Pontifícia Universidade Católica
do Rio de Janeiro

Henrique Helfer Hoeltgebaum

Statistical models with parameters changing through an adaptive mechanism

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Cristiano Fernandes
Co-Advisor: Prof. Niall Adams

Rio de Janeiro
May 2019

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Henrique Helfer Hoeltgebaum

Statistical models with parameters changing through an adaptive mechanism

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the undersigned Examination Committee.

Prof. Cristiano Fernandes
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Niall Adams
Co-Advisor
Imperial College London

Prof. Alexandre Street
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Álvaro Veiga
Departamento de Engenharia Elétrica – PUC-Rio

Dr. Din-Houn Lau
Imperial College London

Prof. Rodrigo Targino
FGV-Rio

Rio de Janeiro, May 3rd, 2019

Henrique Helfer Hoeltgebaum

Graduated in Statistics at the Federal University of Rio Grande do Sul in 2012 and obtained his M.Sc. Degree in Electrical Engineering from PUC-RIO in 2014. From March 2018 to March 2019 was a visiting student at the Mathematics Department of Imperial College London under the supervision of Professor Niall Adams.

Bibliographic data

Hoeltgebaum, Henrique Helfer

Statistical models with parameters changing through an adaptive mechanism / Henrique Helfer Hoeltgebaum; advisor: Cristiano Fernandes; co-advisor: Niall Adams. – 2019.
122 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2019.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Generalized Autoregressive Scores. 3. Cópula dinâmica. 4. Streaming data. 5. Aprendizado de máquina. 6. Filtragem adaptativa. I. Fernandes, Cristiano Augusto Coelho. II. Adams, Niall M. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Acknowledgments

First and foremost I would like to thank both my supervisors Cristiano and Niall. Without their guidance, patience, insights, and long hours of meetings, none of the papers that compose this thesis would be possible. Thank you very much to both of you for believing in my work.

I would like to thank Marcelo Medeiros for giving me the opportunity to work on very interesting research projects in D-LAB during part of my PhD.

I would like to thank my mother, sister and father who gave me their love and emotional support when I needed the most during my PhD.

Also a special thanks to all my friends from Rio de Janeiro and London that made my PhD one of the best experiences of my life.

All this research would not be possible without financial support from PUC-Rio, National Council for Research and Development (CNPq) and Lojas Americanas. Thank you very much.

Abstract

Hoeltgebaum, Henrique Helfer; Fernandes, Cristiano (Advisor); Adams, Niall (Co-Advisor). **Statistical models with parameters changing through an adaptive mechanism**. Rio de Janeiro, 2019. 122p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This thesis is composed of three papers in which the common ground among them is statistical models with time-varying parameters. All of them adopt a framework that uses a data-driven mechanism to update its coefficients. The first paper explores the application of a new class of non-Gaussian time series framework named Generalized Autoregressive Scores (GAS) models. In this class of models the parameters are updated using the score of the predictive density. We motivate the use of GAS models by simulating joint scenarios of wind power generation. In the last two papers, Stochastic Gradient Descent (SGD) is adopted to update time-varying parameters. This methodology uses the derivative of a user specified cost function to drive the optimization. The developed framework is designed to be applied in a streaming data context, therefore adaptive filtering techniques are explored to account for concept-drift. We explore this framework on cyber-security and instrumented infrastructure applications.

Keywords

Generalized Autoregressive Scores; Dynamic copula; Streaming data; Machine learning; Adaptive filtering;

Resumo

Hoeltgebaum, Henrique Helfer; Fernandes, Cristiano; Adams, Niall. **Modelos estatísticos com parâmetros variando segundo um mecanismo adaptativo**. Rio de Janeiro, 2019. 122p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta tese é composta de três artigos em que a ligação entre eles são modelos estatísticos com parametros variantes no tempo. Todos os artigos adotam um arcabouço que utiliza um mecanismo guiado pelos dados para a atualização dos parâmetros dos modelos. O primeiro explora a aplicação de uma nova classe de modelos de séries temporais não Gaussianas denominada modelos Generalized Autegressive Scores (GAS). Nessa classe de modelos, os parâmetros são atualizados utilizando o *score* da densidade preditiva. Motivamos o uso de modelos GAS simulando cenários conjuntos de fator de capacidade eólico. Nos últimos dois artigos, o gradiente descentente estocástico (SGD) é adotado para atualizar os parâmetros que variam no tempo. Tal metodologia utiliza a derivada de uma função custo especificada pelo usuário para guiar a otimização. A estrutura desenvolvida foi projetada para ser aplicada em um contexto de fluxo de dados contínuo, portanto, técnicas de filtragem adaptativa são exploradas para levar em consideração o *concept-drift*. Exploramos esse arcabouço com aplicações em segurança cibernética e infra-estrutura instrumentada.

Palavras-chave

Generalized Autoregressive Scores; Cópula dinâmica; Streaming data; Aprendizado de máquina; Filtragem adaptativa;

Table of contents

List of figures

List of tables

1
Introduction

Many modern statistical problems feature data for which typical statistical assumptions are unjustified. When considering the big data hype, for example, statistical properties of the studied data, such as stationarity, might change over the observation period. These problems of unknown temporal variation paved the way to the development of statistical models that can cope with such phenomena. Statistical models with time-varying parameters have a wide range of applications, including fraud detection (1), cyber-security (2) and mortality rates forecast (3) to name a few.

One way to capture the structure of such processes is by introducing, in the statistical model, parameters that evolve in time. In (4), a classification of models with time-varying parameters was proposed, namely parameter and observation driven models. All papers that compose this thesis deal with the development and implementation of observation driven framework, in which the parameter updating mechanism is a function of the observed data via an optimization criterion. Such mechanisms should adequately cope with some sources of alteration to the statistical properties of the data generating processes. The effectiveness of both were assessed considering real-world applications and extensive simulation studies.

The use of the density *score*, namely the second derivative of the log-likelihood function, as the driving force to update the parameters is explored in Chapter 2. This framework is known in the time series literature as *Generalized Autoregressive Scores* (GAS) model (5) or *Dynamic Conditional Scores* (DCS) model (6)[1]. More specifically, in GAS models, conditional on past observations,

[1] Although there are two acronyms, this thesis uses the acronym GAS.

an appropriate probability model is chosen for the response variable. The use of the *score* for updating time-varying parameters is intuitive, given that it defines the steepest ascent direction for improving the model's local fit in terms of the likelihood or density, given the current parameter position. By construction, such an updating mechanism uses information from the entire density to track the evolution of time-varying parameters, not only first- or second-order moments. This modelling feature allows GAS models to capture relevant nonlinear information of the time series dynamics.

The aforementioned aspects motivate the use of this framework in one real world application. In Chapter 2 we simulate wind power generation considering both time- and spatial-dependent scenarios, as published in (7). Our case study shows, based on real data from the Brazilian power system, that the proposed methodology is capable of producing predictive scenarios with coherent temporal and spatial dependence that are needed for power system studies.

The second updating mechanism considered in this thesis is *Stochastic Gradient Descent* (SGD) (8, 9), used in Chapters 3 and 4. More specifically, this framework is explored on *streaming data* sets, which consist of potentially unending sequences of data values arriving at high frequency. When analysing big data streams, one needs to provide algorithms with fixed memory and processing speed. Moreover, there are challenges related to constructing procedures that can handle *concept drift* – the tendency of future data to have different underlying properties to current and historic data.

The issue of handling temporal structure, such as trend and periodicity in an online manner, remains a difficult problem for streaming estimation. In Chapter 3 we propose *RAC* (Real-Time Adaptive Component), a penalized-regression modelling framework which satisfies the computational constraints of streaming data, and provides capability for dealing with concept drift. To handle this collection of issues, techniques from adaptive filtering (9) were used. Finally, experiments with simulated data suggest the procedure has merit for a variety of scenarios, and an illustration with real cyber-security data further demonstrates the promise of the method.

SGD methods are further adapted to a class of problem motivated by a data set collected from an instrumented railway bridge. The instrumentation on the bridge consists of a network of spatially distributed fibre-optic sensors, from which data is recorded at high-frequency (10). The application objective of Chapter 4 is to detect train passage events in this bridge data. Methodologically, this required the development of a novel time-varying and incremental Principal Component Analysis (PCA) method, again based on adaptive filtering techniques. This estimation procedure is complemented by an anomaly detection method based on conformal prediction (11, 12). Finally, the performance of this method is evaluated, for both estimation accuracy and train event detection, using simulated and real data set.

The reminder of this thesis is divided in three chapters. Each chapter makes reference to one article. The mathematical notation for each chapter is described within each and are not the same across chapters. Moreover, the first paper was produced under the supervision of Professor Cristiano Fernandes while the last two under the supervision of Professor Niall Adams.

2
Generating joint scenarios for renewable generation: The case for non-Gaussian models with time-varying parameters

Keywords: Dynamic copula, probabilistic forecasting, score-driven models, time-varying parameters, wind power.

Abstract: The development of medium/long-term studies for power-system operation and planning under the uncertainty of renewable generation is a key challenge faced by power-system agents worldwide. There is a vast literature on stochastic optimization models devoted to addressing the relevant issues on both operation and planning applications. Notwithstanding, few papers focus on addressing the gaps within the subject of joint scenario generation despite the high sensibility of stochastic optimization models with regard to their input scenarios. Characterizing wind power generation (WPG) stochastic processes to devise time- and spatial-dependent scenarios, based on simulation procedures, for time horizons of one to a few years is a difficult task. Multiple regimes and non-Gaussian distributions are two of the main issues that significantly change the risk described through generated scenarios. In this paper, a new methodology to simulate long-term joint scenarios for multivariate WPG time series is presented. The proposed framework, known as Generalized Auto Regressive Score (GAS) models, is derived based on a new class of time-series model with time-varying parameters and an arbitrary non-Gaussian distribution. Our case study shows, based on real data from the Brazilian power system, that the proposed methodology is capable of producing scenarios with coherent temporal and spatial dependence that are needed in power system studies.

2.1 Introduction

Renewable energy expansion has been growing worldwide, mainly in response to governmental incentives for reducing greenhouse gas emissions. In particular, wind power generation (WPG) is one of the largest sources of renewable energy, and according to the International Energy Agency, it will account for 18% of global power by 2050 (13). However, the uncertainty associated with its nondispatchable nature may jeopardize the reliability of electricity supply. In attempting to minimize this type of risk, it is highly desirable to produce reliable probabilistic forecasts for WPG time series suitable for scenario generation procedures such as Monte Carlo-based methods. The importance of such scenarios emerges in many instances where complex optimization-based decision models are used, e.g., (i) energy trading, (ii) unit commitment, (iii) grid expansion planning, and (iv) investment decisions (see (14, 15, 16, 17) and references therein). There is a vast literature on stochastic optimization models devoted to addressing the relevant issues on the aforementioned power-system applications. Notwithstanding, few papers focus on addressing the gaps within the subject of joint scenario generation despite the high sensibility of the stochastic optimization models with regard to their input scenarios.

For example, (18) and (19) use importance sampling to model tail dependencies under a reduced number of scenarios for stochastic unit commitment and transmission planning, respectively. While in (18) a transformed wind speed time series is modeled by an autoregressive process to generate wind power scenarios, in (19), an interesting clusterization method is applied to historical data to capture empirical information of the wind power variability. Furthermore, (20) focus on two-stage transmission planning using a sample of hourly data to empirically characterize the correlations in demand and wind power generation among different regions. Within the subject of two-stage robust optimization models applied to planning and operation, in (14) a robust model is proposed to address reliability in co-optimized generation and transmission planning, while in (21), a unit commitment problem is addressed under wind power uncertainty. Both (14) and (21) apply Monte Carlo simulation techniques based on static and Gaussian distributions to evaluate the performance of the developed robust strategies. It is worth noting that, except for (18) that used a linear autoregressive model, all previously reported works make use of scenario generation approaches based on either empirical or Gaussian static models to address relevant operation and planning problems affected by renewable variability. Hence, the development of novel time series models applied to generate accurate scenarios for renewable energy constitutes a relevant and timely research topic that might be of interest for many power system applications.

In contrast to the main objective of WPG simulation methods, which aim to produce scenarios that characterize all the conditional density and its fit to the observed quantile data, much research has been devoted to devise short-term forecast methods for wind speed and WPG time series (for instance, see (22)). In some applications on medium- and long-term forecasting, model construction is based on conventional ARMA models with seasonal lags, or SARIMA models, under a Gaussian distribution (see (23) for a high-dimension estimation process based on LASSO). In its original formulation, SARIMA models present a fixed conditional variance. To add extra flexibility to these models, (24) proposes an ARIMA model with a GARCH effect, allowing the conditional variance of the WPG distribution to vary over time. In addition, to provide a description of the spatial correlation of wind-speed time series, (25) proposed an Autoregressive Fractionally Integrated–GARCH model (ARFIMA-GARCH). In these and other models, conditional on the past, the distribution of the response variable is assumed to be Gaussian.

The non-Gaussian nature of WPG time series is well reported in the literature (see (26, 27, 28, 29, 18), and references therein). However, only very limited classes of non-Gaussian time-series models are available for modeling WPG. The state-of-the-art literature on short-term probabilistic modeling of renewable energy time series (wind and solar) mainly relies on nonparametric models. For instance, in (26, 29, 30), nonparametric approaches were devised to forecast conditional distributions based on kernel density estimators, quantile regression, and extreme learning machine, respectively. Despite the virtues found in nonparametric methods, these models require a large amount of data to be fitted and are mainly devised for the univariate case. Thus, the development of new parametric models capable of properly characterizing the full multivariate non-Gaussian distribution for WPG time series is a relevant research theme, which has not received much attention so far. Relevant applications arise from risk assessment in both medium- and long-term planning and investment studies, which mainly rely on a few data points to characterize conditional estimates of extreme quantiles (31, 32, 33, 17, 34).

In (5) and (6), a general framework for time-series models with time-varying coefficients was proposed by considering any univariate or multivariate non-Gaussian conditional distribution, either discrete or continuous. Such a model has been named in the recent literature as a Generalized Autoregressive Score (GAS) model (5) or a Dynamic Conditional Score (DCS) model (6). In (5), it has been shown that several well-known time-series models from the econometric literature are a particular case of GAS models[1]. More specifically, a GAS model is built based on a user-defined conditional probability function whose parameters follow a data-driven dynamic equation that uses the score as its driving force. The use of the score function for updating time-varying parameters is an intuitive choice. It is defined as the steepest ascent direction (gradient) for improving the local fit of the model in terms of likelihood. In such an updating mechanism, information from the whole density is used within the

[1] For instance, GARCH models that address heavy-tailed distributions (35), autoregressive conditional duration models (36) to tackle asymmetric distributions, and the Poisson count models of Davis (37) are particular cases of GAS models.

model to track the evolution of time-varying parameters through a nonlinear transformation of past data. This modeling feature allows GAS models to capture relevant nonlinear information of the time-series dynamics, which is not possible through linear models.

To generate joint scenarios for WPG time series of power plants belonging to different geographical areas, it is important to capture the spatial dependence among these units (23, 18, 38). Within the GAS framework, in (39), the authors provided an empirical application of a multivariate Student t density to a panel of daily-equity returns, where both the variance and correlation matrix are updated through a GAS mechanism. They indicate the link between their framework and time-varying copulas, but no results and tests were provided. Furthermore, it is possible to combine linear models with copulas that also have time-varying parameters as shown in (40, 41, 42). However, none of these works have proposed a GAS mechanism to update time-varying parameters of marginal densities and copulas simultaneously.

Most of the applications using GAS models have focused on problems in finance and economics. We refer the interested reader to the main online repository on GAS papers (43). Notwithstanding its virtue in properly addressing nonlinearity and non-Gaussianity of real time series, to the best of the authors' knowledge, there is no publication in the power-system literature using such a framework. Moreover, the inclusion of a seasonal structure in the updating mechanism for the time-varying parameters has not been addressed in the GAS literature so far, despite being a relevant structure shared by a wide range of time series such as WPG (see (44, 45, 31, 23), and references therein).

As previously reported, accurate scenarios characterizing non-Gaussianity and non-linearities found in WPG time series are highly demanded in power system applications such as planning and operations

(18, 19, 20, 14, 15, 16, 17, 21). Hence, the objective of this paper is to propose a new methodology to generate synthetic non-Gaussian and multivariate scenarios for WPG time series using copulas and GAS models. Thus, in our proposed framework, both time and spatial dependence are captured through a new parametric time series model based on GAS updating mechanisms. More specifically, the contributions of our paper are twofold:

- A new framework based on GAS models is proposed to fit and generate a non-Gaussian univariate (predictive) density of individual WPG time series. In such models, parameters are made time varying through an updating equation based on the score vector (the first derivative of the log-predictive density). By construction, the form of the parameter's updating equation will depend on the particular choice of the predictive density. Model diagnostics and multistep ahead prediction procedures are also presented as part of the proposed framework.
- A method for the generation of joint WPG scenarios, where spatial dependence among the different units is captured by a time-varying Student t copula. In our proposed framework, the copula parameters, namely, entries of the correlation matrix, have their evolution in time driven by a second GAS updating mechanism allowing spatial dependence to change along seasons. In this paper, the newly proposed time-varying copula updating mechanism uses the individual predictive density devised in the first contribution to generate the multivariate scenarios. Nevertheless, we highlight that it can also be used with other univariate methods for the same end, thereby constituting a broader contribution to the state-of-the-art literature on renewable-energy scenario generation[2].

[2] Note that many nonparametric models based on quantile regression and other related methods are not directly extensible to the multivariate framework. In such cases, our proposed time-varying copula scheme can be viewed as a second-step procedure used to

The rest of this paper is organized as follows. In Section 2.2, the proposed univariate GAS model framework with seasonal dynamics is presented; estimation, diagnostics, and forecasting procedures are also provided in this section. In Section 2.3, a dynamic elliptical-copula model is presented based on the results of (39). In Section 2.4, a case study using real data from the Brazilian power system is presented to illustrate the application of the proposed framework in simulating joint scenarios for multivariate WPG time series. Finally, Section 2.5 presents the conclusion of this study.

2.2 GAS models

In GAS models, the choice of candidate (conditional) densities for a given response variable is based on the support of the variable being modeled. As WPG time series have normalized support in the range $[0, 100)$, also known as capacity factor (generation in percentage of the maximum power), it is natural to consider a beta density to describe such process. More precisely, in our application, the i-th WPG time series, from a set $\mathcal{K} = \{1, ..., K\}$, is described by a beta probability density function (PDF) where the first shape parameter, f_{it}, varies over time, whereas the minimum (a_{it}) and maximum (b_{it}) parameters varies according to each month. The minimum and maximum are *ex ant* estimated through a standard maximum likelihood method based on static (unconditioned) betas. Thus, the univariate probability models are given by

$$
\begin{aligned}
&p(y_{it}|f_{it},\mathcal{F}_{i,t-1};\theta_i) \\
&\qquad = \frac{\Gamma(\beta_{it}+\alpha_i)}{\Gamma(\beta_{it})\Gamma(\alpha_i)} \cdot \frac{(y_{it}-a_{it})^{\beta_{it}-1}(b_{it}-y_{it})^{\alpha_i-1}}{(b_i-a_{it})^{\beta_{it}+\alpha_i-1}}
\end{aligned}
\tag{2-1}
$$

characterize spatial dependencies and extend existing univariate simulation methods to the multivariate case.

$\forall\ i \in \mathcal{K}$ and $t \in T$. In (2-1), y_{it} represents the WPG injected by unit $i \in \mathcal{K}$ at time t, $\mathcal{F}_{t-1}$ is the past information of time series i up to time $t-1$, β_{it} is the exponential of the underlying time-varying parameter, f_{it}, and α is the second shape parameter, assumed to be fixed. The vector of fixed parameters of the GAS model is θ, and in our methodology, it encompasses α, the coefficients associated with explanatory and dummy variables, when they exist, and the parameters that are part of the updating mechanism that defines f_t, as given in Expression (2-2).

The time-varying parameter GAS(p,q) updating mechanism for parameter f_{it}, $\forall\ i \in \mathcal{K}$ and $t \in T$, is given by

$$f_{i,t+1} = \omega_i + \sum_{l=1}^{p} A_{i,l} s_{i,t-l+1} + \sum_{l=1}^{q} B_{i,l} f_{i,t-l+1}. \tag{2-2}$$

In (2-2), $s_{i,t-l+1}$ is the score of the beta PDF for unit i at time $t-l+1$, and ω_i, $A_{i,l}$, and $B_{i,l}$ are fixed parameters.

To complete the description of the updating mechanism presented in (2-2), it is necessary to define $s_{i,t}$, the scaled score. This is given by the following nonlinear transformation of the data (see (5)):

$$s_{i,t} = \mathcal{I}_{i,t|t-1}^{-d} \cdot \nabla_{i,t}, \tag{2-3}$$

$$\nabla_{i,t} = \frac{\partial \ln p(y_{i,t}|f_{i,t}, \mathcal{F}_{i,t-1}; \theta_i)}{\partial f_{i,t}}, \tag{2-4}$$

$$\mathcal{I}_{i,t|t-1} = E_{t|t-1}[\nabla'_{i,t} \nabla_{i,t}] \tag{2-5}$$

$\forall\ i \in \mathcal{K}$ and $t \in T$, where $p(y_{i,t}|f_{i,t}, \mathcal{F}_{i,t-1}; \theta_i)$ is the conditional PDF chosen to model the time series and $\mathcal{I}_{i,t|t-1}$ is the Fisher information matrix. Such a matrix acts as a scaling factor for the score normalizing the variance of $s_{i,t}$. More details can be found in (5). In practice, the choice of d can be decided empirically: for a given model, one chooses the value of $d \in \{0, 1/2, 1\}$ that produces the best model diagnostics and forecasting.

As mentioned before, because the beta parameter can only assume positive values, it is conveniently reparameterized as the exponential of $f_{i,t}$. In addition, the model can accommodate the effect of explanatory (exogenous) variables through the equation that defines the time-varying beta parameter as follows:

$$\beta_{i,t} = e^{\phi_i' X_{i,t} + f_{i,t}}. \tag{2-6}$$

In (2-6), $X_{i,t}$ represents the vectors of explanatory variables for time series i at period t. Within such framework, the effect of any external factor that might improve the data fit (seasonal dummy variables, weather condition indexes, etc.) can be estimated and used in the simulation step to generate conditional scenarios.

2.2.1 Estimation

The estimation of the vector of fixed parameters θ_i for each time series $i \in \mathcal{K}$ is based on the maximization of the log-likelihood function, that is,

$$\hat{\theta}_i = \underset{\theta_i}{\operatorname{argmax}}\ l_i(\theta_i). \tag{2-7}$$

In our case, where the predictive density is beta (see Equation (2-1)), the log-likelihood is given by

$$\begin{aligned} l_i(\theta_i) = \sum_{t=1}^{T} \Bigg\{ & - \ln(b_{it} - a_{it})(\beta_{it} + \alpha_i - 1) + \ln\Gamma(\beta_{it} + \alpha_i) \\ & - \ln\Gamma(\beta_{it}) - \ln\Gamma(\alpha_i) + \beta_{it}\ln(y_{it} - a_{it}) \\ & - \ln(y_{it} - a_{it}) + \alpha_{it}(b_{it} - y_{it}) - \ln(b_{it} - y_{it}) \Bigg\}. \end{aligned} \tag{2-8}$$

Given that $f_{i,t}$ is a function of θ_i (recall that the coefficients in (2-2) are all part of the vector of fixed parameters, θ_i), to solve the estimation problem (2-7), a numerical nonlinear optimization procedure has to be used. In this work, we apply the Nelder–Mead algorithm to find initial points to the BFGS algorithm (46). Given the recursive nature of the updating equation for f_{it}, one needs initial values for $f_{i,0}, f_{i,-1}, \ldots, f_{i,1-q}$ and $s_{i,0}, s_{i,-1}, \ldots, s_{i,1-p}$. These are obtained by fitting individual beta distributions, one for each month, under the assumption that both beta parameters are fixed in time. More precisely, each WPG time series i is split into 12 monthly series, i.e., $\{Y_{Jan}^{(i)}, \ldots, Y_{Dec}^{(i)}\}$, where $Y_v^{(i)}$ stands for the vector comprising the data for month v and unit i. Then, for each of these time series, $\{Y_v^{(i)}\}_{v=1}^{12}$, a fixed beta density is estimated via maximum likelihood, resulting in the estimates $\{\alpha_v^{(i)}\}_{v=1}^{12}$ and $\{\beta_v^{(i)}\}_{v=1}^{12}$. These are then used to initialize the recursion in (2-2) for each evaluation of the log-likelihood function. The minimum and maximum parameters are also obtained in this step based on the monthly static estimation. Thus, the function that is passed to the BFGS algorithm is a numerical function that receives as input a trial value for vector θ_i, calculates the value of $f_{i,t}$ according to the recursive rule in (2-2), and returns as output the value of Expression (2-7). Based on multiple randomly generated initial values for θ_i, the BFGS algorithm iterates until a local maximum is found and the best solution is used. Although customarily used in non-linear statistical estimation schemes, it should be noticed that this technique does not guarantee a global optima, but an improved local solution.

2.2.2 Diagnostics

Diagnostics in GAS models can be obtained using quantile residuals, an appropriate type of residuals for nonlinear and non-Gaussian time-series models that has been defined in (47). The observed quantile residual is given by

$$r_{t,\hat{\theta}_i} = \Phi^{-1}[F(y_{i,t}|f_{i,t}, \mathcal{F}_{i,t-1}; \hat{\theta}_i)] \tag{2-9}$$

$\forall\ i \in \mathcal{K}$ and $t \in T$, where $F(\cdot)$ is the cumulative distribution function (CDF) associated with the proposed beta density, i.e., $p(y_{i,t}|f_{i,t}, \mathcal{F}_{i,t-1}; \theta_i)$, and $\Phi^{-1}[\cdot]$ is the quantile function of a standard Gaussian distribution. Under correct model specification, these residuals should be normally distributed and show no temporal dependence. These can be checked, for example, using the standard Jarque–Bera test for normality, the Ljung–Box test for absence of serial correlation, and the Ljung–Box (on squared residuals) test to test for the absence of nonlinear dependence, such as an ARCH effect.

2.3 Time-varying spatial dependence model

Once the proposed beta GAS models are individually fitted to each of the K WPG time series, the observed dependence among the different units can be captured by applying a Student t copula to the set of probability integral transforms (PITs) derived from the individual beta PDF's. The PITs are uniformly distributed values, $u_{1t}, u_{2t}, \ldots, u_{Kt}$, whose components, associated with each unit $i \in \mathcal{K}$ and period $t \in T$, can be obtained from the observed data and the respective univariate CDFs as follows:

$$u_{i,t} = F(y_{i,t}|f_{i,t}, \mathcal{F}_{i,t-1}, \hat{\theta}_i). \tag{2-10}$$

A conditional copula is defined as a multivariate distribution, namely, $C(u_{1t}, \ldots, u_{Kt}|\mathcal{F}_{t-1})$, of the PITs values, conditional on the available set of information $\mathcal{F}_{t-1}$ (40). To that end, if we let G_ν be the CDF of an univariate Student t distribution with ν degrees of freedom, we can transform the PITs to Student t variables, $\tilde{y}_{it} = G_\nu^{-1}(u_{it})$, and define our copula as a multivariate Student t CDF with a time-varying correlation matrix. The associated copula density function, which will be used in the estimation process, assumes the following form $\forall$ $i \in \mathcal{K}$ and $t \in T$:

$$g(\tilde{\mathbf{y}}_t|\Sigma_t;\nu) = \frac{\Gamma\left(\frac{\nu+K}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)[(\nu-2)\pi]^{K/2}|\Sigma_t|^{1/2}} \cdot \left[1 + \frac{\tilde{\mathbf{y}}_t'\Sigma_t^{-1}\tilde{\mathbf{y}}_t}{(\nu-2)}\right]^{-\frac{\nu+K}{2}}, \tag{2-11}$$

where $\tilde{\mathbf{y}}_t = [\tilde{y}_{1t}, ..., \tilde{y}_{Kt}]'$ and Σ_t is a time varying-correlation matrix, assumed to follow a GAS updating mechanism similar to that proposed in (48).

To apply the GAS updating mechanism to the correlation matrix, it is key to use a factorization scheme for Σ_t that allows always generating positive definite matrices. Such a factorization scheme, proposed in (48), is used in this work as follows:

$$\Sigma_t = \text{diag}(Q_t)^{-1/2} \cdot Q_t \cdot \text{diag}(Q_t)^{-1/2}, \tag{2-12}$$

where Q_t is a symmetric positive definite matrix that guarantees a symmetric positive definite matrix Σ_t with elements outside the main diagonal lying inside the $[-1, 1]$ range. The matrix Q_t carries all the information regarding the spatial dependence structure among different WPG time series. The idea of adopting a time-varying mechanism to update the elements of Q_t is to consider the temporal variation of this dependence when generating joint scenarios for WPG time series.

Updating the quantities of the time-varying Q_t, is accomplished through a second GAS mechanism as proposed in (39), i.e.,

$$\text{vech}(Q_{t+1}) = \Omega + \Pi s_t + \Upsilon \text{vech}(Q_t), \tag{2-13}$$

where

$$s_t = E[\nabla_t \nabla_t^{'}]^{-d} \cdot \nabla_t, \tag{2-14}$$

$$\nabla_t = \frac{\partial \ln g(\tilde{\mathbf{y}}_t | \Sigma_t; \nu)}{\partial (\text{vech}(Q_t))}. \tag{2-15}$$

In (2-13), $\text{vech}(Q_{t+1})$ is the half-vectorization of Q_{t+1}, which is a linear transformation used to convert the lower triangular portion of $K \times K$ symmetric matrices to $K(K+1)/2$ vectors (recall $K = |\mathcal{K}|$). In addition, Π and Υ are diagonal matrices, and Ω is a $K \times 1$ vector with elements equal to one for updating the elements of the main diagonal of Q_t to make the model identifiable. Hence, the elements of Ω corresponding to the diagonal elements of Q_t can be multiplied by any arbitrary positive number without changing the decomposition. Details and proofs regarding the Expressions (2-14)-(2-15) for the multivariate Student t distribution can be found in (39).

