Detection of Changes in Multivariate Time Series With Application to EEG data *

Claudia Kirch†, Birte Muhsal‡, Hernando Ombao§

December 19, 2013

Abstract

The primary contributions of this paper are rigorously developed novel statistical methods for detecting change points in multivariate time series. We extend the class of score type change point statistics considered by [Hušková et al., 2007] to the vector autoregressive (VAR) case and the epidemic change alternative. Our proposed procedures do not require the observed time series to actually follow the VAR model. Instead, following the strategy implicitly employed by practitioners, our approach takes model misspecification into account so that our detection procedure uses the model background merely for feature extraction. We derive the asymptotic distributions of our test statistics and show that our procedure has asymptotic power of 1. The proposed test statistics require the estimation of the inverse of the long-run covariance matrix which is particularly difficult in higher-dimensional settings (i.e., where the dimension of the time series and the dimension of the parameter vector are both large). Thus we robustify the proposed test statistics and investigate their finite sample properties via extensive numerical experiments. Finally, we apply our procedure to electroencephalograms and demonstrate its potential impact in identifying change points in complex brain processes during a cognitive motor task.

Keywords: Change points; multivariate time series; epidemic change; vector autoregressive model; EEG data;

AMS Subject Classification 2010:

1 Introduction

The primary contributions of this paper are rigorously developed novel statistical methods for detecting change points in multivariate time series. Change point detection is crucial in analyzing many time series

*This work was supported by DFG grant KI 1443/2-2, US National Science Foundation (Division of Mathematical Sciences) and the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust which also financed the position of the first author.

†Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Kaiserstr. 89, 76133 Karlsruhe, Germany; claudia.kirch@kit.edu

‡Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Kaiserstr. 89, 76133 Karlsruhe, Germany; birte.muhsal@kit.edu

§University of California at Irvine, Department of Statistics, Irvine CA 92697, USA; hombao@uci.edu

data – ignoring them leads to incorrect conclusions and forecasts. Thus, experts in finance keenly watch out for changes in several market indicators because these could indicate impact of government policy or mark a need to revise long term investment strategies. Moreover, neuroscientists study the effect of external stimuli on neuronal responses by tracking changes in brain signals such as electroencephalograms. Brain signals are highly dynamic since these are realizations of complex cognitive processes. In this context detection of change points is necessary in order to understand how a cognitive process unfolds in response to a stimulus.

Consider two background processes, denoted Π_1 and Π_2, that run in parallel but for each time point only one of these two is activated. In other words, when Π_1 is activated at time t then Π_2 is turned off. These two processes are identifiable in the sense that they have different second order structure and thus our goal is to search for the point(s) in time when the covariance structure changes. Here, we shall investigate two scenarios: the at-most-one-change (AMOC) and the epidemic change. In the AMOC situation, the goal is to estimate the time t_1 when Π_1 is turned off (equivalently, the time when Π_2 is activated). In the epidemic change situation, change point t_1 corresponds to the time when Π_1 is turned off (equivalently, when Π_2 is activated) and change point t_2 corresponds to the time Π_2 is turned off thus returning to Π_1. Here, the goal is to estimate the two change points t_1 and t_2. To identify change points, or to analyze time series data in general, there are many possible stochastic representations or time series models that could be utilized. Based on both theoretical and practical considerations, we shall develop methods that use autoregressive (AR) models. As studied by [Bickel and Bühlman, 1999] in the context of bootstrap methods, linear stationary processes can be well approximated by autoregressive (AR) processes that have sufficiently large order. Moreover, the AR model has been utilized to analyze many real-life time series data such as electroencephalograms ([Pfurtscheller and Haring, 1972]) seismic signals ([Lesage et al., 2002]) and speech signals ([Vermaak et al., 2002]). While AR models do not fully capture the features of many complex time series data, here, we do not claim that the observed time series data is a realization of some AR model. Rather, the AR model will be utilized as a tool for feature extraction whose parameters capture changes in the second order structure of the time series.

We provide the theory for a large class of statistics, which will be of interest in many different areas. However, in practice it is of utter importance to adapt the test statistic to the actual data being analyzed. We will demonstrate how to do this throughout the paper using the data example that will be examined in detail in Section 4. Our data consists of electroencephalograms (EEG) recorded from a single participant in a visual-motor task experiment. In this experiment, the participant was instructed to move the hand-held joystick either towards the left of center or towards the right of center. Each trial consists of one second (total of $n = 512$ time points) whose midpoint is approximately the point in time when the stimulus was presented (at $t = 250$). Thus, each trial consists of half a second of EEG recording before stimulus presentation and another half a second post stimulus presentation. Change points are expected on the post-stimulus recording as the brain integrates and processes information. Indeed the proposed method clearly captures these changes in brain dynamics. We note that our EEG data example serves only as an illustration – the methodology has broad applicability and is by no means limited to EEG data.

To describe our proposed procedure, we first give a brief introduction on the vector autoregressive model. Let $\mathbf{Y}_t = [Y_t(1), \ldots, Y_t(d)]^\top$ be a $d-$dimensional vector recorded at time t across all d channels. Then $\mathbf{Y}_t$ is said to follow a vector autoregressive model of order p, denoted VAR(p), if it can be expressed as $\mathbf{Y}_t = \boldsymbol{\alpha}_1\mathbf{Y}_{t-1} + \ldots + \boldsymbol{\alpha}_p\mathbf{Y}_{t-p} + \boldsymbol{\epsilon}_t$ where $\boldsymbol{\alpha}_\ell,\ \ell = 1, \ldots, p,$ are the $d \times d$ autoregressive matrices

corresponding to the lagged values of $\mathbf{Y}_t$; and $\boldsymbol{\epsilon}_t$ is white noise. An equivalent formulation of the VAR(p) model is given as follows. Define

$$\mathbb{Y}_{t-1} = [\mathbb{Y}_{t-1}(1)^\top, \ldots, \mathbb{Y}_{t-1}(d)^\top]^\top \quad \text{where} \quad \mathbb{Y}_{t-1}(j) = [Y_{t-1}(j), \ldots, Y_{t-p}(j)]^\top, \;\; j = 1, \ldots, d.$$

Here, $\dim(\mathbb{Y}_{t-1}) = pd \times 1$ and $\dim(\mathbb{Y}_{t-1}(i)) = p \times 1$. Next define the matrix $\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}^\top(1) \\ \ldots \\ \boldsymbol{\Phi}^\top(d) \end{pmatrix}$ with $\dim(\boldsymbol{\Phi}) = d \times dp$ whose i-th row is given by

$$\boldsymbol{\Phi}^\top(i) = [\phi_{1,1}(i), \ldots, \phi_{1,p}(i), \ldots, \phi_{d,1}(i), \ldots, \phi_{d,p}(i)].$$

The conditional mean of the i-th channel $Y_t(i)$ is a linear combination of values at all channels from time $t-1$ through $t-p$ and $\phi_{\ell,j}(i)$ is the coefficient corresponding to the lag j value of the ℓ-th channel $Y_{t-\ell}(j)$. Thus, the VAR(p) model can be written as

$$\mathbf{Y}_t = \boldsymbol{\Phi}\mathbb{Y}_{t-1} + \boldsymbol{\epsilon}_t.$$

The least squares estimator for the unknown matrix $\boldsymbol{\Phi}$ based on the observed time series $\{\mathbf{Y}_t,\ t = 1, \ldots, n\}$ is the minimizer

$$\widehat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{d}\sum_{t=1}^{n}[Y_t(i) - \boldsymbol{\Phi}^\top(i)\mathbb{Y}_{t-1}]^2.$$

The vector of estimators for the coefficients in the conditional mean of the i-th channel is then given by

$$\widehat{\boldsymbol{\Phi}}^\top(i) = \left(\sum_{t=1}^{n} Y_t(i)\mathbb{Y}_{t-1}^\top\right)\left(\sum_{t=1}^{n} \mathbb{Y}_{t-1}\mathbb{Y}_{t-1}^\top\right)^{-1}.$$

Let us define the residual vector to be $\mathbf{R}_t = \mathbf{Y}_t - \widehat{\boldsymbol{\Phi}}\mathbb{Y}_{t-1}$. If the null hypothesis of no change across $t = 1, \ldots, n$ is true and that VAR model has a sufficiently large order providing a good approximation to the covariance structure of $\mathbf{Y}_t$, then the expectation of $\mathbf{R}_t$ is (almost) zero and we can expect the partial sum process to be "small". Strong deviations of the partial sum process from zero, on the other hand, indicate that the estimator of $\boldsymbol{\Phi}$ does not provide a good fit to the data and could be evidence for a change point. In fact, the first statistics considered for univariate AR time series were based on these partial sum processes (see [Kulperger, 1985] and [Horváth, 1993]). However, such statistics can only detect certain alternatives because they ignore other properties of the residuals such as the uncorrelatedness with past observations. In particular, these statistics are meaningless if the time series are centered.

Consequently, [Davis et al., 1995] investigate the maximum likelihood statistic in the univariate AR model, while [Hušková et al., 2007] consider the corresponding score type test and variations thereof. The score type test is based not only on partial sums of the estimated residuals (in the non-centered situation) but also on partial sums of the residuals multiplied with past observations which – under the null hypothesis and correct specification – has also expectation zero, so that the partial sums should be small. If this is not the case, the global estimator $\widehat{\boldsymbol{\Phi}}$ does not fit the full data set well possibly due to the existence of one or more change points and statistical inference based on this estimator is rendered meaningless. All of the above statistics are particularly designed for the at-most-one-change situation.

While they do have some power against multiple changes, a fact binary segmentation procedures make use of, their power is best in the at-most-one-change situation.

Tests particularly designed for the multiple change situation are only recently being developed but most of them require at least an upper bound for the number of change points (see [Bai and Perron, 1998] and [Marušiaková, 2009]. Exceptions are MOSUM based procedures discussed in [Hušková and Slabý, 2001] and [Kirch and Muhsal, 2011]. On the other hand, there are established procedure available not for testing but for estimation in a multiple change scenario that are based on the minimization of an appropriate information criteria such as AutoParm developed by [Davis et al., 2006] which minimizes the minimum description length in the context of autoregressive time series. However, these methods do not provide measures of uncertainty (e.g., confidence levels) on the change point procedures. Similarly, tests particularly designed for multivariate time series are only currently becoming of interest, some recent examples are [Aue et al., 2009] and [Preußs et al., 2013].

In this paper, we extend the class of score type change point statistics considered by [Hušková et al., 2007] to the vector autoregressive case and the epidemic change alternative. Furthermore, we no longer require that the observed time series actually follows this model but take misspecification into account leading to a detection procedure that uses the model background merely for feature extraction, something that is implicitly done by many practitioners.

The remainder of the paper is organized as follows. In Section 2, we develop our statistical procedure for testing and identification of change points in the at most one change (AMOC) and in the epidemic settings. We derive the asymptotic distributions of our test statistics and show that the corresponding tests have asymptotic power of 1. The proposed test statistics require the estimation of the inverse of the (long-run) auto-covariance matrix a task particularly difficult in higher-dimensional settings. Consequently, in Section 3, we describe methods of robustifying the proposed test statistics and investigate their finite sample properties via extensive numerical experiments. We demonstrate the potential broad impact of our procedure by analyzing electroencephalogram data set in Section 4.

2 Proposed Methods for Change Points Analysis

2.1 Test Statistics and Null Asymptotics

2.1.1 Background and Notation

Recall that the i-th channel in the VAR model is $Y_t(i) = \mathbf{\Phi}^\top(i)\mathbb{Y}_{t-1} + \epsilon_t(i)$, where $\mathbf{\Phi}(i)$ are the coefficients for past data at all channels $\mathbb{Y}_{t-1}$. The number of coefficients for the i-th component is dp which is typically large and consequently performing a large number of tests of hypotheses could lead to severe size distortions or reductions in power (see [Aston and Kirch, 2013]) both of which can possibly be avoided in situations where changes in the time series can be captured by a subset of the dp parameters. Of course, there is always the danger of excluding a parameter that is sensitive to changes in the process. Since our focus is on change point detection, we need to delicately balance between two competing aims: (a.) ensuring a good approximation to the covariance structure which may require a VAR model with a sufficiently high order and (b.) controlling the size of the test and ensuring sufficient power which requires some careful reduction of the parameter space. Thus, our goal here is to provide a flexible procedure that allows user-input by removing the VAR parameters that are determined, from prior knowledge,

to be not meaningful in change point detection. Moreover, since the $d \times dp$ matrix of parameters $\mathbf{\Phi}$ contains information on all lags of the covariance function $\text{cov}(\mathbf{Y}_{t+h}, \mathbf{Y}_t)$, $h = 0, \pm 1, \ldots$, there will be some redundant information shared by the elements of $\mathbf{\Phi}$. Hence, it is sensible to construct a test statistic based only on a subset of $\mathbf{\Phi}$ that captures the covariance features in the data well enough for change detection purposes.

