. \tag{C.88}$$

□

The proof of Theorem 4.13 follows directly from the results of Lemmas C.1 and

C.2.

Proof of Theorem 4.13. From Lemma C.1, we establish that $\lim_{T\to\infty}\mathbb{E}\left[\hat{\vartheta}_m\right] = \vartheta_m$. Turning to the result of Lemma C.2,

$$\lim_{T\to\infty}\text{Var}\left(\hat{\vartheta}_m\right) = \frac{\tau_\omega^2}{\varphi} + \frac{\tau_r^2}{\varphi\mathcal{R}} \tag{C.89}$$

$$\lim_{\mathcal{R}\to\infty}\lim_{T\to\infty}\text{Var}\left(\hat{\vartheta}_m\right) = \lim_{\mathcal{R}\to\infty}\frac{\tau_\omega^2}{\varphi} + \frac{\tau_r^2}{\varphi\mathcal{R}} = \frac{\tau_\omega^2}{\varphi}. \tag{C.90}$$

When we observe the data without error ($\tau_\omega^2 = 0$), the limiting variance is zero, and the estimate is consistent.

In the alternative scenario, we no longer require error free data observation.

$$\lim_{\varphi\to\infty}\lim_{T\to\infty}\text{Var}\left(\hat{\vartheta}_m\right) = \lim_{\varphi\to\infty}\frac{\tau_\omega^2}{\varphi} + \frac{\tau_r^2}{\varphi\mathcal{R}} = 0. \tag{C.91}$$

□

The proof of Theorem 4.14 is a direct result of Condition 4.11.

Proof of Theorem 4.14. From Condition 4.11, we see that $\text{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right) \overset{i.i.d.}{\sim} \left(\vartheta_m, \tau_\omega^2\right)$. For an individual realization, we have shown in the proofs of Lemma C.1 and C.2 that

$$\frac{1}{\mathcal{R}K}\sum_{r=1}^{\mathcal{R}}\sum_{k=1}^{K} T_k^{-1}\text{tr}\left(\mathbf{R}_{m,\omega,r,k}^\top\mathbf{R}_{m,\omega,r,k}\right)|\boldsymbol{\Sigma}_{m,\omega} \xrightarrow{P} \text{tr}\left(\boldsymbol{\Sigma}_{m,\omega}\right) \tag{C.92}$$

as the number of observations $T \to \infty$. The variation parameter is the average of the individual realization-specific covariance matrices, and the Central Limit Theorem arises from the application of Slutsky's Theorem. □

The proof of Theorem 4.15 begins with the empirical distribution function, and

employs the result of the Glivenko-Cantelli to show convergence in distribution to a Uniform$(0, 1)$ random variable.

Proof of Theorem 4.15. Under the null hypothesis, $\vartheta_1 = \cdots = \vartheta_{T!} = \vartheta$. We define the quantile estimate $\hat{Q}_M$, shown in Equation 4.13, as an evaluation of the empirical distribution function $\hat{\mathcal{H}}_M(s)$ at $\hat{\vartheta}_1$. As the number of permutations M approaches the total number of possible permutations $T!$, the empirical distribution $\hat{\mathcal{H}}_M(s)$ trivially converges to $\hat{\mathcal{H}}(s)$ defined in Equation 4.11.

From the results of Lemmas C.1 and C.2, we know that as $T \to \infty$, $\hat{\vartheta}_1 \xrightarrow{D} \mathcal{F}$, where the continuous distribution $\mathcal{F}$ has expectation ϑ and variance $\varphi^{-1}\tau_\omega^2 + (\varphi\mathcal{R})^{-1}\tau_r^2$. Under the null, each realization $\hat{\vartheta}_m$ is independently drawn from $\mathcal{F}$. From the Glivenko-Cantelli Theorem as $T \to \infty$,

$$\sup_s \left| \hat{\mathcal{H}}(s) - \mathcal{F}(s) \right| \xrightarrow{a.s.} 0. \tag{C.93}$$

We define $\mathcal{U} \sim$ Uniform$(0, 1)$ and note that if $s \sim \mathcal{G}$, we can write $s \sim \mathcal{G}^{-1}(\mathcal{U})$. We state that $\mathcal{F}\left(\lim_{T\to\infty} \hat{\vartheta}_1\right)$ follows the same distribution as $\mathcal{F}\left(\mathcal{F}^{-1}\left(\mathcal{U}\right)\right) \sim \mathcal{U}$.