2.3.1 Estimation of the copula parameters

We now briefly describe the method used to estimate the fixed parameters of the GAS mechanism associated with the Student t copula, $\Theta = (\nu, \Omega, \Pi, \Upsilon)$. For this we applied the inference for margins (IFM) optimization procedure (see (49), (50)). In IFM estimation, the vector of fixed parameters, $\{\hat{\theta}_i\}_{i\in\mathcal{K}}$, associated with the conditional density of each WPG time series is first estimated individually using the optimization procedure described in Section 2.2.1. Then, the copula parameters are estimated via maximum likelihood in a second step optimization procedure. By considering the K-dimensional vector $\tilde{\mathbf{y}}_t = [G_\nu^{-1}(u_{1,t}), \ldots, G_\nu^{-1}(u_{K,t})]'$, the log-likelihood of the multivariate Student t copula can be written as

$$\begin{aligned}\mathcal{L}(\Theta) =& \sum_{t=1}^{T}\left\{ \ln\left[\Gamma\left(\frac{\nu+K}{2}\right)\right] - \ln\left[\Gamma\left(\frac{\nu}{2}\right)\right]\right.\\ & -\frac{1}{2}\ln|\Sigma_t| - \frac{K}{2}\ln[(\nu-2)\pi]\\ & \left. -\frac{\nu+K}{2}\ln\left[1+\frac{\tilde{\mathbf{y}}_t'\Sigma_t^{-1}\tilde{\mathbf{y}}_t}{\nu-2}\right]\right\}\end{aligned} \tag{2-16}$$

$\forall\ i \in \mathcal{K}$, where Σ_t follows a GAS updating mechanism given by (2-13). Maximization of this log-likelihood function $\mathcal{L}$ is analogous to the univariate optimization described in Section 2.2.1. The nonlinear optimization problem is solved by using a Nelder–Mead algorithm to find initial values for the BFGS algorithm. The degrees of freedom ν was estimated by using profile likelihood. For the interested reader, we referrer to (51).

2.3.2

Generating joint scenarios

After the estimation procedure, the joint scenario generation process proposed in this work is based on two steps that are repeated for every period within the simulation horizon. In the first step, M individually simulated scenarios, one for each WPG time series, are sampled from the conditional beta univariate PDFs, where the time-varying parameters are updated according to (2-2). In the second step, the spatial dependences between different units is introduced applying the estimated Student t copula with time-varying parameters being updated according to (2-13). After performing these two steps for the whole simulation horizon, one multivariate path (scenario) is generated. Then, this process is repeated (or parallelized) as many times as necessary, according to the number of samples needed. Assuming a simulation horizon set $\mathcal{H}$ with H periods, i.e., $\mathcal{H} = \{1, ..., H\}$, the simulation procedure that generates M multivariate samples of WPG according to the proposed methodology is as follows:

Part I: Univariate

1. Given $\{\hat{\theta}_i\}_{i\in\mathcal{K}}$, obtained by the estimation process described in Section (2.2.1) and the filtered vector $\{\hat{f}_{i,T+1}\}_{i\in\mathcal{K}}$ obtained by applying recursion (2-2), for each unit $i \in \mathcal{K}$, draw M scenarios[3], $y^{(1)}_{i,T+1}, \dots, y^{(M)}_{i,T+1}$, from the estimated conditional density (2-1) for period $T+1$.

2. Use $\{y^{(1)}_{i,T+1}, \dots, y^{(M)}_{i,T+1}\}_{i\in\mathcal{K}}$ and the updating Equation (2-2) to obtain $\{\hat{f}^{(1)}_{i,T+2}, \dots, \hat{f}^{(M)}_{i,T+2}\}_{i\in\mathcal{K}}$ based on the values of $\{\hat{\theta}_i\}_{i\in\mathcal{K}}$ and $\{\hat{f}_{i,T+1}\}_{i\in\mathcal{K}}$.

3. Repeat steps 1 and 2, for periods $h = T+2, \dots, T+H$, generating scenarios $y^{(1)}_{i,T+h}, \dots, y^{(M)}_{i,T+h}$ for each unit $i \in \mathcal{K}$ and period $h \in \mathcal{H}$.

[3] In the case where explanatory variables are considered, conditioned scenarios will be generated based on future values for the exogenous variables $X_{i,T+h}$ that should be provided for all $h \in \mathcal{H}$.

In the sequel, the spatial-dependence structure is introduced on the aforementioned individually-generated scenarios through the proposed copula model.

Part II: Multivariate

1. Given the previously generated M time series scenarios (temporal paths) $\{y_{i,T+1}^{(m)}, \ldots, y_{i,T+H}^{(m)}\}_{m=1}^{M}$, apply (2-10) to obtain $\{u_{i,T+1}^{(m)}, \ldots, u_{i,T+H}^{(m)}\}_{m=1}^{M}$ and apply $\tilde{y}_{it} = G_{\nu}^{-1}(u_{it})$ to find $\{\tilde{y}_{i,T+1}^{(m)}, \ldots, \tilde{y}_{i,T+H}^{(m)}\}_{m=1}^{M}$ for each unit $i \in \mathcal{K}$.

2. Using as initial value the estimated, or filtered, Σ_T and sampled quantities $\{\tilde{y}_{i,T+1}^{(m)}, \ldots, \tilde{y}_{i,T+H}^{(m)}\}_{m=1}^{M}$ for each unit $i \in \mathcal{K}$, apply recursion (2-13) to find a set of M temporal scenarios of vectors $\text{vech}(\hat{Q}_{T+h}^{(1)}), ..., \text{vech}(\hat{Q}_{T+h}^{(M)})$. Using (4-10), find M samples for the correlation process, $\Sigma_{T+h}^{(1)}, ..., \Sigma_{T+h}^{(M)}$, for all periods h within the simulation horizon $\mathcal{H}$.

3. To find the joint (multivariate) scenarios for the WPG time series, apply the conditional sampling technique (see (52)) to produce conditional vectors of PITs, $\{\mathbf{u}_{T+h}^{(m)*} \in \mathbb{R}^{K}\}_{m=1}^{M}$. The aforementioned conditional sampling technique consists of (i) generating a sample of M multivariate Student t scenarios, $\{\tilde{\mathbf{y}}_{T+h}^{(m)} \in \mathbb{R}^{K}\}_{m=1}^{M}$, with the estimated $\hat{\nu}$ degrees of freedom and correlation matrices, $\Sigma_{T+h}^{(1)}, ..., \Sigma_{T+h}^{(M)}$, and (ii) using the univariate Student t CDF to find the PITs, $\mathbf{u}_{T+h}^{(m)*} = [G_{\hat{\nu}}(\tilde{y}_{1,T+h}^{(m)*}), \ldots, G_{\hat{\nu}}(\tilde{y}_{K,T+h}^{(m)*})]'$.

4. Finally, use the produced univariate temporal scenarios of time varying parameter, $\{\hat{f}_{i,T+1}^{(m)}, \ldots, \hat{f}_{i,T+H}^{(m)}\}_{m=1}^{M}$ – obtained for each unit $i \in \mathcal{K}$, to calculate the set of multivariate WPG scenarios (K-dimensional vectors accounting for spatial dependencies), $\{\mathbf{y}_{T+1}^{(m)*}, \ldots, \mathbf{y}_{T+H}^{(m)*}\}_{m=1}^{M}$, where each component is obtained as $y_{i,T+h}^{*(m)} = F^{-1}(u_{i,T+h}^{(m)*} | \hat{f}_{i,T+h}^{(m)}, \mathcal{F}_{i,T+h-1}, \hat{\theta}_i)$.

By following this Monte Carlo-based procedure applied to the proposed GAS model, one obtains a set of M joint (multivariate) time series scenarios for the future WPG of the K units, $\{\mathbf{y}_{T+1}^{(m)*}, \ldots, \mathbf{y}_{T+H}^{(m)*}\}_{m=1}^{M}$. Hereinafter, we will refer to the two-step simulation procedure presented in this section as t-GAS methodology.

2.4 Case study: simulating medium- and long-term joint scenarios

Our data comprise monthly WPG time series, from January 1981 to December 2011, measured at three wind plants located in northeast Brazil, namely in Rio do Fogo (RF), Icaraizinho (IC), and Enacel (EN). The last three years were removed from the estimation process for out-of-sample evaluation. The intrinsic non-Gaussian nature of WPG time series can be checked by observing the positive skewness suggested by the shape of the qq plots (see Figure 2.1) and by the results of the Jarque–Bera test for normality, which has been rejected for all WPG time series (p values < 0.001). From these, one can conclude the inadequacy of adopting Gaussian models for such series.

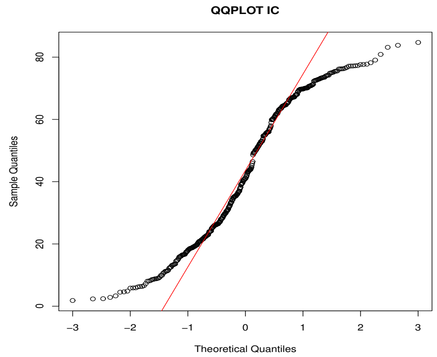

Figure 2.1: QQ plot of IC monthly WPG time series ranging from January 1981 to December 2011.

In this application, when fitting the beta GAS model, the same lag structure was used for the three WPG time series, according to (2-2). Estimated parameters are presented in Table 2.1. Dummy variables were considered to remove the effect of outliers. To do that, an explanatory variable was considered with $X_{i,t} = 1$ for those periods whose residuals were considered outliers and $X_{i,t} = 0$ for all other periods. In our application, an outlier is defined by a residual whose absolute value is larger than three times its standard deviation. For the analysed dataset, only one outlier was found in the IC and EN time series as reported in the last row of Table 2.1.

Parameter	RF		IC		EN	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
A_1	0.336	0.058	0.379	0.049	0.401	0.053
A_2	0.282	0.056	0.155	0.066	0.374	0.068
A_3	0.175	0.072	-0.079^*	0.081	0.083^*	0.057
A_{11}	-0.012^*	0.039	-0.031^*	0.043	0.023^*	0.038
A_{12}	-0.146	0.038	0.048^*	0.050	0.213	0.053
B_1	0.127^*	0.105	0.638	0.098	-0.055^*	0.092
B_2	-0.207	0.063	0.516	0.053	0.418	0.068
B_3	0.574	0.073	-0.311	0.074	0.240	0.058
B_{11}	0.398	0.046	0.389	0.043	-0.383	0.049
B_{12}	-0.197	0.076	-0.426	0.041	0.400	0.071
ω	0.349^*	0.203	0.229	0.063	0.357	0.143
α	3.448	0.250	3.705	0.273	2.855	0.207
$\phi_{t=201}^{(Sep,1998)}$	—	—	-0.899^*	0.453	-0.601^*	0.490

SE stands for the estimated standard errors for each parameter and * stands for non-statistically significant values at 0.05 a significance level. We refer to (5) for further details.

Table 2.1: Maximum likelihood estimation of the beta-GAS model applied to three Brazilian wind farms.

In the following, to investigate the correct model specification, a fully detailed analysis based on quantile residuals was undertaken and reported in Table 2.2. A Jarque–Bera test for normality (in which, under the correct specification of the beta density, the residuals should be normally distributed), a Ljung–Box test for the absence of autocorrelation, and a Ljung–Box on the squared residuals to check for ARCH effects were considered (both were conducted using until lag 30) (53, 54). As reported in Table 2.2, there were no rejections of the null hypothesis in any of the tests. Hence, it can be concluded that the proposed beta GAS models are adequate to describe the three WPG time series.

In this study, the standard SARIMA model (23, 55) was used as benchmark. Table 2.3 presents, for both beta GAS(12,12) and SARIMA

Test	RF	IC	EN
Normality	0.731	0.272	0.075
Autocorrelation	0.604	0.836	0.641
ARCH effect	0.477	0.145	0.390

Table 2.2: p values of standard diagnostic tests.

models, some measures of out-of-sample forecasting accuracy by considering a forecasting horizon of 36 months. In all three measures of accuracy the beta GAS(12,12) has been shown to be superior to the SARIMA model. In addition to better point forecasts as given by the conditional density h steps ahead mean, GAS models have much to offer in this context. First, they deliver a sound conditional density, which respects the support of the variable being modeled. Given that, extreme quantiles can be better estimated through GAS models. Second, through the use of a Student t copula, the spatial dependence among the WPG time series is duly captured (and generally located at different geographical areas).

Model	Fit	RF	IC	EN
beta GAS(12,12)	RMSE	6.438	8.976	8.789
	MAE	5.527	7.573	7.358
	Pseudo R^2	0.676	0.856	0.773
SARIMA	RMSE	8.473	10.391	11.746
	MAE	6.850	9.059	10.218
	Pseudo R^2	0.460	0.812	0.677

Table 2.3: Forecasting evaluation between beta GAS(12,12) and SARIMA models 36 months ahead.

In what follows, we will present the results of the multivariate modeling from the fitting of a dynamic Student t copula to the three WPG time series as described in the steps 1–4 of Part II of Section 2.3.2. The estimated values of the copula parameters are displayed in Table 2.4 and include their degrees of freedom, which have been estimated by using the profile likelihood. The estimated value of $\hat{\nu} = 340$ indicates that the Student t copula can be fairly well approximated by a Gaussian copula.

From the third column of Table 2.4 (p-value), one can conclude that

Parameter	t-GAS		
	Estimate	S.E.	p-value
Ω	1.674	0.356	<0.001
Π	0.143	0.005	<0.001
Υ	0.994	0.001	<0.001

Table 2.4: Maximum likelihood estimation of the second-step GAS model applied to the PIT variables of three Brazilian wind plants.

all parameters are statistically significant, at 5%, or less. In particular, the significant value of the parameter Π plays a special role here, because it gives empirical support for the adoption of a time-varying correlation matrix among the three WPG time series. In Figure 2.2, the estimated (filtered) correlation processes is shown. Interestingly, note that the correlation pattern between IC and RF exhibits a relevant drop between 1990 and 1995, which may constitute a significant information for portfolio investment applications such as (33, 17).

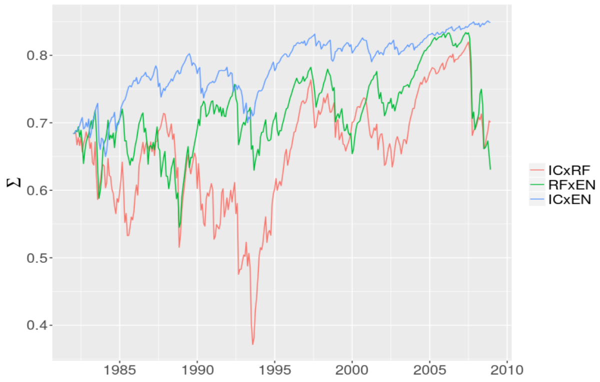

Figure 2.2: Correlation matrix updated by the GAS mechanism for $\{\tilde{y}_{RF,t}, \tilde{y}_{IC,t}, \tilde{y}_{EN,t}\}_{t\in T}$

Following the steps given in Section 2.3.2, 2000 multivariate scenarios were generated for the whole out-of-sample period (the last 36 months of the dataset). To evaluate the quality of the simulated scenarios, we have compared the empirical quantiles, $Q^{(\alpha\%)}\,\forall\,\alpha \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$, of the WPG time series (both in the original scale)[4] with those obtained from the simulated scenarios produced by both the t-GAS and SARIMA models. The results are depicted in Figures 2.3 and 2.4.

[4] The monthly quantiles were calculated by segmenting the WPG time series into 12 monthly time series.

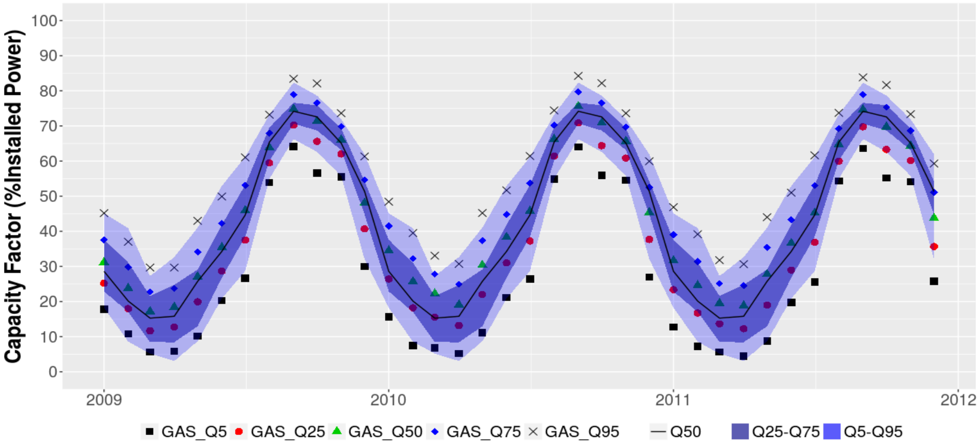

Figure 2.3: Scenario evaluation of IC, for years 2009, 2010 and 2011, through the beta t-GAS model against the real data set. Limits of dashed areas represent the quantiles 5%, 25%, 75%, and 95% of the real data while the black solid line shows the median.

From these figures, one can conclude that the quantiles obtained from scenarios produced by the t-GAS model are closer to the historical quantiles when compared with those obtained by the SARIMA model. In addition, it is worth mentioning that the SARIMA model generated ~1.24% of negative scenarios for IC WPG and 0.061% of scenarios above 100%, which is the physical limit of production.

To quantify this findings, we evaluate the distance between historical quantiles and those obtained by both GAS and SARIMA models. We used a simple forecasting accuracy metric such as the mean absolute percentage error (MAPE) defined as

$$\text{MAPE}^{\alpha} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{q_t^{\alpha} - \hat{y}_t^{\alpha}}{q_t^{\alpha}}\right|, \qquad (2\text{-}17)$$

where q_t^{α} is the α-quantile from historical data and $\hat{y}_t^{\alpha}$ is the quantile from the scenarios produced by the competing models. Such results are presented in Table 2.5.

The results of Table 2.5 indicate that the scenarios generated by our proposed framework, especially on the extreme quantiles, are well estimated. Such outcome was expected due to the fact that the copula model acts mainly

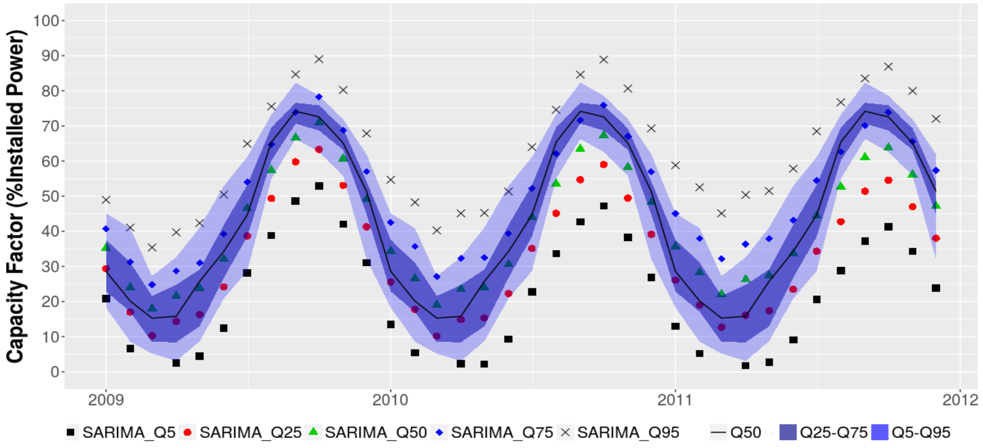

Figure 2.4: Scenario evaluation of IC, for years 2009, 2010 and 2011, through the SARIMA model against the real data set. Limits of dashed areas represent the quantiles 5%, 25%, 75%, and 95% of the real data while the black solid line shows the median.

	GAS			SARIMA		
α	RF	EN	IC	RF	EN	IC
5%	0.1020	0.1163	0.1471	0.2417	0.3788	0.4030
25%	0.0877	0.1220	0.1860	0.1338	0.3740	0.2263
50%	0.0615	0.0509	0.1005	0.0995	0.2388	0.1692
75%	0.0346	0.0518	0.0523	0.0867	0.1826	0.1136
95%	0.0317	0.0296	0.0466	0.1140	0.1534	0.1583

Table 2.5: Mean absolute errors.

on the behaviour of the extreme quantities of the proposed joint density.

2.5 Conclusion

In this paper, we proposed a framework to simulate joint (multivariate) non-Gaussian scenarios of wind power generation (WPG) time series taking into account the spatial dependence among different units. Such scenarios are highly demanded for the development of many studies in power-system planning, operations, and generation investment under the uncertainty of renewable energy generation. The proposed framework was derived by making use of a recently introduced class of time-series models with time-varying parameters and arbitrary non-Gaussian distributions, known as Generalized Autoregressive Score (GAS) models.

Our framework is customized for the range of values that WPG time series can assume, allowing the user to select a distribution that better fits the data. The proposed copula scheme made it possible to capture the spatial dependence among the WPG time series, producing a new tool for the power system community to generate joint scenarios taking into account the time-varying nature of the copula (correlation) parameters. It is worth emphasizing that the proposed time-varying copula model is an independent model and can be used to account for spatial dependencies regardless of the univariate scenario-generation methodology. Hence, it constitutes a broader contribution to the subject of renewable energy forecasting that can be explored in future works. Finally, our framework delivers accurate extreme scenarios for WPG time series (accounting for seasonal effects in both correlation and average), which are of great importance in risk analysis for power system applications such planning, operations, and energy trading.

3
Estimation, forecasting and anomaly detection for nonstationary streams using adaptive estimation

Keywords: Adaptive filtering, Data stream, Time-varying sparsity, Forgetting factor

Abstract: Streaming data provides substantial challenges for data analysis. From a computational standpoint, these challenges arise from constraints related to computer memory and processing speed. Statistically, the challenges relate to constructing procedures that can handle so-called *concept drift* – the tendency of future data to have different underlying properties to current and historic data. The issue of handling structure, such as trend and periodicity, remains a difficult problem for streaming estimation. We propose *RAC* (Real-Time Adaptive Component), a penalized-regression modelling framework which satisfies the computational constraints of streaming data, and provides capability for dealing with concept drift. At the core of the estimation process are techniques from adaptive filtering. The RAC procedure adopts a specified basis to handle local structure, along with a LASSO-like penalty procedure to handle over-fitting. We enhance the RAC estimation procedure with a streaming anomaly detection capability. Experiments with simulated data suggest the procedure has merits for a variety of scenarios, and an illustration with real cyber-security data further demonstrates the promise of the method.

3.1 Introduction

Streaming data – an unending sequence of data values arriving at high frequency – is becoming ubiquitous due to advances in data acquisition technology (56, 57). There is a clear demand for the development of streaming statistical methods, considering applications in diverse areas such as cyber-security (2), finance (58), fraud detection (1) and structural health monitoring (10). Such data brings significant challenges, related to both demands arising from sequential computation and the design of suitable estimators (59, 60).

An outstanding, open, problem relates to handling structure, such as trend or seasonality, in the data stream. For the batch case, there is a large arsenal of time series and related tools available to address such issues. However, in the streaming case the data is revealed sequentially, with the risk that statistical properties of the data may vary over time. This is known as concept drift (56), which invalidates the use of methods that assume various modes of stationarity. The contribution of this paper is to develop a forecasting procedure and associated *local* anomaly detector, capable of dealing with the many challenges of streaming data.

In (61) was considered the objective of characterising and forecasting an *arbitrary* streaming data sequence. These authors made use of a partially observed Markov process, where the evolution of the latent state is governed by a continuous-time Markov process, which allows modelling of irregularly spaced observations. The justification for irregular spacing arises from the sampling frequency of sensors which are constantly interrupted and re-started. (61, Sec. 3) described a very elegant way to combine models, to capture the dynamics of all possible latent state variables that constitute the underlying structure of the stream. As an example, they used a composition of a Negative Binomial model with two seasonal models, considering daily and weekly seasonality effects respectively, to fit vehicle traffic monitoring data. However, a potential shortcoming of their approach, is that the analyst must use prior knowledge, for instance to specify seasonal components (daily and weekly). This means that any behaviour outside such specification cannot be accommodated. This potentially limits the application of the method for tracking an infinite data stream subject to random fluctuations and concept drift.

One approach which can, in principle, cope with arbitrary structure in the data stream is based on estimation with Adaptive Forgetting Factors (AFF) e.g. (62, 9, 63). The use of stochastic gradient descent (9) to update a Forgetting Factor (FF) enables models to handle arbitrary changes in the data generating process. The FF is intended to down weight historic data in the estimation process. To cope with the concept drift, we will utilise adaptive estimation methods in a number of ways, with the intention of constructing a streaming regression model for a univariate response, where the explanatory variables can be regarded as local auto-regressive terms. This model is well suited to the streaming context, in terms of data storage and computational requirements, and offers a choice of basis for the auto-regressive regression.

The proposed method, which we refer to as RAC (Real-time Adaptive Component), relies on a penalised streaming regression-based framework. A simple form of the streaming setting, described in (64), for a linear regression model is

$$y_t = x_t^{'} \beta_t + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$ are independent and identically distributed (*i.i.d.*) Gaussian random variables with known variance σ^2. Data processing typically proceeds as follows: Consider the time step t, acquire basis vector $x_t \in \mathbb{R}^d$ and use it to forecast $\hat{y}_t \in \mathbb{R}$, the one step ahead forecast, using the sequentially estimated weight vector $\hat{\beta}_t \in \mathbb{R}^d$. Later, with the acquisition of the true value y_t, $\hat{\beta}_t$ is updated to $\hat{\beta}_{t+1}$. In our case, the basis vector represents lagged and transformed values of the response.

In the streaming context, the choice of basis for the regression, or equivalently, the set of transformations and lagged variables is critical for successful forecasting, and hence anomaly detection. However, since streaming data is subject to unpredictable temporal variation, the construction of a suitable basis is challenging. There is the usual problem of ensuring the model has sufficient complexity to capture underlying structure, while not affording the opportunity for over-fitting.

This work proposes the use of a relatively large basis (the maximum dimension defined by computational constraints), and then appeals to sparsity inducing approaches to manage over-fitting. We use a scheme based on the Least Absolute Shrinkage Operator (LASSO) technique, proposed by (65) and further detailed in (66, 67). A vast literature showing that the LASSO attains good performance under various assumptions on the basis is available (see (68, 69, 70, 71, 72, 73) and references therein). In our context, the appealing aspect of the LASSO is that it induces sparsity, which will provide a means for managing over-fitting. To achieve this, we require a method for sequentially determining the LASSO penalisation parameter. This issue was addressed in (74). They proposed an adaptive extension of a LASSO Vector Autoregressive (VAR) model to perform hourly wind power forecasts considering several wind farms. Similar to AFF, the authors also used FF to handle the nonstationary of the signal, albeit in their version it is a fixed quantity. Unfortunately, despite the autoregressive coefficients of the VAR being updated sequentially, the penalty term is not. By default, these models are fitted using a grid of penalty parameters which are computationally unfeasible in streaming data context. Similarly, (75) proposed an updating rule to the penalty term designed to estimate the parameters as soon as new data arrives and assuming the underlying distribution is nonstationary. Similar to AFF, (75) framework also adopts adaptive filtering estimation based on stochastic gradient descent.

In this work we extend the results of (75) and deploy them in the context of a penalised streaming regression model, to provide a temporally-adaptive estimation procedure and corresponding anomaly detection tool.

In summary, this work has two main contributions. First we extend (75) to perform forecasting and anomaly detection, using a basis construction strategy, that accurately tracks a time-varying quantity of interest. Second, we developed a new method to sequentially perform anomaly detection. This is achieved using a streaming method based on an approximation of a sum of weighted *i.i.d.* chi-squared random variables (76).

3.2 Methodology

In this section the basic components of RAC are introduced. Specifically the LASSO, adaptive estimation, optimization procedures, basis construction and extension to anomaly detection are discussed.

3.2.1 The LASSO procedure

The batch LASSO estimator (65) was initially proposed as a variable selection procedure. Considering the pair (X, y), where y denotes a T-dimensional response vector and X be a $T \times d$ basis, with rows composed by the vectors $x_t \in \mathbb{R}^d$, define the simple linear regression model

$$y = X^{'}\beta + \varepsilon, \tag{3-1}$$

with weight vector $\beta \in \mathbb{R}^d$ and ε being *i.i.d.* Gaussian random variables with known variance σ^2. Then the LASSO estimator is defined as

$$\hat{\beta}(\gamma) = \arg\min_{\beta} \sum_{t=1}^{T} (y_t - x_t^{'}\beta)^2 + \gamma||\beta||_1, \tag{3-2}$$

where $\gamma \geq 0$ is the penalty parameter and $||\cdot||_1$ denotes the ℓ_1 norm. Note that we write $\hat{\beta}(\gamma)$ to emphasise the dependence on γ in the estimation of β. Denote the set of variables as $\mathcal{J} = \{1, ..., d\}$, define the active set of variables, i.e., which variables are selected, as $\mathcal{A}(\gamma) = \{j \in \mathcal{J} : \hat{\beta}^{(j)}(\gamma) \neq 0\}$, where $\hat{\beta}^{(j)}(\gamma)$ makes reference to the jth element of the vector. In practice the solutions $\hat{\beta}(\gamma)$ are estimated on a grid of γ values, ranging from 0, where no shrinkage is applied, to

$$\gamma^{(max)} = \max_{j \in \mathcal{J}} \left| \frac{1}{T} X_j' y \right|, \tag{3-3}$$

for which all values of $\hat{\beta}(\gamma)$ will be exactly zero, except the intercept and denote the jth column of X as X_j. Selection of the penalty parameter is often made through data reuse methods, for example cross-validation (CV), however this is not feasible for streaming data analysis due to computational speed requirements.

