Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística

DFA and DCCA estimation in the presence of missing data

Alisson Silva Neimaier

Porto Alegre, Fevereiro de 2024.

CIP - Catalogação na Publicação

Silva Neimaier, Alisson
DFA and DCCA estimation in the presence of missing data / Alisson Silva Neimaier. -- 2024.
98 f.
Orientadora: Taiane Schaedler Prass.

Coorientador: Guilherme Pumi.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Programa de Pós-Graduação em Estatística, Porto Alegre, BR-RS, 2024.

1. time series. 2. missing data. 3. DFA. 4. DCCA. 5. decision trees. I. Schaedler Prass, Taiane, orient. II. Pumi, Guilherme, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Dissertação submetida por Alisson Silva Neimaier como requisito parcial para a obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade Federal do Rio Grande do Sul.

Orientador(a):

Profa. Dra. Taiane Schaedler Prass

Co-orientador(a):

Prof. Dr. Guilherme Pumi

Comissão Examinadora:

Prof. Dr. Cleiton Guollo Taufemback (PPGEst - UFRGS)

Prof. Dr. Flávio Augusto Ziegelmann (PPGEst - UFRGS)

Prof. Dr. Gilney Figueira Zebende (PPGM - UEFS)

Data de Apresentação: 22 de Fevereiro de 2024

"The sun...
I had forgotten how it feels..."
(The Fountain, ROGUE LEGACY - 2013)

Agradecimentos

Este trabalho é dedicado à

Meus pais, Elvis e Tatiane, fico feliz de ter sido o acidente mais feliz da vida de ambos. Agradeço pelo amor e confiança que tive desde que nasci e pelo apoio nas minhas decisões. Tenho orgulho de ser filho de vocês.

Meus avós, Maria e Nelson, por todo amor de vó e vô que recebi e recebo de vocês até hoje. Pelas férias intermináveis em Mariluz, pelas músicas nas festas de final de ano, pelos docinhos carinhosamente enrolados antes das festas de aniversário e por todos os atos de carinho, obrigado.

Meus tios, pela presença em todos os momentos da minha vida, me apoiando e torcendo por mim. Quando tomaram conta de mim quando meus pais estavam cansados demais, quando me ensinaram a mexer no computador, quando me acompanharam em campeonatos de futebol e até hoje quando me ouvem falar dessas baboseiras matemáticas que não fazem o menor sentido, muito obrigado.

Meus amigos, colegas, primos. Seria injusto citar nominalmente a qualquer um, pois a cada leitura lembro de outra pessoa essencial em minha vida. Seja no dia-a-dia, em noites não dormidas jogando (videogame, jogo de tabuleiro, carta, etc.) ou me acompanhando em qualquer bobagem. Sou grato a todos que riram e choraram comigo.

Maria Bethânia e Frida, por terem me aceitado e compartilhado o amor infindável que cabe nessas quatro patas, corpinho roliço e cabeça desproporcional, padrasto ama, padrasto cuida. Diego e Feijão por terem sido meus irmãos e crescido junto comigo. Bandite, por ter sido o melhor alazão que uma criança de 3 anos poderia pedir.

Bacharela em Estatística Martha, minha companheira de vida. Tua presença faz tudo ser mais fácil, esse trabalho inclusive. Fico muito feliz por encontrar alguém que compartilha das minhas idiotices e abraça o que me faz eu. Tenho muito orgulho de ti, de mim, de nós e tenho certeza que o futuro é maravilhoso. Uma vida é suficiente, mas do teu lado é completa.

Elenita, minha sogra, pelas partidas de canastra, cervejas meio geladas e por ter criado a oitava maravilha do mundo. pt saudações.

Professores do instituto de matemática e estatística, pela formação de qualidade, por terem me apresentado esse mundo incrível da estatística e pelas trocas dentro e fora de sala.

Taiane, pela disponibilidade e amizade. Tua dedicação desde os encontros como teu aluno de IC, o trabalho de conclusão e a dissertação foi impressionante e me inspira a ir mais longe.

Eu mesmo, obrigado eu.

Resumo

Técnicas tradicionais de análise de associação não se aplicam ou produzem resultados pouco confiáveis quando aplicadas a séries temporais não estacionárias. Portanto, técnicas alternativas que possam abordar efetivamente as limitações dos métodos convencionais e fornecer resultados mais precisos e robustos nesse tipo de dado são de extrema importância. Duas dessas técnicas são a Análise de Flutuação Destendenciada (DFA) e a Análise de Correlação-Cruzada Destendenciada (DCCA), que são meios indiretos de quantificar variância e correlação-cruzada em séries temporais estacionárias com tendência e são comumente empregadas para estudar propriedades de séries temporais no contexto de longa dependência. Os resultados obtidos para as funções DFA e DCCA são válidos apenas quando as séries temporais estão completas. No entanto, comumente séries temporais observadas podem conter dados faltantes. Este trabalho concentra-se no estudo do comportamento da DFA e DCCA em cenários com um volume considerável de valores ausentes, utilizando uma variedade de métodos clássicos de imputação. Contribuímos ainda com uma adaptação inovadora das Árvores de Regressão Probabilísticas para o preenchimento de séries temporais com dados faltantes. Adicionalmente, um resultado assintótico para a matriz de covariância correspondente às séries temporais preenchidas com imputação de média é derivado, e seu impacto nos valores esperados das funções DFA e DCCA é analisado empiricamente.