To develop our testing procedure, we first change the notation of the entries of $\mathbf{\Phi}$ as follows

$$\mathbf{\Phi}(i) \;=\; [\phi(i,1), \ldots, \phi(i,p), \ldots, \phi(i, d(p-1)+1), \ldots, \phi(i,dp)]^\top$$

to simplify the description of the algorithm. Then, we define the indicator set $\mathcal{I}(i) = \{r : \phi(i,r) = 0\}$ to be the coefficients, corresponding to the past values at all channels that are associated with the current value for the i-th channel, that we set 0 for the change detection procedure and denote $|\mathcal{I}(i)|$ to be the cardinality of this set. Thus, there remain $dp - |\mathcal{I}(i)|$ coefficients that are unknown and need to be estimated. To obtain the specific past values of $\mathbb{Y}_{t-1}$ which correspond to the non-zero coefficients $\phi(i,r)$, we introduce the following operator $\mathcal{P}_{\mathcal{I}}(\cdot)$: for any dp-dimensional vector $\mathbb{A} = [A_1, A_2, \ldots, A_{dp}]^\top$, let

$$\mathcal{P}_{\mathcal{I}}(\mathbb{A}) = [A_r,\ r \notin \mathcal{I}]^\top$$

denote the vector, that only contains the elements not set 0 in the change detection procedure. Introducing the notation

$$\mathbf{a}(i) = \mathcal{P}_{\mathcal{I}(i)}\left(\mathbf{\Phi}(i)\right) \quad \text{and} \quad \mathbb{X}_{t-1}(i) = \mathcal{P}_{\mathcal{I}(i)}\left(\mathbb{U}_{t-1}\right)$$

with $\dim(\mathbb{X}_{t-1}(i)) = dp - |\mathcal{I}(i)| = \dim(\mathbf{a}(i))$, the AR model (used in the change detection procedure) for the i-th channel becomes $Y_t(i) = \mathbf{a}(i)^\top \mathbb{X}_{t-1}(i) + e_t(i)$. To emphasize the fact that we do not expect this to be the true model, we rename the residuals $\{e_t\}$ and allow them to be dependent. Then, the least squares estimator for $\mathbf{a}(i)$ is given by

$$\widehat{\mathbf{a}}_n^\top(i) = \left(\sum_{t=1}^{n} Y_t(i)\mathbb{X}_{t-1}(i)^\top\right)\left(\sum_{t=1}^{n} \mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}(i)^\top\right)^{-1}.$$

Example 2.1. One possible subset of coefficients that can be used to detect change points in multivariate time series are the coefficients that relate $Y_t(i)$ with its *own* past observations $\mathbb{Y}_{t-1}(i)$. This is appealing because of interpretability as this is precisely what one would do if only the univariate information is available for model building. It may *appear* at first glance that the auto-regressive parameters ignore cross-dependence between channels of the multivariate time series. However, we demonstrate below that this is not entirely true. For channel 1, the indicator set is $\mathcal{I}(1) = \{(p+1), \ldots, pd\}$ and the set of non-zero coefficients and corresponding past values are, respectively,

$$\mathbf{a}(1) = [\phi(1,1), \ldots, \phi(1,p)]^\top \quad \text{and} \quad \mathbb{X}_{t-1}(1) = [Y_{t-1}(1), \ldots, Y_{t-p}(1)]^\top$$

where $|\mathcal{I}(1)| = dp - p = (d-1)p$; $\dim(\mathbb{X}_{t-1}(1)) = p$ and $\dim(\mathbf{a}(1)) = p$. The AR model for channel 1 and for any channel $i = 2, \ldots, d$ are, respectively,

$$\begin{aligned}
Y_t(1) &= \phi_{1,1}(1)Y_{t-1}(1) + \ldots + \phi_{1,p}(1)Y_{t-p}(1) + e_t(1) \\
&= \phi(1,1)Y_{t-1}(1) + \ldots + \phi(1,p)Y_{t-p}(1) + e_t(1), \\
Y_t(i) &= \phi_{1,1}(i)Y_{t-1}(i) + \ldots + \phi_{1,p}(i)Y_{t-p}(i) + e_t(i) \\
&= \phi(i,(i-1)p+1)Y_{t-1}(i) + \ldots + \phi(i,ip)Y_{t-p}(i) + e_t(i).
\end{aligned}$$

Note that the cross-dependence structure across channels is actually not being ignored in this change point detection scheme because the within-channel (auto-) regression parameters capture information concerning the cross-dependence structure and any left-over information will be absorbed by the residuals.

2.1.2 The At-Most-One-Change and Epidemic Change Point Settings

We shall develop a testing procedure for two types of change point situations: (1.) at-most-one-change situation (AMOC) and (2.) epidemic change model which consists of two change points where the process reverts back to the original regime after the second change point. Denote $Y_t^{(1)}(i)$ to be the process at regime Π_1 and $Y_t^{(2)}(i)$ at Π_2. The AMOC model with a change point at $\tilde{k}$ is given by

$$Y_t(i) = \begin{cases} Y_t^{(1)}(i) = \mathbf{a}_1(i)^\top \mathbb{X}_{t-1}^{(1)}(i) + e_t^{(1)}(i), & 1 \leqslant t \leqslant \widetilde{k}, \\ Y_t^{(2)}(i) = \mathbf{a}_2(i)^\top \mathbb{X}_{t-1}^{(2)}(i) + e_t^{(2)}(i), & \widetilde{k} < t \leqslant n. \end{cases} \tag{2.1}$$

The epidemic change model is given by

$$Y_t(i) = \begin{cases} Y_t^{(1)}(i) = \mathbf{a}_1(i)^\top \mathbb{X}_{t-1}^{(1)}(i) + e_t^{(1)}(i), & 1 \leqslant t \leqslant \widetilde{k}_1, \widetilde{k}_2 < t \leqslant n \\ Y_t^{(2)}(i) = \mathbf{a}_2(i)^\top \mathbb{X}_{t-1}^{(2)}(i) + e_t^{(2)}(i), & \widetilde{k}_1 < t \leqslant \widetilde{k}_2, \end{cases} \tag{2.2}$$

where the two change points are given by $\widetilde{k}_1$ and $\widetilde{k}_2$.

2.1.3 Assumptions on the Processes

Before we introduce the corresponding test statistics and their null asymptotics we first need to set the stage by posting some assumptions on the underlying processes. In the correctly specified model the innovations are i.i.d., but here we allow for misspecification which introduces some dependency to the residuals. For example, if the true underlying process follows a full VAR-model but we only use the restricted model of Example 2.1, the errors are no longer independent but contain the remaining structural information of the time series. The assumptions below allow for this but also for other kinds of misspecification.

Assumption $\mathcal{A}$. 1. Let $\{\mathbf{Y}_t^{(1)} = (Y_t^{(1)}(1), \ldots, Y_t^{(1)}(d))^\top : t \geqslant 1\}$, be an $\mathbb{R}$-valued strictly stationary sequence of non-degenerate random vectors with

$$\mathrm{E}\,\mathbf{Y}_1^{(1)} = 0, \qquad \mathrm{E}\,\|\mathbf{Y}_1^{(1)}\|^{4+\nu} < \infty \quad \text{for some } \nu > 0,$$

satisfying a strong mixing condition with mixing rate

$$\alpha_1(n) = O\left(n^{-\beta}\right) \quad \text{for some } \beta > \max\left(3, \frac{4+\nu}{\nu}\right).$$

We merely make this classic mixing assumption for simplicity of presentation because it ensures invariance principles for functionals of the time series. However, our results carry over to different more recent weak dependency concepts as long as the necessary invariance principles are still available. In the correctly specified case, this assumption is not necessary (e.g. one can allow for discrete errors) as other tools to derive invariance principles are available (for details we refer to [Hušková et al., 2007]).

A large class of these weak dependent processes such as ARMA-models or linear processes (in the strict sense) can well be approximated by a large enough linear autoregressive model as given in (2.1). Define

$$\mathbf{a}_1(i) = \boldsymbol{C}_1^{-1}(i)\boldsymbol{c}_1(i), \qquad e_t^{(1)}(i) = Y_t^{(1)}(i) - \mathbf{a}_1^\top(i)\mathbb{X}_{t-1}^{(1)}(i),$$
$$\boldsymbol{C}_1(i) = \mathrm{E}\left(\mathbb{X}_{t-1}^{(1)}(i)(\mathbb{X}_{t-1}^{(1)}(i))^\top\right), \qquad \boldsymbol{c}_1(i) = E\left(\mathbb{X}_{t-1}^{(1)}(i)Y_t^{(1)}(i)\right). \tag{2.3}$$

By Assumption 1 and Davydov's covariance inequality (see [Kuelbs and Philipp, 1980], Lemma 2.1) the autocovariance function $\gamma(h)$ of $\mathbb{Y}_t^{(1)}$ converges to zero as $h \to \infty$, hence $\boldsymbol{C}_1(i)$ is invertible (see [Brockwell and Davis, 1991], Proposition 5.1.1). The parameter vector $\mathbf{a}_1(i)$ is best approximating in the following sense:

$$\mathbf{a}_1(i) = \arg\min_{\mathbf{b}_1(i)} \mathrm{E}\left(\left(Y_t^{(1)}(i) - \mathbf{b}_1(i)^\top\mathbb{X}_{t-1}^{(1)}(i)\right)^2\right),$$

so that the parameters in (2.3) are equal to the true causal parameters in the correctly specified case.

Under alternatives, we need to make the following assumptions:

Assumption $\mathcal{A}$. 2. Let $\{\mathbf{Y}_t^{(2)} = (Y_t^{(2)}(1), \ldots, Y_t^{(2)}(d))^\top : t \geqslant 1\}$, be an $\mathbb{R}$-valued ergodic sequence of random vectors with existing second moments and

$$\frac{1}{n-\tilde{k}}\sum_{j=\tilde{k}+1}^{n} \mathrm{E}(\mathbb{X}_{t-1}^{(2)}(i)(\mathbb{X}_{t-1}^{(2)}(i))^\top) \to \mathbf{C}_2(i), \quad \frac{1}{n-\tilde{k}}\sum_{j=\tilde{k}+1}^{n} \mathrm{E}(\mathbb{X}_{t-1}^{(2)}(i)Y_t^{(2)}(i)) \to \mathbf{c}_2(i)$$

in case of the AMOC alternative respectively

$$\frac{1}{\tilde{k_2}-\tilde{k_1}}\sum_{j=\tilde{k}_1+1}^{\tilde{k}_2} \mathrm{E}(\mathbb{X}_{t-1}^{(2)}(i)(\mathbb{X}_{t-1}^{(2)}(i))^\top) \to \mathbf{C}_2(i), \quad \frac{1}{\tilde{k}_2-\tilde{k}_1}\sum_{j=\tilde{k}_1+1}^{\tilde{k}_2} \mathrm{E}(\mathbb{X}_{t-1}^{(2)}(i)Y_t^{(2)}(i)) \to \mathbf{c}_2(i)$$

in case of an epidemic change alternative.

This assumption obviously holds for stationary sequences, however it also allows for starting values from the stationary distribution of $\mathbf{Y}^{(1)}$. Similarly, in case of the epidemic change alternative, we can allow the process after the second change point to have starting values from the distribution of $\mathbf{Y}^{(2)}$.

2.1.4 Test Statistics for the At-Most-One-Change Situation

For the AMOC situation, we propose the following class of test statistics:

$$M_n^{(1)} = \max_{1\leqslant k\leqslant n} \frac{w^2(k/n)}{n}\mathbf{Z}_k^\top\widehat{\mathbf{H}}\mathbf{Z}_k, \tag{2.4}$$

$$M_n^{(2)} = \sum_{1\leqslant k\leqslant n} \frac{w^2(k/n)}{n^2}\mathbf{Z}_k^\top\widehat{\mathbf{H}}\mathbf{Z}_k \tag{2.5}$$

$$\mathbf{Z}_k^\top = \left(\mathbf{S}_k^\top(1), \ldots, \mathbf{S}_k^\top(d)\right), \qquad \mathbf{S}_k(i) = \sum_{t=1}^{k}\widehat{\boldsymbol{\xi}}_t(i) = \sum_{t=1}^{k}\mathbb{X}_{t-1}\widehat{e}_t(i),$$
$$\widehat{e}_t(i) = Y_t(i) - \widehat{\mathbf{a}}_n^\top(i)\mathbb{X}_{t-1}(i),$$
$$\widehat{\mathbf{a}}_n(i) = \widehat{\mathbf{C}}_n^{-1}(i)\,\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_{t-1}(i)Y_t(i), \qquad \widehat{\mathbf{C}}_n(i) = \frac{1}{n}\sum_{t=1}^{n}\mathbb{X}_{t-1}(i)\left(\mathbb{X}_{t-1}(i)\right)^\top. \tag{2.6}$$

The matrix $\widehat{\mathbf{H}}$ is $pd \times pd$ symmetric and positive semi-definite and fulfills $\widehat{\boldsymbol{H}} \xrightarrow{P} \boldsymbol{H}$ for some symmetric, positive semi-definite matrix $\boldsymbol{H}$ under H_0. Possible choices of $\boldsymbol{H}$ and $w(\cdot)$ are discussed in Section 2.1.6 below.

The above test statistics also have some power against multiple change point situations including epidemic changes which is the basic idea behind binary segmentation algorithms. The point, where the maximum is attained, is a consistent estimator for the change point in rescaled time (see Section 2.2.2 for details). In the EEG data example, we apply this estimator for each of the independent trials and plot the corresponding density estimates, which appear to be bimodal (see Figure 4.2). Therefore, we suspect that the EEG process returned to the "waiting" state after the task (moving the joystick) has been completed, which is why the following test statistics for epidemic changes are of particular interest for that data example.