Typically, prior to the estimation, one should first center the columns of the basis X ($\frac{1}{T}\sum_{t=1}^{T} X_{tj} = 0$) and fix unit variance ($\frac{1}{T}\sum_{t=1}^{T} X_{tj}^2 = 1$). This is done to prevent the LASSO solution from depending on the predictor's units of measurement. In addition, the response values y_t are also assumed to be centred ($\frac{1}{T}\sum_{t=1}^{T} y_t = 0$). These centering conditions allows one to omit the intercept term $\beta^{(0)}$ when optimizing (3-2). Given the optimal LASSO solution $\hat{\beta}(\gamma)$ on the centred data, it is possible to recover the optimal solutions for the uncentred data, $\hat{\beta}(\gamma)$ remains the same and the intercept is

$$\hat{\beta}^{(0)} = \bar{y} - \sum_{j=1}^{d} \bar{X}_j \hat{\beta}^{(j)}(\gamma), \tag{3-4}$$

where $\bar{y}$ and $\{\bar{X}_j\}_{j=1}^{d}$ denotes the mean of the referred variables and are calculated in the original scale. In the context of streaming data, computing these values are challenging. Despite having useful properties, the batch LASSO estimator is not feasible in streaming data environment. In the next section we introduce adaptive estimation (9), which can be used to adapt the results of batch LASSO to a streaming data environment, respecting memory and speed constraints.

3.2.2
Adaptive estimation

The task of filtering corresponds to controlling the rate at which past information is discarded while avoiding storing all the data in memory. The most common filtering strategy discards information at a constant rate, fixing the value of FF – denoted here by λ. Adaptive filtering methods do not require the data to be stationary (9). As an alternative to a fixed value of λ, which may be difficult to set, much interest has focused on sequentially selecting an adaptive forgetting factor (AFF) – λ_t, using an updating mechanism based on stochastic gradient descent (9, 63, 62). Such methods are called adaptive because the quantity of data discarded is not constant over time. Particularly, the benefits of using such a strategy are highly relevant in nonstationary environment.

Suppose we have a univariate stream

$$y_1, y_2, ..., y_{T-1}, y_T, ...,$$

then our goal is to accurately estimate the mean at time T. This estimator will be used to detect anomalous behaviour in the stream. One way to estimate the mean of the stream $E[Y_t]$ is

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t. \tag{3-5}$$

This estimator makes sense only if $E[Y_t] = \mu$, a constant for all time points. However, denoting τ^* as the change point instant, if there was a change at $\tau^* < T$ such that

$$E[Y_t] = \begin{cases} \mu^{'},\ t = 1, 2, ..., \tau^* \\ \mu,\ t = \tau^* + 1, \tau^* + 2, ..., T, ... \end{cases} , \mu^{'} \neq \mu$$

the arithmetic mean $\hat{\mu} = \bar{y}_T$ cannot estimate μ accurately if there is a big difference between $\mu^{'}$ and μ. In order to improve the estimation, one could take the mean of those observations that occur only after the change point τ^*,

$$\hat{\mu} = \frac{1}{T - \tau^*}\left[y_{\tau^*+1} + y_{\tau^*+2} + ... + y_{\tau^*+T}\right].$$

However, the point τ^* is unknown, which makes this estimation unfeasible in sequential settings.

Such drawback, motivates the use of adaptive estimation to calculate the current mean process at time t, in which more weight is placed on more recent observations, and do not store all data in memory. Using such methods results in improved estimation of the data stream after the change point τ^* (62). This is achieved by introducing an FF, $\lambda \in (0,1)$, in Equation (3-5) and using a normalizing constant ($w_{t,\lambda}$) to weight the estimation process,

$$\bar{y}_{t,\lambda} = \frac{1}{w_{t,\lambda}} \sum_{i=1}^{t} \lambda^{t-i} y_i, \qquad w_{t,\lambda} = \sum_{i=1}^{t} \lambda^{t-i}.$$

The advantage of this formulation is that it leads to a sequential formulation for streaming contexts by defining the following updating mechanism for $t \geq 1$,

$$m_{t,\lambda} = \lambda m_{t-1,\lambda} + y_t \tag{3-6}$$

$$w_{t,\lambda} = \lambda w_{t-1,\lambda} + 1 \tag{3-7}$$

$$\bar{y}_{t,\lambda} = \frac{m_{t,\lambda}}{w_{t,\lambda}}, \tag{3-8}$$

with $m_{0,\lambda} = w_{0,\lambda} = 0$. Note that setting $\lambda = 0$ corresponds to forgetting all previous observations, and only using the most recent observation, i.e. $\bar{y}_{t,\lambda} = y_t$. On the other hand, $\lambda = 1$ corresponds to no forgetting, and then the forgetting factor mean, $\bar{y}_{t,\lambda}$, is simply the usual arithmetic mean given in Equation (3-5). Note that practical algorithms restrict the range of λ to prevent it becoming too small, see for example (62).

As pointed in (62), the updating mechanism of Equations (3-6)-(3-8) bears some resemblance to the Exponential Weighted Moving Average (EWMA) equations. Indeed, they are related; using the above equations it is possible to rewrite

$$\begin{aligned}\bar{y}_{t,\lambda} &= \left(1 - \frac{1}{w_{t,\lambda}}\right)\bar{y}_{t-1,\lambda} + \frac{1}{w_{t,\lambda}} y_t \\ &= \lambda\left(\frac{1-\lambda^{t-1}}{1-\lambda^t}\right)\bar{y}_{t-1,\lambda} + \left(\frac{1-\lambda}{1-\lambda^t}\right) y_t,\end{aligned}$$

and then if $\lambda \in (0,1)$, as $t \to \infty$, this becomes

$$\bar{y}_{t,\lambda} = \lambda\bar{y}_{t-1,\lambda} + (1-\lambda)y_t,$$

which is equivalent to the EWMA scheme.

The previous updating mechanism extends readily to the linear regression framework of Equation (3-1). To achieve this, we require FF estimates of both mean and covariance of response y_t and basis vector x_t at each time t. In (62, 63, 75) an adaptive estimation framework for both sample mean vector, and sample covariance matrix was used. Define $\Pi_t = \left(y_t, x_t^{(1)}, x_t^{(2)}, ..., x_t^{(d)}\right)' \in \mathbb{R}^{d+1}$ as the data vector, $\bar{\Pi}_{t,\lambda}$ as the sample mean vector and $\Sigma_{t,\lambda} \in \mathbb{R}^{(d+1)\times(d+1)}$ the sample covariance matrix. Considering a fixed FF λ, the sample mean vector is sequentially updated as

$$\bar{\Pi}_{t,\lambda} = \left(1 - \frac{1}{w_{t,\lambda}}\right)\bar{\Pi}_{t-1,\lambda} + \frac{1}{w_{t,\lambda}}\Pi_t, \tag{3-9}$$

with $w_{t,\lambda}$ a normalizing constant defined in Equation (3-7) and $\Pi_{0,\lambda} = (0,0,...,0)'$ a vector of zeros. Note that this is equivalent to applying recursions (3-6)-(3-8) to y_t and each element of the d-dimensional vector x_t individually. Further, the covariance matrix $\Sigma_{t,\lambda}$ is updated as

$$\Sigma_{t,\lambda} = \left(1 - \frac{1}{w_{t,\lambda}}\right)\Sigma_{t-1,\lambda} + \frac{1}{w_{t,\lambda}}(\Pi_t - \bar{\Pi}_{t,\lambda})'(\Pi_t - \bar{\Pi}_{t,\lambda}). \tag{3-10}$$

adopting as $\Sigma_{0,\lambda}$ an identity matrix. The effect of these initial values will vanish when adopting a *burn-in* period of $\mathcal{B}$ observations. Similar to the sequential algorithms proposed in (62, 63, 75), assuming that observations $y_1, ...y_{\mathcal{B}}$ will not face any change in the underlying structure, this set of observations will be used to estimate the initial values of the sequential quantities.

3.2.3 Streaming LASSO

RAC relies on a penalised streaming regression-based framework. However, as pointed out in the introduction, should the stream manifest concept drift then weighted estimation procedures, as described in Section 3.2.2, are clearly appropriate.

The formulation of linear regression in this context would feature coefficients analogous to autoregressive weights in a time series model, called *tap* weights in adaptive filtering (9). Restricting to linear forms for the regression is limited and cannot readily handle trend or seasonality. Thus, we will design bases for streaming regression with the potential to capture these phenomena. These bases are potentially overparametrized and hence streaming penalization methods, adapted from procedures such as LASSO, are required. In this case, the choice of an optimal regularization parameter may itself be time-varying (75). In order to sequentially fit the underlying structure of y_t, a time-varying penalty parameter $\gamma_t \in \mathbb{R}^+$ is introduced (c.f. Equation (3-2)).

The regularization parameter is iteratively updated as

$$\gamma_{t+1} = \gamma_t - \eta_\gamma \frac{\partial \mathcal{C}_{t+1}}{\partial \gamma_t}, \tag{3-11}$$

where $\mathcal{C}_{t+1} = ||y_{t+1} - x'_{t+1}\hat{\beta}_t(\gamma_t)||_2^2$ is the designated cost function to update the stochastic gradient associated with γ_t, while $\eta_\gamma > 0$ is the step size. Here – also in (75) – a quadratic cost function is adopted since the future mean behaviour of y_t is currently being tracked. Nevertheless, other cost functions could also be used. For example, (63) define efficient update equations, based on maximum likelihood, for the exponential family of distributions.

To calculate the derivative $\frac{\partial \mathcal{C}_{t+1}}{\partial \gamma_t}$ from Equation (3-11) using implicit differentiation,

$$\frac{\partial \mathcal{C}_{t+1}}{\partial \gamma_t} = \frac{\partial \mathcal{C}_{t+1}}{\partial \hat{\beta}_t(\gamma_t)} \frac{\partial \hat{\beta}_t(\gamma_t)}{\partial \gamma_t}. \tag{3-12}$$

While the first term of the right hand side of Equation (3-12) is straightforward to obtain by direct differentiation, the second is more involving. (77) show that for the LASSO, under a squared error loss function and ℓ_1-norm of β, the optimal coefficient path is piecewise linear, which implies that $\partial \hat{\beta}_t(\gamma_t)/\partial \gamma_t$ is piecewise constant. A closed-form solution for this derivative, adapted from (78), is presented in (75, Proposition 1) as

$$\frac{\partial \hat{\beta}_t(\gamma_t)}{\partial \gamma_t} = -(x_t^{'} x_t)^{-1} \text{sign}(\hat{\beta}_t(\gamma_t)) \tag{3-13}$$

$$= -(\Sigma_{t,\lambda})^{-1} \text{sign}(\hat{\beta}_t(\gamma_t)). \tag{3-14}$$

This result is equivalent to calculating the gradient of the LASSO solution as suggested by the LARS formulae in (77, Equations (2.4) - (2.6)).

In addition, one should note that similar to the LARS gradient update, Equation (3-14) is only nonzero over the active set $\mathcal{A}_t(\gamma_t) = \{j \in \mathcal{J} : \hat{\beta}_t^{(j)}(\gamma_t) \neq 0\}$ of regression weights, and zero elsewhere. Note that the subscript t was added to emphasize that this is a time-varying active set. Therefore, at each update of $\partial \hat{\beta}_t(\gamma_t)/\partial \gamma_t$, one should consider the two scenarios

- Non-empty active set, $\mathcal{A}_t(\gamma_t) \neq \emptyset$, which in this case, as proved in (75) Equation (3-14) is well-defined;
- The active set is empty, $\mathcal{A}_t(\gamma_t) = \emptyset$, then both algorithms, LARS and the one of (75), take a step in the direction of the most correlated predictor $\hat{j} = \underset{j \in \mathcal{J}}{\arg\max} \left\{ | \sum_{t=1}^{T} y_t x_t^{(j)} | \right\}$. Hence define the gradient as

  $$\left(\frac{\partial \hat{\beta}_t(\gamma_t)}{\partial \gamma_t} \right)^{(l)} = \delta_{\hat{j}}^{(l)} \text{sign} \left(\sum_{t=1}^{T} y_t x_t^{(l)} \right),$$

  where δ is the Kronecker delta function. All entries of $\partial \hat{\beta}_t(\gamma_t)/\partial \gamma_t$ will be zero with exception of the corresponding to the most correlated predictor.

Note that one could also include an adaptive forgetting factor for the parameter estimates, Π_t and Σ_t, which can be concurrently updated with γ_t just by calculating

$$\lambda_{t+1} = \lambda_t - \eta_\lambda \frac{\partial \mathcal{L}_{t+1}}{\partial \lambda_t},$$

where $\mathcal{L}_{t+1}$ can be the squared loss as assumed in (62). However, non reported experiments suggests that the forecasting performance is dramatically degraded when compared to the fixed FF approach. This is related to the fact that λ_t and γ_t interact, as described in (74). The penalization parameter may increase to adapt the regression to, for example, a new regime while the AFF value will be reduced to give weight to only recent observations. Both parameters updates are attempting to adapt the model to the new regime. Therefore, we opt to keep λ fixed and update only γ_t as a time-varying quantity in RAC framework.

Adopting the concepts of adaptive filtering discussed in Section 3.2.2, both Equations (3-9) and (3-10) are suitable for streaming data as they require storing only a few parameters and data points in computer memory, instead of all historical values. These concepts of adaptive filtering provide grounds to propose sequential updates for Equations (3-3) and (3-4), which represent the maximum value of the penalty parameter and the intercept, respectively. Considering the operators $\max\{\cdot\}$ and $\min\{\cdot\}$ that returns the maximum and minimum values, the maximum value of the penalty can be defined as

$$\gamma_t^{(\max)} = \max\left\{\left|\bar{\Pi}_{t,\lambda}^{(1)}\bar{\Pi}_{t,\lambda}^{(2)}\right|, \left|\bar{\Pi}_{t,\lambda}^{(1)}\bar{\Pi}_{t,\lambda}^{(3)}\right|, ..., \left|\bar{\Pi}_{t,\lambda}^{(1)}\bar{\Pi}_{t,\lambda}^{(d)}\right|\right\}, \qquad (3\text{-}15)$$

with $|\cdot|$ denoting the absolute value. Moreover, to sequentially ensure that $\gamma_t \in [0, \gamma_t^{(\max)}]$, the following rule must be adopted after the update of both γ_t and $\gamma_t^{(\max)}$,

$$\gamma_t = \max\{\min\{\gamma_t, \gamma_t^{(\max)}\}, 0\}.$$

Regarding the intercept, Equation (3-4) is rewritten in terms of the elements in $\bar{\Pi}_{t,\lambda}$,

$$\hat{\beta}_t^{(0)}(\gamma_t) = \bar{\Pi}_{t,\lambda}^{(1)} - (\bar{\Pi}_{t,\lambda}^{(2)}, ..., \bar{\Pi}_{t,\lambda}^{(d)})' \hat{\beta}_t(\gamma_t), \tag{3-16}$$

where different from Equation (3-2), the subscript t denotes that the updating of $\hat{\beta}_t(\gamma_t)$ is sequential and $\bar{\Pi}_{t,\lambda}^{(j)}$ makes reference to the jth element of the vector.

3.2.4 Cyclic coordinate descent

Having defined the update of the penalty parameter, the next step is to calculate the values of $\hat{\beta}_t(\gamma_t)$ sequentially conditional to γ_t. Unfortunately, there is no analytical solution to find the minimum in Equation (3-2). Several approaches are already available to optimize such problems, as pointed out in (75, 74, 66), when updating multiple regression weights sequentially, however the most appropriate is the Cyclical Coordinate Descent (CCD) algorithm (79, 80). The main advantage of CCD is computational efficiency, exploring an analytic expression for Equation (3-2) taking into account a partial optimum conditional to one specific weight, while the others remain fixed. Hence the name *cycles*, because the analytical expression is considered to update all the weights successively until convergence is reached.

As described by (66, Sec. 2.4.2), the algorithm will propose an arbitrary order for the predictors and cycle trough them. In RAC, CCD is adopted to update the regression weights using the columns of $\Sigma_{t,\lambda}$. At each step j the weights $\hat{\beta}_t^{(j)}(\gamma_t)$ are updated by minimizing the analytic expression for Equation (3-2) in this coordinate, maintaining the values of remaining variables $\beta_t^{(l)}(\gamma_t), l \neq j$ fixed. To remove the effect of the other variables, CCD makes use of a partial residual $r_t^{(j)} = \Sigma_{t,\lambda}^{(1)} - \sum_{l \neq j} \Sigma_{t,\lambda}^{(l)} \hat{\beta}_t^{(l)}(\gamma_t)$.

Hence, the update is given by

$$\hat{\beta}_t^{(j)}(\gamma_t) \leftarrow S\left(\hat{\beta}_t^{(j)}(\gamma_t) + \left\langle \Sigma_{t,\lambda}^{(j)}, r_t^{(j)} \right\rangle, \gamma_t\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $S(\cdot)$ makes reference to the soft threshold operator defined by

$$\begin{aligned} S(z,\gamma) &= \operatorname{sign}(z)(|z|-\gamma)_+ \\ &= \begin{cases} z-\gamma, & \text{if } z>0 \text{ and } \gamma<|z| \\ z+\gamma, & \text{if } z<0 \text{ and } \gamma<|z| \\ 0, & \text{if } \gamma>|z|, \end{cases} \end{aligned}$$

with $(\cdot)_+$ denoting the positive part.

Similar to the update proposed in (74) for the VAR coefficients, the calculations of $\hat{\beta}_t(\gamma_t)$ in RAC also involves a FF λ, used in Equations (3-9) and (3-10).

3.2.5 Basis construction

At the core of our method is an autoregressive regression with basis designed to capture local structure, overlayed with LASSO-inspired complexity control to prevent overfitting. Of course, there are many approaches to constructing a suitable basis which embodies such features. The features of particularly interest relates to trend and simple curvature.

The two proposed specifications of basis X in this section are designed to capture the streaming local structure. Such bases try to mimic several well known time series unobserved components, namely trend and seasonality.

3.2.5.1

Trig basis

Consider the basis based on a Fourier coefficient expansion for $X = [\psi^{\sin}\ \psi^{\cos}]$, with

$$\psi^{cos} = \begin{bmatrix} | & | & \cdots & | \\ \cos\omega_1 t & \cos\omega_2 t & & \cos\omega_j t \\ | & | & \cdots & | \end{bmatrix} \quad \psi^{sin} = \begin{bmatrix} | & | & \cdots & | \\ \sin\omega_1 t & \sin\omega_2 t & & \sin\omega_j t \\ | & | & \cdots & | \end{bmatrix},$$

where the frequencies $\omega_1, ..., \omega_j$ are defined by the user. This is closely related to the so called ℓ_1 trend filtering from (81) where the authors proposed a slight variation on the Hodrick-Prescott filter.

This basis is strongly recommended when the local structure of the stream has smooth curvatures, typical of a stationary periodic signal. For both simulations and real data application, the fixed quantities were defined as $\omega_j = e^{-2\pi} + 0.2(j-1)\ \forall j \in \mathcal{J}$ as an attempt to mimic a Fourier transform.

3.2.5.2 Cycle basis

Define the vector

$$\Lambda_{(n,i)} := ((1/n)^i, (2/n)^i, ..., (n/n)^i)',$$

and for some $\upsilon \in \mathbb{R}^n$, denote the indefinite concatenation operator of the elements of υ by

$$\upsilon^R = (\upsilon_1, \upsilon_2, ..., \upsilon_n, \upsilon_1, \upsilon_2, ..., \upsilon_n, \upsilon_1, \upsilon_2 ...)'.$$

Let Ξ_{min} and Ξ_{max} denote the minimum and maximum sequence lengths, respectively. The basis can be defined as the concatenation of the vectors

$$
\begin{aligned}
X = [&\Lambda^R_{(\Xi_{min},1)} \; \Lambda^R_{(\Xi_{min},2)} \\
&\Lambda^R_{(\Xi_{min+1},1)} \; \Lambda^R_{(\Xi_{min+1},2)} \\
&\vdots \\
&\Lambda^R_{(\Xi_{max-1},1)} \; \Lambda^R_{(\Xi_{max-1},2)} \\
&\Lambda^R_{(\Xi_{max},1)} \; \Lambda^R_{(\Xi_{max},2)}],
\end{aligned}
$$

where the number of columns in X is a function of the quantities Ξ_{min} and Ξ_{max}. The user needs to specify Ξ_{min} and Ξ_{max}, the minimum and maximum sequence lengths in the columns of X. To avoid scaling problems, the jth column X_j is scaled by its maximum value. As a consequence, all columns of X are scaled to the interval $[0, 1]$. After this scaling, the X matrix contains in the first column the sequence from 1 to Ξ_{min}, the second column is the square of the first, and so on until the last sequence from 1 to Ξ_{max} and its square.

For illustration purposes, consider a small example with stream observed up to time $T = 8$ and where the user chooses $\Xi_{min} = 6$ and $\Xi_{max} = 7$, the *Cycle* basis X used in this case is

$$
X = \begin{bmatrix}
1/6 & 1/36 & 1/7 & 1/49 \\
2/6 & 4/36 & 2/7 & 4/49 \\
3/6 & 9/36 & 3/7 & 9/49 \\
4/6 & 16/36 & 4/7 & 16/49 \\
5/6 & 25/36 & 5/7 & 25/49 \\
6/6 & 36/36 & 6/7 & 36/49 \\
1/6 & 1/36 & 7/7 & 49/49 \\
2/6 & 4/36 & 1/7 & 1/49
\end{bmatrix}.
$$

A promising feature of this basis is the possibility to capture piecewise trends and quadratic structures caused by seasonal fluctuations. To avoid an extensive discussion over the choices of these hyperparameters, henceforth in all sections, $\Xi_{min} = 6$ and $\Xi_{max} = 40$. These choices are arbitrary, selected for illustration. In practice we may set the latter to a large value based on available processing resources.

3.2.6 Anomaly detection using a weighted sum of chi-squared random variables

To perform anomaly detection with a conditional model such as the RAC, the most straightforward approach is to look for anomalous values in the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, ...$ at each time t. Intuitively, if the stream is well behaved, the model should be able to fit the local structure of $y_1, y_2, ...$, which will result in residuals close to zero. On the other hand, if the stream is poorly behaved, the residuals will exhibit anomalous values.

After estimation of $\hat{\beta}_t(\gamma_t)$ and the actual value y_{t+1} is observed, the residual one-step ahead forecasting error is

$$\hat{\varepsilon}_{t+1} = \hat{y}_{t+1} - y_{t+1}$$

where $\hat{y}_{t+1} = x'_{t+1}\hat{\beta}_t(\gamma_t)$ denotes the one-step ahead forecast. In the streaming context this generates an unending sequence of residuals, upon which we will make inference on the quantity

$$\xi_{t+1} = \left(\frac{\hat{y}_{t+1} - y_{t+1}}{\sqrt{\phi_{t+1}}}\right)^2,$$

where ϕ_{t+1} is a scalar value associated with the variance of y_t, i.e., the element in the position $[1,1]$ of $\Sigma_{t+1,\lambda}$ matrix (see Equation (3-10)). Assuming *i.i.d.* data, this standardised quantity behaves as

$$\left(\frac{\hat{y}_{t+1} - y_{t+1}}{\sqrt{\phi_{t+1}}}\right) \sim N(0,1) \implies \xi_{t+1} \sim \chi^2_1.$$

Since we are tracking a moving target it is convenient to estimate the squared residual sequence using a forgetting mechanism, as described earlier

(Section 3.2.2). This also provides a better means for calibration of the residuals, by computing

$$\kappa_{t,\theta} = \theta\kappa_{t-1,\theta} + \xi_t \tag{3-17}$$

$$\nu_{t,\theta} = \theta\nu_{t-1,\theta} + 1 \tag{3-18}$$

$$\bar{\xi}_{t,\theta} = \frac{\kappa_{t,\theta}}{\nu_{t,\theta}}, \tag{3-19}$$

and using the quantity $\bar{\xi}_{t,\theta}$ as the object for inference. Of course, this is the same recursive formulation as Equations (3-6) - (3-8). The random formulation of this quantity is a weighted mixture of chi-square distributions, described by (76), does not have closed form but can be well approximated in streaming contexts by the Hall-Buckley-Eagleson method (HBE) (82). Using such result, a sequential anomaly detector is constructed using the HBE method. The user here needs to control the value of θ, which states how many observations will be averaged to detect a change at each time t. This is essentially a second stage of smoothing, using a fixed FF in Equations (3-17)-(3-19). Our experiments suggest that performance is robust for $0.9 < \theta < 1$.

To evaluate if there is an anomaly at each time t, the p-value is computed as

$$F_{HBE}(\bar{\xi}_{t,\theta}, \xi_t)$$

where $F_{HBE}(\cdot)$ is the cumulative distribution function (cdf) of a positively-weighted sum of chi-squared random variables using the HBE method with coefficient vector $\bar{\xi}_{t,\theta}$ evaluated at quantile ξ_t.

3.2.6.1 False positives

Having developed an anomaly detection method, it is worthwhile to consider if the expected number of false positive is close to the observed one. To address this aspect of detection performance, 100 replicates of length 1000 from the following data generation processes (DGP) were used,

$$y_t = 6\cos\left(2\pi\frac{1}{24}t + 0.1\right) + \varepsilon_t, \text{ with } \varepsilon_t \sim N(0,4). \tag{3-20}$$

Noting that there are no changes in this process, the results explore the average number of false positive for different levels of significance, which we denote as $\alpha \in \{0.001, 0.01\}$. A burn-in period of $\mathcal{B} = 100$ observations was adopted. The results for the average number of observed false positive are displayed in Table 3.1 using the *Cycle* basis. This table suggests that, for well chosen η_γ and λ the expected number of false positive is well-calibrated with respect to the selected significance level, which gives support for the use of RAC for anomaly detection.

α	$(\eta_\gamma.\lambda)$	0.6	0.7	0.8	0.9	0.95	0.995
0.001	0.1	5.59	5.23	2.36	0.93	0.23	0.01
	0.01	3.91	3.40	1.68	0.97	0.19	0.00
	10^{-3}	1.80	1.46	0.93	0.58	0.15	0.00
	10^{-4}	0.46	0.28	0.33	0.53	0.18	0.01
	10^{-5}	0.14	0.11	0.13	0.18	0.11	0.06
	10^{-6}	0.16	0.10	0.15	0.08	0.11	0.02
0.01	0.1	14.04	15.06	9.14	7.33	5.11	2.67
	0.01	10.05	10.13	6.34	5.17	4.41	2.81
	10^{-3}	7.47	6.42	4.17	4.47	3.62	2.56
	10^{-4}	6.17	5.51	5.20	4.98	4.26	2.78
	10^{-5}	6.55	6.35	6.27	4.44	4.39	2.96
	10^{-6}	6.78	7.43	6.73	4.22	4.41	2.72

Table 3.1: Monte Carlo estimates of the observed number of false positive using the method in Section 3.2.6 over 100 replicates of the DGP in Equation (3-20), using the *Cycle* basis.

3.2.7 Algorithm

An illustration on how RAC is sequentially performed is presented in Algorithm 1.

3.3 Simulation study

In this section we assess the performance of RAC in two respects, *Estimation* and *Detection*. The first is concerned with the method's forecasting ability to track an arbitrary signal with underlying structure varying over time. The second is concerned with the detection capabilities of the method when

Algorithm 1 Real-Time **A**daptive **C**omponent

```
Require: $\lambda \in [0.6, 1)$, $\eta_\gamma \in \mathbb{R}^+$, $\theta \in [0.9, 1)$, $\gamma_0$, $\mathcal{B}$
for $t \leftarrow 1, ..., t, ...$ do
    Receive $(y_t, x_t^{'})$
    Update $\Pi_{t,\lambda}$ and $\Sigma_{t,\lambda}$
    With $\hat{\beta}_t(\gamma_t)$ calculated at time $t-1$
        $\beta_{0t}(\Pi_{t,\lambda}, \hat{\beta}_t(\gamma_t))$
        $\hat{y}_t = \beta_{0t} + x_t^{'}\hat{\beta}_t(\gamma_t)$
        $\varepsilon_t = y_t - \hat{y}_t$
        $p$-value($\varepsilon_t^2$, $\theta$)
    Update penalty term $\gamma_{t+1}$
        $\frac{\partial \mathcal{C}_{t+1}}{\partial \gamma_t}(x_t^{'}, \Sigma_{t,\lambda}, \hat{\beta}_t(\gamma_t), \varepsilon_t)$
        $\gamma_t^{(max)}(\Pi_{t,\lambda})$
        Calculate $\gamma_{t+1}$
    Update regression weights $\hat{\beta}_{t+1}(\gamma_{t+1})$
        Calculate $\hat{\beta}_{t+1}(\gamma_{t+1})(\gamma_{t+1}, \Sigma_{t,\lambda})$
```

facing a change in the signal's underlying structure. RAC will be compared to other methods discussed in the literature, such as AFF (62) and online VAR[1] (74). A simple benchmark, which we will refer to as the NAIVE benchmark is based on using y_t as the forecast $\hat{y}_{t+1}$ is also included.

Our experiments are based on the data generated by the following process

$$y_t = \begin{cases} 2\cos\left(2\pi\frac{1}{100}t + 0.1\right) + \varepsilon_t, & \text{if } t \leq 5000 \\ 10\cos\left(2\pi\frac{1}{200}t + 0.1\right) + \varepsilon_t, & \text{if } 5001 \leq t \leq 10000 \end{cases} \tag{3-21}$$

where $\varepsilon_t \sim N(0,1)$. The results obtained in Sections 3.3.1 and 3.3.2 are based on adopting $\mathcal{B} = 1000$ observations, $\eta_\gamma \in \{0.1,\ 0.01,\ 10^{-3},\ 10^{-4},\ 10^{-5},\ 10^{-6}\}$ and $\lambda \in \{0.6,\ 0.7,\ 0.8,\ 0.9,\ 0.95,\ 0.995\}$.