Abstract

Traditional association analysis techniques do not apply or yield unreliable results when applied to non-stationary time series, therefore alternative techniques that can effectively address the limitations of conventional methods and provide more accurate and robust results under non-stationarity are of utmost importance. Two widely applied techniques in this context are the Detrended Fluctuation Analysis (DFA) and Detrended Cross-Correlation Analysis (DCCA), which are indirect means to quantify variance and cross-correlation in trend-stationary time series, commonly employed in studying properties of time series in the context of long-range dependence. The results derived for the DFA and DCCA functions are only valid when the time series are complete. However, in practice, often observed time series can contain missing data. This work is focused on studying the behavior of DFA and DCCA in time series with short-range dependence with a considerable volume of missing values using a diverse array of classical imputation methods, regression trees, and a novel adaptation of the Probabilistic Regression Trees. Additionally, an asymptotic result for the covariance matrix corresponding to the time series imputed using the observed mean is derived, and its impact on the expected values of the DFA and DCCA functions is empirically analyzed.

Index

Chapter 1

Introduction

Classical methods used for association analysis are not suitable for or may generate unreliable outcomes when employed with non-stationary time series data. For instance, it is easy to show that, if $\boldsymbol{X}_t = (X_{1,t}, X_{2,t})'$, $t \geq 1$, is a bivariate random walk, then $\mathrm{Corr}(X_{1,t}, X_{2,t}) = \rho_X$, for all $t \geq 1$, but

$$\hat{\rho}_X = \frac{\sum_{t=1}^{n} X_{1,t}X_{2,t}}{\sqrt{\sum_{t=1}^{n} X_{1,t}^2}\sqrt{\sum_{t=1}^{n} X_{2,t}^2}} \xrightarrow{d} \frac{\int_0^1 M_{1,s}M_{2,s}ds}{\sqrt{\int_0^1 M_{1,s}^2 ds}\sqrt{\int_0^1 M_{2,s}^2 ds}}, \quad \text{as } n \to \infty,$$

where $\hat{\rho}_X$ is the sample cross-correlation coefficient, $\boldsymbol{M}_s = (M_{1,s}, M_{2,s})' := \Sigma^{-1/2}\boldsymbol{W}_s$, with $\boldsymbol{W}_s = (W_{1,s}, W_{2,s})'$ a bivariate Brownian motion and Σ is a covariance matrix. Hence, in this context, the sample correlation coefficient $\hat{\rho}_X$ is not a consistent estimator for ρ_X. This issue is further compounded in the presence of missing data, which is a common occurrence in real-life data. Hence, techniques capable of addressing the limitations of conventional methods in the context of non-stationary data are of utmost importance.

The Detrended Fluctuation Analysis (DFA) was initially proposed by Peng et al. (1994) as a method for identifying long-term correlations within DNA sequences. It is sometimes informally described as an indirect way to quantify variation in a trend-stationary time series. An extension of the DFA, suitable when the interest is on the joint behavior of two time series, is the Detrended Cross-Correlation Analysis (DCCA). This method was introduced by Podobnik and Stanley (2008) as a way to investigate cross-correlation between two non-stationary time series. Later on, Zebende (2011) proposed the detrended cross-correlation coefficient ρ_{DCCA} as an alternative to indirectly quantify the cross-correlation between two time series. In a certain way, ρ_{DCCA} is similar to Pearson's correlation coefficient, being a ratio between the detrended cross-covariance function and the detrended fluctuation function. The reasons why the interpretation of this coefficient differs from Pearson's will become clear in the course of this dissertation.

The DFA and DCCA have been successfully applied in many fields, such as medicine, physiology, meteorology, geophysics, economics, and physics - see Kantelhardt et al. (2001), Marinho et al. (2013) and references therein. Typically the goal of using DFA and DCCA is to identify long-range dependence in non-stationary time series, even though the theoretical properties of these methodologies in this context are generally unknown. Some theoretical results for the DFA and DCCA were derived under restrictions on the underlying process. For instance, Bardet and Kammoun (2008) presents large sample results for the DFA and DCCA in the context of fractional Gaussian noise and fractional Brownian Motion whereas Blythe (2013) and Blythe et al. (2016) derive some asymptotic results for the DCCA in the context of long-range dependent trend-stationary time series that can be decomposed

as a sum of a polynomial trend plus a fractional Gaussian noise (FGN). Prass and Pumi (2021) improves numerous findings presented in the literature, developing the asymptotic theory of the DFA and DCCA for general trend stationary processes and a collection of law of large numbers related results. It is important to note that the results derived in these works are only valid when the time series are complete.