2.1.5 Test Statistics for the Epidemic Situation

The following test statistics are designed to detect epidemic changes. See [Csörgő and Horváth, 1997] (Section 2.8.4) for corresponding statistics for mean changes.

$$M_n^{(3)} = \max_{\lfloor \tau_1 n \rfloor \leqslant k_1 < k_2 \leqslant n - \lfloor \tau_2 n \rfloor} \frac{1}{n} (\mathbf{Z}_{k_2} - \mathbf{Z}_{k_1})^\top \widehat{\mathbf{H}} (\mathbf{Z}_{k_2} - \mathbf{Z}_{k_1}), \tag{2.7}$$

$$M_n^{(4)} = \frac{1}{n^3} \sum_{\lfloor \tau_1 n \rfloor \leqslant k_1 < k_2 \leqslant n - \lfloor \tau_2 n \rfloor} (\mathbf{Z}_{k_2} - \mathbf{Z}_{k_1})^\top \widehat{\mathbf{H}} (\mathbf{Z}_{k_2} - \mathbf{Z}_{k_1}), \tag{2.8}$$

where $1 \leqslant \tau_1 \leqslant 1 - \tau_2 \leqslant 1$ and the rest of the notation is as in(2.6). Here, we only allow for very simple weight functions

$$w(t_1, t_2) = 1_{\{\tau_1 \leqslant t_1, t_2 \leqslant 1 - \tau_2\}}$$

because the asymptotics in the epidemic setting with more general weight functions can become more involved. Similarly, to the at most one change situation, we can estimate both change points (in rescaled time) by the points where the maximum is attained (for details we refer to Section 2.2.2).

2.1.6 Some Remarks on the Tuning Parameters H and w

To understand the influence of $\mathbf{H}$ better, we define

$$\boldsymbol{\xi}_t = (\boldsymbol{\xi}_t^\top(1), \ldots, \boldsymbol{\xi}_t^\top(d))^\top = \left((\mathbb{X}_{t-1}^{(1)}(1))^\top e_t^{(1)}(1), \ldots, (\mathbb{X}_{t-1}^{(1)}(d))^\top e_t^{(1)}(d) \right)^\top \tag{2.9}$$

and note that the asymptotic covariance of the partial sum process $\mathbf{Z}_k$ is given by the covariance of $\boldsymbol{\xi}$ in the correctly specified model and by its long-run covariance in the misspecified case (confer Theorem 2.1 for details).

Now, the choice of the weight matrix $\mathbf{H}$ allows us to look for changes in particular directions more intensely. The following choice, e.g., leads to the univariate statistics, introduced by [Hušková et al., 2007], that looks for changes only in the first channel of the observed multivariate time series:

$$\mathbf{H} = \begin{pmatrix} \operatorname{cov}^{-1}(\boldsymbol{\xi}_0(1)) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{2.10}$$

The standard choice for $\mathbf{H}$ in the correctly specified case is given by

$$\mathbf{H} = \operatorname{cov}^{-1}(\boldsymbol{\xi}_0). \tag{2.11}$$

However, if the channels of the observations are correlated the estimation of this matrix let alone its inverse is very difficult in higher dimensional settings (even for moderate dimensions if there are not that many time points available). Things become even worse if we are in the misspecified situation where the inverse of the long-run variance needs to be estimated. Furthermore, in the change point situation possible alternatives should be taken into account in the estimation of $\mathbf{H}$ in order to improve the power of the corresponding test, making things even more complicated. So, some practical solutions to this problem will be discussed in detail in Section 3.

The weight function $w(\cdot)$ essentially determines where we look closest for a change since the statistic has a higher power for changes close to k if $w(k/n)/c_w$ is larger where c_w is the critical value associated with $w(\cdot)$. While the matrix $\boldsymbol{H}$ can be used to incorporate a priori information about the direction of the expected changes, we can use the weight function $w(\cdot)$ to incorporate a priori information about the location of the change points. In the EEG data example we do not expect a change before the signal was sent, so we will choose a weight function that only looks for changes in the second half of the data set.

To derive the asymptotics for the AMOC statistic, we need to formulate the following assumptions:

Assumption $\mathcal{A}$. 3. The weight function $w : [0,1] \to [0,\infty)$ is a non-negative function, $w \not\equiv 0$. Furthermore, it has only a finite number of discontinuities $a_1, \ldots, a_K$, where it is either left or right continuous with existing limits from the other side. Additionally, the following regularity conditions hold

$$\lim_{t\to 0} t^\alpha w(t) < \infty, \quad \lim_{t\to 1}(1-t)^\alpha w(t) < \infty \qquad \text{for some } 0 \leqslant \alpha < 1/2,$$
$$\sup_{\eta \leqslant t \leqslant 1-\eta} w(t) < \infty \qquad \text{for all } 0 < \eta \leqslant \frac{1}{2}.$$

Typical symmetric weight functions are given by

$$w_1(t) = 1_{\{\epsilon < t < (1-\epsilon)\}}(t(1-t))^{-1/2},$$
$$w_2(t) = (t(1-t))^{-\beta} \quad \text{for some } 0 \leqslant \beta < 1/2.$$

2.1.7 Asymptotic Results under the Null

We are now ready to state the asymptotic distribution of the proposed test statistics under the null hypothesis.

Theorem 2.1. *Under the null hypothesis let the process* $\{\mathbf{Y}_t = \mathbf{Y}_t^{(1)}\}$ *fulfill* $\mathcal{A}$.1.*Let the weight function fulfill* $\mathcal{A}$.3 *and* $\widehat{\mathbf{H}} \xrightarrow{P} \mathbf{H}$ *(under* H_0*), then*

(a) $M_n^{(1)} \xrightarrow{\mathcal{D}} \sup_{0\leqslant t\leqslant 1} w^2(t) \sum_{j=1}^{pd} B_j^2(t)$,

(b) $M_n^{(2)} \xrightarrow{\mathcal{D}} \int_0^1 w^2(t) \sum_{j=1}^{pd} B_j^2(t)\, dt$,

where $\{\mathbf{B}(t) = (B_1(t), \ldots, B_{pd}(t))^\top : t \in [0,1]\}$ *denotes a pd-dimensional Brownian bridge with covariance matrix* $\mathbf{H}^{\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{\frac{1}{2}}$. $\mathbf{\Sigma}$ *is the long-run autocovariance matrix of* $\{\boldsymbol{\xi}_t\}$ *as in* (2.9)*, i.e.* $\mathbf{\Sigma} = \sum_{k\in\mathbb{Z}} \mathrm{E}\,\boldsymbol{\xi}_0\boldsymbol{\xi}_k^\top$.

Theorem 2.2. *Under the null hypothesis let the process* $\{\mathbf{Y}_t = \mathbf{Y}_t^{(1)}\}$ *fulfill* $\mathcal{A}$.1.*Let* $\widehat{\mathbf{H}} \xrightarrow{P} \mathbf{H}$ *(under* H_0*), then*

$$a) \quad M_n^{(3)} \xrightarrow{\mathcal{D}} \sup_{\tau_1 \leqslant t_1 < t_2 \leqslant 1-\tau_2} \sum_{j=1}^{pd} \left(B_j(t_2) - B_j(t_1)\right)^2,$$

$$b) \quad M_n^{(4)} \xrightarrow{\mathcal{D}} \int\limits_{\tau_1 \leqslant t_1 < t_2 \leqslant 1-\tau_2} \int w^2(t_1,t_2) \sum_{j=1}^{pd} (B_j(t_2) - B_j(t_1))^2 dt_1\, dt_2,$$

where $\{\mathbf{B}(t) = (B_1(t), \ldots, B_{pd}(t))^\top : t \in [0,1]\}$ *denotes a pd-dimensional Brownian bridge with covariance matrix* $\mathbf{H}^{\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{\frac{1}{2}}$. $\mathbf{\Sigma}$ *is the long-run autocovariance matrix of* $\{\boldsymbol{\xi}_t\}$ *as in* (2.9)*, i.e.* $\mathbf{\Sigma} = \sum_{k\in\mathbb{Z}} \mathrm{E}\,\boldsymbol{\xi}_0\boldsymbol{\xi}_k^\top$.

2.2 Tests and Estimators under Alternatives

In this section, we will show that the above statistics detect a large class of alternatives asymptotically. In fact, in the correctly specified case, all alternatives are detectable if a positive definite weight matrix $\mathbf{H}$ is used. In this situation the corresponding estimators are consistent.

We make the following standard assumptions on the locations of the change point:

Assumption $\mathcal{A}$**. 4.** For the at most one change situation we assume that $\tilde{k} = \lfloor \lambda n \rfloor$, $0 < \lambda < 1$, for the epidemic change alternative we assume $\tilde{k}_j = \lfloor \lambda_j n \rfloor$, $0 < \lambda_j < 1$, $j = 1, 2$.

In order to classify detectable changes under misspecification we need to understand the behavior of the estimator $\widehat{\mathbf{a}}_n$ as in (2.6) under the alternative. In this situation it can no longer be expected to converge to the true or best approximating value before the change but it will be contaminated by the alternative. In fact, Lemma 6.1 shows that $\widehat{\mathbf{a}}_n(i)$ converges to

$$\begin{aligned}
&\widetilde{\mathbf{a}}(i) = \mathbf{Q}^{-1}(i)\mathbf{q}(i), \quad i = 1, \ldots, d,\\
&\text{where} \quad \mathbf{Q}(i) = \tilde{\lambda}\mathbf{C}_1(i) + (1-\tilde{\lambda})\mathbf{C}_2(i), \quad \mathbf{q}(i) = \tilde{\lambda}\boldsymbol{c}_1(i) + (1-\tilde{\lambda})\boldsymbol{c}_2(i),\\
&\qquad \tilde{\lambda} = \begin{cases} \lambda, & \text{at most one change,} \\ 1 - (\lambda_2 - \lambda_1), & \text{epidemic change.} \end{cases}
\end{aligned} \tag{2.12}$$

The matrix $\mathbf{Q}(i)$ is the weighted average of the autocovariances of the time series, before and after the change, with weights proportional to the duration of time the series remains in that state. Here, $\tilde{\lambda}$ is the duration time in state 1. We now make the additional assumption:

Assumption $\mathcal{A}$**. 5.** Let $\mathbf{Q}(i)$, $i = 1, \ldots, d$, be regular matrices.

This assumption guarantees that $\widetilde{\mathbf{a}}(i)$, $i = 1, \ldots, d$, above are well defined and the unique minimizer of

$$(1-\tilde{\lambda})\,\mathrm{E}(Y_1^{(1)}(i) - \boldsymbol{a}^\top(i)\mathbb{X}_1^{(1)}(i))^2 + \tilde{\lambda}\,\mathrm{E}(Y_1^{(2)}(i) - \boldsymbol{a}^\top(i)\mathbb{X}_1^{(2)}(i))^2.$$

2.2.1 Asymptotic Power One under Alternatives

Theorem 2.3. Let Assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, $\mathcal{A}.3$, as well as $\mathcal{A}.5$ hold true. Furthermore, assume that there exists $\epsilon > 0$ such that

$$P(\Delta^2_{\widehat{H}} \geqslant \epsilon) \to 1, \tag{2.13}$$

where $\Delta^2_{\widehat{H}} = (\boldsymbol{c}_1 - \mathrm{diag}(\boldsymbol{C}_1(1), \ldots, \boldsymbol{C}_1(d))\widetilde{\boldsymbol{a}})^\top \widehat{\boldsymbol{H}}^{-1} (\boldsymbol{c}_1 - \mathrm{diag}(\boldsymbol{C}_1(1), \ldots, \boldsymbol{C}_1(d))\widetilde{\boldsymbol{a}})$.

a) Under the at most one change alternative, it holds

$$\text{(i)} \quad M_n^{(1)} \xrightarrow{P} \infty, \qquad \text{(ii)} \quad M_n^{(2)} \xrightarrow{P} \infty.$$

b) Under the epidemic change alternative, it holds

$$\text{(i)} \quad M_n^{(3)} \xrightarrow{P} \infty, \qquad \text{(ii)} \quad M_n^{(4)} \xrightarrow{P} \infty.$$

Remark 2.1. Condition (2.13) is the heart of the theorem explaining which changes are detectable. If $\widehat{\boldsymbol{H}} \xrightarrow{P} \mathbf{H}_A$ for a positive definite matrix, then all changes are detectable for which $\mathbf{C}_1^{-1}(i)\boldsymbol{c}_1(i) \neq \mathbf{C}_2^{-1}(i)\boldsymbol{c}_2(i)$ for at least one of the channels $i = 1, \ldots, d$. Hence, asymptotically all changes in the correctly specified model are detectable. If $\mathbf{H}_A$ has a block diagonal form with block lengths of p such that the i-th block matrix is positive definite, then we will detect all changes that appear in channel i, i.e. for which $\mathbf{C}_1^{-1}(i)\boldsymbol{c}_1(i) \neq \mathbf{C}_2^{-1}(i)\boldsymbol{c}_2(i)$.

2.2.2 Estimators for the Change Point(s) under Alternatives

If the test statistics are consistent we also obtain consistent estimators as follows: Under the AMOC model, we define the change point estimator as

$$\hat{k}_n := \operatorname*{arg\,max}_{1<k<n} w^2(k/n)\mathbf{S}_k^T \widehat{\mathbf{H}} \mathbf{S}_k \tag{2.14}$$

and for the epidemic model we define

$$\left(\hat{k}_{1,n}, \hat{k}_{2,n}\right) := \operatorname*{arg\,max}_{\tau_1 \leqslant k_1 < k_2 \leqslant \tau_2} (\mathbf{S}_{k_2} - \mathbf{S}_{k_1})^\top \widehat{\mathbf{H}} (\mathbf{S}_{k_2} - \mathbf{S}_{k_1}). \tag{2.15}$$

Theorem 2.4. Let the assumptions of Theorem 2.3 be fulfilled in addition to $\widehat{\boldsymbol{H}} \to \boldsymbol{H}_A$, where $\boldsymbol{H}_A$ can be different from $\boldsymbol{H}$.