3.3.1 Estimation performance

A straightforward way to evaluate prediction performance is averaging forecasting error metrics, such as Mean Square Error (MSE) and Mean Absolute Error (MAE), over the 100 replicates for all methods. These metrics

[1] Note that here, as we are only considering one stream at a time, the VAR model is actually a univariate autoregressive one.

are defined as

$$MSE = \sum_{t=\mathcal{B}+1}^{T} \frac{1}{T-\mathcal{B}} (\hat{y}_{t+1} - y_{t+1})^2, \tag{3-22}$$

$$MAE = \sum_{t=\mathcal{B}+1}^{T} \frac{1}{T-\mathcal{B}} |\hat{y}_{t+1} - y_{t+1}|, \tag{3-23}$$

where T is the total length of the stream.

The averaged forecasting metrics are displayed in Table 3.2. It is clear that RAC can provide good forecasting performance provided the control parameters are selected appropriately. Notably, both NAIVE benchmark and AFF have reasonably good performances. However, its performance for anomaly detection is low, pointing to the dichotomy between forecasting accuracy and anomaly detection capability. Considering the online VAR model of (74), the computation became infeasible for $\lambda < 0.9$ because their method is not appropriate for streaming data. Also, the values of MSE and MAE for AFF and NAIVE are displayed in only one column. Regarding AFF, the analyst just chooses η_λ, the step of gradient descent, to update the forgetting factor λ_t, while NAIVE no choices are needed. Figures 3.1 and 3.2 show the averages of penalty term γ_t, p-value and number of non zero coefficients for $\hat{\beta}(\gamma_t)$ using *Cycle* and *Trig* basis respectively. These plots illustrates the combination of η_γ and λ that minimises the mean square error (MSE) of one step ahead forecasting error.

Some observations about Figures 3.1 and 3.2. First, RAC is able to continuously track an arbitrary target after its latent underlying structure faces a change. Second, around the change point, $\tau^* = 5000$, the average penalty parameter, γ_t, increases as the average number of non zero weights decreases. This is expected, because after the change point, RAC will re-estimate new weights to filter the new underlying structure of the signal. This is consistent with the findings of (75) that the optimal regularization parameter is time-varying when the underlying distribution is nonstationary. Also, we note in

			MSE							MAE						
Basis	Method	(η_γ, λ)	0.6	0.7	0.8	0.9	0.95	0.995	-	0.6	0.7	0.8	0.9	0.95	0.995	-
-	AFF	0.1	-	-	-	-	-	-	1.6223	-	-	-	-	-	-	1.0166
		0.01	-	-	-	-	-	-	1.4580	-	-	-	-	-	-	0.9620
		10^{-3}	-	-	-	-	-	-	1.4415	-	-	-	-	-	-	0.9558
		10^{-4}	-	-	-	-	-	-	1.4634	-	-	-	-	-	-	0.9636
		10^{-5}	-	-	-	-	-	-	1.8974	-	-	-	-	-	-	1.1074
		10^{-6}	-	-	-	-	-	-	3.4723	-	-	-	-	-	-	1.4458
	NAIVE	-	-	-	-	-	-	-	2.0379	-	-	-	-	-	-	1.1379
Cycle	RAC	0.1	0.7597	1.2010	3.2935	6.3154	14.8134	33.8847	-	0.6139	0.7530	1.0195	1.6035	2.6562	4.4573	-
		0.01	0.7387	1.1245	2.6802	5.9772	14.3902	33.3531	-	0.6144	0.7437	0.9852	1.5846	2.6263	4.4169	-
		10^{-3}	0.7686	1.1173	2.2985	5.6565	13.8614	31.0243	-	0.6305	0.7556	0.9803	1.5685	2.6027	4.2791	-
		10^{-4}	0.7940	1.1192	2.2305	7.3103	14.8228	30.7455	-	0.6733	0.8021	1.0572	1.6811	2.6729	4.2630	-
		10^{-5}	0.9705	1.3857	2.6985	8.6459	21.7764	31.0420	-	0.7580	0.9082	1.1953	1.8728	3.1087	4.2850	-
		10^{-6}	1.2111	1.8296	3.3745	12.5211	32.7619	35.1512	-	0.8477	1.0394	1.3295	2.2478	3.7446	4.5673	-
	VAR	-	-	-	-	8.2948	9.1316	8.2312	-	-	-	-	1.7803	1.9708	1.8042	-
Trig	RAC	0.1	0.6531	1.0815	1.8610	3.2311	7.2563	8.0548	-	0.6154	0.7571	0.9736	1.4339	2.1468	2.1674	-
		0.01	0.6347	0.9621	1.5317	3.1070	7.2131	8.0203	-	0.6175	0.7428	0.9295	1.4113	2.1358	2.1562	-
		10^{-3}	0.6408	0.9184	1.3027	2.7414	6.5646	8.1231	-	0.6340	0.7574	0.9018	1.3253	2.0184	2.1750	-
		10^{-4}	0.6897	0.9893	1.3444	1.8909	3.7482	8.7487	-	0.6614	0.7927	0.9213	1.0878	1.4542	2.2491	-
		10^{-5}	0.7556	1.0880	1.5119	2.0017	2.2885	7.9473	-	0.6933	0.8315	0.9776	1.1163	1.1734	2.1110	-
		10^{-6}	0.8323	1.1941	1.7840	2.4928	2.1407	6.2601	-	0.7271	0.8704	1.0594	1.2404	1.1576	1.8774	-
	VAR	-	-	-	-	8.7856	8.3085	8.0287	-	-	-	-	1.8632	1.7905	1.7484	-

Table 3.2: Average forecasting accuracy measures, MAE and MSE, over 100 replicates of the process displayed by Equation (3-21). Note that the results of AFF are the same across columns because the forgetting factors are adaptive. Also, the online VAR model for values of $\lambda < 0.9$ were not feasible to compute.

Figures 3.1 and 3.2 the average behaviour of the p-values around the change point. Considering a 5% significance level, on average RAC rejects the null hypothesis that the stream is not experiencing anomalous behaviour around $t = \tau^*$. Finally, regarding the average non zero weights plots, a *LOESS* curve (83, 84) is fitted to provide additional intuition. For both bases, on average, the number of non zero weights in each segment appears reasonably stable, though subject to substantial variability. The *Trig* basis, on average, seems better suited to this specific signal. A possible explanation is that with the *Cycle* basis, RAC selects quadratic terms to fit local curvatures. This is evident after $t = 5000$, when the amplitude of the curve is higher, fewer quadratic terms are selected – implying the slight decay in the *LOESS* curve.

3.3.2 Detection performance

The AFF (62) method has its own detection method while online VAR, NAIVE and RAC will use the one proposed in Section 3.2.6. Allowing ϑ observations after a change was proposed in (85), and also adopted in this work, since there may be a slight delay after a change occurs. The average false detection (FD) rate and average correct detection (CD) rate over 100

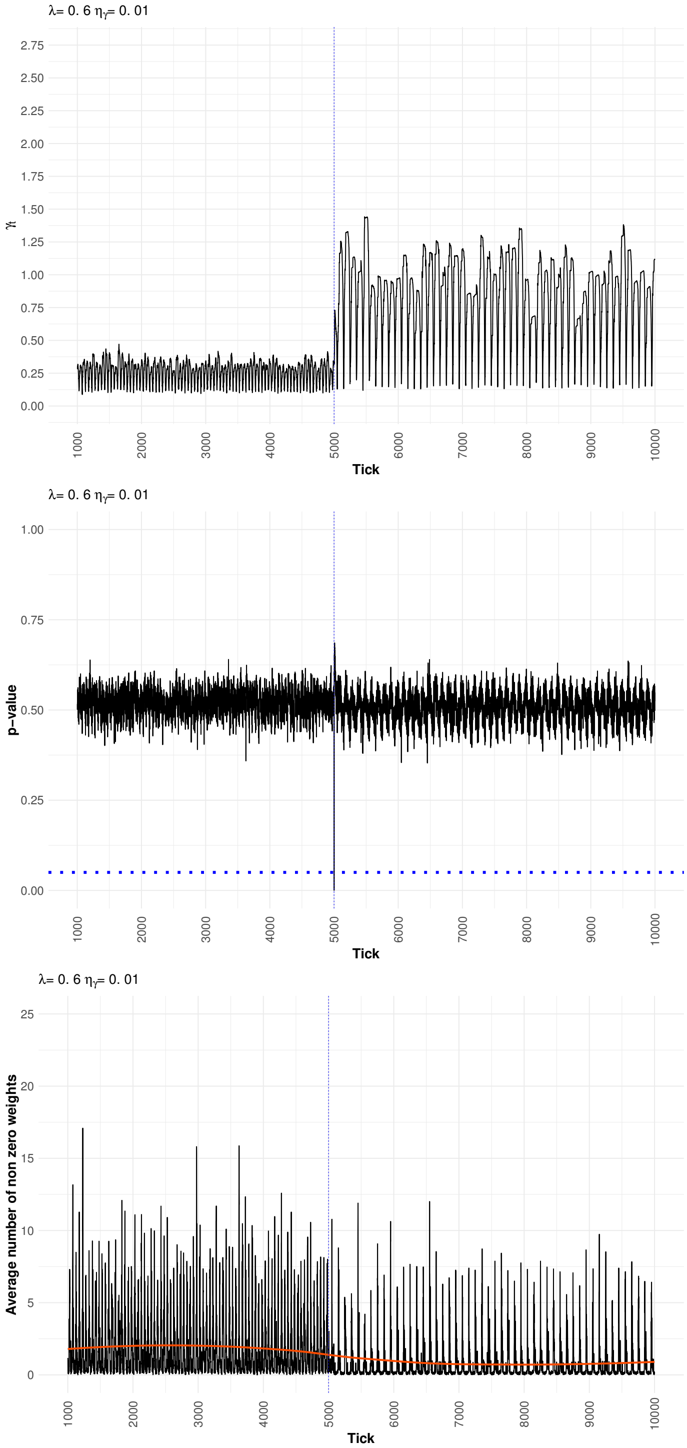

Figure 3.1: Average quantities of the penalty term γ_t, p-value with a dashed horizontal line at the 5% significance level and number of non zero weights $\hat{\beta}(\gamma_t)$ with a smoothed *LOESS* curve over 100 Monte Carlo simulations using *Cycle* basis. The title of the figures shows the parameters η_γ and λ that produces the minimum Mean Square Error.

replicates are calculated as:

- For each replicate, false detection (FD) is calculated as the proportion

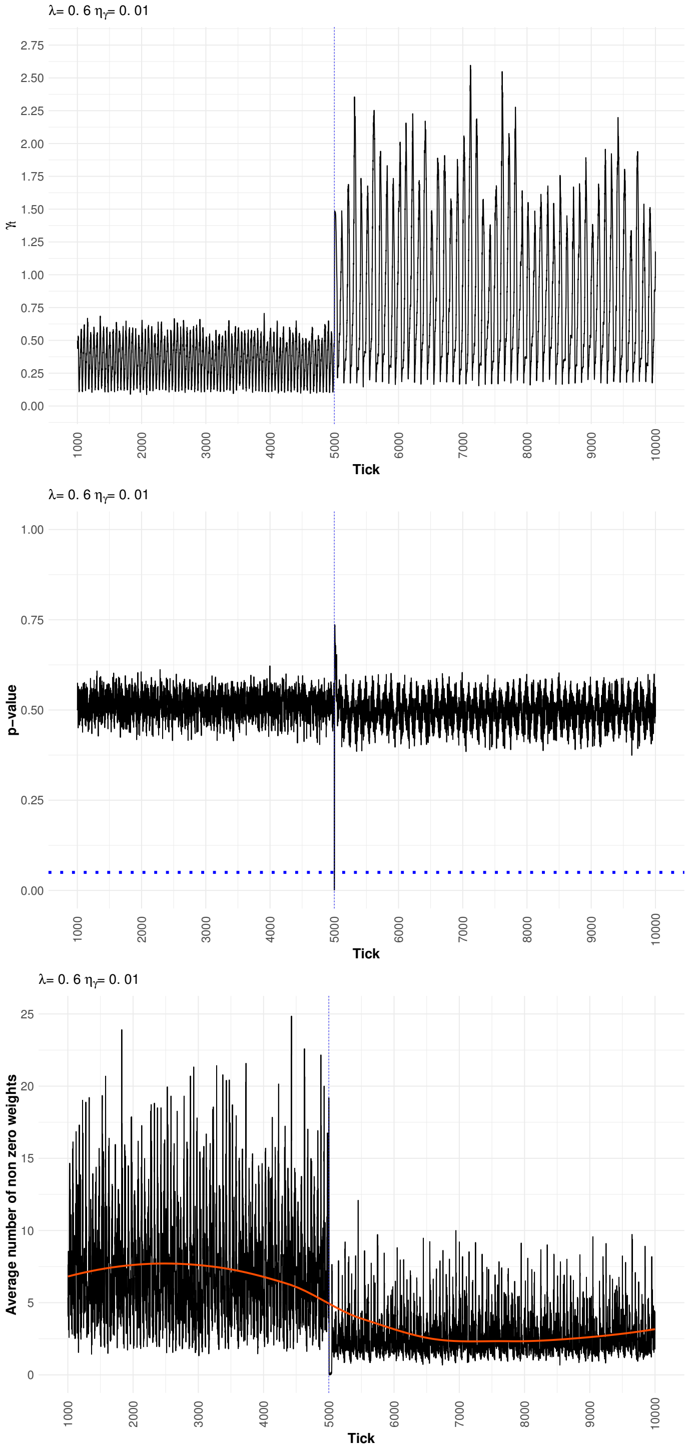

Figure 3.2: Average quantities of the penalty term γ_t, p-value with a dashed horizontal line at the 5% significance level and number of non zero weights $\hat{\beta}(\gamma_t)$ with a smoothed *LOESS* curve over 100 Monte Carlo simulations using *Trig* basis. The title of the figures shows the parameters η_γ and λ that produces the minimum Mean Square Error.

of events that are defined as anomalous when $y_t \notin (\tau^*, \tau^* + \vartheta)$;

– For each replicate, correct detection (CD) is a indicator function that assumes 1 if an anomaly is detected when $y_t \in (\tau^*, \tau^* + \vartheta)$ and zero otherwise.

Tables 3.3 and 3.4 show the CD and FD rates of 100 replicates when adopting $\vartheta = 20$ for both basis *Cycle* and *Trig*.

							CD			
α	Basis	Method	(η_γ, λ)	0.6	0.7	0.8	0.9	0.95	0.995	-
0.0001	-	AFF	0.1	-	-	-	-	-	-	0.00 (0.0000)
			0.01	-	-	-	-	-	-	0.00 (0.0000)
			10^{-3}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-4}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-5}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-6}	-	-	-	-	-	-	0.27 (0.4461)
		NAIVE	-	-	-	-	-	-	-	0.00 (0.0000)
	Cycle	RAC	0.1	0.31 (0.4648)	0.30 (0.4605)	0.20 (0.4020)	0.04 (0.1969)	0.00 (0.0000)	0.00 (0.0000)	-
			0.01	0.22 (0.4163)	0.23 (0.4229)	0.24 (0.4292)	0.11 (0.3144)	0.00 (0.0000)	0.01 (0.1000)	-
			10^{-3}	0.21 (0.4093)	0.25 (0.4351)	0.17 (0.3775)	0.08 (0.2726)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-4}	0.16 (0.3684)	0.11 (0.3144)	0.10 (0.3015)	0.02 (0.1407)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-5}	0.11 (0.3144)	0.09 (0.2876)	0.04 (0.1969)	0.01 (0.1000)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-6}	0.04 (0.1969)	0.02 (0.1407)	0.06 (0.2386)	0.01 (0.1000)	0.00 (0.0000)	0.01 (0.1000)	-
		VAR	-	-	-	-	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
	Trig	RAC	0.1	0.17 (0.3775)	0.12 (0.3265)	0.08 (0.2726)	0.02 (0.1407)	0.03 (0.1714)	0.31 (0.4648)	-
			0.01	0.17 (0.3775)	0.14 (0.3487)	0.22 (0.4163)	0.12 (0.3265)	0.08 (0.2726)	0.35 (0.4793)	-
			10^{-3}	0.19 (0.3942)	0.20 (0.4020)	0.21 (0.4093)	0.17 (0.3775)	0.08 (0.2726)	0.28 (0.4512)	-
			10^{-4}	0.20 (0.4020)	0.10 (0.3015)	0.11 (0.3144)	0.05 (0.2190)	0.10 (0.3015)	0.15 (0.3588)	-
			10^{-5}	0.15 (0.3588)	0.13 (0.3379)	0.08 (0.2726)	0.04 (0.1969)	0.12 (0.3265)	0.21 (0.4093)	-
			10^{-6}	0.11 (0.3144)	0.07 (0.2564)	0.06 (0.2386)	0.05 (0.2190)	0.04 (0.1969)	0.22 (0.4163)	-
		VAR	-	-	-	-	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
0.01	-	AFF	0.1	-	-	-	-	-	-	0.00 (0.0000)
			0.01	-	-	-	-	-	-	0.00 (0.0000)
			10^{-3}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-4}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-5}	-	-	-	-	-	-	0.00 (0.0000)
			10^{-6}	-	-	-	-	-	-	0.45 (0.5000)
		NAIVE	-	-	-	-	-	-	-	0.04 (0.1969)
	Cycle	RAC	0.1	0.99 (0.1000)	1.00 (0.0000)	0.99 (0.1000)	0.99 (0.1000)	0.84 (0.3684)	1.00 (0.0000)	-
			0.01	0.99 (0.1000)	1.00 (0.0000)	0.99 (0.1000)	0.97 (0.1714)	0.77 (0.4229)	1.00 (0.0000)	-
			10^{-3}	0.98 (0.1407)	0.98 (0.1407)	0.99 (0.1000)	1.00 (0.0000)	0.88 (0.3265)	1.00 (0.0000)	-
			10^{-4}	0.98 (0.1407)	1.00 (0.0000)	0.97 (0.1714)	0.94 (0.2386)	0.76 (0.4292)	1.00 (0.0000)	-
			10^{-5}	0.97 (0.1714)	0.98 (0.1407)	0.98 (0.1407)	0.93 (0.2564)	0.81 (0.3942)	1.00 (0.0000)	-
			10^{-6}	0.88 (0.3265)	0.93 (0.2564)	1.00 (0.0000)	0.93 (0.2564)	0.87 (0.3379)	1.00 (0.0000)	-
		VAR	-	-	-	-	0.03 (0.1714)	0.01 (0.1000)	0.02 (0.1407)	-
	Trig	RAC	0.1	0.98 (0.1407)	0.98 (0.1407)	0.94 (0.2386)	0.97 (0.1714)	1.00 (0.0000)	1.00 (0.0000)	-
			0.01	0.96 (0.1969)	0.98 (0.1407)	0.94 (0.2386)	0.97 (0.1714)	1.00 (0.0000)	1.00 (0.0000)	-
			10^{-3}	1.00 (0.0000)	0.99 (0.1000)	1.00 (0.0000)	0.97 (0.1714)	1.00 (0.0000)	1.00 (0.0000)	-
			10^{-4}	0.99 (0.1000)	1.00 (0.0000)	0.98 (0.1407)	1.00 (0.0000)	1.00 (0.0000)	1.00 (0.0000)	-
			10^{-5}	1.00 (0.0000)	1.00 (0.0000)	0.99 (0.1000)	0.97 (0.1714)	0.98 (0.1407)	1.00 (0.0000)	-
			10^{-6}	1.00 (0.0000)	0.97 (0.1714)	0.96 (0.1969)	0.87 (0.3379)	0.97 (0.1714)	1.00 (0.0000)	-
		VAR	-	-	-	-	0.02 (0.1407)	0.08 (0.2726)	0.04 (0.1969)	-

Table 3.3: Average of Correct Detection (CD) over 100 replicates of the process displayed by Equation (3-21). Note that the online VAR model for values of $\lambda < 0.9$ were not feasible to compute.

The main points from Tables 3.3, 3.4, first, RAC appears accurate at detect anomalies due to its high CD rates, and corresponding low FD rates, second the performance of RAC is somehow dependent on the choice of control parameters, third in this simulation both choices of basis perform comparably. Finally, the effect of the significance on detection performance is as expected.

The poor CD performance of AFF is due to the fact that it cannot cope with underlying structure in the data. It was designed to calculate a dynamic

				FD						
α	Basis	Method	(η_γ, λ)	0.6	0.7	0.8	0.9	0.95	0.995	-
0.0001	-	AFF	0.1	-	-	-	-	-	-	0.35 (0.0078)
			0.01	-	-	-	-	-	-	0.27 (0.0103)
			10^{-3}	-	-	-	-	-	-	0.28 (0.0105)
			10^{-4}	-	-	-	-	-	-	0.32 (0.0087)
			10^{-5}	-	-	-	-	-	-	0.41 (0.0048)
			10^{-6}	-	-	-	-	-	-	0.44 (0.1437)
		NAIVE	-	-	-	-	-	-	-	0.00 (0.0000)
	Cycle	RAC	0.1	0.00 (0.0003)	0.00 (0.0003)	0.00 (0.0004)	0.00 (0.0003)	0.00 (0.0001)	0.00 (0.0000)	-
			0.01	0.00 (0.0003)	0.00 (0.0003)	0.00 (0.0004)	0.00 (0.0003)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-3}	0.00 (0.0002)	0.00 (0.0003)	0.00 (0.0003)	0.00 (0.0003)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-4}	0.00 (0.0001)	0.00 (0.0001)	0.00 (0.0002)	0.00 (0.0002)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-5}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0001)	0.00 (0.0001)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-6}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
		VAR	-	-	-	-	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
	Trig	RAC	0.1	0.00 (0.0001)	0.00 (0.0002)	0.00 (0.0002)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
			0.01	0.00 (0.0000)	0.00 (0.0001)	0.00 (0.0002)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-3}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0001)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-4}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-5}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
			10^{-6}	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
		VAR	-	-	-	-	0.00 (0.0000)	0.00 (0.0000)	0.00 (0.0000)	-
0.01	-	AFF	0.1	-	-	-	-	-	-	0.40 (0.0057)
			0.01	-	-	-	-	-	-	0.36 (0.0078)
			10^{-3}	-	-	-	-	-	-	0.37 (0.0069)
			10^{-4}	-	-	-	-	-	-	0.39 (0.0056)
			10^{-5}	-	-	-	-	-	-	0.44 (0.0035)
			10^{-6}	-	-	-	-	-	-	0.43 (0.1206)
		NAIVE	-	-	-	-	-	-	-	0.00 (0.0005)
	Cycle	RAC	0.1	0.00 (0.0008)	0.01 (0.0008)	0.01 (0.0009)	0.01 (0.0010)	0.01 (0.0007)	0.00 (0.0001)	-
			0.01	0.00 (0.0006)	0.00 (0.0008)	0.01 (0.0008)	0.00 (0.0009)	0.00 (0.0007)	0.00 (0.0001)	-
			10^{-3}	0.00 (0.0007)	0.00 (0.0007)	0.00 (0.0008)	0.00 (0.0009)	0.00 (0.0008)	0.00 (0.0001)	-
			10^{-4}	0.00 (0.0007)	0.00 (0.0007)	0.00 (0.0008)	0.00 (0.0010)	0.00 (0.0007)	0.00 (0.0001)	-
			10^{-5}	0.00 (0.0007)	0.00 (0.0007)	0.00 (0.0008)	0.00 (0.0009)	0.00 (0.0007)	0.00 (0.0001)	-
			10^{-6}	0.00 (0.0008)	0.00 (0.0007)	0.00 (0.0008)	0.00 (0.0009)	0.00 (0.0006)	0.00 (0.0001)	-
		VAR	-	-	-	-	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0005)	-
	Trig	RAC	0.1	0.00 (0.0008)	0.01 (0.0009)	0.01 (0.0009)	0.00 (0.0006)	0.00 (0.0004)	0.00 (0.0003)	-
			0.01	0.00 (0.0007)	0.00 (0.0008)	0.00 (0.0007)	0.00 (0.0007)	0.00 (0.0003)	0.00 (0.0004)	-
			10^{-3}	0.00 (0.0005)	0.00 (0.0006)	0.00 (0.0005)	0.00 (0.0006)	0.00 (0.0004)	0.00 (0.0003)	-
			10^{-4}	0.00 (0.0005)	0.00 (0.0004)	0.00 (0.0006)	0.00 (0.0006)	0.00 (0.0006)	0.00 (0.0003)	-
			10^{-5}	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0006)	0.00 (0.0005)	0.00 (0.0003)	-
			10^{-6}	0.00 (0.0004)	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0006)	0.00 (0.0004)	-
		VAR	-	-	-	-	0.00 (0.0005)	0.00 (0.0005)	0.00 (0.0005)	-

Table 3.4: Average of False Detection (FD) over 100 replicates of the process displayed by Equation (3-21). Note that the online VAR model for values of $\lambda < 0.9$ were not feasible to compute.

average under the assumption of *i.i.d.* Gaussian observations. If seasonality is present in the data, filtering the unconditional mean in this case will end up resulting in p-values oscillating with the stream's unconditional mean. The good performance of RAC's CD and FD rate, when comparing to NAIVE and online VAR, provides evidence that this method has merits regarding *Detection* performance.

3.4 Real data: Cyber-security

Cyber-crime is an increasing burden to society, costing around \$600 billions per annum (86). (87) mention that operational anomaly detection methods in this area predominantly rely on signature-based methods, which seek to identify events and behaviour of known form. Such methods are incapable of dealing with so-called “zero-day” attacks – activities that have not previously been seen (88). Moreover, the increasing intensity and consequences of cyber-attacks suggests that existing signature-based methods are insufficient. There is a burgeoning research area using automatically collected computer and network data in conjunction with statistical and machine learning methods which is focused on anomaly detection. Specifically, detecting departures from “normal” behaviour (89, 90). These data-driven methods are intended to complement existing signature-based tools.

There are a number of challenges in developing statistical methods for cyber-security. Some relate to the volume and velocity of the data – typical data sources can be vast in a large enterprise, and as demonstrated later, data can arrive at very high rates. Another type of challenge relates to practical usage. Any data-driven anomaly detection method will suffer from false positives, and these create a misleading and potentially costly false signal. In this section we deploy RAC against cyber-security data sets provided by Los Alamos National Laboratory (LANL). The data set is described in detail in (87) and is available online[2].

We focus on the *host Log* data, which comprises a subset of computer event logs collected from all computers running the Microsoft Windows operating system on LANL’s enterprise network. Events from the *host log* included in the data set are all related to authentication and process activity on each machine. We focus on this data source, rather than the more widely studied *Network Flow* data because, as noted in (87), remote attackers and malicious insiders increasingly use encryption, hence reducing the effectiveness of communication-based detection mechanisms.

There are several kinds of events and log on-types described by (87). We only consider those which are driven only by human behaviour, i.e., which do

[2] https://csr.lanl.gov/data/2017.html

not run any automated process periodically. The selected events were

- 4624: An account was successfully logged on;
- 4625: An account failed to logon;
- 4634: An account was logged off;
- 4800: The workstation was locked;
- 4801: The workstation was unlocked;
- 4802: The screensaver was invoked;
- 4803: The screensaver was dismissed.

For each of these event types, we construct a data stream, $y_1, y_2, ...y_t, y_{t+1}, ...$, that counts the number of events per minute. For the LANL data, this yields seven time series, each consisting of 129600 data points. A one-minute window is a plausible size in this context, given constraints on data collection and the requirement for timely detection. We then want to assess RAC's performance, in terms of *Estimation* and *Detection*, compared to AFF, online VAR and NAIVE. For forecasting accuracy, similar to the simulations in Section 3.3, MSE and MAE are adopted.

The streams are formed from count data, and hence strictly non-negative. RAC was not designed to specifically deal with this case – modifications are possible, though challenging and hence left to future work. Here, we use a simple rule whereby negative predictions from RAC are set to zero. While this is perhaps the crudest possible fix, it can be applied consistently without computational overhead.

To explore the capabilities of RAC, the results are considered in three different contexts:

- $\mathcal{W}$: The whole data set for each event will be used, i.e., $t = 1, ..., 129600$;
- $\mathcal{A}_t(\gamma_t) \neq \emptyset$: Only predictions from those points where at least one weight of $\hat{\beta}_t(\gamma_t)$ is non zero will be considered;

– $\mathcal{A}_t(\gamma_t) = \emptyset$: Only predictions from those points where all the weights of $\hat{\beta}_t(\gamma_t)$ are equals to zero will be considered.

The results for *Estimation* metrics considering one step ahead forecasting error over the sets $\mathcal{W}$, $\mathcal{A}_t(\gamma_t) \neq \emptyset$ and $\mathcal{A}_t(\gamma_t) = \emptyset$ are depicted in Tables 3.5, 3.6 and 3.7. The initial values were fixed as follows. For burn-in, we use $\mathcal{B} = 1440$ observations, i.e., one day of data. The remaining control parameters are selected as $\gamma_0 = 10$, $\eta_\gamma = 0.01$, and $\lambda = 0.6$. The choices of η_γ and λ are the values that produces minimum MSE in the simulations of Section 3.3, while γ_0 was fixed based on empirical evidence while running the simulations. Regarding online VAR, we also fixed $\lambda = 0.995$ which was the best results regarding MSE performance in the simulations.

Considering the forecasting errors when $y_t \in \mathcal{W}$, displayed in Table 3.5, RAC is able to accurately track a complicated target with underlying structure varying over time. Considering MSE and MAE, RAC outperform the other methods in all of the 7 cyber-security examples, regardless of the selected basis. The complex structure of RAC allows for different combination of basis functions, and notably allows for all regression weights to be forced to zero. This provides some modelling advantage, as the results in Table 3.6 clearly illustrate. Such advantage takes into account, for the calculation of both forecasting metrics, only events where $\mathcal{A}_t(\gamma_t) \neq \emptyset$, i.e., ticks of y_t where RAC estimates weights $\hat{\beta}_t(\gamma_t) \neq 0$. In comparing bases, on average, the *Trig* basis tends to select more coefficients than the *Cycle* basis, producing smaller errors in $\mathcal{A}_t(\gamma_t) \neq \emptyset$. This can be verified in the last two columns of Table 3.6, the total number of points in which the weights $\hat{\beta}_t(\gamma_t)$ are different from zero and the proportion compared to the total number of observations $T = 129000$.