Missing values occur when there is a lack of associated values for one or more observations of a variable, posing a common problem with substantial implications for statistical analysis. For Molenberghs et al. (2020), the presence of missing data leads to information loss and a decrease in estimation precision that is directly related to the amount of missing data and is influenced (to some extent) by the method of analysis. Missing data can also introduce bias and lead to misleading inferences about the parameters of interest. According to Pratama et al. (2016), methods for handling missing values can be categorized into three main groups: ignoring or discarding data, estimation and imputation. Nakagawa and Freckleton (2008) argues that, in the context of time series, ignoring missing values can lead to bias in parameter estimation and loss of relevant information about the dependence structure. Moreover, complete time series are necessary for calculating DFA and DCCA functions, leaving us with only estimation and imputation as options. This work will focus on the latter.

Several solutions have been suggested to handle missing data in DFA and DCCA applications. For instance, Wilson et al. (2003) considered ordinary imputation methods (mean, linear interpolation, and random) and studied their effects in parameter estimation in the context of FGN. The simulations conducted indicated that applying gap-filling techniques introduces significant deviations from the expected scaling behavior. The authors also report that, for persistent time series, interpolation methods provide a reliable estimation of long memory for scales longer than the largest likely gap. Zebende et al. (2020) analyzed behavior of the DFA and ρ_{DCCA} calculated after removing parts of simulated ARFIMA time series. Imputation methods were not considered, instead the time series pieces were merged and the analyses were carried out as usual. It was reported that for up to 50% of removed parts, compared to the original time series, there is no change in the final results for detrended auto and cross-correlation. Furthermore, Løvsletten (2017) proposed a modification of the DFA fluctuation function which can handle missing data. If there are no missing values, the proposed function coincides with the traditional DFA and, under some regularity conditions, it has the same expected values with or without gaps.

In this study, we investigate the behavior of the DFA, DCCA, and ρ_{DCCA} under the presence of missing data, in the context of short-range dependent stationary time series. Following the approach in Wilson et al. (2003), the detrended analyses are performed after reconstructing the time series. An imputation method based on probabilistic decision trees is proposed and then compared, through a Monte Carlo simulation study, to traditional imputation methods. The approach considered here draws inspiration from Neimaier and Prass (2023), where traditional decision trees were employed for imputation purposes.

Decision trees are chosen for their non-parametric and flexible nature. In the literature, various algorithms are employed to construct decision trees, differing in their methodologies and the types of variables they support. The Iterative Dichotomiser 3 (ID3) (Quinlan, 1986), for instance, was one of the pioneering decision tree algorithms. In each interaction, the algorithm chooses the covariate that provides the most information about the response variable, measuring it using entropy and information

theory concepts. It supports binary response and categorical explanatory variables. This method is prone to overfitting, however, meaning it might create a very complex tree that fits the training data almost perfectly but does not generalize well to a new dataset. The chi-squared automatic interaction detection (CHAID) was proposed by Kass (1980) as an alternative decision tree algorithm. It supports categorical response and explanatory variables. The tree is built using a chi-square test to identify statistically significant differences between categories and selecting the most significant split among the covariates. This allows CHAID to capture interactions between variables and create a tree structure that reflects these interactions.

While both ID3 and CHAID have their merits, the most widely recognized and extensively studied decision tree algorithm is the so-called Classification and Regression Trees (CART), proposed by Leo Breiman and Olshen (1984). CART is a supervised, non-parametric statistical learning method that recursively partitions the input space using a greedy algorithm, determining the best split at each step until a stopping criterion is met. As highlighted by İrsoy et al. (2012) and Linero and Yang (2018), the previously discussed decision tree algorithms share a common limitation in that their piecewise responses may not adapt well to the smoothness of the relationship between covariates and the response variable. To address this issue, some adaptations of the traditional decision trees were proposed in the literature incorporating in different forms this complexity into the analysis.

Soft Trees (İrsoy et al., 2012) modify the traditional structure of decision trees by introducing "soft" decisions at internal nodes. Instead of binary splits, where a child node is chosen or not, in soft trees, each child node is selected with a probability determined by a sigmoid function. This means that all leaf nodes in the tree contribute to the final decision with different probabilities. The main advantages of soft trees include providing a continuous response at split points, resulting in smoother predictions and smaller bias. Furthermore, the sigmoid function enables soft trees to make oblique splits (splits based on linear combinations of the covariates), in contrast to the axis-orthogonal splits made by traditional decision trees. However, one important disadvantage to consider is that this method is likely to get stuck at local minima during the optimization process. To overcome this problem İrsoy et al. (2012) initializes the method using the same splitting points as the traditional decision trees.

Smooth Transition Regression Trees (STR-Tree) proposed by Correa da Rosa et al. (2008) is a tree-based model that combines aspects of CART and Smooth transition regression (STR), to capture non-linear relationships by estimating a parametric non-linear model through a tree structure. The key concept behind the STR-Tree model is to use the structure of CART while introducing elements that allow for standard inferential methods to be used, maintaining the interpretability of the model whenever possible. In the same way that soft trees use a sigmoid function instead of binary splits, STR-Trees employs a logistic function with a slope parameter that controls the smoothness of the function to determine the probability of selecting each child node. Replacing sharp splits with smooth ones enables the application of standard inferential theory to test hypotheses regarding the location of the splits. An adaptation of the Lagrange Multiplier test presented in Luukkonen et al. (1988) is used for determining whether a node should be split or not.