(a) Under the AMOC change alternative assume additionally that $\sup_{0 \leqslant t \leqslant 1} w(t) < \infty$ and let the function

$$h(t) = w(t) \begin{cases} t(1-\lambda), & t \leqslant \lambda, \\ (1-t)\lambda, & t \geqslant \lambda \end{cases}$$

has a unique supremum at λ, which holds e.g. for the weight functions $w(t) = (t(1-t))^{-\beta}$, $0 \leqslant \beta < 1/2$. Then

$$\frac{\hat{k}_n}{n} \xrightarrow{P} \lambda.$$

(b) Under the epidemic alternative assume additionally that $\tau_1 < \lambda_1 < \lambda_2 < \tau_2$, then

$$\left(\frac{\hat{k}_{1,n}}{n}, \frac{\hat{k}_{2,n}}{n}\right) \xrightarrow{P} (\lambda_1, \lambda_2)\,.$$

Remark 2.2. If we want to use more general weight functions as given in Assumption $\mathcal{A}.3$ for the at most one change alternative then we need to make the additional assumption that the time series after the change $\{Y_t^{(2)}(l), l = 1, \ldots, d\}$ is also strong mixing with a rate as given in Assumption $\mathcal{A}.1$.

3 Robustifying the Test Statistics and Empirical Study

In order to use asymptotic critical values in the above test statistics we need to choose $\boldsymbol{H}$ (resp. $\widehat{\boldsymbol{H}}$) in such a way that the limit becomes pivotal. The standard solution is to choose $\boldsymbol{H} = \boldsymbol{\Sigma}^{-1}$ as in (2.11) where $\boldsymbol{\Sigma}$ is the long-run covariance matrix of $\{\boldsymbol{\xi}_t\}$ as in (2.9). For most applications the true underlying long-run variance (under H_0) will not be known so it needs to be estimated. However, this raises several problems:

$\mathcal{P}.1$ When designing the estimation procedure, we do not only need it to be consistent under the null hypothesis, but also to behave reasonably under alternatives e.g. converging to some contaminated limit matrix $\mathbf{H}_A$. In fact, the small sample power behavior will depend crucially on this choice.

$\mathcal{P}.2$ Even in the univariate situation it is a much harder statistical problem to estimate the long-run variance than the variance. Consequently, the estimation error for the long-run variance is much bigger than for the variance (for fixed sample size).

$\mathcal{P}.3$ Increasing the number of entries of $\boldsymbol{H}$ that need to be estimated leads to more imprecise estimates.

Typically, we need to estimate some statistical quantity such as the covariance or long-run covariance matrix and then invert the estimated matrix to obtain $\widehat{\boldsymbol{H}}$ which leads to even less precise estimates and additional numerical errors. As a consequence, the size can be very distorted (see Figure 3.1 below). A block diagonal structure and many zeros in $\mathbf{H}$ can help overcome these estimation issues and stabilize the empirical size.

3.1 Robustifying the Univariate Statistic

3.1.1 Stabilizing the Power

The univariate situation is not only included by $d = 1$ but also by choosing $\mathbb{X}_{t-1}(1) = (Y_{t-1}(1), \ldots, Y_{t-p}(1))^\top$ and letting

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

In this section, we choose the second approach because it makes generalizations to the multivariate situation in the next section more transparent.

Ideally, we choose $\mathbf{H}(1)$ in such a way that (a.) the limit becomes pivotal under the null hypothesis and (b.) all parameter changes in the correctly specified model respectively all changes leading to different

best approximating parameters under misspecification should be asymptotically detectable. We achieve both goals by choosing the inverse of the covariance matrix of $\{\boldsymbol{\xi}_t(1)\}$ in the correctly specified and of the long-run covariance matrix of $\{\boldsymbol{\xi}_t(1)\}$ in the misspecified case. However, $\{\boldsymbol{\xi}_t(1)\}$ is not observable, hence we need to use estimated versions which because of $\mathcal{P}.1$ should be chosen in such a way that they take alternatives into account. Under the at most one change alternative typically the following approach is taken (see, e.g., [Hušková et al., 2007]): Use a preliminary estimator $\widetilde{\boldsymbol{H}}$ for $\boldsymbol{H}$ which does not take the alternative into account. Estimate the change point $\widehat{k}(1)$ as

$$\widehat{k}(1) = \arg\max \frac{w^2(k/n)}{n} \mathbf{Z}_k^\top \widetilde{\mathbf{H}} \mathbf{Z}_k,$$

then estimate the residuals before and after the change by

$$\widehat{\xi}_t^{(H1)}(1) = \mathbb{X}_{t-1}(1)\widehat{e}_t(1), \quad t = 1, \ldots, n$$
$$\text{where } \widehat{e}_t(1) = \begin{cases} Y_t(1) - \mathbb{X}_{t-1}(1)^\top \widehat{\mathbf{a}}_{\widehat{k}(1)}(1), & t \leqslant \widehat{k}(1), \\ Y_t(1) - \mathbb{X}_{t-1}(1)^\top \widehat{\mathbf{a}}^\circ_{\widehat{k}(1)}(1), & t > \widehat{k}(1). \end{cases}$$

$\widehat{\mathbf{a}}_{\widehat{k}(1)}(1)$ is the least squares estimator for $\widehat{\mathbf{a}}^{(1)}(1)$ based on $Y_1(1), \ldots, Y_{\widehat{k}(1)}(1)$ while $\mathbf{a}^\circ_{\widehat{k}(1)}(1)$ is the least squares estimator for $\mathbf{a}^{(2)}(1)$ based on $Y_{\widehat{k}(1)+1}(1), \ldots, Y_n(1)$. For the epidemic change alternative an analogous version can be used. This approach works quite well, if the alternative is correctly specified and the residuals before and after the change have similar statistical properties. Otherwise, e.g. in the presence of multiple changes, a systematic estimation error is introduced. While this systematic error is typically asymptotically negligible as long as the limit is sufficiently nice (see Section 2.2), it usually leads to some power loss in small samples.

3.1.2 Stabilizing the Power for the EEG Data

By design, we do not expect pre-stimulus changes in the EEG time series (i.e., during the first half of each trial). Following stimulus presentation, there may be gradual changes or multiple changes. First, there could be brief changes in the EEG due to the visual stimulus followed by changes that accompany brain processing of the visual information. Consequently, the approach discussed in the previous section does not seem ideal here. However, because the data set is designed in such a way that the first half is stationary, we can and will only use the data set up to time point 250 in the estimation thus automatically taking possible change points after the visual signal into account. Precisely, we will use

$$\widehat{\xi}_t^{(250)}(1) = \mathbb{X}_{t-1}(1)\widehat{e}_t(1), \quad t = 1, \ldots, 250, \qquad \text{where } \widehat{e}_t(1) = Y_t(1) - \mathbb{X}_{t-1}(1)^\top \widehat{\mathbf{a}}_{250}(1),$$

and $\widehat{\mathbf{a}}_{250}(1)$ is the least squares estimator for $\mathbf{a}^{(1)}(1)$ based only on $Y_1(1), \ldots, Y_{250}(1)$.

3.1.3 Stabilizing the Size with Respect to Possible Misspecification

Based on the estimated residuals from Sections 3.1.1 or 3.1.2 the inverse of the long-run covariance matrix can be estimated e.g., using flat-top kernel estimators with an automatic bandwidth selection criteria (as proposed in [Politis, 2003]). However, even in the one-dimensional case where only the long-run variance

needs to be estimated the estimation error can become quite large. Therefore, we propose to trade this possibly large estimation error for a hopefully smaller model error. To this end, we propose to increase the length p of the fitted AR-model which can approximate a large class of time series quite well so that the corresponding errors behave more like i.i.d. time series even in the misspecified case. This allows us to use an estimator for the covariance matrix rather than the long-run covariance matrix.

A possible drawback is that this approach now increases the unknown number of parameters of the weight matrix $\boldsymbol{H}$ associated with problem $\mathcal{P}.3$ above. Additionally, we now need to look for changes in a larger number of parameters which in combination with the choice $\widehat{\boldsymbol{H}}(1) = \operatorname{cov} \boldsymbol{\xi}_1(1)$ leads to a power loss if this change is actually already quite visible in the smaller model. This is why, we propose to only look for particular changes by choosing

$$\boldsymbol{H}(1) = \begin{pmatrix} \operatorname{cov}\left[e_t^{(1)}(1)\,(Y_{t-1}^{(1)}(1), \ldots, Y_{t-p'}^{(1)}(1))^\top\right] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{3.1}$$

for some $p' < p$. This choice also leads to a pivotal limit.

3.1.4 Investigating the Performance of the Robustified Univariate Procedures

To study the effect of these robustifications, we generate data from the following AR(6) time series (with i.i.d. standard normally distributed errors and a change at $t = 330$):

- *Case AR1.* $\mathbf{a}_1(1) = (-0.1, 0.1, 0, 0, -0.1, -0.1)^\top$, $\mathbf{a}_2(1) = (-0.4, 0.1, 0, 0, -0.1, -0.1)^\top$, i.e. the change occurs only in the first parameter. Moreover, the last four channels of $\mathbf{a}_1(1)$ are small. Thus, by fitting an AR(2) model to the data, our procedure only suffers from mild model misspecification.

- *Case AR2.* $\mathbf{a}_1(1) = (-0.1, 0.1, 0, 0.2, -0.3, -0.2)^\top$, $\mathbf{a}_2(1) = (-0.2, 0.3, 0, 0.2, -0.3, -0.2)^\top$, i.e. the change occurs only in the first two parameters but there is stronger misspecification when using the AR(2) model.

- *Case AR3.* $\mathbf{a}_1(1) = (-0.1, 0.1, 0, 0, -0.3, -0.2)^\top$, $\mathbf{a}_2(1) = (-0.1, 0.2, 0, 0.4, -0.1, -0.2)^\top$, i.e. there is only a small change in the second parameter in addition to changes in the fourth and fifth parameter again under a stronger misspecification when using the AR(2) model.

In all cases, we used the following weight function adapted to the EEG data example $w(t) = \mathcal{I}_{[\frac{250}{512}, \frac{490}{512}]}(t)(1-t)^{-0.25}$, $t \in [0, 1]$.

Figure 3.1 gives the results for several choices of p and p' and the maximum type statistic $M_{512}^{(1)}$ as in (2.4) as well as the sum type statistic $M_{512}^{(2)}$ as in (2.5) with $\boldsymbol{H}$ as described in Sections 3.1.2 and 3.1.3. The first row shows the empirical size on the y-axis for the nominal one on the x-axis, the second row shows the size-corrected power. For $p = p' = 2$ we additionally use both the covariance estimator based on the first 250 observations ($C250$) as well as the flat-top kernel estimator in [Politis, 2003] with automatic bandwidth selection choice ($S250$). Clearly, the size behavior is best in all three cases for $p = 6$, $p' = 2$ as we expected and this is the only choice with an acceptable size behavior. For $p = 2$, $p' = 2$ the size behavior is not acceptable due to a too large an estimation error (when estimating the long-run covariance matrix) and too large a model error (when estimating the covariance matrix) and is worse the stronger

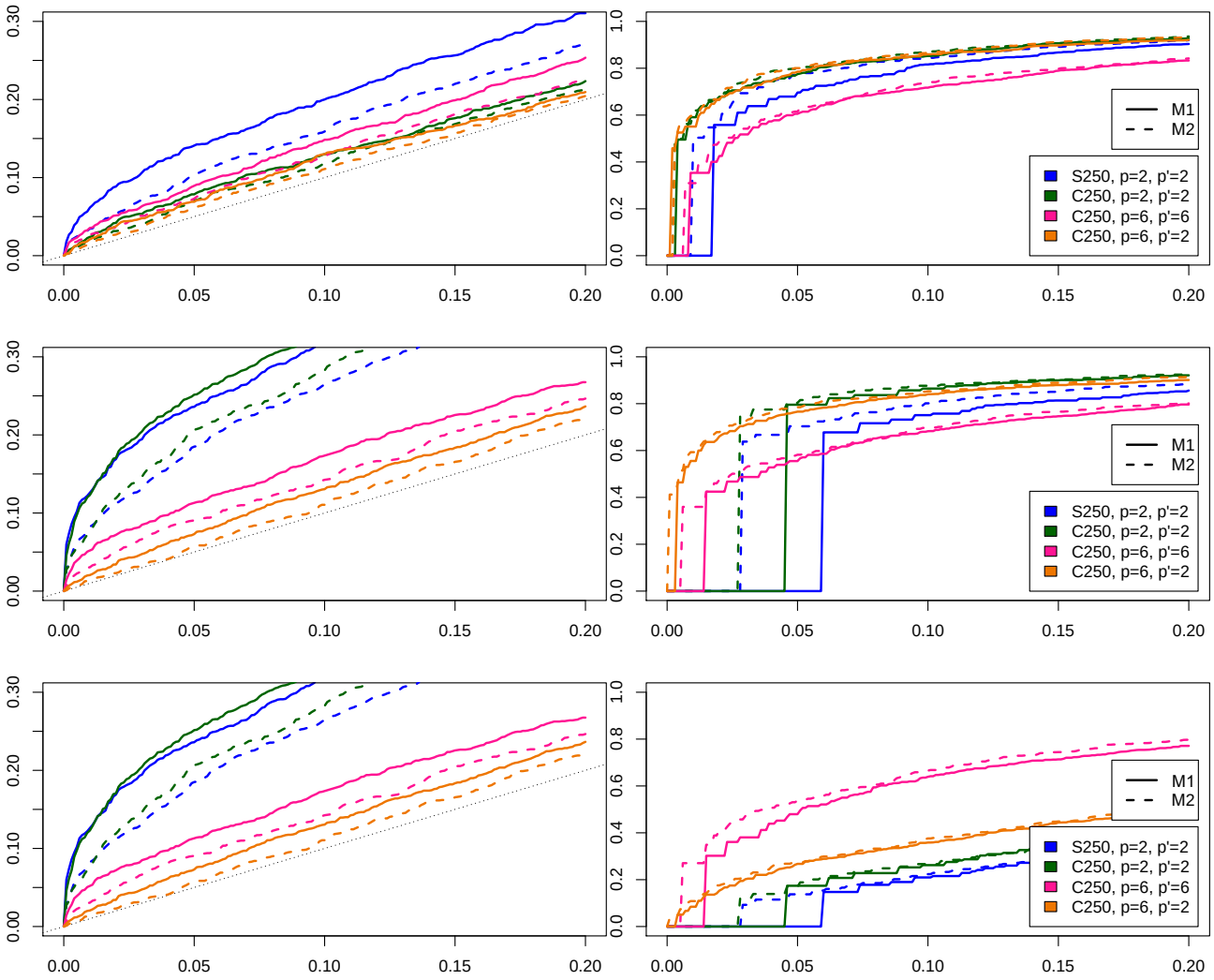

Figure 3.1: Empirical size and size-corrected power for AR1 – AR3 (top to bottom). $C250$ uses the covariance estimate; $S250$ uses the long run covariance estimate based on the first 250 observations.

the misspecification. For $p = 6$, $p' = 6$ there is no model error but nevertheless the estimation error due to $\mathcal{P}.3$ makes the tests far too liberal (although not as badly so as for the choice $p = p' = 2$). Using the sum-type statistic is in all cases more robust and gives a better size while not losing actual power. In the first two cases, where the main change occurs in the first two parameters, the power for the choice $p = 6$, $p' = 2$ is best, even better than for $p = 2$, $p' = 2$, while also being good for this second choice. As expected we do lose some power in this case for $p = 6$, $p' = 6$. However, if the change mainly occurs in the last four parameters as in the third case above, the power for $p = 6$, $p' = 6$ is better than for the other cases while they still have some power against this alternative. The bootstrap procedure of Section 3.2.1 will further improve the size.