Finally, consider $\mathcal{A}_t(\gamma_t) = \emptyset$, the case in which all RAC weights $\hat{\beta}_t(\gamma_t)$ equal zero. In this case, RAC outperformed the benchmarks in terms of forecasting performance. This suggests that RAC is capable of handling periods that are locally constant, in addition to periods exhibiting high non-linearity.

This also provides evidence about the proposed updates in Equations (3-15) and (3-16), since the forecast is produced only by the dynamic intercept $\hat{\beta}_t^{(0)}(\gamma_t)$ in this set.

Basis	Error Metric	EVENT	AFF	VAR	NAIVE	RAC
Cycle	MSE	4624	103.3800	113.5808	249.8860	62.6507
		4625	2.1279	1.8457	4.6593	1.1388
		4634	102.9293	113.2251	249.1380	62.4926
		4800	15.7501	19.7395	40.7777	10.6797
		4801	15.2402	17.0619	37.8169	9.7710
		4802	5.5410	5.9006	12.8318	3.4723
		4803	22.9722	21.8406	46.7114	14.5741
	MAE	4624	4.8119	5.1336	7.6086	3.7678
		4625	0.6953	0.6575	1.0016	0.5079
		4634	4.8084	5.1311	7.6050	3.7660
		4800	1.9622	2.2392	3.1527	1.5969
		4801	1.9327	2.0996	3.0590	1.5265
		4802	1.3198	1.3481	1.9611	1.0011
		4803	1.3916	1.3624	2.0127	1.0252
Trig	MSE	4624	103.3800	113.1731	249.8860	61.5966
		4625	2.1279	3.1763	4.6593	1.1086
		4634	102.9293	112.8646	249.1380	61.9998
		4800	15.7501	20.3095	40.7777	10.8447
		4801	15.2402	17.5381	37.8169	9.1979
		4802	5.5410	8.0557	12.8318	3.2608
		4803	22.9722	25.8474	46.7114	13.9264
	MAE	4624	4.8119	5.2245	7.6086	3.7522
		4625	0.6953	0.8533	1.0016	0.5074
		4634	4.8084	5.2258	7.6050	3.7546
		4800	1.9622	2.3142	3.1527	1.5933
		4801	1.9327	2.1636	3.0590	1.5190
		4802	1.3198	1.6012	1.9611	0.9927
		4803	1.3916	1.6308	2.0127	1.0175

Table 3.5: One step ahead estimation accuracy in cyber-security events by ID using the whole set $y_t \in \mathcal{W}$. Response variable is the count of events by minute over 90 days, which end up with 129600 observations. Both forecasting metrics MSE and MAE are calculated as defined by Equations (3-22) and (3-23) with a burn-in period of one day, i.e., $\mathcal{B} = 1440$.

Regarding the *Detection* performance, it is nearly impossible to evaluate on real cyber-security data sets since it is unlabelled, i.e., there is no information regarding when, or if, an anomaly happened. We perform anomaly detection using the approach described in Section 3.2.6, fixing a p-value of $1/\mathcal{B}$ observations, i.e., we intend to capture 1 anomaly per day and forgetting factor $\theta = 0.95$. Figures 3.3 and 3.4 show 6 days of the first week of LANL's *host log* data set for the 7 described events, excluding the first day, which was used as burn-in. The vertical lines make reference to detected anomalies. Some of them seem to capture unusual patterns, but as argued we cannot prove these are actual anomalies due to the absence of labelled data.

Basis	Error Metric	EVENT	AFF	VAR	NAIVE	RAC	Size	Prop
Cycle	MSE	4624	149.6860	172.0855	238.2412	101.1185	26251	20.26%
		4625	6.0015	5.3002	8.5379	3.6542	12408	9.57%
		4634	147.0458	170.0318	235.8500	100.4786	26108	20.15%
		4800	19.3835	25.1502	34.9095	16.2884	26111	20.15%
		4801	18.8977	21.9409	30.6779	14.4425	26249	20.25%
		4802	8.9511	10.1686	14.1310	6.8977	28277	21.82%
		4803	61.4953	51.5657	135.1770	44.4220	27721	21.39%
	MAE	4624	4.2259	4.4351	5.1185	3.5842	26251	20.26%
		4625	1.4401	1.3992	1.6096	1.1908	12408	9.57%
		4634	4.1452	4.3580	5.0305	3.5313	26108	20.15%
		4800	1.6654	1.8705	2.0532	1.4700	26111	20.15%
		4801	1.6841	1.8371	2.0411	1.4510	26249	20.25%
		4802	1.3897	1.4118	1.5990	1.1433	28277	21.82%
		4803	1.5060	1.5084	1.7897	1.2306	27721	21.39%
Trig	MSE	4624	57.6816	66.8097	99.7589	39.4957	51842	40.00%
		4625	0.9168	1.1874	1.5913	0.5393	70446	54.36%
		4634	57.0664	66.4187	101.5935	40.7172	51798	39.97%
		4800	7.3827	10.3185	17.0442	7.0982	51580	39.80%
		4801	6.5142	8.1232	13.0563	4.6416	51966	40.10%
		4802	3.8455	4.9307	6.8386	2.5562	56723	43.77%
		4803	30.3047	30.0738	38.3748	14.5867	54464	42.02%
	MAE	4624	2.0361	2.3369	2.8391	1.7498	51842	40.00%
		4625	0.2971	0.3400	0.3834	0.2360	70446	54.36%
		4634	2.0818	2.3708	2.8927	1.7980	51798	39.97%
		4800	0.9357	1.1258	1.3295	0.8301	51580	39.80%
		4801	0.8732	1.0062	1.2190	0.7532	51966	40.10%
		4802	0.8459	0.9323	1.0866	0.6683	56723	43.77%
		4803	0.9059	0.9825	1.1523	0.7037	54464	42.02%

Table 3.6: One step ahead estimation accuracy in cyber-security events by ID adopting $\mathcal{A}_t(\gamma_t) \neq \emptyset$. Response variable is the count of events by minute over 90 days, which end up with 129600 observations. Both forecasting metrics MSE and MAE are calculated only considering point where $\mathcal{A}_t(\gamma_t) \neq \emptyset$. The last two columns states the total number of points in which the weights $\hat{\beta}_t(\gamma_t)$ are different from zero and the proportion compared to the total number of observations $T = 129000$.

Basis	Error Metric	EVENT	AFF	VAR	NAIVE	RAC
Cycle	MSE	4624	91.0496	98.5110	251.9971	52.6347
		4625	1.7107	1.4753	4.2392	0.8761
		4634	91.2466	98.6927	251.6752	52.7576
		4800	14.7294	18.3550	42.0677	9.2123
		4801	14.2279	15.8053	39.4890	8.5523
		4802	4.5636	4.6923	12.4191	2.4991
		4803	12.2102	13.6369	22.0932	6.2720
	MAE	4624	4.9507	5.3135	8.2301	3.8090
		4625	0.6147	0.5779	0.9345	0.4347
		4634	4.9656	5.3288	8.2437	3.8206
		4800	2.0308	2.3335	3.4235	1.6256
		4801	1.9908	2.1672	3.3116	1.5431
		4802	1.2995	1.3300	2.0608	0.9605
		4803	1.3572	1.3220	2.0698	0.9668
Trig	MSE	4624	133.0916	144.6666	348.8498	75.9926
		4625	3.5565	5.6040	8.2942	1.7866
		4634	132.7266	144.3692	346.2823	75.8284
		4800	21.1436	27.0388	56.2152	13.2486
		4801	20.9698	23.9592	54.2028	12.1999
		4802	6.8249	10.5370	17.4267	3.7890
		4803	17.2771	22.7239	52.0130	13.2162
	MAE	4624	6.6488	7.1859	10.7738	5.0791
		4625	1.1662	1.4797	1.7324	0.8286
		4634	6.6102	7.1623	10.7280	5.0490
		4800	2.6325	3.1145	4.3493	2.0935
		4801	2.6354	2.9529	4.2828	2.0279
		4802	1.6874	2.1321	2.6399	1.2439
		4803	1.7393	2.1099	2.6315	1.2419

Table 3.7: One step ahead estimation accuracy in cybersecurity events by ID adopting $\mathcal{A}_t(\gamma_t) = \emptyset$. Response variable is the count of events by minute over 90 days, which end up with 129600 observations. Both forecasting metrics MSE and MAE are calculated only considering point where $\mathcal{A}_t(\gamma_t) = \emptyset$.

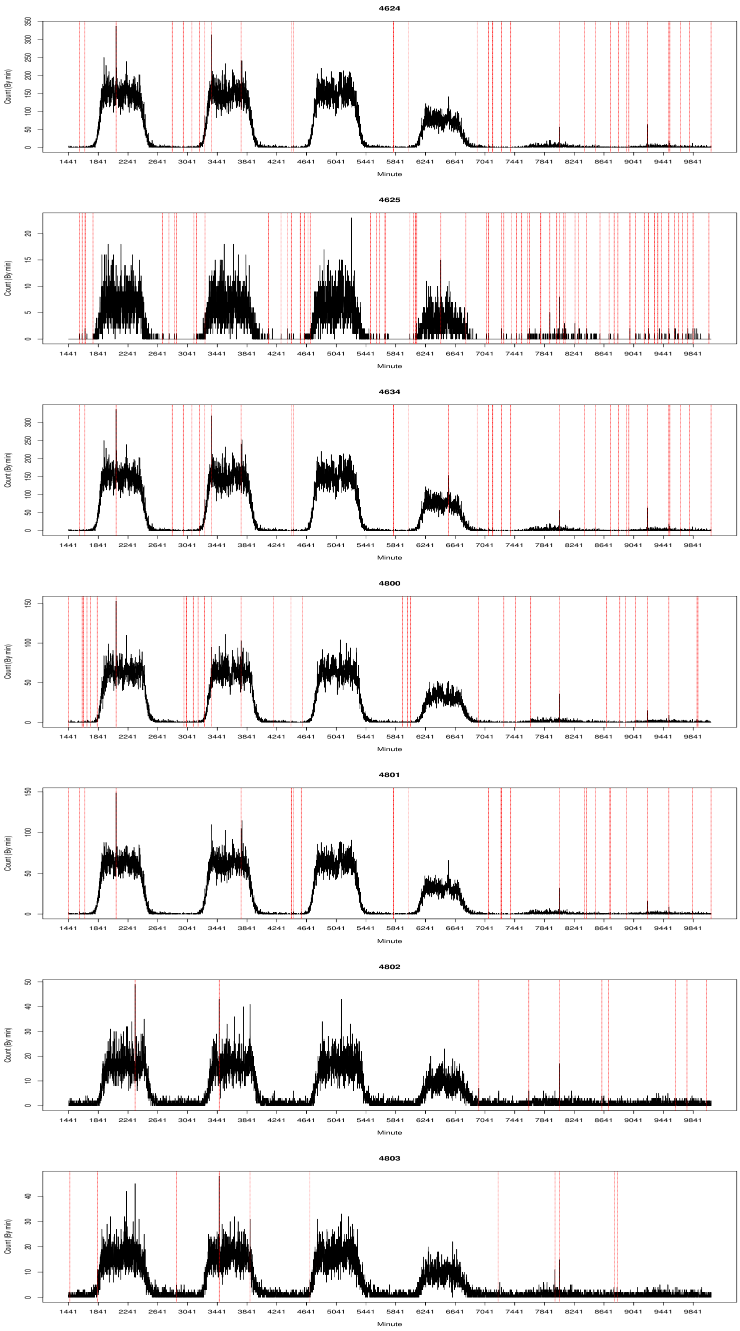

Figure 3.3: Detection performance of RAC using *Cycle basis* and the approach described in Section 3.2.6 during the first 6 days, discarding the burn-in day. The p-value used to detect anomalous behaviour was fixed in $1/\mathcal{B}$.

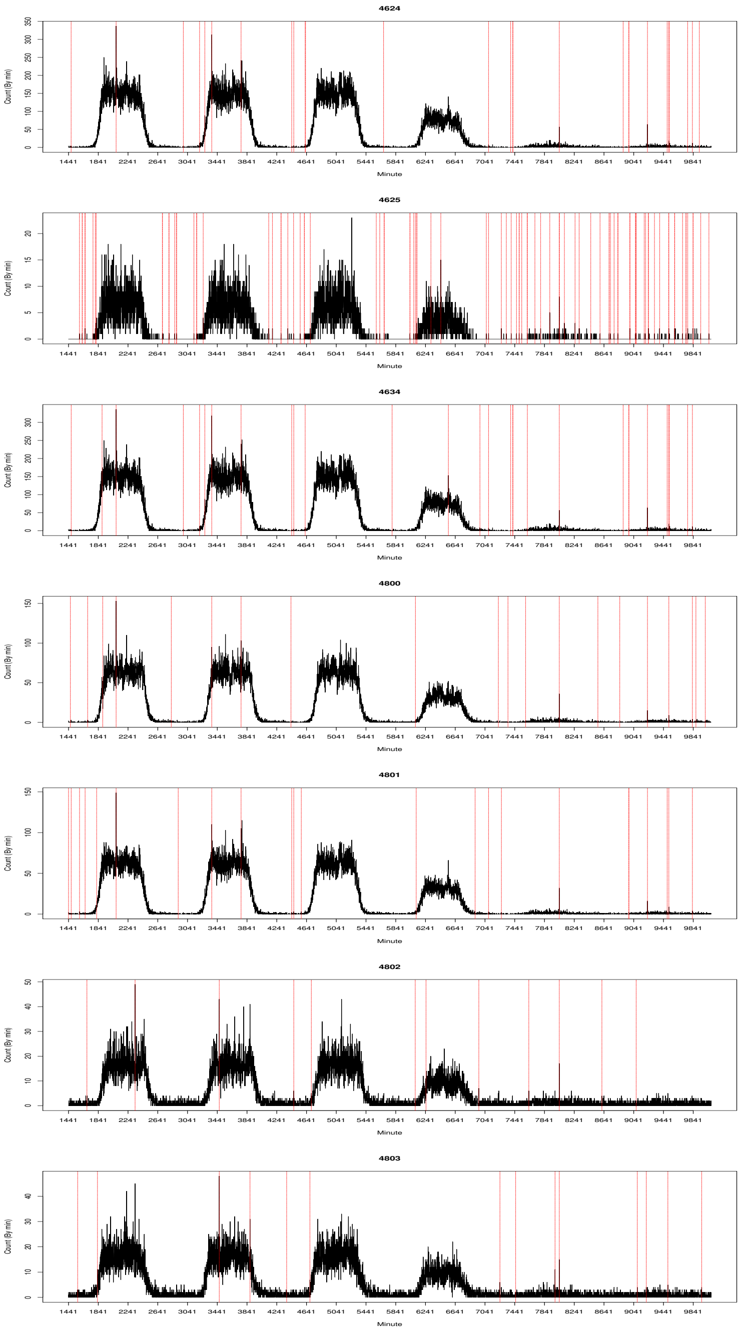

Figure 3.4: Detection performance of RAC using *Trig basis* and the approach described in Section 3.2.6 during the first 6 days, discarding the burn-in day. The p-value used to detect anomalous behaviour was fixed in $1/\mathcal{B}$.

3.5 Conclusion

This paper propose RAC, a framework to estimate, forecast and perform anomaly detection in streaming data environments. Empirical results regarding *Estimation* and *Detection* performance, both for simulated and real data sets, provides evidence that our proposed framework is able to track a moving target, and identify changes in local structure.

In the cyber-security example, the forecasting accuracy of RAC was better than existing methods for almost all of the ID events. Additionally, both bases proposed to sequentially fit the underlying structure of the signal, namely *Trig* and *Cycle* had excellent performance. Regarding the *Detection* performance, we could only evaluate it in simulation studies, due to the fact that LANL's cyber-security data set is unlabelled. Still, RAC performed well when comparing correct and false detection rates against the benchmarks.

For future work, we mention a few possible extensions of this framework, (i) sequential prediction intervals might be calculated adapting the results of (91) to a time-varying penalty term γ_t, (ii) extensions respecting the range of response, essentially extending the framework to the coverage of GLMs.

4
Unsupervised streaming methods for anomaly structure detection in instrumented infrastructure

Keywords: Structural Health Monitoring, streaming PCA, stochastic gradient descent, adaptive estimation, conformal prediction.

Abstract: Structure Health Monitoring often involves instrumenting structures with distributed sensor networks. These networks typically provide high frequency data describing the spatio-temporal behaviour of the assets. A main objective of SHM is to reason about changes in structures' behaviour using sensor data. We construct a streaming anomaly detection method for data from a railway bridge instrumented with a fibre-optic sensor network. The data exhibits trend over time, which may be attributed to environmental factors, calling for a temporally adaptive estimation. Exploiting a latent structure present in the data motivates a quantity of interest. This quantity is estimated sequentially and adaptively using a new formulation of streaming Principal Component Analysis. Anomaly detection for this quantity is then provided using Conformal Prediction. Like all streaming methods, the proposed method has free control parameters which are set using simulations based on bridge data. Experiments demonstrate that this method can operate at data rate while providing accurate tracking of the target quantity. Further, the anomaly detection is able to detect train passage events. Finally the method reveals a previously unreported cyclic structure present in the data.

4.1 Introduction

Modern civil engineering is increasingly leveraging sensor technology to understand and monitor physical assets, such as bridges and pipe networks (92, 93). This sensor technology is becoming cheaper to deploy and hence more widely used. The ultimate ambition for such sensor networks is to better understand the behaviour and degradation of physical assets. However, there are significant data processing and analysis challenges prior to this.

We are concerned with the analysis of sensor data from railway bridges instrumented with a fibre-optic sensor network. The sensor system is used to quantify the stress behaviour of the asset, typically at high frequency. Taking bridges as an example, the sensor network is distributed spatially over the bridge and in some sense captures spatio-temporal response. Our experience suggests that civil engineers do not yet fully appreciate that these sensor systems manifest a noise process independent of the behaviour of the bridge and moreover the measurements represent only a partial characterisation of the physics assumed to define bridge behaviour.

We seek to develop a streaming method for monitoring the bridge, *as a whole*, for train passage events and other events which manifest similar responses. This is challenging for a number of reasons. First, the data is recorded at high frequency, often 250 Hz. Constructing sequential methods that can update at this rate is challenging. Second, the data is subject to explainable temporal *drift*, for example relating to temperature, and other less understood processes. This calls for temporally adaptive methodology. Third, since the sensor system is physically distributed, it provides a view of the whole bridge at every time instance (hereafter, *tick*). A train passage event has a direction with relation to the sensors and hence manifests differentially over the network at any tick. Analysing sensors separately will lead to definitional problems (e.g. when did the event start) and issues of multiple testing.

In this paper we propose a novel streaming methodology that accurately detects the start of train passage events over the entire bridge. Exploratory data analysis reveals a striking low dimensional latent structure arising from these sensor networks. This structure is observed when the bridge is “at rest”, that is when a train is not interacting with the bridge. Further, the structure is a linear combination of all sensors in the network, and hence represents the collective behaviour of the bridge. We develop a streaming estimation procedure for tracking components of this structure at data rate. Data analysis further reveals a collapse of this latent structure during train passage events. Our estimation procedure forms the basis of an anomaly detection procedure designed to capture the start of train passage events.

The structure of this paper is as follows. Section 4.2 provides a detailed overview of the bridge sensor system and a specific data set. Principal Component Analysis (PCA) reveals a specific low dimensional latent structure during at rest periods. In Section 4.3 a streaming PCA procedure is developed. This procedure synthesises ideas of adaptive estimation (9) with ideas for sequential eigendecomposition (see (94, 95) and references therein). A collection of numerical problems arise with sequential eigendecomposition, which we address. Additionally, like all streaming estimation methods a number of input parameters need to be set. In Section 4.4 we use ideas from Conformal Prediction (11, 12) to deal with the challenge of statistical anomaly detection for eigendecomposition. The detection performance of the method relates to its estimation performance, which in turn depends on input parameters, as noted above. Section 4.5 reports a simulation study designed to both compare variants of the methodology and determine input parameters for practical deployment. Finally, the preferred methodology is deployed against a large amount of bridge data in Section 4.6. In addition to demonstrating the detection capabilities of the methodology, the proposed method also reveals interesting and previously unreported properties of the sensor system.

4.2 Instrumented infrastructure and data

In this section, a brief review of Structure Health Monitoring (SHM) and a detailed description of the bridge data set are provided.

4.2.1 Structural health monitoring

The area of SHM has changed in recent times due to the automated collection of data pertaining to physical assets. Historically and currently most integrity evaluation is performed on a manual basis, which is costly and time-consuming. Structures such as railway bridges and pipe networks are now being instrumented with a variety of sensor systems in an effort to better understand their behaviour and reduce the cost burden of monitoring (96, 97, 98, 99, 100, 101, 102, 103, 104). Asset operators are using these data to reason about SHM questions. There are significant challenges related to the development of statistical methods for handling such data. Physical assets have both spatial and temporal extent and the data exhibits rich idiosyncratic structure which raises challenges for spatio-temporal modelling. The high frequency recording rate of the data poses a challenge for data curation and statistical analysis. Detection of train passage events at data rate requires both acquiring the data and performing a statistical procedure. Some of these statistical aspects have been addressed by (105). Data generated by the sensor system is a combination of physical response of the bridge, noise arising from the sensor system and environmental factors at each sensor location.

4.2.2 Fiber bragg grating sensors

Fiber Bragg Grating (FBG) sensors are commonly used to measure strain, that is, for example, the vertical deflection of a bridge under load. In this work, the sensor system consists of a distributed network of fibre-optic strain sensors. These sensors used inscribed Bragg Gratings (106) within the fibre-optic cable which refract light at a particular wavelength. When subjected to strain, the cable is deformed resulting in a change of wavelength.

A statistical challenge arises because the FBG cable, and hence data, also responds to environmental factors such as temperature changes. This is just one of several possible factors that contribute to temporal variations observed in the data. A sophisticated algorithm is used to transform the refracted light in the optical cable to wavelength (more details in (106)).

The raw data is measured in wavelength (nano-meters), although engineers prefer the physical quantity, strain. By construction, each sensor's wavelength is offset by a fixed constant. Wavelength and strain are linearly related and hence we will report results on the wavelength scale. Denote the wavelength measurement at tick N from sensor j as $x_{j,N}$ for which we form the vector over d sensors

$$\mathbf{x}_N = (x_{1,N}, \ldots, x_{d,N}).$$

4.2.3 Data

We are concerned with data from a steel-concrete railway bridge located in Staffordshire, UK. This bridge is a 26.8 metre composite instrumented with 80 FBG sensors. In this case, the sensor network was installed during construction, though retrofitting is possible. The acquisition rate of the data is 250 Hz, i.e., each sensor records 250 ticks a second. The sensors are organised as in Figure 4.1, spaced one metre along the fibre-optic cable. There are four cables located in the main girders, which are each 20 metres long.

Examining data from an individual sensor reveals unusual structures. For example, Figure 4.2 presents a sequence of observations from a single sensor which exhibits a "banding" structure. This is an artefact of the sensor recording algorithm mentioned above. Further, Figure 4.3 presents a smoothed version of multiple sensors over time. This shows that there is temporal variation which

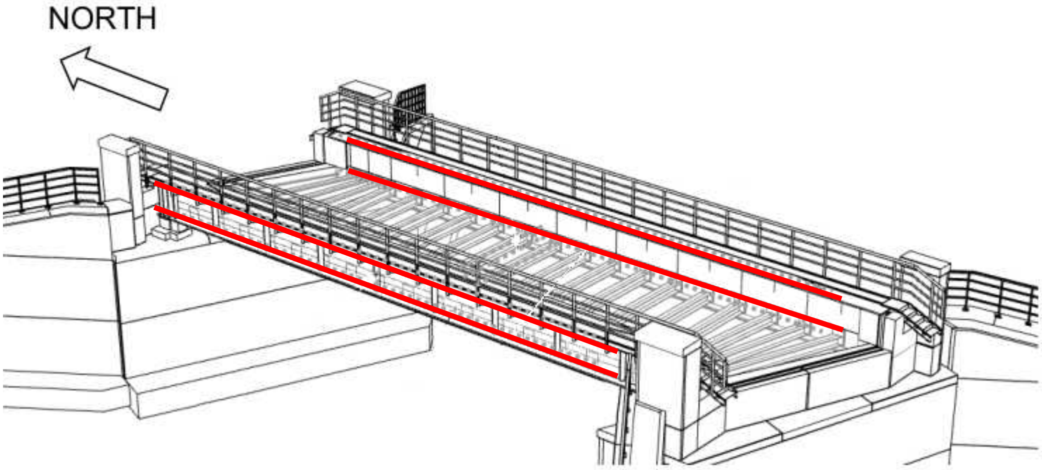

Figure 4.1: Configuration of sensor cables along the bridge. Each of the four sensors contains 20 fibre Bragg gratings.

is different by sensor. Figure 4.4 presents measurements from multiple sensors including a train passage event. Note that sensor measurements during the event differ according to their spatial location.

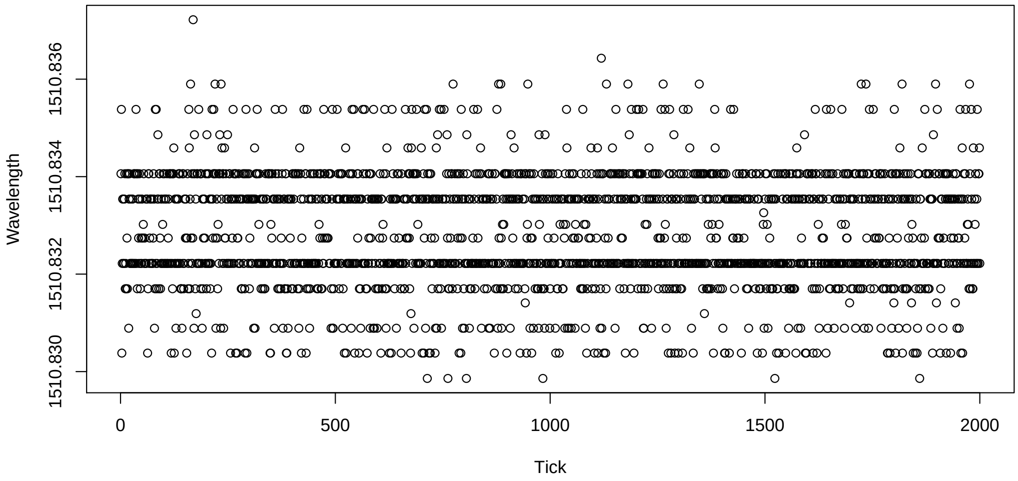

Figure 4.2: Wavelength measurements of a randomly selected sensor during an at rest period.

It is important to distinguish the two types of periods present in the data: periods *at rest* and *train passage events*. Both types capture the space-time response of the sensor system, however at rest it is reasonable to assume that

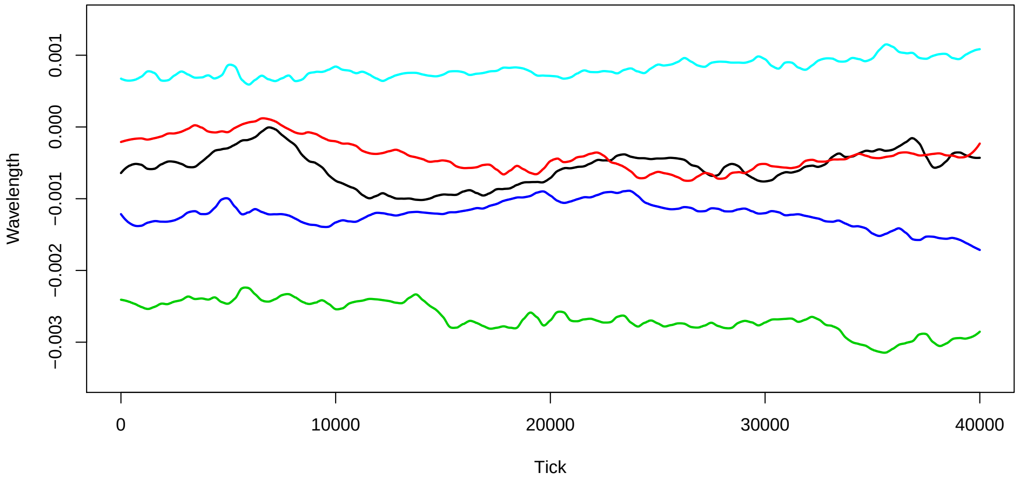

Figure 4.3: LOESS curves fitted to five sensors during an at rest period. For illustration, the first value of each sensor was subtracted from each stream.

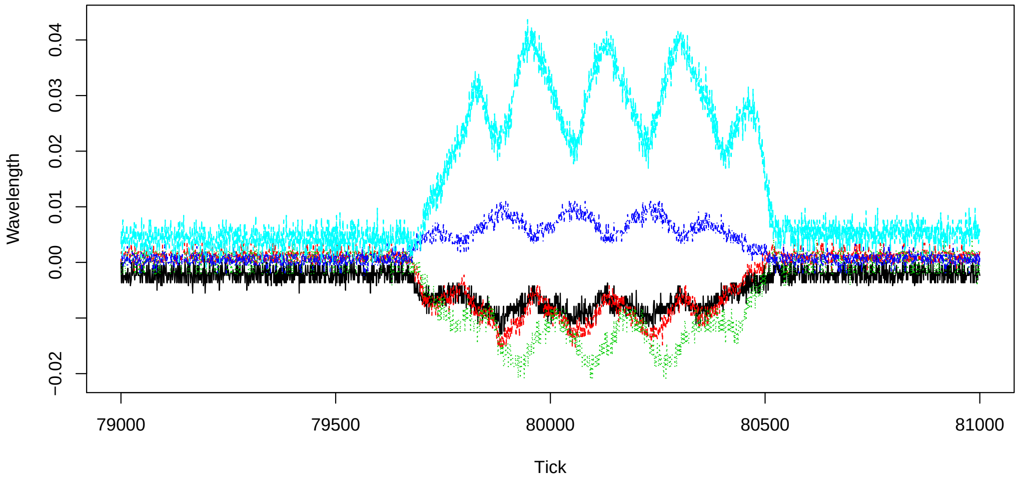

Figure 4.4: Measurement for five sensors including a train passage event. For illustration, the first value of each sensor was subtracted from each stream.

the variation in the sensor system is simply noise. Identifying train passage events is a step toward SHM, since changes in response during train passage events might indicate degradation.