Combining these results and Theorem 4.13 with the continuous mapping theorem in Equation C.96 and Slutsky's Theorem,

$$\lim_{M\to T!} \hat{\mathcal{H}}_M\left(\hat{\vartheta}_1\right) \to \hat{\mathcal{H}}\left(\hat{\vartheta}_1\right) \tag{C.94}$$

$$\lim_{T\to\infty} \hat{\mathcal{H}}\left(\hat{\vartheta}_1\right) \to \mathcal{F}\left(\hat{\vartheta}_1\right) \tag{C.95}$$

$$\lim_{T\to\infty} \mathcal{F}\left(\hat{\vartheta}_1\right) \xrightarrow{D} \mathcal{F}\left(\lim_{T\to\infty} \hat{\vartheta}_1\right) \tag{C.96}$$

$$\mathcal{F}\left(\lim_{T\to\infty} \hat{\vartheta}_1\right) \sim \mathcal{U}, \tag{C.97}$$

so under Conditions 4.6 - 4.12,

$$\lim_{T\to\infty}\lim_{M\to T!}\hat{Q}_M = \lim_{T\to\infty}\lim_{M\to T!}\hat{\mathcal{H}}_M\left(\hat{\vartheta}_1\right) \xrightarrow{D} \mathcal{U}. \tag{C.98}$$

□

The proof of Theorem 4.16 follows from the result of Theorem 4.13.

Proof of Theorem 4.16. Under the alternative, we have $\vartheta_1 < \vartheta_i$ for all possible permutations $i = 1, \ldots, T!$. We observe M permutations of the total possible. Define δ_M as the minimum difference between the underlying parameter ϑ_1 and another permutation in $m = 1, \ldots, M$, and η_M as the maximum estimation error over all permutations $m = 1, \ldots, M$.

$$\delta_M = \min_{1<m\leq M} |\vartheta_m - \vartheta_1| \geq \delta \quad \text{where } \delta = \lim_{M\to T!}\delta_M \tag{C.99}$$

$$\eta_M = \max_{1\leq m\leq M} \left|\hat{\vartheta}_m - \vartheta_m\right| \leq \eta \quad \text{where } \eta = \lim_{M\to T!}\eta_M \tag{C.100}$$

If $2\eta_M < \delta$, then the quantile estimate $\hat{Q}_M$ will return the true value in the limit. Writing the minimum using the definition in Equation 4.13,

$$\min_{2\eta_M<\delta}\hat{Q}_M = \min_{2\eta_M<\delta}\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{\hat{\vartheta}_m \leq \hat{\vartheta}_1\right\} \tag{C.101}$$

$$= \min_{2\eta_M<\delta}\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{\hat{\vartheta}_m - \hat{\vartheta}_1 \leq 0\right\} \tag{C.102}$$

$$= \min_{2\eta_M<\delta}\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{(\hat{\vartheta}_m - \vartheta_m) + (\vartheta_1 - \hat{\vartheta}_1) + (\vartheta_m - \vartheta_1) \leq 0\right\}. \tag{C.103}$$

The quantity in Equation C.103 will reach its minimum when the estimation errors are maximized, leading to a larger value inside the indicator and fewer pairs that

meet the criteria.

$$\min_{2\eta_M<\delta} \hat{Q}_M \geq \min_{2\eta_M<\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{2\eta_M + (\vartheta_m - \vartheta_1) \leq 0\right\} \tag{C.104}$$

$$\geq \min_{2\eta_M<\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{2\eta_M + \vartheta_m \leq \vartheta_1\right\} \tag{C.105}$$