3.2 Robustifying the Multivariate Statistics

In general, if a change is present in (almost) all channels, one obtains higher power when the information is pooled and the multivariate statistic is used. The statistic at hand is a quadratic form so that the

changes in each channel add up quadratically, while the variance of the noise (if reasonably uncorrelated) adds up only linearly, leading to a better signal-to-noise ratio of the multivariate statistic. On the other hand, if there is no change in many channels then we merely add noise and thus decreasing the power. If additional information about the direction of the change is available appropriate projections (which are included in the above theory by an appropriate choice of $\boldsymbol{H}$) can further help increase the power. A detailed analysis of this effect in the simpler mean change model can be found in [Aston and Kirch, 2013].

3.2.1 Size Correction: A Bootstrap Approach

In the multivariate procedure, the estimation problem becomes even worse because the number of entries of $\boldsymbol{H}$ that need to be estimated increases polynomially with the number of dimensions. Since we already have massive estimation problems in the univariate case, estimating the inverse of the covariance or even long-run covariance matrix of $\{\boldsymbol{\xi}_t\}$ is statistically infeasible. To overcome this problem, we first choose to relate $Y_t(i)$ only to its own past as in Example 2.1 and secondly to choose

$$\boldsymbol{H} = \operatorname{diag}(\boldsymbol{H}(1), \boldsymbol{H}(2), \ldots, \boldsymbol{H}(d)) \tag{3.2}$$

with $\boldsymbol{H}(j)$ analogously to (3.1). In case the channels were independent this would lead to a pivotal limit. However, we do not believe that the channels are independent, hence the limit distribution depends on that underlying covariance structure which we cannot estimate in a stable way. As a solution, we propose to use the following regression bootstrap (which is related to the one investigated in [Horváth, L. et al., 1999]). We also implemented a version of the pair bootstrap that was discussed in that paper, but the regression bootstrap was always superior in simulations. The bootstrap works as follows:

(1) Calculate the residuals $\hat{e}_1(j), \ldots, \hat{e}_{250}(j)$, $j = 1, \ldots, 250$:

$$\hat{e}_t(j) \;=\; Y_{t-1}(j) - \mathbb{X}_{t-1}^T(j)\widehat{\mathbf{a}}_{250}(j) - \frac{1}{250}\sum_{i=1}^{250}\left(Y_i(j) - \mathbb{X}_{i-1}^T(j)\widehat{\mathbf{a}}_{250}(j)\right), \qquad 1 \leqslant t \leqslant 250,$$

where $\widehat{\mathbf{a}}_{250}(j)$ is the least squares estimator based on $Y_1(j), \ldots, Y_{250}(j)$.

(2) Draw n i.i.d. random variables $U_1, \ldots, U_n$ such that $P(U_1 = i) = 1/250$, $i = 1, \ldots, 250$.

(3) Let for $k = 1, \ldots, n$, $j = 1, \ldots, d$

$$e_k^*(j) \;:=\; \hat{e}_{U_k}(j)$$

and $\boldsymbol{\xi}_t^* := \left(\mathbb{X}_t^T(1)e_t^*(1), \ldots, \mathbb{X}_t^T(d)e_t^*(d)\right)^T$, $t = 1, \ldots, n$.

(4) Calculate the multivariate statistics from the previous sections, but with $\mathbf{Z}_k$ replaced by $\mathbf{Z}_k^*$ and $\widehat{\mathbf{H}}$ as in (3.2) by $\boldsymbol{H}^*$ where

$$\mathbf{Z}_k^* = \sum_{i=1}^{k}\boldsymbol{\xi}_i^* - \widehat{\mathbf{C}}_k\widehat{\mathbf{C}}_n^{-1}\sum_{i=1}^{n}\boldsymbol{\xi}_i^*,$$

$$\widehat{\mathbf{C}}_k = \operatorname{diag}(\widehat{\mathbf{C}}_k(1), \ldots, \widehat{\mathbf{C}}_k(d)), \quad \widehat{\mathbf{C}}_k(i) = \frac{1}{k}\sum_{t=1}^{k}\mathbb{X}_{t-1}(i)\mathbb{X}_{t-1}^T(i)$$

and $\boldsymbol{H}_n^*$ is a block diagonal matrix with $j = 1, \ldots, d$ blocks $(\boldsymbol{H}_n^*(j))^{-1}$, where

$$\boldsymbol{H}_n^*(j) = \begin{pmatrix} \frac{1}{250}\sum_{l=1}^{250}\left(e_l^*(j) - \frac{1}{250}\sum_{s=1}^{250} e_s^*(j)\right)^2 \frac{1}{250}\sum_{i=1}^{250}\tilde{\mathbb{X}}_{i-1}(j)\tilde{\mathbb{X}}_{i-1}^T(j) & 0 \\ 0 & 0 \end{pmatrix},$$

$$\tilde{\mathbb{X}}_{i-1}(j) = (X_{i-1}(j), \ldots, X_{i-p'}(j))^T.$$

(5) Repeat steps (2)-(4) M times ($M = 2000$).

(6) The critical value $c^*(\alpha)$ is obtained as the upper α-quantile of the M statistics.

(7) Reject H_0 if the corresponding statistic based on the original sample exceeds the critical value $c^*(\alpha)$.

3.2.2 Investigating the Performance of the Robustified Multivariate Procedure

To investigate the influence of dependency between panels on the asymptotic procedure and the performance of the proposed bootstrap method, we consider channelwise AR-models, where there is some correlation between the errors among channels. In each channel the following models are used: $\boldsymbol{a}_1 = (0.5, -0.2, 0.1, 0, 0, 0.2)^\top$ before the change and $\boldsymbol{a}_2 = (0.5, -0.3, 0.1, 0, 0, 0.2)^\top$ after the change. The multivariate errors are normal with mean zero and variance 1 and given as follows:

E.1 The errors are uncorrelated.

E.2 There is a correlation of 0.2 between any two channels.

E.3 The correlation is 0.6 between any two channels.

E.4 There is a correlation of 0.1 between any two channels except for two pairs of channels where the correlation is 0.5.

E.5 We additionally have three pairs with a correlation of 0.3.

The empirical results can be found in Figure 3.2. In all cases the bootstrap performs very well in terms of size and only slightly inferior in terms of power showing that it is clearly robust against deviations from independence between channels. The size of the asymptotic tests on the other hand are not so good and in particularly strongly distorted for case 3 where there is a strong correlation between all panels.

3.3 Sensitivity Study under Alternatives

In this section, we study the sensitivity of the procedures with respect to the location of the change, the epidemic duration length, the size of changes and the number of channels affected by change points. With a view to the EEG data we choose AR(8) time series of length 512 with various epidemic changes, i.e. the univariate time series are given by

$$Y_t(1) = \mathbf{a}_1(1)^\top (Y_{t-1}(1), \ldots, Y_{t-1}(8)) + 1_{\{t_1 \leqslant t < t_2\}} \boldsymbol{d}(1)^\top (Y_{t-1}(1), \ldots, Y_{t-1}(8)) + e_t, \tag{3.3}$$

where $\{e_t\}$ are i.i.d. standard normally distributed and model parameters $\boldsymbol{a}_1 = (0.5, 0, 0.1, 0, 0, 0.2, 0.1, -0.2)^\top$, $\boldsymbol{a}_1 + \boldsymbol{d}(1) = (0.5, a_{22}, 0.1, 0, 0, 0.2, 0.1, -0.2)^\top$. We apply the epidemic change statistics $M_{512}^{(3)}$ and $M_{512}^{(4)}$ as

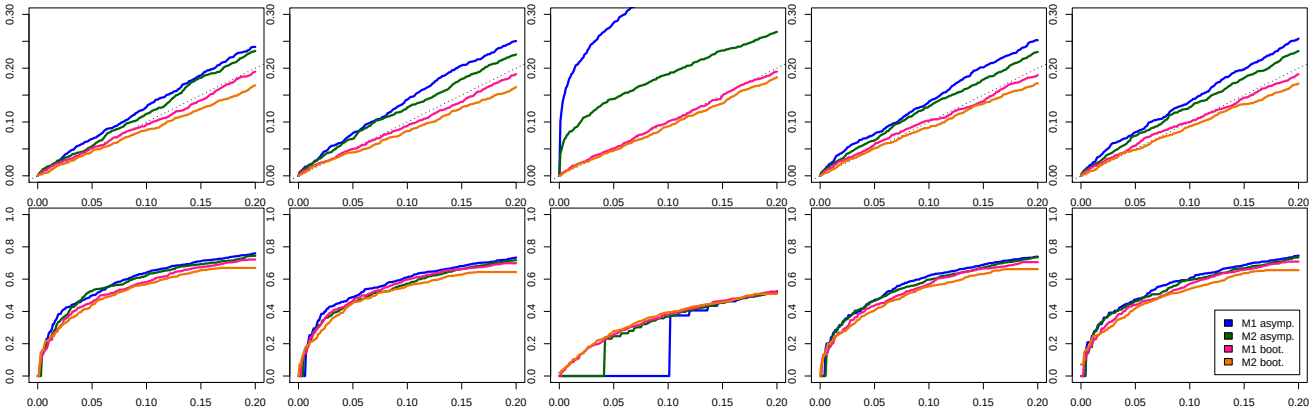

Figure 3.2: Empirical size (top) and size-corrected power (bottom) with dependent errors according to E.1 – E.5 (left to right)

in (2.7) and (2.8) with weight function $1_{\{250/512 \leqslant t \leqslant 490/512\}}$ (as will be used in the EEG data due to the prior knowledge that there should not be a change before point 250 or after 490). The weight matrix $\boldsymbol{H}$ will be chosen according to (3.2) with $p = 6$, $p' = 2$ (allowing for some misspecification). We do not apply the bootstrap due to computational time but report the size corrected power to take size variations into account.

The empirical size corrected power to the 5% level for a change in a single parameter based on 2000 repetitions is reported in Table 3.1 Obviously, the larger the change and the longer the epidemic duration, the better the power, where changes to a negative parameter are apparently more difficult to detect. Nevertheless, the power is quite good even for small epidemic durations (30 out of 512 observations are in the epidemic state) and can further be increased by using the full multivariate information.

To demonstrate the power behavior under the multivariate setting, we simulate i.i.d. channels as in (3.3) with $a_{22} = -0.4$. We allow the change locations to slightly vary across channels by independently drawing the change points for each channel. To this end we choose uniform distributions centered around 300 and 350 respectively with several ranges ± 0, ± 5 and ± 10. Table 3.2 reports the empirical size corrected power for the asymptotic 5% level. As expected, Table 3.2 shows clearly that the power increases if the multivariate statistics are used and changes are present in several channels. Furthermore, it shows that the power decreases if only noise is added. Interestingly, the power does not seem to be affected by a slight variation of change locations across channels, a situation which one would expect in many applications.

4 Data Analysis

We analyze data consisting of electroencehpalograms (EEG) recorded from a single participant in a visual-motor task experiment. Our neuroscientist collaborator is broadly interested in studying brain dynamics during sensory information and execution of the motor task. In this study, we tackle two primary statistical goals. The first is to identify the change points in brain signals *within a trial*. The second is to study how the change point structure varies across *many trials* across the experiment. The neuroscience community is currently highly interested in characterizing variations in brain responses. To the best of our knowledge, our procedure is the first to systematically investigate between-trial variation

5%	$a_{22}=0.6$	$a_{22}=0.4$	$a_{22}=0.2$	$a_{22}=-0.2$	$a_{22}=-0.4$	$a_{22}=-0.6$
$t_1=300$, $t_2=330$	0.973	0.7350	0.152	0.055	0.1465	0.3875
	0.970	0.7285	0.148	0.067	0.1590	0.4325
$t_1=300$, $t_2=350$	0.9995	0.9245	0.257	0.0725	0.292	0.7245
	0.9990	0.9235	0.267	0.0910	0.338	0.7655
$t_1=300$, $t_2=400$	1	0.9980	0.515	0.1635	0.7130	0.9875
	1	0.9985	0.557	0.2070	0.7905	0.9935

Table 3.1: Size-corrected power for 5% level and univariate versions of the statistics $M_{512}^{(3)}$ (upper value) and $M_{512}^{(4)}$ (lower value).

in change points.