The collective response of the sensor system can be understood using various statistical procedures. In this case Principal Component Analysis (PCA), as used in (107), reveals an unexpected structure. Using PCA on an at rest period of 5000 ticks of data, we find that the first 2 principal components accounts for almost 40% of the variation in the data. The *scores* of PCA are computed as

$$T_q = \mathbf{U}_q X$$

where X is the centred matrix where each column has its mean subtracted and $\mathbf{U}_q = [\mathbf{u}_1 \dots \mathbf{u}_q]$ is the loading matrix consisting of the eigenvectors of the first q principal components. Denote the covariance matrix of X as Σ and the eigenvalue corresponding to eigenvector $\mathbf{u}_j$ as γ_j. Figure 4.5 shows that the distribution of the first $q = 2$ principal components scores has an annular support. In fact, this latent structure is observed for all at rest periods in this particular data and we observe it for other FBG instrumented bridges. This annular structure is seldom observed in multivariate data analysis, although has been noted on occasion (108, 109).

As far as we are aware, this latent structure is not fully appreciated in the SHM community. Again, we stress that this structure does not characterise the physics of the bridge but rather the innate properties of the sensor system. The train passage event has a marked effect on this latent structure, see Figure 4.6. More variation is attributed to the first component rather than being shared approximately equally across the first two. This property is observed over many data examples and is exactly this difference that we exploit to construct an anomaly detection system concerned with the collective spatio-temporal response of the bridge.

Figure 4.9 is a schematic that shows an illustrative data stream from a single sensor during an at rest period, followed by a train passage event, followed by another at rest period. The tick in which the train passage event starts is denoted by J_S while J_E denotes the end of the event. Our anomaly detection system will attempt to detect J_S. In practice we would want to identify the whole event but it is more convenient to simply collect a fixed amount of data, κ, following the detection. These blocks of data would then be subject to further offline analysis. The period δ is an allowance for measuring

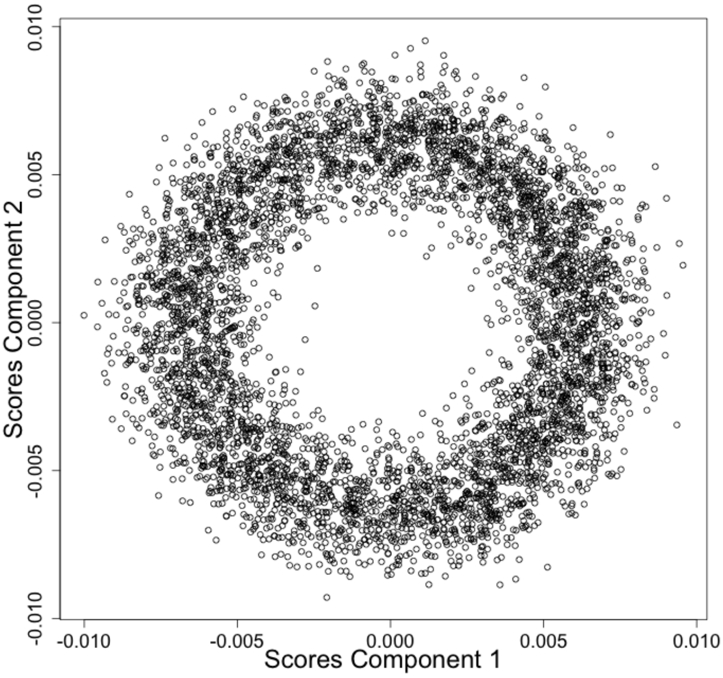

Figure 4.5: First two scores during at rest period based on a batch of 5000 measurements.

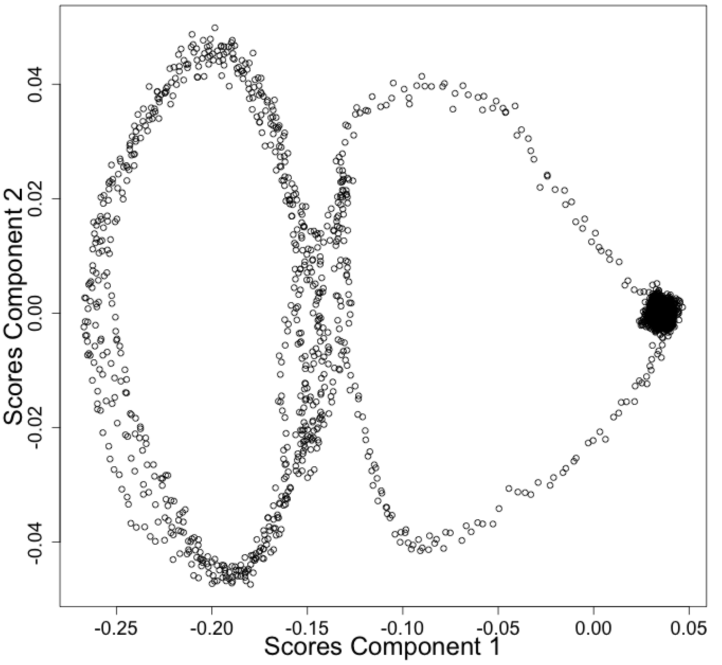

Figure 4.6: First two scores, computed using a sliding window of 5000 measurements including a train passage event. The left “lobe” corresponds to at rest periods.

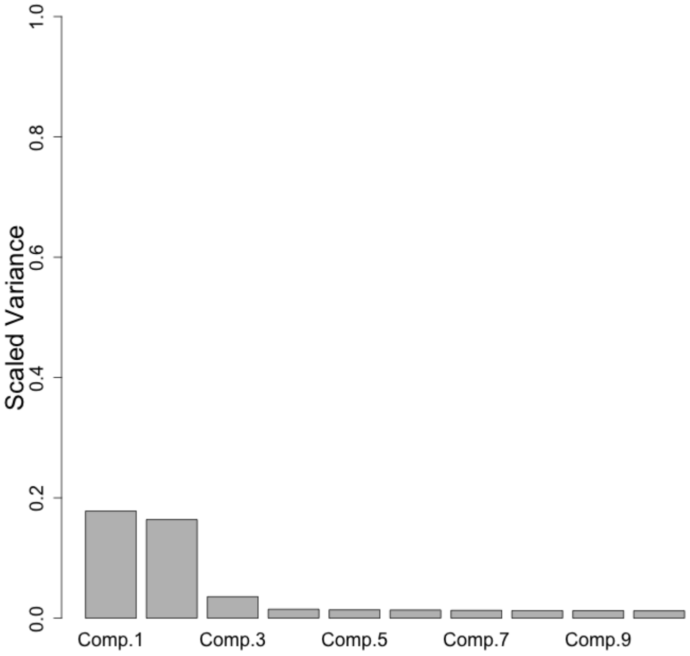

Figure 4.7: Scaled screeplot during at rest period.

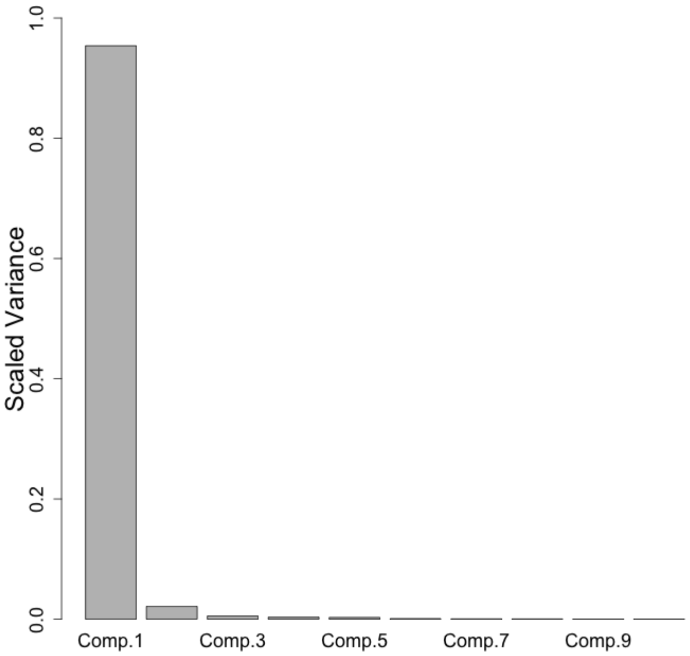

Figure 4.8: Scaled screeplot during train passage event.

the detection delay, see Section 4.4. Detection problems of this type raise the unresolved question of whether to continue estimation following detection or to restart. This is addressed further in Section 4.5.

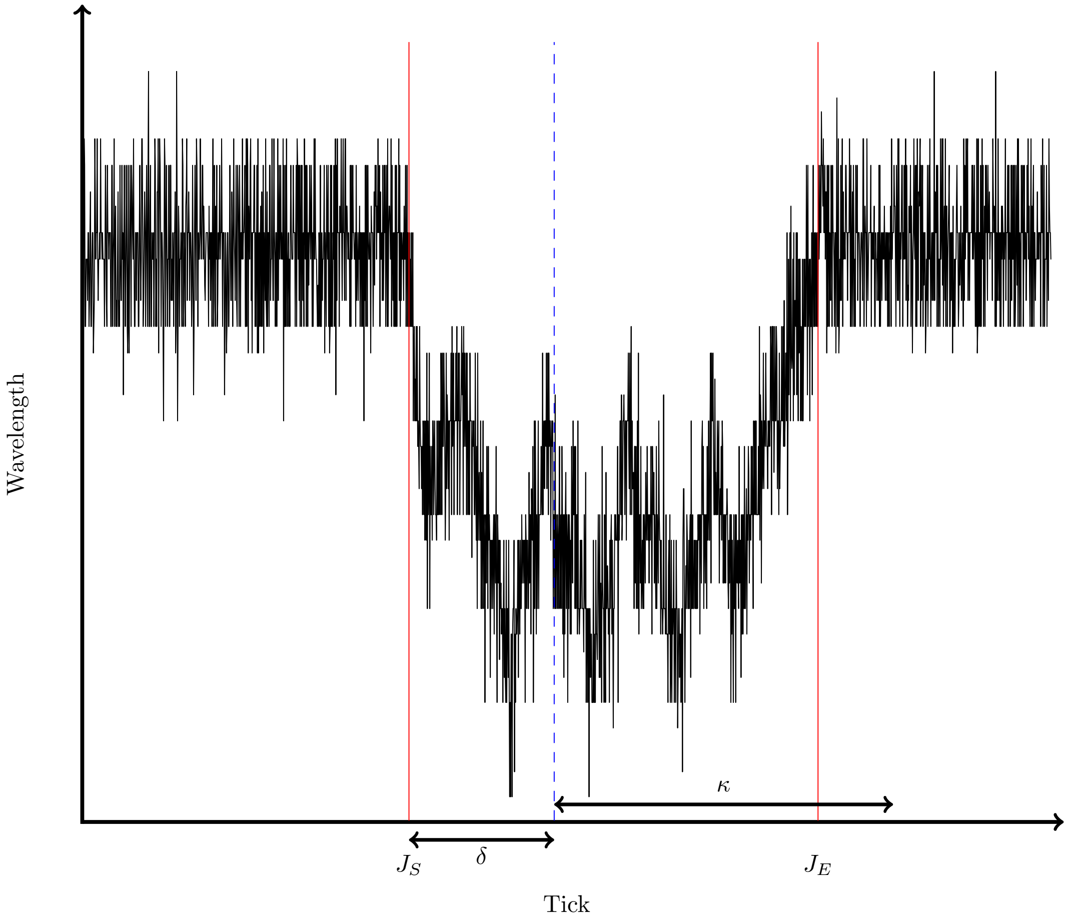

Figure 4.9: Illustration of detection when the train passage event happens. The events starts at tick J_S and ends at tick J_E. We aim to capture the event between J_S and the tolerance period δ, which we deem correct detection in simulation studies.

4.3 Streaming PCA

Computing PCA on batches of data is generally straightforward. However, computing PCA sequentially and with some capacity for temporal adaption is challenging, as described in the conclusions of (94). We develop a sequential estimation method in Section 4.3.2 through the use of adaptive filtering techniques and extend it to a streaming PCA context in Section 4.3.3.

4.3.1

Principal component analysis

Denote $X \in \mathbb{R}^{N \times d}$ as a column-centred data matrix whose rows consists of $\mathbf{x}_1, ..., \mathbf{x}_N$, where $\mathbf{x}_i$ is the d-dimensional multivariate sensor measurement at tick i. This centering is achieved using the mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]$ where $\mu_j = \frac{1}{N}\sum_{i=1}^{N} X_{i,j}$. Our methodology will later reformulate this estimator. PCA seeks an accurate representation of the original data set in a lower-dimensional subspace $\mathbb{R}^q$, $q < d$, which maximizes the explained variance. PCA seeks a projection matrix $\hat{\mathbf{U}}\hat{\mathbf{U}}^T$ which is approximated by

$$\hat{\mathbf{U}} = \underset{\mathbf{U} \in \mathbb{R}^{d \times q}, \mathbf{U}^T\mathbf{U} = I_q}{\operatorname{argmin}} ||X - \mathbf{U}\mathbf{U}^T X||_F^2 \tag{4-1}$$

where $||\cdot||_F$ denotes the *Frobenius norm*, I_q denotes the q dimensional identity matrix and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, ..., \hat{\mathbf{u}}_q]$ are the q largest eigenvectors of the sample covariance matrix Σ. These largest q eigenvectors, associated with the q leading eigenvalues $\gamma_{(1)} \geq \gamma_{(2)} \geq ... \geq \gamma_{(d)}$, are the principal components. For reasons explained in the previous section, we are interested in $q = 2$.

A streaming PCA framework needs both a sequential updating mechanism and a method for reweighting estimators as the data process changes. Recalling the quantities involved in the eigendecomposition, sequential updates are required for the mean vector $\boldsymbol{\mu}$, to achieve centering, and the covariance matrix Σ. To accommodate trend and changes in the covariance structure over time, we require time-dependent estimators, $\boldsymbol{\mu}_N$ and Σ_N. The following sections outline the construction of adaptive and sequential estimates of $\boldsymbol{\mu}_N$ and Σ_N and the subsequent sequential construction of the eigendecomposition. In this latter step, there is some scope for reducing computational burden by evaluating a partial eigendecomposition.

4.3.2

Adaptive estimation

Adaptive filtering (9) provides suitable tools for both sequential and time-dependent estimation of $\boldsymbol{\mu}_N$ and Σ_N. In practice, when dealing with streaming data, methods need to cope with (i) memory efficiency, i.e., the entire data set cannot be stored, and (ii) sequential estimation of model parameters (56, 57, 59). Temporal adaptation is provided by incorporating a forgetting factor (FF), that controls the contribution of each data point to the estimator. Practically setting this FF parameter, $\lambda \in (0, 1)$, is challenging in a streaming data context. Therefore much interest has focused on sequentially selecting an adaptive forgetting factor (AFF) – λ_N, using an updating mechanism based on stochastic gradient descent (9, 63, 62).

Consider the univariate data stream

$$\langle y_1, y_2, ..., y_{N-1}, y_N, ... \rangle,$$

from which the objective is to accurately estimate $E[Y_j]$ sequentially at each tick j. If $E[Y_j]$ is the same constant for all j then the sample mean

$$\bar{y}_j = \frac{1}{j} \sum_{i=1}^{j} y_i, \tag{4-2}$$

is a sensible estimate which admits a sequential formulation. On the other hand, if $E[Y_j]$ varies over j, then the estimate would be inappropriate. This limitation motivates the use of adaptive estimation to calculate the mean at time N, in which more weight is placed on more recent measurements. These methods result in improved estimation for time-varying processes (62). A *fixed* FF λ is introduced into Equation (4-2) with normalizing constant ($w_{N,\lambda}$) (sometimes called the *effective sample size*) to weight the estimation process as follows

$$\bar{y}_{N,\lambda} = \frac{1}{w_{N,\lambda}} \sum_{i=1}^{N} \lambda^{N-i} y_i, \qquad w_{N,\lambda} = \sum_{i=1}^{N} \lambda^{N-i}.$$

This formulation leads to a sequential computation for streaming contexts by defining the following updating mechanism for $N \geq 1$,

$$m_{N+1,\lambda} = \lambda m_{N,\lambda} + y_{N+1} \tag{4-3}$$

$$w_{N+1,\lambda} = \lambda w_{N,\lambda} + 1 \tag{4-4}$$

$$\bar{y}_{N,\lambda} = \frac{m_{N,\lambda}}{w_{N,\lambda}}, \tag{4-5}$$

with $m_{1,\lambda} = y_1$ and $w_{1,\lambda} = 1$. Setting $\lambda = 0$ corresponds to forgetting all previous measurements and only using the most recent measurement, i.e. $\bar{y}_{N,\lambda} = y_N$. On the other hand, $\lambda = 1$ corresponds to no forgetting, and then the forgetting factor mean, $\bar{y}_{N,\lambda}$, is simply the arithmetic mean given in Equation (4-2).

A more flexible approach is based on AFF, which results in a sequence of FFs $\overrightarrow{\lambda} = (\lambda_1, ..., \lambda_N)$ over time. As shown later this sequence can be selected using sequential stochastic gradient descent (SGD) approaches. Practical algorithms restrict the range of λ to prevent it becoming too small, see for example (62). We use these adaptive filtering techniques to update $\boldsymbol{\mu}_N$ and Σ_N in the eigendecomposition as described next.

Consider the following system of sequential update equations for a mean vector

$$\mathbf{m}_{N+1,\overrightarrow{\lambda}} = \lambda_N \mathbf{m}_{N,\overrightarrow{\lambda}} + \mathbf{x}_{N+1} \tag{4-6}$$

$$w_{N+1,\overrightarrow{\lambda}} = \lambda_N w_{N,\overrightarrow{\lambda}} + 1 \tag{4-7}$$

$$\boldsymbol{\mu}_{N,\overrightarrow{\lambda}} = w^{-1}_{N,\overrightarrow{\lambda}} \mathbf{m}_{N,\overrightarrow{\lambda}}, \tag{4-8}$$

with $\mathbf{m}_{1,\overrightarrow{\lambda}} = \mathbf{x}_1$ and $w_{1,\overrightarrow{\lambda}} = 1$. As pointed out by (62) and shown in Appendix

4-24, it is possible to rewrite Equation (4-8) as

$$\boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}} = \left[1 - w^{-1}_{N+1,\overrightarrow{\lambda}}\right]\boldsymbol{\mu}_{N,\overrightarrow{\lambda}} + \left[w^{-1}_{N+1,\overrightarrow{\lambda}}\right]\mathbf{x}_{N+1}, \tag{4-9}$$

which, following further manipulation, results in the following updating mechanism for the covariance matrix

$$\begin{aligned}\Sigma_{N+1,\overrightarrow{\lambda}} = &\left[1 - w^{-1}_{N+1,\overrightarrow{\lambda}}\right]\Sigma_{N,\overrightarrow{\lambda}} \\ &+ \left[w^{-1}_{N+1,\overrightarrow{\lambda}}\right](\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}})(\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}})^T,\end{aligned} \tag{4-10}$$

with $\boldsymbol{\mu}_{1,\overrightarrow{\lambda}} = \mathbf{x}_1$ and $\Sigma_{1,\overrightarrow{\lambda}} = I_d$.

The fixed forgetting version of updates (4-9) and (4-10) are familiar in the streaming PCA literature (see (94, 110, 95) and references therein). Our contribution is the introduction of an *adaptive* forgetting factor mechanism to provide time-dependent estimation.

To select λ_N, we use SGD which requires a cost function, $L_{N+1,\overrightarrow{\lambda}}$. The update of λ_N is defined as

$$\lambda_{N+1} = \lambda_N - \eta\frac{\partial}{\partial\overrightarrow{\lambda}}L_{N+1,\overrightarrow{\lambda}}, \tag{4-11}$$

where η is the step size and $\lambda_1 = 1$. We will set the step size based on performance in realistic simulations experiments.

In this paper, similar to (62), the cost function is

$$L_{N+1,\overrightarrow{\lambda}} = \left[\boldsymbol{\mu}_{N,\overrightarrow{\lambda}} - \mathbf{x}_{N+1}\right]^T\left[\boldsymbol{\mu}_{N,\overrightarrow{\lambda}} - \mathbf{x}_{N+1}\right]. \tag{4-12}$$

There is some flexibility in the choice of cost function. (63) shows that differentiable and likelihood-based cost functions yield efficient update equations for the exponential family of distributions. For example, Equation (4-9) can be motivated by an *i.i.d.* Gaussian argument.

The derivative in (4-11) depends on time-varying quantities which are evident after some manipulation, resulting in

$$\frac{\partial}{\partial \overrightarrow{\lambda}} L_{N+1,\overrightarrow{\lambda}} = 2 \left[\frac{\partial}{\partial \overrightarrow{\lambda}} \boldsymbol{\mu}_{N,\overrightarrow{\lambda}} \right]^T \left[\boldsymbol{\mu}_{N,\overrightarrow{\lambda}} - \mathbf{x}_{N+1} \right]. \tag{4-13}$$

(62) compute $\frac{\partial}{\partial \overrightarrow{\lambda}} \boldsymbol{\mu}_{N,\overrightarrow{\lambda}}$ from first principles. From their computation

$$\frac{\partial}{\partial \overrightarrow{\lambda}} \boldsymbol{\mu}_{N,\overrightarrow{\lambda}} = \left[\boldsymbol{\Delta}_{N,\overrightarrow{\lambda}} w_{N,\overrightarrow{\lambda}} - \mathbf{m}_{N,\overrightarrow{\lambda}} \Omega_{N,\overrightarrow{\lambda}} \right] / w^2_{N,\overrightarrow{\lambda}},$$

the following two auxiliary quantities appear

$$\Omega_{N+1,\overrightarrow{\lambda}} = \lambda_N \Omega_{N,\overrightarrow{\lambda}} + w_{N+1,\overrightarrow{\lambda}},$$
$$\boldsymbol{\Delta}_{N+1,\overrightarrow{\lambda}} = \lambda_N \boldsymbol{\Delta}_{N,\overrightarrow{\lambda}} + \mathbf{m}_{N+1,\overrightarrow{\lambda}},$$

which must also be updated sequentially. Note that $\boldsymbol{\Delta}_{N,\overrightarrow{\lambda}}$ is a vector and $\Omega_{N,\overrightarrow{\lambda}}$ is a scalar valued quantity where $\Omega_{1,\overrightarrow{\lambda}} = 1$ and $\boldsymbol{\Delta}_{1,\overrightarrow{\lambda}} = \mathbf{x}_1$.

This section has outlined a complete sequential updating mechanism for estimating the mean vector and covariance matrix for a time-varying process. We now turn to the construction of principal components using these time-varying estimates.

4.3.3 Streaming PCA

A naive approach would evaluate the full eigendecomposition of Σ_N or the Singular Value Decomposition (SVD). However for streaming problems this will be computationally burdensome at each tick. A number of approaches for sequentially updating an eigendecomposition have been proposed (111, 112, 113, 114, 115). In the case of a *known* covariance matrix Ψ, (116, 117), proposed a gradient ascent update for the full projection matrix of the form

$$\mathbf{U}_{N+1} = \mathbf{U}_N + \xi_N \Psi \mathbf{U}_N, \tag{4-14}$$

where ξ_N is the step size scaled by tick N, so that $\xi_N \to 0$ as $N \to \infty$. The columns of the matrix, $\mathbf{U}_N$, are the sequential updates for the q largest eigenvectors.

As pointed out by (116) and (117), Ψ is in fact unknown and so a random approximation must be adopted. As a new vector $\mathbf{x}_{N+1}$ arrives, the update is

$$\tilde{\mathbf{U}}_{N+1} = \mathbf{U}_N + \xi_N(\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1})(\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1})^T\mathbf{U}_N \tag{4-15}$$

$$\mathbf{U}_{N+1} = \Pi(\tilde{\mathbf{U}}_{N+1}), \tag{4-16}$$

where $\Pi(\cdot)$ denotes an orthonormalization operator. Orthonormalization can be achieved, for instance, using a Gram-Schmidt (GS) procedure. In (117), a sequential version of GS was proposed by combining Equations (4-15)-(4-16) (detailed in Equations (4-29)-(4-31) of Appendix 4.8.2). This operator is necessary to guarantee orthonormality as is required by PCA. The proposal included Robbins-Monro conditions on the sequence of ξ_N to ensure convergence (117, 116), i.e., $\sum_{N\geq 1}\xi_N^2 < \infty$ and $\sum_{N\geq 1}\xi_N = \infty$. This condition is clearly satisfied in the case when ξ_N is scaled by tick N. This procedure, while theoretically well justified for *i.i.d.* data, requires modification to match the requirements of streaming data analysis. In particular, we do not scale the sequence of ξ_N to decrease with time as is required to satisfy the Robbins-Monro conditions. Future data may not be generated by the same process as current data and hence it is undesirable to suppress the learning capabilities of the estimator. Bearing this in mind, we modified the results in (117), using the adaptive estimators $\boldsymbol{\mu}_{N,\overrightarrow{\lambda}}$ and $\boldsymbol{\Sigma}_{N,\overrightarrow{\lambda}}$, to update the j-th column of the matrix $\mathbf{U}_{N+1}$ as

$$\begin{aligned}\mathbf{u}_{j,N+1} &= \mathbf{u}_{j,N} + \xi\phi_{j,N} \\ &\quad \left[(\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1,\vec{\lambda}}) - \phi_{j,N}\mathbf{u}_{j,N} - 2\sum_{i=1}^{q-1}\phi_{i,N}\mathbf{u}_{i,N}\right]\end{aligned} \tag{4-17}$$

$$\phi_{j,N} = (\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1,\vec{\lambda}})^T\mathbf{u}_{j,N} \tag{4-18}$$

$$\gamma_{j,N+1} = \gamma_{j,N} + \xi(\phi_{j,N}^2 - \gamma_{j,N}), \tag{4-19}$$

denoting $\gamma_{j,N}$ as the j-th largest eigenvalue at time N and $\phi_{j,N}$ is an auxiliary quantity. The subscript of ξ was removed to denote that this is a fixed value henceforth. An advantage of this scheme is that only the q largest eigenvectors and eigenvalues are updated, which introduces a *trade-off* between accuracy and computational speed (94).

4.3.4 Computational implementation

In this section we present two implementations of streaming PCA, which we refer to as Multivariate Adaptive Forgetting Factor (MAFF) PCA, which uses adaptive estimation, and Multivariate Fixed Forgetting Factor (MFFF) PCA. The main difference between the two methods is that MAFF uses time-varying forgetting factors, tuned as in Equation (4-11). Later we prefer MAFF since it alleviates the need to select, and rely on, a single forgetting factor in perpetuity.

The step by step implementation of MFFF is implemented in Algorithm 2 while MAFF is detailed in Algorithm 3. Note that both algorithms require a burn-in period, based on $\mathcal{B}$ consecutive measurements. Both algorithms thus feature a burn-in phase, followed by adaptive and sequential updating. Note also that both algorithms include a sequential anomaly detection stage, described in Section 4.4.

In Algorithm 2, first select the value for the FF and step size $\xi \in \mathbb{R}^+$. There is no principled method to select the FF in advance. Then initial values for the recursive estimators are set as $\mathbf{m}_{1,\lambda} = \mathbf{x}_1$, $w_{1,\lambda} = 1$, $\Sigma_1 = I_d$. Note that the updates of $\mathbf{m}_{N+1,\lambda}$, $w_{N+1,\lambda}$, $\boldsymbol{\mu}_{N+1,\lambda}$ and $\Sigma_{N+1,\lambda}$ are the same as Equations (4-6), (4-7), (4-9), (4-10). The only difference is that $\lambda_N = \lambda$. Other input parameters, such as ξ, will be determined by simulation studies based on properties of the real data.