Probabilistic Regression Trees (PRTrees), proposed by Alkhoury et al. (2020), represent another generalization of decision trees constructed from a probabilistic standpoint. This method proposes modifying hard splits to achieve smooth decisions and a continuous response by incorporating probability functions that associate each data point with different regions of the tree. PRTrees maintain the

interpretability of predictions and stand out as the only consistent method among the three probability-based methods mentioned. In this work, we propose a modification of the algorithm so that it can also be applied when one or more covariates are missing.

Objective

The objectives of this master's thesis are as follows. To conduct a literature review on missing data and gather existing methods for imputing missing values, evaluating their applicability in the context of stationary time series. To propose an adaptation of the Probabilistic Regression Trees algorithm capable of handling missing values, and to create an `R` package implementing it. To derive theoretical results about the autocovariance and cross-covariance matrices corresponding to the reconstructed processes when missing values are imputed with the mean. To perform an empirical analysis considering AR and MA processes to complement these findings. To conduct Monte Carlo simulations to explore the behavior of Detrended Fluctuation Analysis (DFA) and Detrended Cross-Correlation Analysis (DCCA) functions in the presence of missing data.

Novelties in this work

This study introduces a modification of the standard Probabilistic Regression Trees algorithm capable of handling missing values. The implementation is carried out in `R`, utilizing `FORTRAN` and `C` functions for computational efficiency. The original algorithm, developed by Alkhoury et al. (2020), does not inherently address missing values and is exclusively available in `Python`.

Furthermore, this work derives an asymptotic result for the covariance and cross-covariance functions corresponding to processes with missing values imputed using the mean. Such findings enable the derivation of properties regarding the expected value of DFA and DCCA in this context. To the best of our knowledge, no theoretical results for DFA and DCCA functions in the context of missing values are available in the existing literature.

Finally, the study explores the behavior of the DFA, DCCA, and ρ_{DCCA} in the context of short-range dependent processes with missing data, varying the proportion of missing data from 10% to 80%. Existing literature predominantly focuses on processes with long-range dependence and none have considered such a high proportion of missing data in their studies.

Computational Support

All simulations, analyses, and graphics in this work were generated using `R` (version 4.3.0). The `PRTree` package includes codes made in `FORTRAN` and `C` programming languages in addition to `R`, which drastically speeds up its performance compared to the same algorithm written solely on `R`. Since January 16, 2024, the package is available from CRAN (Comprehensive R Archive Network) as an `R` package, or for download at https://cran.r-project.org/package=PRTree. As of the end of the day on April 8, 2024, the package was installed 1426 times.

Outline

Chapter 2 provides a literature review on missing data and describes some imputation methods commonly used in the literature. Chapter 3 presents the main concepts related to the CART and PRTree algorithms, describes the procedure adopted in this work to fill missing values using decision trees, and introduces the `R` package developed. Chapter 4 presents the definitions of the DFA and DCCA functions, summarizes some theoretical results regarding these quantities, and describes the theoretical result derived in this work for time series imputed with the mean. Chapter 5 showcases Monte Carlo simulations considering different scenarios, imputation methods, and proportions of missing data. Finally, Chapter 6 describes the conclusions and outlines future work.

Chapter 2

Missing Data

In what follows, we delve into the mechanisms of missing data and identify which one applies to the study conducted in this work. The concepts presented in this subsection are discussed in greater detail in Molenberghs et al. (2020). Additionally, we provide a comprehensive description of classical methods employed in the processing of missing data.

2.1 Missing data mechanisms

To obtain valid statistical results from incomplete data, the nature of the missing data mechanisms must be considered. As researchers often lack control over the occurrence of missing data, its nature is not well understood. Therefore, it is necessary to formulate assumptions regarding the missing data mechanism and the validity of the analysis depends on the reasonableness of these assumptions. To further this work's discussion, a formal definition of the missing data mechanisms is provided in the sequel.

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$ denote the vector containing n values of the response variable, while X represents an $n \times p$ matrix of covariates associated with $\boldsymbol{Y}$. Additionally, let $\boldsymbol{R} = (R_1, \cdots, R_n)'$, where $R_i = 1$ if Y_i is observed and 0 otherwise, for $i \in \{1, \cdots, n\}$. Given $\boldsymbol{R}$, the vector $\boldsymbol{Y}$ can be partitioned into two subvectors: $\boldsymbol{Y}^o$ and $\boldsymbol{Y}^m$, corresponding to the observed and missing observations of $\boldsymbol{Y}$, respectively. The subvector $\boldsymbol{Y}^o$ is commonly referred to as the "observed data" while $\boldsymbol{Y}^m$ represents the "missing data". The hypothetical vector in the absence of missing data is denoted as the "complete data", denoted by $\boldsymbol{Y}$.

The missing data mechanism describes the probability that a response is observed or non-observed. More precisely, it establishes a probabilistic model governing the distribution of the response indicators $\boldsymbol{R}$ conditional on $\boldsymbol{Y}^o$, $\boldsymbol{Y}^m$, and X. As postulated by Rubin (1976), considering how the indicator $\boldsymbol{R}$ is related to the response variable $\boldsymbol{Y}$ and the covariates X, it is possible to classify missing values into three categories, namely, Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR).