For the maximum,

$$\max_{2\eta_M<\delta} \hat{Q}_M = \max_{2\eta_M<\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{(\hat{\vartheta}_m - \vartheta_m) + (\vartheta_1 - \hat{\vartheta}_1) + (\vartheta_m - \vartheta_1) \leq 0\right\} \tag{C.106}$$

$$\leq \max_{2\eta_M<\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{-2\eta_M + (\vartheta_m - \vartheta_1) \leq 0\right\} \tag{C.107}$$

$$\leq \max_{2\eta_M<\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\left\{\vartheta_m \leq \vartheta_1 + 2\eta_M\right\}. \tag{C.108}$$

To show consistent behavior with the original ordering, if $2\eta_M < \delta$ and $\vartheta_m \leq \vartheta_1$,

$$2\eta_M + \vartheta_m \leq 2\eta_M + \vartheta_1 - \delta \leq \vartheta_1 \tag{C.109}$$

$$\text{and} \quad \vartheta_m \leq \vartheta_1 - \delta + 2\eta_M \leq \vartheta_1 + 2\eta_M, \tag{C.110}$$

or if $2\eta_M < \delta$ and $\vartheta_m > \vartheta_1$,

$$2\eta_M + \vartheta_m > 2\eta_M + \vartheta_1 + \delta > \vartheta_1 \tag{C.111}$$

$$\text{and} \quad \vartheta_m > \vartheta_1 + \delta > \vartheta_1 + 2\eta_M. \tag{C.112}$$

Thus, the ordering is preserved.

$$\lim_{M\to T!}\min_{2\eta_M<\delta}\hat{Q}_M \geq \lim_{M\to T!}\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{\vartheta_m \leq \vartheta_1\right\} \geq Q \tag{C.113}$$

$$\lim_{M\to T!}\max_{2\eta_M<\delta}\hat{Q}_M \leq \lim_{M\to T!}\frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\left\{\vartheta_m \leq \vartheta_1\right\} \leq Q \tag{C.114}$$

$$\implies \lim_{M\to T!}\hat{Q}_M = Q \;\text{ when } 2\eta_M < \delta \tag{C.115}$$

We now need to show the probability of $2\eta_M < \delta$ goes to one in the limit. Pick any $\delta > 0$.

$$\mathbb{P}\left(\eta_M < \frac{\delta}{2}\right) = \mathbb{P}\left(\max_{1\leq m\leq M}\left|\hat{\vartheta}_m - \vartheta_m\right| < \frac{\delta}{2}\right) \tag{C.116}$$

From the result of Theorem 4.13, when $\tau_\omega^2 = 0$,

$$\lim_{\mathcal{R}\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\left|\hat{\vartheta}_m - \vartheta_m\right| < \frac{\delta}{2}\right) = 1. \tag{C.117}$$

Similarly,

$$\lim_{\varphi\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\left|\hat{\vartheta}_m - \vartheta_m\right| < \frac{\delta}{2}\right) = 1. \tag{C.118}$$

Returning to the maximum,

$$\mathbb{P}\left(\max_{1\leq m\leq M}\left|\hat{\vartheta}_m - \vartheta_m\right| < \frac{\delta}{2}\right) = \mathbb{P}\left(\left|\hat{\vartheta}_1 - \vartheta_1\right| < \frac{\delta}{2}, \ldots, \left|\hat{\vartheta}_M - \vartheta_M\right| < \frac{\delta}{2}\right) \tag{C.119}$$

$$\geq 1 - \sum_{m=1}^{M}\mathbb{P}\left(\left|\hat{\vartheta}_m - \vartheta_m\right| \geq \frac{\delta}{2}\right), \tag{C.120}$$

and taking the limit under the two scenarios,

$$\lim_{\mathcal{R}\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\max_{1\leq m\leq M}\left|\hat{\vartheta}_m-\vartheta_m\right|<\frac{\delta}{2}\right)\geq 1-\sum_{m=1}^{M}\lim_{\mathcal{R}\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\left|\hat{\vartheta}_m-\vartheta_m\right|\geq\frac{\delta}{2}\right)=1, \tag{C.121}$$

when $\tau_\omega^2=0$, and

$$\lim_{\varphi\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\max_{1\leq m\leq M}\left|\hat{\vartheta}_m-\vartheta_m\right|<\frac{\delta}{2}\right)\geq 1-\sum_{m=1}^{M}\lim_{\varphi\to\infty}\lim_{T\to\infty}\mathbb{P}\left(\left|\hat{\vartheta}_m-\vartheta_m\right|\geq\frac{\delta}{2}\right)=1. \tag{C.122}$$