Algorithm 2 **M**ultivariate **F**ixed **F**orgetting **F**actor **PCA**

```
Require: $\xi$, $\mathbf{x}_1$, $\mathbf{m}_{1,\lambda}$, $w_{1,\lambda}$, $\Sigma_1$, $\lambda$, $\mathcal{B}$, $\delta$, $\mathcal{W}$
 1: for $N \leftarrow 1, ..., \mathcal{B}$ do
 2:     Receive $\mathbf{x}_{N+1}$
 3:     Update $\mathbf{m}_{N+1,\lambda}$, $w_{N+1,\lambda}$, $\boldsymbol{\mu}_{N+1,\lambda}$, $\Sigma_{N+1,\lambda}$
 4: if $N = \mathcal{B}$ then
 5:     $[\mathbf{U}_N, \boldsymbol{\gamma}_N] = SVD(\Sigma_{N,\lambda})$
 6: for $N \leftarrow \mathcal{B} + 1, ...$ do
 7:     Receive $\mathbf{x}_{N+1}$
 8:     Update $\mathbf{m}_{N+1,\lambda}$, $w_{N+1,\lambda}$, $\boldsymbol{\mu}_{N+1,\lambda}$, $\Sigma_{N+1,\lambda}$
 9:     Update $\mathbf{U}_{N+1}$, $\boldsymbol{\phi}_N$, $\boldsymbol{\gamma}_{N+1}$
10:     if $N \geq \mathcal{B} + \mathcal{W}$ then
11:         $\ell_N = (\gamma_{1,N} - \gamma_{2,N})^2$
12:         $\hat{p}_N = \frac{1}{\mathcal{W}+1} \sum_{j=N-\mathcal{W}}^{N} 1(\ell_j \geq \ell_N)$
13:         Flag if $\hat{p}_N < \frac{1}{\mathcal{W}}$ is satisfied 3 times in a row
14:         Save the set $\{\ell_{N-\mathcal{W}}, \ell_{N-\mathcal{W}+1}, ..., \ell_N\}$ in memory.
```

For the implementation of MAFF, Algorithm 3, first select the values for both step sizes $\xi \in \mathbb{R}^+$ and $\eta \in \mathbb{R}^+$. Then initial values for the recursive estimators are set as $\mathbf{m}_{1,\overrightarrow{\lambda}} = \mathbf{x}_1$, $\boldsymbol{\Delta}_{1,\overrightarrow{\lambda}} = \mathbf{x}_1$, $w_{1,\overrightarrow{\lambda}} = 1$, $\Omega_{1,\overrightarrow{\lambda}} = 1$, $\Sigma_1 = I_d$, $\lambda_1 = 1$, $\lambda_{min} = 0.6$. After updating λ_N, the equation $\lambda_{N+1} = \max\{\min\{\lambda_{N+1}, 1\}, \lambda_{min}\}$ in Line 4, is to guarantee that $\lambda_N \in (\lambda_{min}, 1)$.

The first $\mathcal{B}$ ticks, corresponding to a burn-in, are used to estimate $\mathbf{U}_N$ and $\boldsymbol{\gamma}_{N+1}$ using SVD (Lines 1-5 in Algorithm 2 and Lines 1-6 in Algorithm 3). After this burn-in the estimation of the eigendecomposition is sequential and adaptive. Note that the vectors $\boldsymbol{\phi}_N$ and $\boldsymbol{\gamma}_N$ in Lines 9 (Algorithm 2) and 11 (Algorithm 3) are updated after the burn-in using Equation (4-18) and Equation (4-19), respectively.

This instrumented infrastructure application involves data measured at high frequency, and so it is reasonable to set the burn-in period of both

Algorithm 3 **M**ultivariate **A**daptive **F**orgetting **F**actor **PCA**

```
Require: $\xi$, $\eta$, $\mathbf{x}_1$, $\mathbf{m}_{1,\overrightarrow{\lambda}}$, $w_{1,\overrightarrow{\lambda}}$, $\mathbf{\Delta}_{1,\overrightarrow{\lambda}}$, $\Omega_{1,\overrightarrow{\lambda}}$, $\Sigma_1$, $\lambda_1$, $\lambda_{min}$, $\mathcal{B}$, $\delta$, $\mathcal{W}$
 1: for $N \leftarrow 1, ..., \mathcal{B}$ do
 2:     Receive $\mathbf{x}_{N+1}$
 3:     Update $\mathbf{m}_{N+1,\overrightarrow{\lambda}}$, $w_{N+1,\overrightarrow{\lambda}}$, $\boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}}$, $\Sigma_{N+1,\overrightarrow{\lambda}}$, $\mathbf{\Delta}_{N+1,\overrightarrow{\lambda}}$, $\Omega_{N+1,\overrightarrow{\lambda}}$, $\lambda_{N+1}$
 4:     With $\lambda_{N+1} = \max\{\min\{\lambda_{N+1}, 1\}, \lambda_{min}\}$
 5: if $N = \mathcal{B}$ then
 6:     $[\mathbf{U}_N, \gamma_N] = SVD(\Sigma_{N,\overrightarrow{\lambda}})$
 7: for $N \leftarrow \mathcal{B} + 1, ...$ do
 8:     Receive $\mathbf{x}_{N+1}$
 9:     Update $\mathbf{m}_{N+1,\overrightarrow{\lambda}}$, $w_{N+1,\overrightarrow{\lambda}}$, $\boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}}$, $\Sigma_{N+1,\overrightarrow{\lambda}}$, $\mathbf{\Delta}_{N+1,\overrightarrow{\lambda}}$, $\Omega_{N+1,\overrightarrow{\lambda}}$, $\lambda_{N+1}$
10:     With $\lambda_{N+1} = \max\{\min\{\lambda_{N+1}, 1\}, \lambda_{min}\}$
11:     Update $\mathbf{U}_{N+1}$, $\boldsymbol{\phi}_N$, $\boldsymbol{\gamma}_{N+1}$
12:     if $N \geq \mathcal{B} + \mathcal{W}$ then
13:         $\ell_N = (\gamma_{1,N} - \gamma_{2,N})^2$
14:         $\hat{p}_N = \frac{1}{\mathcal{W}+1} \sum_{j=N-\mathcal{W}}^{N} 1(\ell_j \geq \ell_N)$
15:         Flag if $\hat{p}_N < \frac{1}{\mathcal{W}}$ is satisfied 3 times in a row
16:         Save the set $\{\ell_{N-\mathcal{W}}, \ell_{N-\mathcal{W}+1}, ..., \ell_N\}$ in memory.
```

algorithms $\mathcal{B}$ to 500 ticks. Post burn-in, anomaly detection is performed by retaining a set of $\mathcal{W}$ derived values, see Lines 10-14 (Algorithm 2) and 12-16 (Algorithm 3). This set is used to calculate a p-value which provides the basis of the anomaly detector.

4.4 Anomaly detection

Given the estimation methodology introduced in the previous section, we now develop an anomaly detection mechanism. In general, inference for the parameters of an eigendecomposition is difficult, and there are few results on the distribution of such estimators (118, 119). One can reason about tracking the largest eigenvalue in PCA (118), or even the ratio of the largest eigenvalue to the sum of all eigenvalues (119), using the Tracy-Widom distribution. However, these results require assumptions based on asymptotic theory which are not valid for streaming data. A different approach is required.

To overcome these issues and avoid making distributional assumptions, we adapt Conformal Prediction (CP) to propose a streaming PCA anomaly detector. CP was developed by the machine learning community (11, 120, 12, 121, 122) and is now receiving substantial attention from the statistics community (123, 124, 125, 126, 127). CP methods require storage of $\mathcal{W}$ derived quantities. These derived quantities are called *non-conformity* measures, which makes reference to a distance metric specified by the user. Selecting an appropriate measure remains an open problem in the literature (11). We propose a particular measure that is designed for this instrumented infrastructure application, motivated by the behaviour depicted in Figures 4.7 and 4.8.

4.4.1 Non-conformity measure

Given the multivariate data stream $\mathbf{x}_N$ and the eigenvalues derived from it, $\boldsymbol{\gamma}_N$, we seek to construct a non-conformity measure with which to perform inference.

To monitor for anomalies, the sequence of eigenvalues is converted into the stream of *distances*

$$\langle \ell_2, \ell_3, \ldots, \ell_{N-1}, \ell_N, \ldots \rangle, \qquad \ell_N \equiv \mathrm{D}(\gamma_{1,N}, \gamma_{2,N}), \tag{4-20}$$

using any valid distance metric $\mathrm{D}(\cdot,\cdot)$. Inspection of extensive bridge data shows that during at rest periods the eigendecomposition is consistent with that depicted in Figures 4.5 and 4.7. In contrast during train passage events, there is a significant difference between the two first eigenvalues, which are approximately equal during periods of rest. Hence Equation (4-20) restricts attention to the first two eigenvalues. This motivates the use of a non-conformity measure

$$\ell_N = (\gamma_{1,N} - \gamma_{2,N})^2. \tag{4-21}$$

Of course, other measures are possible, but as noted above there is no principled way to select a measure.

Following CP we sequentially estimate a p-value as

$$\hat{p}_N = \frac{1}{\mathcal{W}+1}\sum_{j=N-\mathcal{W}}^{N} 1(\ell_j \geq \ell_N), \tag{4-22}$$

where $1(\cdot)$ is the indicator function. Generally CP would allow the set of derived quantities, $\mathcal{W}$, to increase with data arrival. This is not suitable for streaming data contexts because both the computational and memory demand would increases over time. Instead we restrict this set to a sliding window of fixed size. By construction $\hat{p}_N \in [\frac{1}{\mathcal{W}+1}, 1]$.

Equation (4-22) defines the p-value, $\hat{p}_N$, which forms the basis of our anomaly detection system. As is often the case in practical sequential analysis, it is useful to incorporate a "run rule" (128), which both reduces false positives and provides some resistance to outliers. We will adopt a run rule in which an anomaly is flagged at tick j if $\hat{p}_j$, $\hat{p}_{j-1}$ and $\hat{p}_{j-2}$ are all less than $\alpha = \frac{1}{\mathcal{W}}$. The latter implies that at least one anomaly will be detected, on average, after observing a set of $\mathcal{W}$ measurements. Theoretically, setting $\mathcal{W}$ large is desirable. Practically however, a balance must be struck between memory and computing constraints. For the bridge data, experimentation suggests that $\mathcal{W} = 10000$, 40 seconds of data, provides good results.

4.5 Simulation

In this section, the performance of these streaming PCA methods, MFFF (Algorithm 2) and MAFF (Algoritm 3), are evaluated. The scaled version of MFFF (as discussed in Section 4.3.3), in which the Robins-Monroe conditions are satisfied, does not perform well in simulations and is not considered in detail. This latter approach, which makes little sense in a streaming context, is denoted as MFFFS.

These three methods are compared based on *Estimation Accuracy* and *Detection Performance.* The first is intended to evaluate how well the algorithm tracks properties of a time-varying eigendecomposition. The second measures the effectiveness of the method proposed in Section 4.4, at detecting anomalies.

It is well known that such computational methods are sensitive to the choice of input parameters. Thus, these simulations are also used to determine input parameters for the bridge data analysis of Section 4.6. Details of the search grid for input parameters are presented in Appendix 4.8.3.

We simulate data, $X \in \mathbb{R}^{N \times d}$, that reproduces important characteristics reported in Section 4.2, with $d = 80$ for consistency with the bridge data and $N = 40500$ ticks. The key features to be replicated are the annular structure observed during at rest periods and its collapse during train passage events. A detailed description of the data generation processes is given in Appendix 4.8.4.

Considering Estimation Accuracy, recall that $\boldsymbol{\gamma}_j$ is a vector of the 2 leading eigenvalues estimated by a streaming PCA procedure. In the simulated setting the corresponding true eigenvalues are denoted as $\boldsymbol{\Gamma}_j$. Accuracy is measured using

$$\frac{1}{N}\sum_{j=1}^{N}\left\|\boldsymbol{\Gamma}_j - \boldsymbol{\gamma}_j\right\|. \tag{4-23}$$

We will report the average of this error measure, $\mathcal{E}$, over 100 Monte Carlo replicates considering $i.i.d.$ data (data generation described in Appendix 4.8.4.1).

To identify whether a flagged change corresponds to a real anomaly in simulation studies we will use a window around the real anomaly. Recall we use a run rule which flags the anomaly as the first tick in the run. Similar to (85), if a flag is given in a window of $\delta = 125$ measurements following a true anomaly, the detection is deemed a correct detection (CD). A flag outside this window is deemed a false detection (FD). The tolerance period, δ, is equivalent to half a second for these data. Average number of FD and CD rates are calculated over 100 replicates of simulated train passage event data (see Appendix 4.8.4.2).

4.5.1

Simulation results and control parameter selection

We seek to compare MAFF and MFFF according to various measures of performance. Additionally we use the results of the simulation study to identify control parameters for Section 4.6.

η	ξ	$\mathcal{E}$	CD	FD
0.1	0.01	0.17	0.70	3.40
0.01	0.01	0.10	0.61	3.66
0.001	0.01	0.09	0.75	3.45
1e-04	0.01	0.08	0.76	2.99
1e-05	0.01	0.08	0.70	3.76
1e-06	0.01	0.08	0.77	3.58
1e-07	0.01	0.08	0.73	3.33

λ	ξ	$\mathcal{E}$	CD	FD
0.85	0.01	0.22	0.63	3.80
0.9	0.01	0.15	0.67	3.60
0.95	0.01	0.10	0.76	3.37
0.99	0.01	0.08	0.79	3.35

Table 4.1: Average results over 100 Monte Carlo replicates, where the first two columns are control parameters of each method, $\mathcal{E}$ is the average error. CD is expressed as a rate and FD is the average number of points. Left: results for MAFF. Right: results for MFFF.

Table 4.1 reports the results of the performance measures for MAFF and MFFF, left and right, respectively. The first two columns of each table refer to input parameters for the respective methods. Note, these tables have been reduced to report only configurations of input parameters which have $CD > 0.5$ and $FD < 4$. The full tables are available in Appendix 4.8.5. The complete results demonstrate that good performance requires very specific choice of parameters.

The results in Tables 4.1 embody two scenarios. First, *i.i.d.* data which is included to address estimation issues alone. The column denoted $\mathcal{E}$ reports the average estimation accuracy, Equation (4-23). Interestingly, the parameter ξ, which has the same role in both methods, the value 0.01 is very frequently selected as a good choice.

In terms of detection performance, simulated train passage event data was used to assess detection performance, as reported in the CD and FD column of Table 4.1. These tables indicate broadly the same performance over the reported input parameters. Specifically, a CD rate of around 0.7 and FD around 3.5. Without using the run-rule, we expect an average of four false detections.

As always, different performance measures demonstrate different characteristics for fixed choices of parameters, and hence our selection will seek to balance estimation accuracy and detection performance. This balance is biased in favour of detection performance which itself is characterised by two measures, CD and FD.

Given the simulation results, we use MAFF for the bridge data with parameters $\eta = 1e-6$, $\xi = 0.01$. We favour MAFF over MFFF, because fixing λ in perpetuity seems an overcommitment of knowledge.

Finally, note that MFFFS behaves in a predictable and practically useless manner. Specifically, MFFFS is shown by simulation (see Table 4.4 in the Appendix 4.8.5) to be incapable of adaptively revising its estimates following the train passage event. This is unsurprising since the scaling of MFFFS renders adaptive learning impossible after sufficient data have been observed.

4.6 Bridge data

In this section we demonstrate the result of our MAFF approach for the bridge data described in Section 4.2. Input parameters selected in the previous section were used throughout this section.

As noted earlier, the existence of the latent annular structure, illustrated in Figure 4.5, has rarely been documented, a notable exception being (108). This structure is persistent when evaluated over sliding windows during periods of rest. Notably when using MAFF, this structure is preserved, see Figure 4.10.

The MAFF method also reveals a previously unknown feature of this type of distributed sensor system. Figure 4.11 shows the maximum eigenvalue computed by MAFF over a long period. There were four train passage events in this period. The striking feature of this figure does not relate to train passage events but rather is the obvious periodicity. The periodicity is very

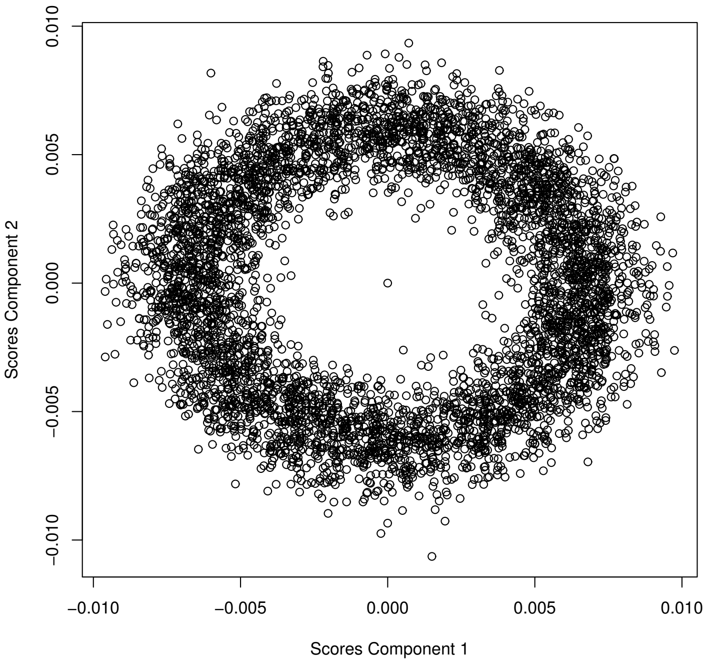

Figure 4.10: *Online annulus* calculated considering 5000 data points, sequentially centred with $\boldsymbol{\mu}_{N+1,\overrightarrow{\lambda}}$, and a selected $\mathbf{U}_N$ during an at rest period.

close to 5 minutes and 40 seconds. Neither we, nor engineering colleagues, are able to explain this beyond attributing it to spatio-temporal proprieties of the sensor system. Worth to mention that one could reason about an anomaly detector that flags a train passage event when the maximum eigenvalue exceeds a threshold. We argue that such a proposal would be *ad-hoc* in nature, which is unfeasible considering that this data is sequentially revealed.

Turning now to anomaly detection, recall that the data set has four train passage events, which were manually identified. Each frame of Figure 4.12 shows $\hat{p}_N$ zoomed-in around the train passage events. The MAFF misses only one of these four events. Excluding the four train event periods the anomaly detector signalled 0.0262%, 157 ticks of 598422, as false positives.

Finally, considering computational burden, we need to address memory and efficiency issues. The MAFF method has constant memory and compute demand per tick, as required by streaming applications. In the example just given, the entire data set consists of 611108 measurements which is equivalent

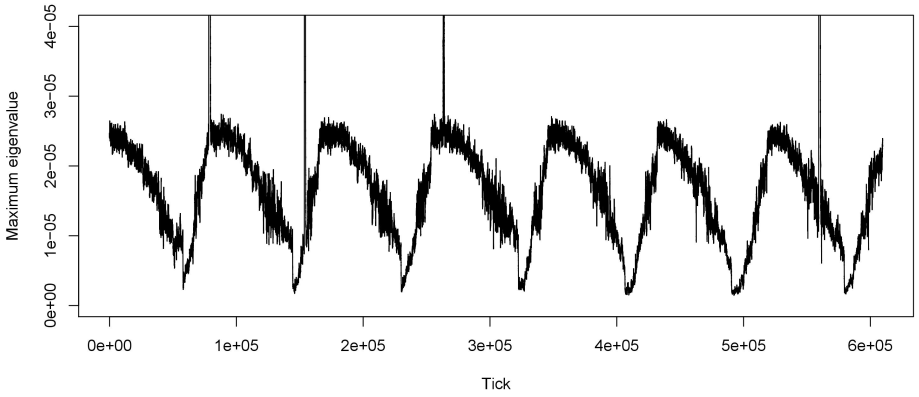

Figure 4.11: Maximum eigenvalue sequentially estimated by MAFF over 611108 data points (approximately 40 minutes and 44 seconds). The four spikes corresponds to train passage events. The maximum value on the vertical axis was selected so the cyclic feature is more evident.

to 40 minutes and 44 seconds on a 250 Hz frequency sample. The code developed in R, not optimized, required 8 minutes and 49 seconds to process the data. Thus the procedure is capable of processing such data in real time.

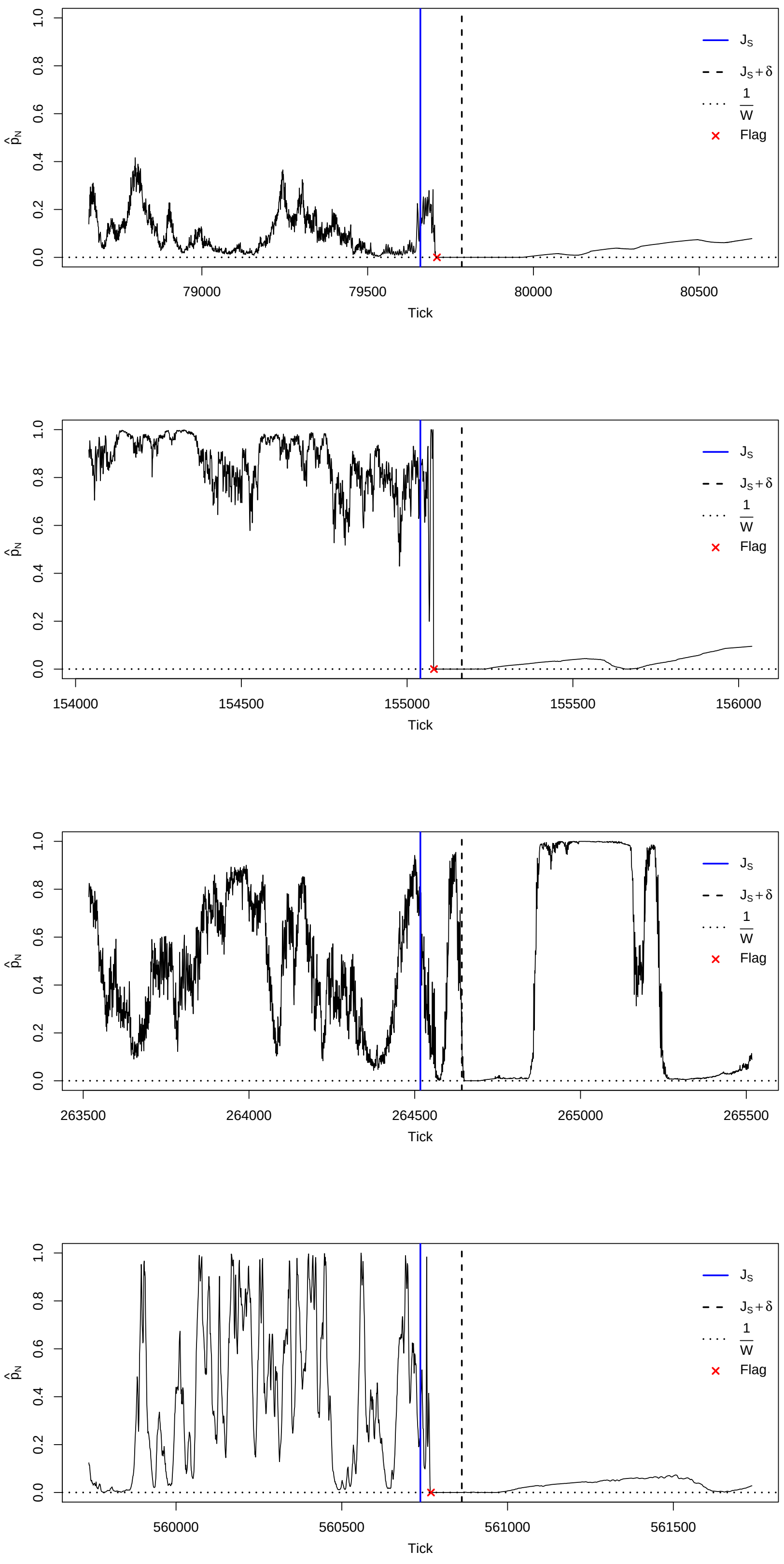

Figure 4.12: Conformal p-values, $\hat{p}_N$, for flagged periods. Solid vertical lines denote the tick in which the train passage event started. The interval between the solid vertical lines and dashed vertical lines denote the tolerance period δ. Horizontal dotted lines indicate $\alpha = \frac{1}{\mathcal{W}}$. The red cross indicates the first tick for which a flag occurred, based on the run rule.

4.7 Conclusion

We have considered data analysis challenges arising from instrumented infrastructure. Specifically, we have developed a novel streaming methodology for multivariate anomaly detection over a spatio-temporal object. The MAFF method has been extended using an adaptive forgetting factor and derived quantities from the method are calibrated using Conformal Prediction in a fixed window. Given appropriate control parameters, and a run-rule, the algorithm provides both effective tracking performance and accurate detection capability.

Deployed against real bridge data, the method has an acceptable detection performance for train passage events. Notably the method reveals a long term cyclic dependence structure that has not been previously reported.

4.8 Appendix

4.8.1 Proof

Equivalence between Equation (4-8) and Equation (4-9) considering the univariate case.

$$\bar{y}_{N,\lambda} = \frac{m_{N,\lambda}}{w_{N,\lambda}} \tag{4-24}$$

$$\bar{y}_{N,\lambda} = \frac{\lambda m_{N-1,\lambda} + y_N}{w_{N,\lambda}} \tag{4-25}$$

$$\bar{y}_{N,\lambda} = \frac{\lambda w_{N-1,\lambda} \bar{y}_{N-1,\lambda}}{w_{N,\lambda}} + \frac{y_N}{w_N} \tag{4-26}$$

$$\bar{y}_{N,\lambda} = \frac{[w_{N,\lambda} - 1]\bar{y}_{N-1,\lambda}}{w_{N,\lambda}} + \frac{y_N}{w_N} \tag{4-27}$$

$$\bar{y}_{N,\lambda} = \left(1 - \frac{1}{w_{N,\lambda}}\right) \bar{y}_{N-1,\lambda} + \frac{y_N}{w_N} \tag{4-28}$$

4.8.2 Stochastic gradient ascent

In order to update the j-th column of the matrix $\mathbf{U}_{N+1}$ under orthonormality conditions, (117) proposed the following

$$\begin{aligned}\mathbf{u}_{j,N+1} &= \mathbf{u}_{j,N} + \xi_N \phi_{j,N} \\ &\qquad \left[(\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1}) - \phi_{j,N}\mathbf{u}_{j,N} - 2\sum_{i=1}^{j-1} \phi_{i,N}\mathbf{u}_{i,N} \right] && \text{(4-29)} \\ \phi_{j,N} &= (\mathbf{x}_{N+1} - \boldsymbol{\mu}_{N+1})^T \mathbf{u}_{j,N} && \text{(4-30)} \\ \gamma_{j,N+1} &= \gamma_{j,N} + \xi_N(\phi_{j,N}^2 - \gamma_{j,N}), && \text{(4-31)}\end{aligned}$$

where $\xi_N = \frac{\xi}{N}$.

4.8.3 Input parameters

The lack of theoretical background on the optimal choice of SGD step size (129) motivates the exploration of a grid of values for the proposed methods. The adopted grids are

$$\xi \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\},$$

for the eigendecomposition step,

$$\eta \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\},$$

for the adaptive forgetting factor step and

$$\lambda \in \{0.6, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99\},$$

for the FF. This simulation study explores all possible combinations of ξ, η and λ.

4.8.4 Data generation process

For both data generation processes, calculate the sample covariance matrix Σ as

$$N\Sigma = X^T X = \mathbf{U}\Lambda\mathbf{U}^T, \tag{4-32}$$

where Λ denotes a diagonal matrix composed of the eigenvalues $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_{80}$.

4.8.4.1 Independent identical distributed data

In order to generate the *i.i.d.* data set, with only resting periods, one should adopt the following steps.

1. Generate one random orthonormal matrix, $\Upsilon \in \mathbb{R}^{80\times 80}$ using QR decomposition (130) or Householder projections (131);

2. Generate one diagonal matrix $A \in \mathbb{R}^{80\times 80}$ with elements $A_{[1,1]} = 1$, $A_{[2,2]} = 0.01$, $A_{[j,j]} \sim Unif[-10^{-6}, 10^{-6}]$, $\forall j = 3, ...80$. Note that the notation $\psi_{[j,j]}$ makes reference to the element in the j-th row, j-th column of the ψ matrix.

3. Fixing 40500 as the sample size, generates the stream $X \in \mathbb{R}^{40500\times 80} \sim \mathcal{N}(\boldsymbol{\mu}, \Upsilon^T A \Upsilon)$ where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a multivariate Gaussian distribution, with mean vector $\boldsymbol{\mu} = [1, 2, ..., 80] \in \mathbb{R}^{80}$ and covariance matrix Σ. Such mean vector was proposed since the sensors are aligned in an increasing sequence equally spaced.

4.8.4.2 Train passage event data

To generate data considering a rest period, followed by a train passage event and followed by another rest period, we adopted the following steps.

1. Generate two random orthonormal matrixes, $\Upsilon \in \mathbb{R}^{80\times 80}$ and $\Xi \in \mathbb{R}^{80\times 80}$ using QR decomposition (130) or Householder projections (131);

2. Generate two diagonal matrixes $A \in \mathbb{R}^{80\times 80}$ and $B \in \mathbb{R}^{80\times 80}$ with elements $A_{[1,1]} = 1$, $A_{[2,2]} = 0.01$, $A_{[j,j]} \sim Unif[-10^{-6}, 10^{-6}]$, $\forall j = 3, ...80$ and $B_{[1,1]} = 5$ $B_{[j,j]} \sim Unif[-10^{-6}, 10^{-6}]$, $\forall j = 2, ...80$. Note that the notation $\psi_{[j,j]}$ makes reference to the element in the j-th row, j-th column of the ψ matrix.

3. Fixing 40500 as the sample size, generates the stream $X \in \mathbb{R}^{40500\times 80}$ as

$$X_N = \begin{cases} \mathcal{N}(\boldsymbol{\mu}, \Upsilon^T A \Upsilon), & N \leq 20000, \\ \mathcal{N}(\boldsymbol{\mu}, \Xi^T B \Xi), & 20001 \leq N \leq 20500, \\ \mathcal{N}(\boldsymbol{\mu}, \Upsilon^T A \Upsilon), & 20501 \leq N \leq 40500 \end{cases}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a multivariate Gaussian distribution, with mean vector $\boldsymbol{\mu} = [1, 2, ..., 80] \in \mathbb{R}^{80}$ and covariance matrix Σ. Such mean vector was proposed since the sensors are align in an increasing sequence equally spaced.

4.8.5 Simulation results

In this subsection, we report the results of Section 4.5 for the whole grid of input parameters. The average error metric of Equation (4-23), considering *i.i.d.* context and also the CD and FD performances on the train passage event scenario for MAFFF (Table 4.2), MFFF (Table 4.3) and MFFFS (Table 4.4) methods.