2.1.1 Missing Completely at Random - MCAR

We say that the data is MCAR, when the probability of an observation being missing is independent of both the observed and missing values of interest, that is, $\boldsymbol{R}$ is independent of $\boldsymbol{Y}^o$ and $\boldsymbol{Y}^m$. As

there is no consensus in the literature about the dependence of the missing values on the covariates X, this work will adopt the same definition used on Little (1995). The term MCAR will be reserved to the case where

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, X) = P(\boldsymbol{R}),$$

and when

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, X) = P(\boldsymbol{R} \mid X),$$

the missing data mechanism will be referred to as "covariate-dependent". An example of MCAR (see Figure 2.1) would be if the researcher left the responded surveys on the table unattended, and his/her dog randomly chewed on some of them. While certainly an unfortunate incident, neither the dog, the surveys, nor any other covariate, such as the weather (sunny weather on the left-hand side, and rain on the right-hand side) or the survey's content (positive feelings are indicated by the heart symbol, while negative feelings are indicated by the thumbs-down symbol) would have any relationship with the missing observations (represented by the crossed documents).

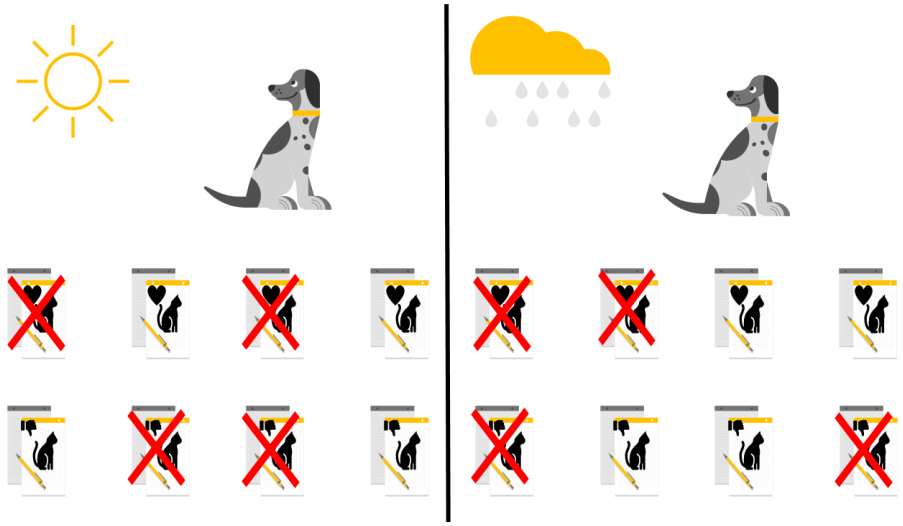

Figure 2.1: Example of Missing Completely at Random.

2.1.2 Missing at Random - MAR

We say that the missing data is MAR when the probability of an observation to be missing depends on the set of observed values and its covariates but is uncorrelated to the missing values, that is,

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, X) = P(\boldsymbol{R} \mid \boldsymbol{Y}^o, X).$$

An example of MAR (see Figure 2.2) would be if the researcher left the responded surveys on the table unattended, and his/her dog had a preference for eating surveys on rainy days to avoid boredom. In this case, the missing observations would be related to the rainy weather but not to the content of the survey itself.

Figure 2.2: Example of Missing at Random.

2.1.3 Not Missing at Random - NMAR

The stronger of the missing mechanism, we say that the data is NMAR when the probability of an observation to be missing depends on the set of observed values and its covariates and also depends on the missing responses, that is,

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, X) = P(\boldsymbol{R} \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m, X).$$

Non-random missing values are also referred as "non-ignorable" because the information about the missing values must be modeled so that the inferences about the distribution of the complete data are valid. An example of NMAR (see Figure 2.3) would be if the researcher left the responded surveys on the table unattended, and his/her dog had a preference for eating surveys with positive responses about cats. In this case, the missing observations would depend on the unobserved responses as well.

Figure 2.3: Example of Not Missing at Random.

In this work, it is assumed that the missing data is generated by the MCAR mechanism. According to Greiner et al. (1997), this is not very feasible in real-life scenarios, however, as seen in Hastie et al. (2009), most imputation methods need this assumption to be valid. With the mechanisms of missing values defined, it is possible to delve further into the consequences of missing data in statistical analysis. The next step is to address the question: how to handle missing data?

2.2 Classical methods for handling missing data

Dealing with missing data is a crucial step in statistical analyses to ensure the accuracy and validity of the results. There are various methods to handle missing data, and the choice of the approach depends on the nature of the data, the amount of missingness, and a trade-off between the simplicity of the method and its ability to introduce as little bias as possible on the data (Salgado et al., 2016). Based on Pratama et al. (2016), methods for handling missing values can be divided into three main categories, namely, ignore or discard data, estimation, and imputation.