This implies $\hat{Q}_M \xrightarrow{P} Q$ in both cases presented in the statement of Theorem 4.16. □

The proof of Theorem 4.17 follows a step by step derivation of the distribution from the known starting point of Condition 4.12.

Proof of Theorem 4.17. We begin with the distribution of the model residuals that is known from Equation 4.27 and Condition 4.12, $\mathbf{U}_{m,\omega,r}\sim\mathcal{MN}_{T\times d}\left(\mathbf{0},\mathbf{I}_T,\mathbf{S}_{m,\omega,r}\right)$. The predicted residuals can be expressed as a function of the model residuals, like in Equation 4.9, and their distribution follows from that of $\mathbf{U}_{m,\omega,r}$ as in Equation C.4. Like in the proof of Lemma C.1, we define

$$\mathbf{\Phi}_{m,\omega,r,k}=\mathbf{H}_{m,\omega,r,k}\left[\mathbf{H}_{m,\omega,r,-k}^{\top}\mathbf{H}_{m,\omega,r,-k}\right]^{-1}\mathbf{H}_{m,\omega,r,k}^{\top}, \tag{C.123}$$

and further define

$$\mathbf{V}_{m,\omega,r,k}\sim\mathcal{MN}_{T_k\times d}(\mathbf{0},\mathbf{I}_{T_k},\mathbf{I}_d) \tag{C.124}$$

such that we can write $\mathbf{R}_{m,\omega,r,k} = \left(\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right)^{1/2} \mathbf{V}_{m,\omega,r,k} \mathbf{S}_{m,\omega,r}^{1/2}$. The distribution has an equivalent vectorized form, with covariance matrix the Kronecker product between the row-wise and column-wise variation.

$$\mathbf{r}_{m,\omega,r,k} = \text{vec}\left(\mathbf{R}_{m,\omega,r,k}\right) \sim \mathcal{N}_{T_K d}\left(\mathbf{0}, \left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right] \otimes \mathbf{S}_{m,\omega,r}\right) \tag{C.125}$$

For simplicity, define $\mathbf{\Delta}_{m,\omega,r,k} = \left[\mathbf{\Phi}_{m,\omega,r,k} + \mathbf{I}_{T_k}\right] \otimes \mathbf{S}_{m,\omega,r}$ and write the equivalent form $\mathbf{r}_{m,\omega,r,k} = \mathbf{\Delta}_{m,\omega,r,k}^{1/2} \mathbf{v}_{m,\omega,r,k}$, where $\mathbf{v}_{m,\omega,r,k} \sim \mathcal{N}_{T_k d}\left(\mathbf{0}, \mathbf{I}\right)$. The covaraince matrix $\mathbf{\Delta}_{m,\omega,r,k} = \left(\delta_{m,\omega,r,k,ij}\right)$ is symmetric and each element in the vector $\mathbf{v}_{m,\omega,r,k} = \left(v_{m,\omega,r,k,i}\right) \sim \mathcal{N}(0,1)$.