4.1 Data Description

The EEG data are electric potentials recorded from the scalp and thus reflect indirect measures of brain electrical activity. In this experiment, the participant was instructed to move the hand-held joystick either towards the left of center or towards the right of center. There were $N_1 = 118$ leftward movement trials and $N_2 = 134$ rightward movement trials. The order of presentation of the left and right conditions was random.

The EEG was recorded at the rate of 512 Hz and band-pass filtered at $(0.02, 100)$ Hz. Each of the $N_1+N_2 = 252$ trials consists of one second (total of $n = 512$ time points) whose midpoint is approximately the point in time when the stimulus was presented (at $t = 250$). Thus, each trial consists of half a second of EEG recording before stimulus presentation and another half a second post stimulus presentation. From the montage of 64 scalp electrodes, our neuroscientist collaborator selected a sub-set of 12 surface leads at: FC3, FC5, C3, P3, and O1 over the left hemisphere; FC4, FC6, C4, P4 and O2 over the right hemisphere; and Cz and Oz over the mid-line. The frontal (FC) leads were presumably placed over the prefrontal cortex, regions previously shown to have involvement in premotor processing. The central (C) leads were placed over structures involved in motor performance, while the parietal (P) and occipital

5%	3/4	2/4	1/4	6/8	4/8	2/8	9/12	6/12	3/12
±0	0.49	0.31	0.15	0.89	0.74	0.22	0.88	0.61	0.27
	0.59	0.41	0.20	0.94	0.83	0.32	0.94	0.78	0.42
±5	0.49	0.32	0.15	0.88	0.73	0.22	0.88	0.61	0.28
	0.59	0.41	0.20	0.94	0.83	0.32	0.94	0.78	0.41
±10	0.48	0.33	0.16	0.87	0.71	0.23	0.86	0.61	0.28
	0.59	0.40	0.20	0.93	0.82	0.31	0.93	0.75	0.39

Table 3.2: Size-corrected power for 5% level and statistics $M_{512}^{(3)}$ (upper value) and $M_{512}^{(4)}$ (lower value) when x out of y (x/y) channels did contain epidemic change points which were located independently for each channel uniformly around 350 and 380 respectively according to $\pm z$.

(O) leads were placed over structures involved in visual sensation and visual motor transformations (see [Marconi et al., 2001]).

4.2 Data Analysis

To determine optimal order of the VAR for this EEG data, we applied both the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for each trial using only the pre-stimulus period because this period, consisting of the first 250 time points, is approximately stationary. Both AIC and BIC produced optimal order d that ranged from 9 to 17 across the many trials. Using the lowest order of $d = 9$ already produced residuals whose auto-correlation and partial auto-correlation plots indicated white noise structure. Thus, we selected the order $d^* = 9$ in the succeeding steps of our analysis.

Since no change was suspected during the pre-stimulus period (before $t = 250$), we used a non symmetrical weight function and thus restricted the change point search interval to $[250, 490]$. The specific weight functions gives somewhat more weight to late (post-stimulus) changes

$$w(t) \quad = \quad \mathcal{I}_{\left[\frac{250}{512}, \frac{490}{512}\right]}(t)(1-t)^{-0.25}, \qquad t \in [0,1].$$

Because we are in a multiple testing situation, where we can assume independence between trials, we apply a false discovery rate correction as proposed by [Benjamini and Hochberg, 1995] at level 5% for all tests below.

4.3 Results

As a first step, we applied the at-most-one-change model to this EEG data. In Table 4.1, we note the following results. The percentage of trials that suggest at least one change is over 70% which is a strong evidence that the brain process is highly dynamic even within a one second interval (which is the length of one trial). As one might expect, there appears to be no significant difference in this percentage between the left and right conditions. Compared to the test statistic based on the average, the test statistic based on the maximum produced a higher percentage of trials with change points. This difference is consistent for both asymptotic and bootstrap-based inference.

We then plotted the estimated density of the locations of the change points for both left and right conditions. We note in Figure 4.2 The primary peak occurs at around $t = 330$ for both the left and right conditions. The secondary peak, which is of smaller magnitude, occurs at roughly $t = 450$. The bimodal feature could simply be due to random variation in the change points or could suggest the presence of a second change point in the brain process within a trial. We investigated this idea further and plotted the densities at each of the 12 channels. We do this primarily to find out which channels best display changes and if all channels do contain changes, we want to identify those changes. The densities obtained using the $M1$-boot statistic in Figure 4.2. We observe that the estimated density plots vary across channels - some channels are more responsive to the stimulus than others. Moreover, some channels indicate the presence of two change points thus suggesting three regimes in the brain dynamics within a trial: the first being the post-stimulus regime, the second is the time period somewhere between instruction and motor response, the third is the post-motor response regime which reverts back to the first pre-stimulus regime.

Motivated by the suggestion of a second change point within a trial, we applied our proposed multivariate epidemic change point test and report the results in Table 4.2. We note that this test also

rejects and even more often than AMOC statistic which is again evidence that, for this EEG dataset, the epidemic change model is closer to reality than the AMOC. Digging further into the behavior at the individual channels, the plots in Figure 4.3 suggest that all channels share similar estimated densities of change points under the epidemic model: the first change point is concentrated around $t = 330$ and the second at around $t = 370$. Moreover, a visual inspection indicates that the first change point for the left condition occurs before that of the right condition in some channels. Of course, this needs to be confirmed via formal statistical testing. There are also a number of studies on handedness and attention. For this right-handed subject, leftward movements might be more novel and more attended than rightward movements and this enhanced directional attention could have slightly influenced the reactivity speed. Our collaborator considers this finding to be very interesting and certainly warrants deeper neuroscientific investigations.

To further study the distribution of the change points under the epidemic model, we produced the contour density plots in Figure 4.4 and noted that while some channels are able to detect changes in the rightward movement better than others, there are also other channels that are more sensitive to changes in the leftward movement. We also observe that the concentration of the first change point occurs at around $t = 330$ and the second occurs at around 40 time points following the first, i.e., $t = 370$. Moreover, for both left and right conditions, a majority of the second change points occur within 60 time points following the first. Moreover, the plots suggest that the variation for the left conditions is slightly higher than that for the right condition. Again, this will need to be formally tested before one can make any strong claims. One could argue that, for right-handed individuals, leftward movements are considered to be more novel than rightward movements which then leads to a greater variation in the brain responses.

4.4 Summary

This analysis demonstrates how our proposed procedures could be helpful in studying the complex brain dynamics during a cognitive motor task. The results were quite interesting for our collaborator because it raises more questions and thus motivates further experimentation. These results are also potentially highly relevant to brain computer interface (BCI) where non-invasive signals such as EEGs serve as inputs to a processor that controls the movement of a robotic arm. When successfully implemented, this technology could eventually allow people with severe disabilities (such as stroke victims and war veterans with limited movement) to control robots that can help them in daily living activities. Our work here provides a small contribution towards understanding the dynamics of brain process during a motor task. This inter-disciplinary collaboration is a demonstration of the importance of statistical tools in generating more scientific hypotheses and thus pushing forward the boundaries of science.

	$M^{(1)}$asymp.	$M^{(2)}$asymp.	$M^{(1)}$boot.	$M^{(2)}$boot.
Left	0.94	0.89	0.83	0.76
Right	0.95	0.87	0.79	0.72

Table 4.1: Relative number of estimated change points under the multivariate AMOC model.

	$M^{(3)}$asymp.	$M^{(4)}$asymp.	$M^{(3)}$boot.	$M^{(4)}$boot.
Left	0.99	0.99	0.89	0.91
Right	0.97	0.95	0.92	0.91

Table 4.2: Relative number of estimated change points under the multivariate epidemic model.

5 Conclusions

Motivated by the current demands within the neuroscientific community to have a deeper understanding of the changing brain dynamics during a cognitive task, we rigorously developed novel statistical methods for detecting change points in multivariate time series. Our proposed method, inspired by [Hušková et al., 2007], is a generalization of the class of score type change point statistics to the vector autoregressive (VAR) case and the epidemic change alternative. Our proposed procedures do not require the observed time series to actually follow the VAR model. In fact, brain processes are unlikely to strictly follow some VAR model. Instead, following the strategy implicitly employed by practitioners, our approach takes model misspecification into account so that our detection procedure uses the model background merely for feature extraction. We derive the asymptotic distributions of our test statistics and show that our procedure has asymptotic power of 1. The proposed test statistics require the estimation of the inverse of the (long run) auto-covariance matrix which is particularly difficult in higher-dimensional settings which is even worse in the multivariate case because the number of entries of the positive definite weight matrix $\mathbf{H}$ increases polynomially with the number of dimensions. We overcome this problem by focusing on changes that are captured by relationships of each channel $Y_t(i)$ with its own past and by judiciously forming $\mathbf{H}$ to be block diagonal. Since the channels of the multivariate time series are highly likely to be dependent, we estimate the underlying covariance structure via a regression bootstrap method which is demonstrated here and in Hušková et al. (2008) to be stable. We investigated the finite sample properties of our proposed method via extensive numerical experiments. Finally, we applied our procedure to electroencephalograms and demonstrated its potential impact in identifying change points in complex brain processes during a cognitive motor task.

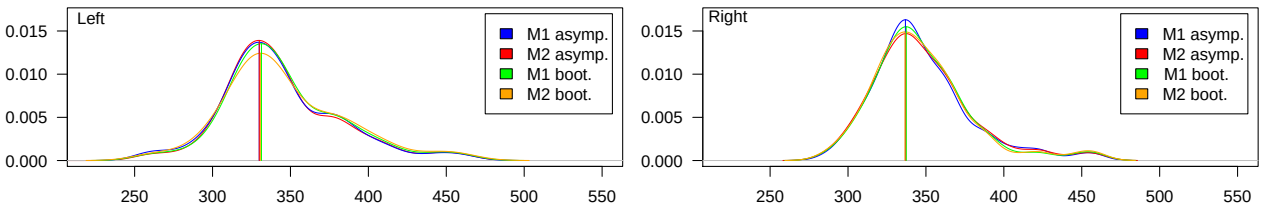

Figure 4.1: Estimated densities of the change point under the multivariate AMOC model.

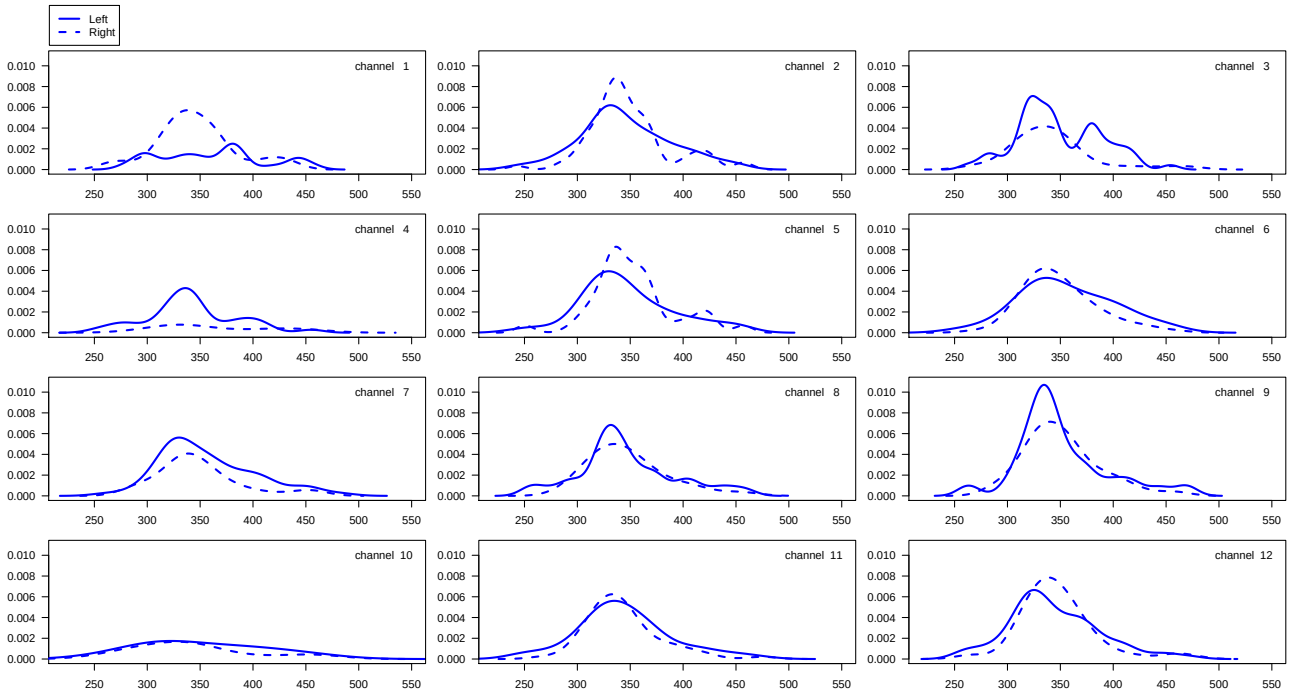

Figure 4.2: Weighted density estimates of the change point for each channel under the AMOC model using the M_1- bootstrap procedure.