2.2.1 Ignore or discard data

This procedure refers to the practice of simply removing observations or variables that have missing values. It can be further divided into two main methods. The first is known as complete case analysis, which involves removing any observations with missing data, and the second is case deletion, which excludes variables depending on the number of missing observations. For more information about this method, see Batista and Monard (2003). While this approach has its merits, it can introduce several problems in the context of time series analysis, for example, the characteristics of the time series such as trend, seasonality, and autocovariance function can be disrupted by the breaks in the temporal structure.

2.2.2 Estimation

Assuming that the complete data follows a particular distribution, procedures are used (usually maximum likelihood) to parametrically estimate a distribution for the complete data. Once the parameters of the distribution are estimated, they can be used to generate imputed values. According to Dempster et al. (1977), some variations of the Expectation-Maximization (EM) algorithm can handle parameter estimation with incomplete data. There are certain advantages in parametric estimation over ignoring data, but its validity heavily depends on the correct specification of the data distribution and may not adapt well to time series data with complex distributional characteristics.

2.2.3 Imputation

Imputation methods are techniques that involve the process of filling missing values based on available information in the data. By not removing the observations with missing data, like ignoring or discarding data, and not relying on assumptions about the distribution of the complete data, like the estimation methods, the imputation methods achieve results that are both more reliable and flexible than the

other two solutions. There are several methods for imputing missing values described in the literature, and many of these methods are implemented in the `R` package `imputeTS` (Moritz and Bartz-Beielstein, 2017).

Neimaier and Prass (2023) consider the use of decision trees to predict missing values. To assess the performance of the proposed algorithm, a simulation study was performed and the results obtained using decision trees were compared to the ones generated by the imputation methods available in the `imputeTS` package. The authors conclude that some of those methods produce nearly identical results. Hence, in this work, we restrict our attention to a subset of methods, which are described in the sequel.

Given a sample of a stochastic process $\{Y_t\}_{t=1}^n$, let T be the set of indexes corresponding to missing observations in the sample and let $\{\hat{Y}_t\}_{t\in T}$ be the imputed values. In the **mean** imputation method, the missing observations are replaced with the arithmetic mean of the observed data, that is,

$$\hat{Y}_t = \frac{1}{\#(T^C)} \sum_{k\in T^C} Y_k, \quad \forall\, t \in T,$$

where T^C is the set of non-missing elements and $\#(T^C)$ is the number of elements on T^C. The **moving averages with exponential weights** method is also an average-based approach that imputes the missing data using a moving average procedure with weights that diminish exponentially through the observations used to compute the moving average. More explicitly,

$$\hat{Y}_t = 2(1-2^{-h}) \sum_{k=1}^{h} 2^{-k}(Y_{l_k} + Y_{n_k}), \quad \forall\, t \in T,$$

where l_k is the index of k-th last non-missing observation and n_k is the index of k-th next non-missing observation.

The **last observation carried forward (LOCF)** method, imputes missing values with the last observed data, that is,

$$\hat{Y}_t = Y_{l_1}, \quad \forall\, t \in T,$$

where l_1 is the index of the last non-missing observation. This can be viewed as a simple interpolation method. A more general approach is the **linear interpolation** method, which fills the missing data by fitting a straight line between the two adjacent observed data, that is,

$$\hat{Y}_t = Y_{l_1} \left(\frac{Y_{n_1} - Y_{l_1}}{n_1 - l_1} \right) (t - l_1), \quad \forall\, t \in T,$$

where l_1 is the index of the last non-missing observation and n_1 is the index of the next non-missing observation. Finally, the **kalman suavization** method, uses a basic structural model and estimates the missing observation via maximum likelihood. More details about this method can be found in Grewal (2011).

In addition to these methods, the Classification and Regression Trees (CART) and Probabilistic Regression Trees (PRTrees) will also be employed for the imputation of time series. As the reconstruction of missing data with tree-based methods is a novel aspect of this work, their detailed discussion is postponed until Chapter 3.

Chapter 3

Decision Trees

A decision tree is a machine learning method that emulates the logical decision-making process of a human being and can be applied in both classification and regression problems. It generates a flowchart of questions and answers, in which the final answer represents the decision to be made. The decision tree algorithm partitions the data into several subspaces, so that the results in each final subspace are as homogeneous as possible. This chapter introduces the fundamental concepts related to decision trees theory. A brief review of the CART algorithm, implemented in the `R` package `rpart` (Therneau and Atkinson, 2019), is provided. The main results regarding the Probabilistic Regression Trees (PRTree) algorithm (Alkhoury et al., 2020) are presented alongside the adaptations proposed in this work, so that the method can be applied in the presence of missing data. Computational issues regarding the implementation of the algorithm in `R` are also discussed. However, before we delve deeper into the forest of algorithms, let us first lay down some roots by answering a fundamental question: What is a tree?

3.1 What is a tree?

A tree is a hierarchical structure consisting of internal and external nodes connected by branches. A node can be classified as a parent or child node depending on its origin: a node that is split into subnodes is called a parent node, and the subnodes are called child nodes. The internal nodes include the root node (or initial node), which receives the entire dataset, and intermediate nodes created from logical tests. On the other hand, the external nodes are the terminal nodes (or leaves) from which no further partitioning is made. Such nodes indicate the final decision (prediction) to be made when the algorithm reaches that point. At each internal node, a logical test is applied to partition the node into two or more subnodes. The branches connecting the nodes to the subnodes represent the possible decisions to be made at each test. Figure 3.1 depicts a simplified decision tree with root and leaf nodes, internal nodes, and branches.