$$\begin{aligned}
\mathbf{r}_{m,\omega,r,k}^{\top} \mathbf{r}_{m,\omega,r,k} &= \mathbf{v}_{m,\omega,r,k}^{\top} \mathbf{\Delta}_{m,\omega,r,k} \mathbf{v}_{m,\omega,r,k} && \text{(C.126)} \\
&= \sum_{i=1}^{T_k d} \left(\delta_{m,\omega,r,k,ii} v_{m,\omega,r,k,i}^2 + \sum_{\substack{j=1 \\ j \neq i}}^{T_k d} \delta_{m,\omega,r,k,ij} v_{m,\omega,r,k,i} v_{m,\omega,r,k,j} \right) && \text{(C.127)} \\
&= \sum_{i=1}^{T_k d} \left(\delta_{m,\omega,r,k,ii} v_{m,\omega,r,k,i}^2 + 2 \sum_{j=1}^{i-1} \delta_{m,\omega,r,k,ij} v_{m,\omega,r,k,i} v_{m,\omega,r,k,j} \right) && \text{(C.128)} \\
&= \sum_{i=1}^{T_k d} \left(\delta_{m,\omega,r,k,ii} v_{m,\omega,r,k,i}^2 + \frac{1}{2} \sum_{j=1}^{i-1} \delta_{m,\omega,r,k,ij} \left(v_{m,\omega,r,k,i} + v_{m,\omega,r,k,j} \right)^2 \right. \\
&\qquad \left. - \frac{1}{2} \sum_{j=1}^{i-1} \delta_{m,\omega,r,k,ij} \left(v_{m,\omega,r,k,i} - v_{m,\omega,r,k,j} \right)^2 \right) && \text{(C.129)}
\end{aligned}$$

We note that

$$v_{m,\omega,r,k,i}^2 \sim \chi_1^2 \tag{C.130}$$

$$\frac{1}{2} (v_{m,\omega,r,k,i} + v_{m,\omega,r,k,j})^2 \sim \chi_1^2 \tag{C.131}$$

$$\frac{1}{2} (v_{m,\omega,r,k,i} - v_{m,\omega,r,k,j})^2 \sim \chi_1^2, \tag{C.132}$$

and $\delta_{m,\omega,r,k,ij}$ depend on the matrices $\mathbf{\Phi}_{m,\omega,r,k} = (\phi_{m,\omega,r,k,ij})$ and $\mathbf{S}_{m,\omega,r} = (s_{m,\omega,r,ij})$. Let $i' = \lceil i/d \rceil$, $j' = \lceil j/d \rceil$, $i^* = i \bmod d$, and $j^* = j \bmod d$. We define the following random variables for all $\omega \in \Omega_{\text{obs}}$, $r = 1, \ldots, \mathcal{R}$, $k = 1, \ldots, K$, $i = 1, \ldots, T_k d$ and $j < i$.

$$X_{m,\omega,r,k,i}, Y_{m,\omega,r,k,ij}, Z_{m,\omega,r,k,ij} \sim \chi^2_1 \tag{C.133}$$

We can write each $\delta_{m,\omega,r,k,ij} = (\phi_{m,\omega,r,k,i'j'} + \mathbf{1}\{i' = j'\})\, s_{m,\omega,r,k,i^*j^*}$ and input the definitions.

$$\begin{aligned}\mathbf{r}^\top_{m,\omega,r,k}\mathbf{r}_{m,\omega,r,k} = \sum_{i=1}^{T_k d} \Bigg[& (\phi_{m,\omega,r,k,i'i'} + 1)\, s_{m,\omega,r,k,i^*i^*} X_{m,\omega,r,k,i} \\ & + \sum_{j=1}^{d\lfloor (i-1)/d \rfloor} \phi_{m,\omega,r,k,i'j'} s_{m,\omega,r,k,i^*j^*} \left(Y_{m,\omega,r,k,ij} - Z_{m,\omega,r,k,ij}\right) \\ & + \sum_{j=d\lfloor (i-1)/d \rfloor + 1}^{i-1} (\phi_{m,\omega,r,k,i'j'} + 1)\, s_{m,\omega,r,k,i^*j^*} \left(Y_{m,\omega,r,j,ij} - Z_{m,\omega,r,k,ij}\right) \Bigg]\end{aligned} \tag{C.134}$$

To obtain the distribution for the variation parameter, we aggregate this generalized chi-square distribution over the test sets $k = 1, \ldots, K$, featurizations $r = 1, \ldots, \mathcal{R}$, and potential realizations $\omega \in \Omega_{\text{obs}}$. □