References

[Aston and Kirch, 2013] Aston, J. and Kirch, C. (2013). Change points in high-dimensional settings. Diskussions-Papier, KIT.

[Aue et al., 2009] Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series model. *Ann. Statist.*, 37:4046–4087.

[Bai and Perron, 1998] Bai, J. and Perron, L. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, J. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc., Ser. B*, 57:289–300.

[Bickel and Bühlman, 1999] Bickel, P. and Bühlman, P. (1999). A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli*, 5:413–446.

[Brockwell and Davis, 1991] Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, second edition.

[Csörgő and Horváth, 1997] Csörgő, S. and Horváth, L. (1997). *Limit Theorems in Change-point Analysis*. Wiley, Chichester.

[Davis et al., 1995] Davis, R., Huang, D., and Yao, Y. (1995). Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.*, 23:282–304.

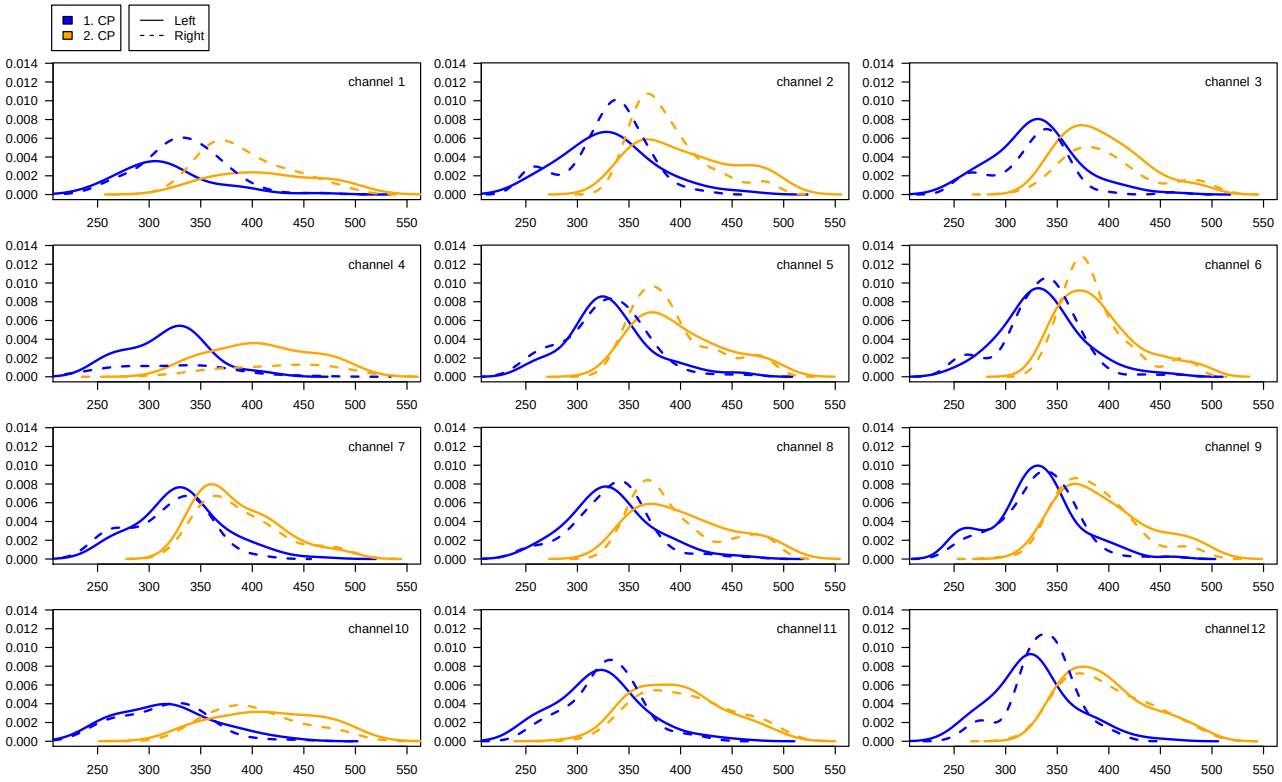

Figure 4.3: Estimated density, weighted to integrate to the relative number of rejections, of the first and second change point under the epidemic model using the T_1-bootstrap procedure.

[Davis et al., 2006] Davis, R., Lee, T., and Rodriguez-Yam, G. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Stat. Assoc.*, 101:223–229.

[Fan and Yao, 2003] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

[Horváth, 1993] Horváth, L. (1993). Change in autoregressive processes. *Stoch. Proc. Appl.*, 5:221–242.

[Horváth, L. et al., 1999] Horváth, L., Kokoszka, P., and Steinebach, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *J. Multivariate Anal.*, 68:96–119.

[Hušková et al., 2007] Hušková, M., Prášková, Z., and Steinebach, J. (2007). On the detection of changes in autoregressive time series, I. Asymptotics. *J. Statist. Plann. Infer.*, 137:1243–1259.

[Hušková and Slabý, 2001] Hušková, M. and Slabý, A. (2001). Permutation test for multiple changes. *Kybernetika*, 37:606–622.

[Kirch and Muhsal, 2011] Kirch, C. and Muhsal, B. (2011). A MOSUM procedure for the estimation of multiple deterministic as well as random change points. Diskussions-Papier, KIT.

[Kuelbs and Philipp, 1980] Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *Ann. Probab.*, 8:1003–1036.

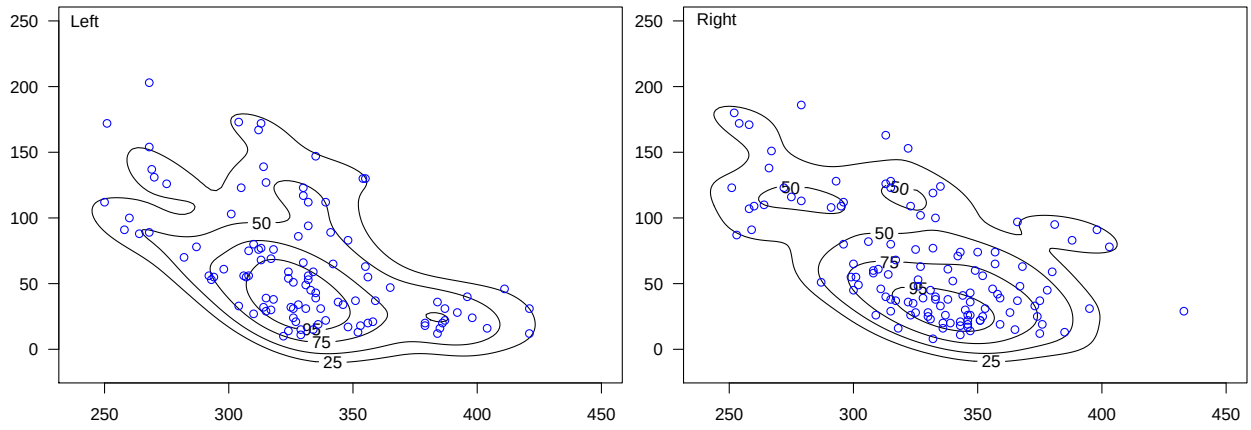

Figure 4.4: Contour plot of the joint density of 1.CP and 2.CP-1.CP

[Kulperger, 1985] Kulperger, R. (1985). On the residuals of autoregressive processes and polynomial regression. *Stoch. Process. Appl.*, 21:107–118.

[Lesage et al., 2002] Lesage, P., Glangeaud, F., and Mars, J. (2002). Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events. *J. Volcan. Geo. Res.*, 115:391–417.

[Marconi et al., 2001] Marconi, B., Genevesio, A., Battaglia-Meyer, A., Ferraina, S., and Caminiti, R. (2001). Eye-hand coordination during reacning: Anatomical relationships between the parietal and frontol cortex. *Cerebral Cortex*, 11:513–527.

[Marušiaková, 2009] Marušiaková, M. (2009). Ph.d. thesis: Tests for multiple changes in linear regression models. Charles University in Prague, Faculty of Mathematics and Physics.

[Pfurtscheller and Haring, 1972] Pfurtscheller, G. and Haring, G. (1972). The use of an eeg autoregressive model for the time-saving calculation of spectral power density distributions with a digital computer. *Elec. Clin. Neurophys.*, 33:113–115.

[Politis, 2003] Politis, D. (2003). Adaptive bandwidth choice. *J. Nonpar. Stat.*, 15:517–533.

[Preußs et al., 2013] Preußs, P., Puchstein, R., and Dette, H. (2013). Detection of multiple structural breaks in multivariate time series. Ruhr-Universität Bochum, Fakultät für Mathematik.

[Vermaak et al., 2002] Vermaak, J., Andrieu, C., Doucet, A., and Godsill, S. (2002). Particle methods for bayesian modeling and enhancement of speech signals. *IEEE Trans. Speech Audio Proc.*, 10:173–185.

6 Appendix

6.1 Proofs of Section 2.1.4

Lemma 6.1. *Let Assumption $\mathcal{A}.1$ hold. Under the alternative additionally assume $\mathcal{A}.2$ and $\mathcal{A}.5$.*

a) Under the null hypothesis as well as both alternatives, it holds

$$\widehat{\boldsymbol{a}}_n(l) \to \widetilde{\boldsymbol{a}}(l) \qquad a.s., \qquad l = 1, \ldots, d,$$

where under the null hypothesis $\widetilde{\boldsymbol{a}}(l) = \boldsymbol{a}_1(l)$.

b) Under the null hypothesis we additionally get

$$\widehat{\mathbf{a}}_n(l) - \widetilde{\mathbf{a}}(l) = O_P(n^{-1/2}), \qquad l = 1, \ldots, d.$$

Remark 6.1. If the time series after the change point fulfills also assumptions as in $\mathcal{A}.1$, then assertion b) also remains true under both alternatives.

Proof. By (2.6) it holds $\widehat{\boldsymbol{a}}_n(l) = \widehat{\mathbf{C}}_n^{-1}(l)\frac{1}{n}\sum_{i=1}^n \mathbb{X}_{i-1}(l)Y_i(l)$. Since mixing sequences are ergodic by the ergodic theorem (see [Fan and Yao, 2003], Prop. 2.8)

$$\widehat{\mathbf{C}}_n(l) \to \boldsymbol{Q}(l) \quad a.s., \qquad \widehat{\boldsymbol{q}}_n(l) := \frac{1}{n}\sum_{i=1}^n \mathbb{X}_{i-1}(l)Y_i(l) \to \boldsymbol{q}(l) \quad a.s.,$$

which proves a) since $\widetilde{\boldsymbol{a}}(i) = \boldsymbol{Q}^{-1}(i)\boldsymbol{q}(i)$ under both the null hypothesis as well a alternatives. For b) we note that Assumptions $\mathcal{A}.1$ imply an invariance principle (see [Kuelbs and Philipp, 1980] Theorem 4) which in turn implies a central limit theorem showing that

$$\sqrt{n}(\widehat{\boldsymbol{q}}_n(l) - \boldsymbol{c}_1(l)) = O_P(1), \qquad \sqrt{n}(\widehat{\mathbf{C}}_n(l) - \boldsymbol{C}_1(l)) = O_P(1).$$

This implies

$$\begin{aligned}
&\widehat{\mathbf{C}}_n^{-1}(l)\,\widehat{\boldsymbol{q}}_n(l) - \boldsymbol{C}_1^{-1}(l)\boldsymbol{c}_1(l) \\
&= \widehat{\mathbf{C}}_n^{-1}(l)(\mathbf{C}_1(l) - \widehat{\mathbf{C}}_n(l))\mathbf{C}_1^{-1}(l)\widehat{\boldsymbol{q}}_n(l) + \mathbf{C}_1^{-1}\left(\widehat{\boldsymbol{q}}_n(l) - \boldsymbol{c}_1(l)\right) = O_P\left(n^{-1/2}\right),
\end{aligned}$$

which concludes the proof. ■

Lemma 6.2. *Let the null hypothesis and Assumption $\mathcal{A}.1$ hold.Then:*

a) The following functional central limit theorem holds:

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} \widehat{\boldsymbol{\xi}}_i, s \in [0,1] \right\} \overset{D[0,1]}{\longrightarrow} \{\mathbf{B}(s) : 0 \leqslant s \leqslant 1\},$$

$$\widehat{\boldsymbol{\xi}}_t = (\widehat{\boldsymbol{\xi}}_t^T(1), \ldots, \widehat{\boldsymbol{\xi}}_t^T(d))^T = ((\mathbb{X}_{t-1}(1))^T\widehat{e}_t(1), \ldots, (\mathbb{X}_{t-1}(d))^T\widehat{e}_t(d))^T,$$

where $\boldsymbol{B}(\cdot)$ is a dp-dimensional Brownian bridge with covariance matrix $\boldsymbol{\Sigma}$.

b) It holds for all $0 \leqslant \beta < \frac{1}{2}$

$$\max_{1 \leqslant k \leqslant n} \frac{n^{2\beta - 1/2}}{k^\beta (n-k)^\beta} \left\| \sum_{i=1}^k \widehat{\boldsymbol{\xi}}_i \right\| = O_P(1).$$

Proof. First, note that $\{\boldsymbol{\xi}_t : t\}$ as well as $\{\mathbb{X}_t(l)\mathbb{X}_t^T(l) : t\}$ are by Assumption $\mathcal{A}.1$ strong mixing with the same mixing rate and $2+\nu/2$ existing moments. Consequently, the assumptions of the invariance principle in [Kuelbs and Philipp, 1980], Theorem 4, are fulfilled. This implies the functional central limit theorem