3.2 CART

Supervised learning methods aim to predict the values of a response (or dependent) variable $Y \in \mathcal{Y}$, or a function of the response variable $g(Y)$, based on a set of explanatory (or independent) variables $\boldsymbol{X} \in \boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p := \otimes_{j=1}^{p} \mathcal{X}_j$. In the context of decision trees, the pair $(\boldsymbol{X}, Y)$ represents

a random vector with a joint distribution $\mathbb{P}$ and the observed data $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ represents a random sample from $\mathbb{P}$.

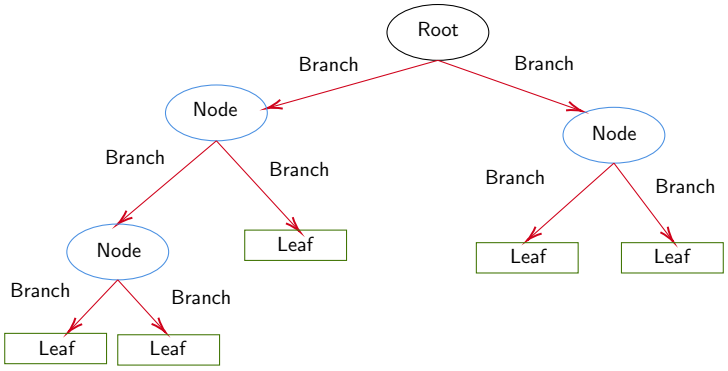

Figure 3.1: A simplified structure of a binary decision tree, i.e., a decision tree in which each node has at most two children.

3.2.1 Growing a tree

The CART algorithm recursively creates a partition of the input space $\boldsymbol{\mathcal{X}}$ and generates predictions in the output space $\mathcal{Y}$. In this work, these spaces are $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}$. A summary of the method is described in the sequel. A detailed description of this algorithm implemented in the `rpart` package can be found in Therneau and Atkinson (2019).

In an informal and resumed manner, the CART algorithm automatically determines which variables and positions will be used to create the partitions in the regions and what is the shape of the tree. Formally, for each existing node $A := \otimes_{j=1}^{p}[\ell_j, r_j] \subset \mathbb{R}^p$, CART determines the best partition (j^*, z^*) from the set of all possible partitions $S = \{(j, z) : \ j \in [1, p] \cap \mathbb{N}, \ z \in [\ell_j, r_j]\}$, where j is the index of the variable in which the partition is made, and z is the position at which the partition occurs. More specifically, given a node A, (j^*, z^*) is one of the possible solutions to the following optimization problem (see Josse et al., 2019, for more details)

$$\begin{aligned}(j^*, z^*) = \underset{(j,z)\in S}{\operatorname{argmin}}\Bigg\{\mathbb{E}\Bigg[&\Big(Y - \mathbb{E}\big[Y|X_j \leq z, \boldsymbol{X} \in A\big]\Big)^2 I(X_j \leq z, \boldsymbol{X} \in A) \\ &+ \Big(Y - \mathbb{E}\big[Y|X_j > z, \boldsymbol{X} \in A\big]\Big)^2 I(X_j > z, \boldsymbol{X} \in A)\Bigg]\Bigg\}.\end{aligned} \tag{3.1}$$

For any node A, the optimization in (3.1) is equivalent to solving the following problem

$$f^* = \underset{f\in\mathcal{P}_c}{\operatorname{argmin}}\Big\{\mathbb{E}\Big[\big(Y - f(\boldsymbol{X})\big)^2 I(\boldsymbol{X} \in A)\Big]\Big\}, \tag{3.2}$$

in which $\mathcal{P}_c$ is the set of piecewise-constante functions on $A \cap [X_j \leq z]$ and $A \cap [X_j > z]$, for $(j, z) \in S$.

The splitting process just described is repeated until a stopping criterion is met, yielding a decision tree $\mathcal{T}$, which consists of M terminal nodes and a partition $R_1, \cdots, R_M$ of $\mathbb{R}^p$. Figure 3.2 illustrates an example of a decision tree with 5 terminal nodes which originate the regions $\{R_1, \cdots, R_5\}$. In this figure, for the root node, the pair that minimizes (3.1) is $(1, t_1)$, i.e., the point t_1 from the variable

X_1. Solving the optimization problem in (3.2) involves addressing a least squares problem within the subset of functions $\mathcal{P}_c$. Thus, by minimizing the mean squared error, the CART procedure targets the quantity $E[Y|\boldsymbol{X}]$. The `rpart` package provides built-in mechanisms to handle missing data during the tree-building process. By default, the function `rpart` removes only those rows in the data set for which either the response or all of the predictors are missing and applies surrogate splits to handle missing values by using the information from other correlated variables to make reasonable decisions. Alternatively, one can consider preprocessing the data to address missing values before fitting the tree. This can involve imputing missing values or removing observations with missing data, depending on the nature of the problem and the amount of missingness.