η	ξ	$\mathcal{E}$	CD	FD
0.1	0.1	0.27	0.41	68.01
0.1	0.01	0.17	0.70	3.40
0.1	0.001	0.16	0.00	2.32
0.1	1e-04	0.17	0.00	20.93
0.1	1e-05	0.23	0.00	98.77
0.1	1e-06	0.29	0.01	213.61
0.1	1e-07	0.28	0.00	126.96
0.01	0.1	0.26	0.34	203.80
0.01	0.01	0.10	0.61	3.66
0.01	0.001	0.08	0.00	2.62
0.01	1e-04	0.10	0.00	14.32
0.01	1e-05	0.14	0.00	85.11
0.01	1e-06	0.15	0.00	97.74
0.01	1e-07	0.19	0.00	92.91
0.001	0.1	0.26	0.35	203.77
0.001	0.01	0.09	0.75	3.45
0.001	0.001	0.04	0.00	2.92
0.001	1e-04	0.05	0.00	11.04
0.001	1e-05	0.10	0.00	32.81
0.001	1e-06	0.10	0.00	52.15
0.001	1e-07	0.11	0.00	44.40
1e-04	0.1	0.26	0.40	271.73
1e-04	0.01	0.08	0.76	2.99
1e-04	0.001	0.03	0.00	3.12
1e-04	1e-04	0.02	0.00	9.11
1e-04	1e-05	0.05	0.00	20.02
1e-04	1e-06	0.07	0.00	24.03
1e-04	1e-07	0.06	0.00	22.69
1e-05	0.1	0.26	0.48	339.58
1e-05	0.01	0.08	0.70	3.76
1e-05	0.001	0.03	0.00	3.08
1e-05	1e-04	0.02	0.00	9.44
1e-05	1e-05	0.04	0.00	20.32
1e-05	1e-06	0.06	0.00	20.51
1e-05	1e-07	0.05	0.00	23.05
1e-06	0.1	0.26	0.50	747.09
1e-06	0.01	0.08	0.77	3.58
1e-06	0.001	0.03	0.00	3.15
1e-06	1e-04	0.02	0.00	8.63
1e-06	1e-05	0.05	0.00	12.25
1e-06	1e-06	0.05	0.00	22.47
1e-06	1e-07	0.05	0.00	22.01
1e-07	0.1	0.26	0.31	135.87
1e-07	0.01	0.08	0.73	3.33
1e-07	0.001	0.03	0.00	3.09
1e-07	1e-04	0.02	0.00	8.16
1e-07	1e-05	0.05	0.00	22.77
1e-07	1e-06	0.05	0.00	25.65
1e-07	1e-07	0.05	0.00	27.85

Table 4.2: Average results for MAFF over 100 Monte Carlo replicates, where the first two columns are control parameters of each method, $\mathcal{E}$ is the average error. CD is expressed as a rate and FD is the average number of points.

λ	ξ	$\mathcal{E}$	CD	FD
0.6	0.1	0.56	0.32	0.07
0.6	0.01	0.56	0.22	4.71
0.6	0.001	0.56	0.00	2.30
0.6	1e-04	0.55	0.00	29.96
0.6	1e-05	0.62	0.01	265.87
0.6	1e-06	0.59	0.01	291.35
0.6	1e-07	0.61	0.00	309.84
0.75	0.1	0.38	0.44	68.04
0.75	0.01	0.36	0.50	4.06
0.75	0.001	0.36	0.00	2.23
0.75	1e-04	0.38	0.00	26.90
0.75	1e-05	0.42	0.00	151.68
0.75	1e-06	0.42	0.00	195.25
0.75	1e-07	0.45	0.01	242.70
0.8	0.1	0.33	0.40	67.94
0.8	0.01	0.29	0.44	4.03
0.8	0.001	0.29	0.00	2.12
0.8	1e-04	0.29	0.00	19.65
0.8	1e-05	0.36	0.00	142.96
0.8	1e-06	0.42	0.01	214.73
0.8	1e-07	0.41	0.00	211.68
0.85	0.1	0.29	0.36	135.89
0.85	0.01	0.22	0.63	3.80
0.85	0.001	0.22	0.00	2.31
0.85	1e-04	0.27	0.00	23.87
0.85	1e-05	0.30	0.01	121.21
0.85	1e-06	0.33	0.00	191.14
0.85	1e-07	0.32	0.00	161.07
0.9	0.1	0.27	0.32	0.05
0.9	0.01	0.15	0.67	3.60
0.9	0.001	0.15	0.00	2.57
0.9	1e-04	0.17	0.00	16.58
0.9	1e-05	0.25	0.01	105.92
0.9	1e-06	0.26	0.00	136.35
0.9	1e-07	0.30	0.00	149.92
0.95	0.1	0.25	0.43	407.55
0.95	0.01	0.10	0.76	3.37
0.95	0.001	0.08	0.00	2.58
0.95	1e-04	0.11	0.00	12.84
0.95	1e-05	0.15	0.00	53.32
0.95	1e-06	0.18	0.00	75.07
0.95	1e-07	0.17	0.00	74.40
0.99	0.1	0.25	0.45	475.43
0.99	0.01	0.08	0.79	3.35
0.99	0.001	0.03	0.00	2.78
0.99	1e-04	0.04	0.00	9.12
0.99	1e-05	0.07	0.00	27.86
0.99	1e-06	0.08	0.00	30.89
0.99	1e-07	0.08	0.00	30.53

Table 4.3: Average results for MFFF over 100 Monte Carlo replicates, where the first two columns are control parameters of each method, $\mathcal{E}$ is the average error. CD is expressed as a rate and FD is the average number of points.

λ	ξ	$\mathcal{E}$	CD	FD
0.6	0.1	0.56	0.00	176.69
0.6	0.01	0.59	0.02	264.79
0.6	0.001	0.58	0.01	284.16
0.6	1e-04	0.57	0.00	301.95
0.6	1e-05	0.60	0.01	324.99
0.6	1e-06	0.58	0.02	282.45
0.6	1e-07	0.61	0.01	358.75
0.75	0.1	0.42	0.00	148.67
0.75	0.01	0.44	0.00	250.38
0.75	0.001	0.49	0.01	287.51
0.75	1e-04	0.41	0.00	184.63
0.75	1e-05	0.42	0.00	236.80
0.75	1e-06	0.46	0.01	211.95
0.75	1e-07	0.45	0.02	201.07
0.8	0.1	0.39	0.00	155.10
0.8	0.01	0.43	0.00	229.62
0.8	0.001	0.38	0.01	194.33
0.8	1e-04	0.43	0.01	243.41
0.8	1e-05	0.38	0.00	189.50
0.8	1e-06	0.39	0.00	200.82
0.8	1e-07	0.37	0.00	172.78
0.85	0.1	0.28	0.00	118.21
0.85	0.01	0.34	0.00	135.31
0.85	0.001	0.34	0.00	183.92
0.85	1e-04	0.34	0.00	169.95
0.85	1e-05	0.37	0.00	231.49
0.85	1e-06	0.32	0.01	181.24
0.85	1e-07	0.37	0.00	214.87
0.9	0.1	0.23	0.00	84.41
0.9	0.01	0.27	0.00	152.84
0.9	0.001	0.28	0.00	169.79
0.9	1e-04	0.24	0.00	123.97
0.9	1e-05	0.25	0.00	112.74
0.9	1e-06	0.29	0.00	169.13
0.9	1e-07	0.29	0.00	150.57
0.95	0.1	0.15	0.00	54.82
0.95	0.01	0.18	0.00	87.02
0.95	0.001	0.18	0.00	77.35
0.95	1e-04	0.19	0.00	82.84
0.95	1e-05	0.18	0.00	82.65
0.95	1e-06	0.18	0.00	80.87
0.95	1e-07	0.17	0.00	72.61
0.99	0.1	0.06	0.00	24.13
0.99	0.01	0.08	0.00	29.55
0.99	0.001	0.08	0.00	30.96
0.99	1e-04	0.08	0.00	35.79
0.99	1e-05	0.08	0.00	27.28
0.99	1e-06	0.08	0.00	34.42
0.99	1e-07	0.08	0.00	29.16

Table 4.4: Average results for MFFFS over 100 Monte Carlo replicates, where the first two columns are control parameters of each method, $\mathcal{E}$ is the average error. CD is expressed as a rate and FD is the average number of points.

References

[1] WESTON, D. J.; HAND, D. J.; ADAMS, N. M.; WHITROW, C.; JUSZCZAK, P. Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification*, v. 2, n. 1, p. 45–62, 2008.

[2] HEARD, N. A.; WESTON, D. J.; PLATANIOTI, K.; HAND, D. J. et al. Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, v. 4, n. 2, p. 645–662, 2010.

[3] NEVES, C.; FERNANDES, C.; HOELTGEBAUM, H. Five different distributions for the Lee–Carter model of mortality forecasting: A comparison using GAS models. *Insurance: Mathematics and Economics*, v. 75, p. 48–57, 2017.

[4] COX, D. R.; GUDMUNDSSON, G.; LINDGREN, G.; BONDESSON, L.; HARSAAE, E.; LAAKE, P.; JUSELIUS, K.; LAURITZEN, S. L. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, p. 93–115, 1981.

[5] CREAL, D.; KOOPMAN, S. J.; LUCAS, A. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, v. 28, n. 5, p. 777–795, 2013.

[6] HARVEY, A. C. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Cambridge University Press, 2013.

[7] HOELTGEBAUM, H.; FERNANDES, C.; STREET, A. Generating joint scenarios for renewable generation: The case for non-Gaussian models with

time-varying parameters. *IEEE Transactions on Power Systems*, v. 33, n. 6, p. 7011–7019, Nov 2018.

[8] ROBBINS, H.; MONRO, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, v. 22, n. 3, p. 400–407, 1951.

[9] HAYKIN, S. S. *Adaptive filter theory*. Pearson, 2008.

[10] LAU, F. D.-H.; BUTLER, L. J.; ADAMS, N. M.; ELSHAFIE, M. Z.; GIROLAMI, M. A. Real-time statistical modelling of data generated from self-sensing bridges. *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, p. 1–42, 2018.

[11] VOVK, V.; GAMMERMAN, A.; SHAFER, G. *Conformal prediction*. Springer, 2005.

[12] VOVK, V. Conditional validity of inductive conformal predictors. *Machine Learning*, v. 92, n. 2-3, p. 349–376, 2013.

[13] International Energy Agency. `https://www.iea.org/newsroom/news/2013/october/wind-power-seen-generating-up-to-18-of-global-power-by-2050.html`. Accessed: 12-May-2018.

[14] MOREIRA, A.; POZO, D.; STREET, A.; SAUMA, E. Reliable renewable generation and transmission expansion planning: Co-optimizing system's resources for meeting renewable targets. *IEEE Transactions on Power Systems*, v. 32, n. 4, p. 3246–3257, Jul 2017.

[15] JABR, R. Robust transmission network expansion planning with uncertain renewable generation and loads. *IEEE Transactions on Power Systems*, v. 28, n. 4, p. 4558–4567, Nov 2013.

[16] ZHAO, C.; GUAN, Y. Data-driven stochastic unit commitment for integrating wind generation. *IEEE Transactions on Power Systems*, v. 31, n. 4, p. 2587–2596, Jul 2016.

[17] PASSOS, A. C.; STREET, A.; BARROSO, L. A. A dynamic real option-based investment model for renewable energy portfolios. *IEEE Transactions on Power Systems*, v. 32, n. 2, p. 883–895, Mar 2017.

[18] PAPAVASILIOU, A.; OREN, S. S. Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network. *Operations Research*, v. 61, n. 3, p. 578–592, 2013.

[19] MUNOZ, F. D.; WATSON, J.-P. A scalable solution framework for stochastic transmission and generation planning problems. *Computational Management Science*, v. 12, n. 4, p. 491–518, 2015.

[20] VAN DER WEIJDE, A. H.; HOBBS, B. F. The economics of planning electricity transmission to accommodate renewables: Using two-stage optimisation to evaluate flexibility and the cost of disregarding uncertainty. *Energy Economics*, v. 34, n. 6, p. 2089–2101, 2012.

[21] COBOS, N. G.; ARROYO, J. M.; STREET, A. Least-cost reserve offer deliverability in day-ahead generation scheduling under wind uncertainty and generation and network outages. *IEEE Transactions on Smart Grid*, v. 9, n. 4, p. 3430–3442, Jul 2018.

[22] ZHANG, Y.; WANG, J.; WANG, X. Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, v. 32, p. 255–270, 2014.

[23] SOUTO, M.; MOREIRA, A.; VEIGA, A.; STREET, A.; GARCIA, J. D.; EPPRECHT, C. A high-dimensional varx model to simulate monthly renewable energy supply. *In: 2014 Power Systems Computation Conference (PSCC)*, p. 1–7, 2014.

[24] LAU, A.; MCSHARRY, P. Approaches for multi-step density forecasts with application to aggregated wind power. *The Annals of Applied Statistics*, p. 1311–1341, 2010.

[25] TAYLOR, J. W.; MCSHARRY, P. E.; BUIZZA, R. Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, v. 24, n. 3, p. 775–782, Sep 2009.

[26] BESSA, R. J.; MIRANDA, V.; BOTTERUD, A.; WANG, J.; CONSTANTINESCU, M. Time adaptive conditional kernel density estimation for wind power forecasting. *IEEE Transactions on Sustainable Energy*, v. 3, n. 4, p. 660–669, Oct 2012.

[27] JEON, J.; TAYLOR, J. W. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, v. 107, n. 497, p. 66–79, 2012.

[28] TAYLOR, J. W.; JEON, J. Forecasting wind power quantiles using conditional kernel estimation. *Renewable Energy*, v. 80, p. 370–379, 2015.

[29] WAN, C.; LIN, J.; WANG, J.; SONG, Y.; DONG, Z. Y. Direct quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems*, v. 32, n. 4, p. 2767–2778, Jul 2017.

[30] GOLESTANEH, F.; PINSON, P.; GOOI, H. B. Very short-term nonparametric probabilistic forecasting of renewable energy generation; with application to solar energy. *IEEE Transactions on Power Systems*, v. 31, n. 5, p. 3850–3863, Sep 2016.

[31] STREET, A.; LIMA, D. A.; VEIGA, A.; FANZERES, B.; FREIRE, L.; AMARAL, B. Fostering Wind Power Penetration into the Brazilian Forward-Contract Market. *IEEE PES General Meeting 2011*, San Diego, California, USA, p. 1–8, Jul. 2011.

[32] SHAPIRO, A.; TEKAYA, W.; DA COSTA, J. P.; SOARES, M. P. Risk neutral and risk averse stochastic dual dynamic programming method. *European Journal of Operational Research*, v. 224, n. 2, p. 375 – 391, 2013.

[33] FANZERES, B.; STREET, A.; BARROSO, L. A. Contracting strategies for renewable generators: A hybrid stochastic and robust optimization approach. *IEEE Transactions on Power Systems*, v. 30, n. 4, p. 1825–1837, Jul 2015.

[34] KARINIOTAKIS, G. *Renewable energy forecasting: From models to applications*. Woodhead Publishing, 2017.

[35] ENGLE, R. F.; BOLLERSLEV, T. Modelling the persistence of conditional variances. *Econometric Reviews*, v. 5, n. 1, p. 1–50, 1986.

[36] ENGLE, R. F.; RUSSELL, J. R. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, p. 1127–1162, 1998.

[37] DAVIS, R. A.; DUNSMUIR, W. T.; STREETT, S. B. Observation-driven models for poisson counts. *Biometrika*, p. 777–790, 2003.

[38] PINSON, P.; GIRARD, R. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, v. 96, p. 12–20, 2012.

[39] CREAL, D.; KOOPMAN, S. J.; LUCAS, A. A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, v. 29, n. 4, p. 552–563, 2011.

[40] PATTON, A. J. Modelling asymmetric exchange rate dependence. *International Economic Review*, v. 47, n. 2, p. 527–556, 2006.

[41] PATTON, A. J. Copula–based models for financial time series. *Handbook of Financial Time Series*, p. 767–785, 2009.

[42] PATTON, A. J. Copula methods for forecasting multivariate time series. *Handbook of Economic Forecasting*, v. 2, p. 899–960, 2012.

[43] GAS on-line repository. `http://www.gasmodel.com/`. Accessed: 17-May-2018.

[44] TSENG, F.-M.; TZENG, G.-H. A fuzzy seasonal arima model for forecasting. *Fuzzy Sets and Systems*, v. 126, n. 3, p. 367–376, 2002.

[45] LEI, M.; SHIYAN, L.; CHUANWEN, J.; HONGLING, L.; YAN, Z. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, v. 13, n. 4, p. 915–920, 2009.

[46] BROYDEN, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, v. 6, n. 1, p. 76–90, Mar 1970.

[47] DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v. 5, n. 3, p. 236–244, 1996.

[48] ENGLE, R. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, v. 20, n. 3, p. 339–350, 2002.

[49] XU, J. J. *Statistical modelling and inference for multivariate and longitudinal discrete response data*. PhD Thesis. University of British Columbia, 1996.

[50] JOE, H. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[51] MURPHY, S. A.; VAN DER VAART, A. W. On profile likelihood. *Journal of the American Statistical Association*, v. 95, n. 450, p. 449–465, 2000.

[52] CHERUBINI, U.; LUCIANO, E.; VECCHIATO, W. *Copula methods in finance*. John Wiley & Sons, 2004.

[53] JARQUE, C. M.; BERA, A. K. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, v. 55, n. 2, p. 163–172, 1987.

[54] LJUNG, G. M.; BOX, G. E. On a measure of lack of fit in time series models. *Biometrika*, v. 65, n. 2, p. 297–303, 1978.

[55] BOZKURT, Ö. Ö.; BIRICIK, G.; TAYŞI, Z. C. Artificial neural network and sarima based models for power load forecasting in turkish electricity market. *PloS one*, v. 12, n. 4, p. 1–24, Apr 2017.

[56] AGGARWAL, C. C. *Data streams: models and algorithms*. Springer Science & Business Media, 2007.

[57] GAMA, J. *Knowledge discovery from data streams*. CRC Press, 2010.

[58] HÄRDLE, W. K.; WANG, W.; YU, L. Tenet: Tail-event driven network risk. *Journal of Econometrics*, v. 192, n. 2, p. 499–513, 2016.

[59] BENCZÚR, A. A.; KOCSIS, L.; PÁLOVICS, R. Online machine learning in big data streams. *arXiv preprint arXiv:1802.05872*, 2018.

[60] FAHY, C.; YANG, S.; GONGORA, M. Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams. *IEEE Transactions on Cybernetics*, v. 49, n. 6, p. 2215–2228, Jun 2019.

[61] LAW, J.; WILKINSON, D. J. Composable models for online bayesian analysis of streaming data. *Statistics and Computing*, v. 28, n. 6, p. 1119–1137, Nov 2018.

[62] BODENHAM, D. A.; ADAMS, N. M. Continuous monitoring for changepoints in data streams using adaptive estimation. *Statistics and Computing*, v. 27, n. 5, p. 1257–1270, Sep 2017.

[63] ANAGNOSTOPOULOS, C.; TASOULIS, D. K.; ADAMS, N. M.; PAVLIDIS, N. G.; HAND, D. J. Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Statistical Analysis and Data Mining*, v. 5, n. 2, p. 139–166, 2012.

[64] STEINHARDT, J.; WAGER, S.; LIANG, P. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

[65] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.

[66] HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[67] FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. Springer series in statistics New York, NY, USA, 2001. v. 1.

[68] MEINSHAUSEN, N.; YU, B. et al. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, v. 37, n. 1, p. 246–270, 2009.

[69] RASKUTTI, G.; WAINWRIGHT, M. J.; YU, B. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, v. 11, p. 2241–2259, 2010.

[70] RASKUTTI, G.; WAINWRIGHT, M. J.; YU, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, v. 57, n. 10, p. 6976–6994, Oct 2011.

[71] VAN DE GEER, S. A. et al. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, v. 36, n. 2, p. 614–645, 2008.

[72] VAN DE GEER, S. A.; BÜHLMANN, P. et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, v. 3, p. 1360–1392, 2009.

[73] ZHAO, P.; YU, B. On model selection consistency of lasso. *Journal of Machine Learning Research*, v. 7, p. 2541–2563, 2006.

[74] MESSNER, J. W.; PINSON, P. Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *International Journal of Forecasting,* in press, 2018.

[75] MONTI, R. P.; ANAGNOSTOPOULOS, C.; MONTANA, G. Adaptive regularization for lasso models in the context of nonstationary data streams. *Statistical Analysis and Data Mining*, v. 11, n. 5, p. 237–247, 2018.

[76] BODENHAM, D. A.; ADAMS, N. M. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, v. 26, n. 4, p. 917–928, Jul 2016.

[77] EFRON, B.; HASTIE, T.; JOHNSTONE, I.; TIBSHIRANI, R. Least angle regression. *The Annals of Statistics*, v. 32, n. 2, p. 407–499, 2004.

[78] ROSSET, S.; ZHU, J. Piecewise linear regularized solution paths. *The Annals of Statistics*, v. 35, n. 3, p. 1012–1030, 2007.

[79] FRIEDMAN, J.; HASTIE, T.; HÖFLING, H.; TIBSHIRANI, R. et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, v. 1, n. 2, p. 302–332, 2007.

[80] FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, v. 33, n. 1, p. 1, 2010.

[81] KIM, S.-J.; KOH, K.; BOYD, S.; GORINEVSKY, D. ℓ_1 trend filtering. *SIAM review*, v. 51, n. 2, p. 339–360, 2009.

[82] BUCKLEY, M.; EAGLESON, G. An approximation to the distribution of quadratic forms in normal random variables. *Australian & New Zealand Journal of Statistics*, v. 30, n. 1, p. 150–159, 1988.

[83] CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, v. 74, n. 368, p. 829–836, 1979.

[84] CLEVELAND, W. S.; DEVLIN, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, v. 83, n. 403, p. 596–610, 1988.

[85] KIM, A. Y.; MARZBAN, C.; PERCIVAL, D. B.; STUETZLE, W. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, v. 89, n. 12, p. 2529–2536, 2009.

[86] LEWIS, J. Economic impact of cybercrime - no slowing down. *Santa Clara: McAfee & CSI (Center for Strategic and International Studies)*, 2018.

[87] Turcotte, M. J. M.; Kent, A. D.; Hash, C. Unified Host and Network Data Set. *arXiv preprint arXiv:1708.07518*, 2017.

[88] PATCHA, A.; PARK, J.-M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, v. 51, n. 12, p. 3448–3470, 2007.

[89] ADAMS, N.; HEARD, N. *Data analysis for network cyber-security*. World Scientific Publishing Co., Inc., 2014.

[90] HEARD, N. A.; ADAMS, N. M. *Dynamic networks and cyber-security*. Imperial College Press, 2016.

[91] HEBIRI, M. Sparse conformal predictors. *Statistics and Computing*, v. 20, n. 2, p. 253–266, Apr 2010.

[92] BUTLER, L. J.; GIBBONS, N.; HE, P.; MIDDLETON, C.; ELSHAFIE, M. Z. Evaluating the early-age behaviour of full-scale prestressed concrete beams using distributed and discrete fibre optic sensors. *Construction and Building Materials*, v. 126, p. 894 – 912, 2016.

[93] BUTLER, L. J.; XU, J.; HE, P.; GIBBONS, N.; DIRAR, S.; MIDDLETON, C. R.; ELSHAFIE, M. Z. Robust fibre optic sensor arrays for monitoring early-age performance of mass-produced concrete sleepers. *Structural Health Monitoring*, v. 17, n. 3, p. 635–653, 2018.

[94] CARDOT, H.; DEGRAS, D. Online principal component analysis in high dimension: Which algorithm to choose? *International Statistical Review*, v. 86, n. 1, p. 29–50, 2018.

[95] BALZANO, L.; CHI, Y.; LU, Y. M. Streaming PCA and subspace tracking: The missing data case. *Proceedings of the IEEE*, v. 106, n. 8, p. 1293–1310, Aug 2018.

[96] FARRAR, C. R.; WORDEN, K. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 365, n. 1851, p. 303–315, 2006.

[97] DAS, S.; SAHA, P.; PATRO, S. Vibration-based damage detection techniques used for health monitoring of structures: a review. *Journal of Civil Structural Health Monitoring*, v. 6, n. 3, p. 477–507, 2016.

[98] TODD, M. D.; NICHOLS, J. M.; TRICKEY, S. T.; SEAVER, M.; NICHOLS, C. J.; VIRGIN, L. N. Bragg grating-based fibre optic sensors in structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 365, n. 1851, p. 317–343, 2006.

[99] BOWERS, K.; BUSCHER, V.; DENTTEN, R.; EDWARDS, M.; ENGLAND, J.; ENZER, M.; PARLIKAD, A. K.; SCHOOLING, J. Smart infrastructure: Getting more from strategic assets. *Centre for Smart Infrastructure and Construction*, 2016.

[100] BUTLER, L. J.; GIBBONS, N.; HE, P.; MIDDLETON, C.; ELSHAFIE, M. Z. Evaluating the early-age behaviour of full-scale prestressed concrete beams using distributed and discrete fibre optic sensors. *Construction and Building Materials*, v. 126, n. Supplement C, p. 894 – 912, 2016.

[101] GLISIC, B.; INAUDI, D.; LAU, J. M.; MOK, Y. C.; NG, C. T. Long-term monitoring of high-rise buildings using long-gauge fibre optic sensors. *7th International Conference on Multi-Purpose High-Rise Towers and Tall Buildings, Dubai, UAM, 10 - 11 December (on conference CD, paper #0416)*, 2005.

[102] LAU, F. D.-H.; BUTLER, L. J.; ADAMS, N. M.; ELSHAFIE, M. Z. E. B.; GIROLAMI, M. A. Real-time statistical modelling of data generated from self-sensing bridges. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, v. 171, n. 1, p. 3–13, 2018.

[103] MEASURES, R. M.; LEBLANC, M.; LIU, K.; FERGUSON, S.; VALIS, T.; HOGG, D.; TURNER, R.; MCEWEN, K. Fiber optic sensors for smart structures. *Optics and Lasers in Engineering*, v. 16, n. 2, p. 127–152, 1992.

[104] HERNANDEZ-GARCIA, M. R.; MASRI, S. F. Application of statistical monitoring using latent-variable techniques for detection of faults in sensor networks. *Journal of Intelligent Material Systems and Structures*, v. 25, n. 2, p. 121–136, 2014.

[105] LAU, F. D.-H.; ADAMS, N. M.; GIROLAMI, M. A.; BUTLER, L. J.; ELSHAFIE, M. Z. E. B. The role of statistics in data-centric engineering. *Statistics & Probability Letters*, v. 136, p. 58 – 62, 2018.

[106] Micron Optics. ENLIGHT User Guide. `http://www.micronoptics.com/download/enlight-user-guide-revision-1-138/#`. Accessed: 06-Apr-2019.

[107] JOLLIFFE, I. *Principal component analysis*. Springer, 2011.

[108] NOVEMBRE, J.; STEPHENS, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, v. 40, n. 5, p. 646–649, 2008.

[109] SCHOLZ, M. Analysing periodic phenomena by circular PCA. *International Conference on Bioinformatics Research and Development*, p. 38–47, 2007.

[110] SCHMITT, E.; RATO, T.; DE KETELAERE, B.; REIS, M.; HUBERT, M. Parameter selection guidelines for adaptive PCA-based control charts. *Journal of Chemometrics*, v. 30, n. 4, p. 163–176, 2016.

[111] SANGER, T. D. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, v. 2, n. 6, p. 459–473, 1989.

[112] WENG, J.; ZHANG, Y.; HWANG, W.-S. Candid covariance-free incremental principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 25, n. 8, p. 1034–1040, Aug 2003.

[113] MITLIAGKAS, I.; CARAMANIS, C.; JAIN, P. Memory limited, streaming PCA. *Advances in Neural Information Processing Systems*, p. 2886–2894, 2013.

[114] WARMUTH, M. K.; KUZMIN, D. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, v. 9, n. Oct, p. 2287–2320, 2008.

[115] BOUTSIDIS, C.; GARBER, D.; KARNIN, Z.; LIBERTY, E. Online principal components analysis. *Proceedings of the twenty-sixth annual ACM-SIAM Symposium on Discrete Algorithms*, p. 887–901, 2015.

[116] OJA, E.; KARHUNEN, J. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, v. 106, n. 1, p. 69–84, 1985.

[117] OJA, E. Principal components, minor components, and linear neural networks. *Neural Networks*, v. 5, n. 6, p. 927–935, 1992.

[118] JOHNSTONE, I. M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, v. 29, n. 2, p. 295–327, 2001.

[119] NADLER, B. On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. *Journal of Multivariate Analysis*, v. 102, n. 2, p. 363–371, 2011.

[120] VOVK, V.; NOURETDINOV, I.; GAMMERMAN, A. On-line predictive linear regression. *The Annals of Statistics*, v. 37, n. 3, p. 1566–1590, 2009.

[121] BURNAEV, E.; VOVK, V. Efficiency of conformalized ridge regression. *Conference on Learning Theory*, p. 605–622, 2014.

[122] BALASUBRAMANIAN, V.; HO, S.-S.; VOVK, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[123] LEI, J.; ROBINS, J.; WASSERMAN, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, v. 108, n. 501, p. 278–287, 2013.

[124] LEI, J.; RINALDO, A.; WASSERMAN, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, v. 74, n. 1-2, p. 29–43, 2015.

[125] LEI, J.; G'SELL, M.; RINALDO, A.; TIBSHIRANI, R. J.; WASSERMAN, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, v. 113, n. 523, p. 1094–1111, 2018.

[126] LEI, J.; WASSERMAN, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v. 76, n. 1, p. 71–96, 2014.

[127] CHERNOZHUKOV, V.; WUTHRICH, K.; ZHU, Y. Exact and robust conformal inference methods for predictive machine learning with dependent data. *arXiv preprint arXiv:1802.06300*, 2018.

[128] CHAMP, C. W.; WOODALL, W. H. Exact results for Shewhart control charts with supplementary runs rules. *Technometrics*, v. 29, n. 4, p. 393–399, 1987.

[129] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. MIT press, 2016.

[130] MEZZADRI, F. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

[131] STEWART, G. W. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, v. 17, n. 3, p. 403–409, 1980.