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}\boldsymbol{\xi}_t, t\in[0,1]\right\} \xrightarrow{D[0,1]} \{\mathbf{W}(t) : 0\leqslant t\leqslant 1\}, \tag{6.1}$$

where $\boldsymbol{W}(\cdot)$ is a dp-dimensional Wiener process with covariance matrix $\boldsymbol{\Sigma}$. Furthermore, in connection with a standard Hájek-Rényi-inequality it implies for all $0\leqslant\beta<1/2$

$$\max_{1\leqslant k\leqslant n}\frac{n^{2\beta-1/2}}{k^\beta(n-k)^\beta}\left|\sum_{i=1}^{k}\left(\zeta_i-\frac{1}{n}\sum_{t=1}^{n}\zeta_t\right)\right| = O_P(1), \tag{6.2}$$

for ζ_t any channel of $\mathbb{X}_{t-1}(l)e_t(l)$, $l=1,\ldots,p$, as well as any channel of $\mathbb{X}_{t-1}(l)\mathbb{X}_{t-1}^T(l)$, $l=1,\ldots,p$. As the minimizer of the least squares equations $\widehat{a}_n(l)$ solve the corresponding score equations, i.e. $\sum_{i=1}^n\widehat{\boldsymbol{\xi}}_i(l)=0$, $l=1,\ldots,d$. Consequently,

$$\begin{aligned}
&\sum_{i=1}^{k}\widehat{\boldsymbol{\xi}}_i(l) = \sum_{i=1}^{k}\left(\widehat{\boldsymbol{\xi}}_i(l)-\frac{1}{n}\sum_{j=1}^{n}\widehat{\boldsymbol{\xi}}_j(l)\right)\\
&=\sum_{i=1}^{k}\left(\boldsymbol{\xi}_i(l)-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\xi}_j(l)\right)-(\widehat{\mathbf{a}}_n(l)-\mathbf{a}_1(l))^T\sum_{i=1}^{k}\left(\mathbb{X}_i(l)\mathbb{X}_i^T(l)-\frac{1}{n}\sum_{j=1}^{n}\mathbb{X}_j(l)\mathbb{X}_j^T(l)\right).
\end{aligned}$$

Now, Assertion a) follows by (6.1) and (6.2) with $\beta=0$ together with Lemma 6.1 a). Assertion b) follows by (6.2) and Lemma 6.1 a). ■

Proof of Theorem 2.1. We will only sketch the proof for the case where $w(\cdot)$ has exactly one discontinuity point $0<u<1$. Define $w_l(t)=w(t-)$ and $w_r(t)=w(t+)$, then

$$\sup_{0\leqslant t\leqslant u}|w_l(\lfloor nt\rfloor/n)-w_l(t)|\to 0, \qquad \sup_{t\geqslant\lceil nu\rceil/n}|w_r(\lfloor nt\rfloor/n)-w_r(t)|\to 0. \tag{6.3}$$

For any $0<\tau<\min(u,1-u)$ it holds $\sup_{\tau\leqslant t\leqslant 1-\tau}w^2(t)<\infty$, hence by Lemma 6.2 a), $\widehat{\boldsymbol{H}}\xrightarrow{P}\boldsymbol{H}$ and (6.3) (for $w(\cdot)$ right continuous in u - otherwise allow for equality in the second supremum)

$$\begin{aligned}
&\max_{\lfloor\tau n\rfloor\leqslant k\leqslant n-\lfloor\tau n\rfloor}\frac{w^2(k/n)}{n}\mathbf{Z}_k^T\widehat{\boldsymbol{H}}\mathbf{Z}_k\\
&=\max\left(\sup_{\tau\leqslant t\leqslant u}\frac{w^2(\lfloor nt\rfloor/n)}{n}\mathbf{Z}_{\lfloor nt\rfloor}^T\widehat{\boldsymbol{H}}\mathbf{Z}_{\lfloor nt\rfloor}, \sup_{\lceil nu\rceil/n<t\leqslant n-\lfloor\tau n\rfloor}\frac{w^2(\lfloor nt\rfloor/n)}{n}\mathbf{Z}_{\lfloor nt\rfloor}^T\widehat{\boldsymbol{H}}\mathbf{Z}_{\lfloor nt\rfloor}\right)\\
&\xrightarrow{\mathcal{D}}\sup_{\tau\leqslant t\leqslant 1-\tau}w^2(t)\sum_{j=1}^{pd}B_j^2(t),
\end{aligned}$$

where we use the $a.s.$ continuity of a Brownian bridge. By Lemma 6.2 b) we get for $1/2>\beta>\alpha$ (as in Assumption $\mathcal{A}.3$)

$$\max_{1\leqslant k\leqslant\lfloor\tau n\rfloor}\frac{w^2(k/n)}{n}\mathbf{Z}_k^T\widehat{\boldsymbol{H}}\mathbf{Z}_k = O_P(1)\left(\sup_{0\leqslant t\leqslant\tau}t^\beta w(t)\right)^2,$$

where the rates are uniform in τ. Assumption $\mathcal{A}.3$ implies for $\beta > \alpha$

$$\sup_{0 \leqslant t \leqslant \tau} t^{\beta} w(t) \to 0 \quad \text{as } \tau \to 0.$$

Similar assertions can be obtained for $k > n - \lfloor \tau n \rfloor$ as well as for the corresponding expressions involving the Brownian bridges in the limit. Standard arguments then conclude the proof of a). Assertion b) can be dealt with analogously. ■

Proof of Theorem 2.2. This follows immediately from Lemma 6.2 a). ■

6.2 Proofs of Section 2.2

The key to understanding the behavior under alternatives is the following lemma:

Lemma 6.3. *Let Assumptions $\mathcal{A}.1$, $\mathcal{A}.2$ and $\mathcal{A}.5$ be fulfilled. Let $\widetilde{\lambda}$ be as in* (2.12) *and the alternative hold.*

a) Then, we get

$$\mathbf{c}_1(l) - \mathbf{C}_1(l)\widetilde{\mathbf{a}}(l) = -\frac{1-\widetilde{\lambda}}{\widetilde{\lambda}} \left(\mathbf{c}_2(l) - \mathbf{C}_2(l)\widetilde{\mathbf{a}}(l)\right).$$

b) Under the AMOC alternative we get

$$\sup_{0 \leqslant t \leqslant 1} \left\| \frac{1}{n} \sum_{j=1}^{\lfloor tn \rfloor} \widehat{\boldsymbol{\xi}}_j(l) - g_A(t) \ \left(\mathbf{c}_1(l) - \mathbf{C}_1(l)\widetilde{\mathbf{a}}(l)\right) \right\|^2 = o_p(1),$$

where

$$g_A(t) = \frac{1}{1-\lambda} \begin{cases} t\,(1-\lambda), & t \leqslant \lambda, \\ \lambda\,(1-t), & t \geqslant \lambda. \end{cases}$$

c) Under the epidemic change alternative we get

$$\sup_{0 \leqslant t_1 < t_2 \leqslant 1} \left\| \frac{1}{n} \sum_{j=\lfloor t_1 n \rfloor}^{\lfloor t_2 n \rfloor} \widehat{\boldsymbol{\xi}}_j(l) - g_B(t_1, t_2) \ \left(\mathbf{c}_1(l) - \mathbf{C}_1(l)\widetilde{\mathbf{a}}(l)\right) \right\|^2 = o_p(1),$$

where $g_B(t_1, t_2) = g_B(t_2) - g_B(t_1)$ and

$$g_B(t) = \frac{1}{\lambda_2 - \lambda_1} \begin{cases} t\,(\lambda_2 - \lambda_1), & t \leqslant \lambda_1, \\ \lambda_1 - t\,(1 - \lambda_2 + \lambda_1), & \lambda_1 \leqslant t \leqslant \lambda_2, \\ (t-1)\,(\lambda_2 - \lambda_1), & t > \lambda_2. \end{cases}$$

Proof. Assertion a) is equivalent to $\mathbf{q}(l) - \mathbf{Q}(l)\widetilde{\mathbf{a}}(l) = 0$ which holds by definition of $\widetilde{\mathbf{a}}(l)$. By Lemma 6.1 a)

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^{\lfloor tn \rfloor} \widehat{\boldsymbol{\xi}}_j(l) &= \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} \mathbb{X}_{j-1}(l)(Y_j(l) - \mathbb{X}_{j-1}^T(l)\widetilde{\mathbf{a}}(l)) - (\widehat{\mathbf{a}}_n(l) - \widetilde{\mathbf{a}}(l)) \ \frac{1}{n} \sum_{j=1}^{\lfloor tn \rfloor} \mathbb{X}_{j-1}(l)\mathbb{X}_{j-1}^T(l) \\ &= \frac{1}{n} \sum_{j=1}^{\lfloor tn \rfloor} \mathbb{X}_{j-1}(l) Y_j(l) - \left(\frac{1}{n} \sum_{j=1}^{\lfloor tn \rfloor} \mathbb{X}_{j-1}(l)\mathbb{X}_{j-1}^T(l) \right) \widetilde{\mathbf{a}}(l) + o_P(1), \end{aligned}$$

where the latter rate is uniform in t. By the ergodic theorem we get

$$\sup_{0\leqslant t\leqslant\lambda}\left|\frac{1}{n}\sum_{j=1}^{\lfloor tn\rfloor}\mathbb{X}_{j-1}(l)Y_j(l)-\mathbf{c}_1(l)t\right|\to 0\qquad a.s.$$

as well as

$$\sup_{\lambda\leqslant t\leqslant 1}\left|\frac{1}{n}\left(\sum_{j=1}^{\lfloor\lambda n\rfloor}\mathbb{X}_{j-1}Y_j+\sum_{j=\lfloor\lambda n\rfloor}^{\lfloor tn\rfloor}\mathbb{X}_{j-1}Y_j\right)-(\lambda\mathbf{c}_1(l)+(t-\lambda)\mathbf{c}_2(l))\right|\to 0\quad a.s.$$

A similar assertion holds for the second sum showing that

$$\sup_{0\leqslant t\leqslant 1}\left|\frac{1}{n}\sum_{j=1}^{\lfloor tn\rfloor}\widehat{\boldsymbol{\xi}}_j(l)-\max(t,\lambda)(\mathbf{c}_1(l)-\mathbf{C}_1(l)\widetilde{\mathbf{a}}(l))-(t-\lambda)_+(\mathbf{c}_2(l)-\mathbf{C}_2(l)\widetilde{\mathbf{a}}(l))\right|\to 0\qquad a.s.,$$

which gives assertion b) using the formula in a). Assertion c) can be obtained analogously. ■

Proof of Theorem 2.3. Let $0<t_0<1$ a continuity point with $w(t_0)>0$, then we get by Lemma 6.3

$$\begin{aligned}M_n^{(1)}&\geqslant n\left(\Delta^2_{\widehat{H}}w^2(t_0)g_A^2(t_0)+o_P(1)\right)\xrightarrow{P}\infty,\\ M_n^{(2)}&=n\left(\Delta^2_{\widehat{H}}\int_0^1 w^2(t)g_A^2(t)\,dt+o_P(1)\right)\xrightarrow{P}\infty,\end{aligned}$$

where for the second statement, we split the sum at possible discontinuity points and use (6.3). The assertions for b) follow analogously. ■

Proof of Theorem 2.4. Define

$$\Delta^2=(\boldsymbol{c}_1-\operatorname{diag}(\boldsymbol{C}_1(1),\ldots,\boldsymbol{C}_1(d))\widetilde{\boldsymbol{a}})^T\,\boldsymbol{H}_A^{-1}\,(\boldsymbol{c}_1-\operatorname{diag}(\boldsymbol{C}_1(1),\ldots,\boldsymbol{C}_1(d))\widetilde{\boldsymbol{a}})\,.$$

We sketch the proof of a) for exactly one (left-continuous – otherwise allow for equality in the second supremum) discontinuity point u. By $\sup_{0\leqslant t\leqslant 1}w^2(t)<\infty$ and (6.3) we get by Lemma 6.3

$$\sup_{0\leqslant t\leqslant u,\lceil nu\rceil/n<t\leqslant 1}\left|\frac{w^2\left(\lfloor nt\rfloor/n\right)}{n^2}\boldsymbol{Z}_{\lfloor nt\rfloor}^T\widehat{\boldsymbol{H}}^{-1}\boldsymbol{Z}_{\lfloor nt\rfloor}-\Delta^2w^2(t)g_A^2(t)\right|=o_P(1).\tag{6.4}$$

Since by assumption $w^2(t)g_A^2(t)$ has a unique supremum at λ, which will eventually be in $\{0\leqslant t\leqslant u,\lceil nu\rceil/n<t\leqslant 1\}$, and the estimator is by definition in that set, standard arguments can be adapted to give assertion a). For b) note that $g_B(t)$ is continuous and has a unique maximum at $t=\lambda_1$ and a unique minimum at $t=\lambda_2$, hence $g_B(t_1,t_2)=g_B(t_2)-g_B(t_1)$ is continuous and has a unique (for $t_1<t_2$) maximum at (λ_1,λ_2). From this assertion b) follows by standard arguments. ■

Proof of Remark 2.2. Similarly as in Lemma 6.2, we obtain for $0\leqslant\beta<1/2$

$$\max_{1\leqslant k\leqslant n}\frac{n^{2\beta+1/2}}{k^\beta(n-k)^\beta}\left\|\frac{1}{n}\sum_{j=1}^{k}\widehat{\boldsymbol{\xi}}_j(l)-\mathbf{g}_A(k/n)\;(\mathbf{c}_1(l)-\mathbf{C}_1(l)\widetilde{\mathbf{a}}(l))\right\|=O_P(1)$$

from which we get (6.4), so that we can conclude as in the proof of Theorem 2.4. ■