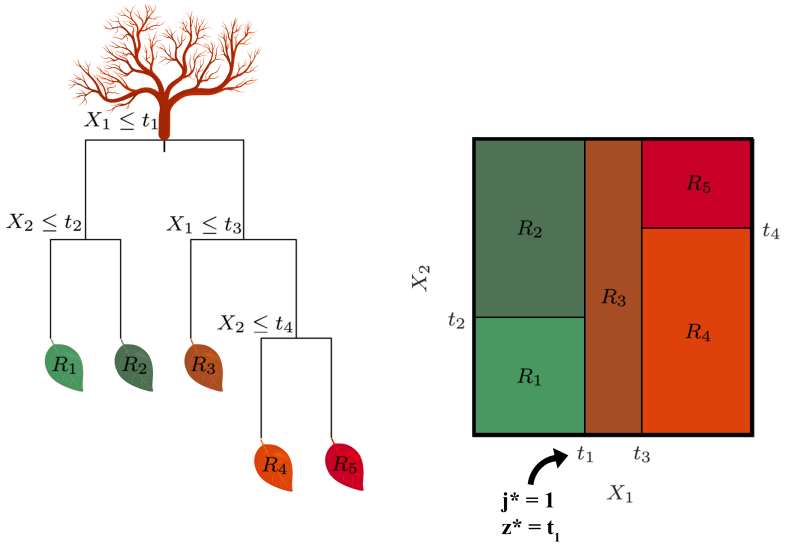

Figure 3.2: Result of a recursive binary partition in a bivariate example (right) and the decision tree that is equivalent to this partition (left).

The prediction function associated with the tree $\mathcal{T}$ is given by

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} c_m I(\boldsymbol{X} \in R_m), \quad c_m = \mathbb{E}[Y|R_m$$

where $I(\cdot)$ denotes the indicator function. The optimal decision tree is obtained by minimizing the expected prediction error of $f(\boldsymbol{X})$. However, an excessively complex tree may lead to overfitting, while a tree with insufficient nodes may fail to capture important information. Therefore, determining the appropriate size of the tree is crucial for achieving good predictive performance. So, how big should the final tree be?

3.2.2 Pruning a tree

A common approach to determining the size of a decision tree involves first growing an overfitted tree $\mathcal{T}_0$ and then pruning it to obtain the optimal tree size based on its performance on a validation set. Pruning is removing some of the branches from the tree and replacing them with leaf nodes, reducing

the model's complexity. Typically, the optimal tree size is chosen using a criterion such as minimum cross-validated error. Since computing cross-validation for every possible sub-tree is computationally intensive, a method called cost complexity pruning (or weakest link pruning) is used to select a subset of sub-trees. The algorithm's general idea is described in the sequel.

Consider a sequence of trees indexed by a non-negative adjust parameter α. For each α, there exists a sub-tree $\mathcal{T}_\alpha \subset \mathcal{T}_0$ that minimizes

$$C_\alpha(\mathcal{T}) = \sum_{m=1}^{\#(\mathcal{T})} \sum_{i:\, \boldsymbol{X}_i \in \mathscr{X}_m} (Y_i - \hat{Y}_m)^2 + \alpha \#(\mathcal{T}), \quad \hat{Y}_m = \frac{1}{\#(\mathscr{X}_m)} \sum_{i:\, \boldsymbol{X}_i \in \mathscr{X}_m} Y_i,$$

in which $\#(\mathcal{T})$ is the number of terminal nodes $\mathcal{T}$, $\mathscr{X}_m = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\} \cap R_m$, R_m is the subset of the input space correspondent of the m-th leaf and $\#(\mathscr{X}_m)$ is the number of observations in the m-th leaf. The adjust parameter α controls the trade-off between the complexity of the sub-tree $\mathcal{T}$ and the quality of fit to the training set. When $\alpha = 0$, the sub-tree $\mathcal{T}_\alpha$ will be $\mathcal{T}_0$. The higher the value of α, the higher the price to pay for having a tree with many terminal nodes, therefore, $C_\alpha(\mathcal{T})$ will be minimized by smaller trees.

As α increases, the tree branches are pruned in a nested and predictable manner (James et al., 2013; Hastie et al., 2009): internal nodes are aggregated pairwise until a single node remains. This produces a sequence of sub-trees indexed by α that includes $\mathcal{T}_\alpha$. The choice of α is then made using a validation set or cross-validation. Once α is determined, the complete dataset is used to obtain the corresponding subtree for α. The process for constructing and pruning trees can be summarized by the following algorithm (James et al., 2013).

Algorithm 1: Constructing and pruning a tree

1. Construct a large tree $\mathcal{T}_0$ using the training set and recursive binary splitting method, stopping only when each terminal node has a certain amount of observations less than or equal to a predetermined minimum.
2. Apply cost complexity pruning to obtain the sequence of best sub-trees as a function of α.
3. Use K-fold cross-validation to select α. That is, divide the observations in the testing set into K subsets. For each $k = 1, \cdots, K$:
   - Repeat steps 1 and 2 on all but the k-th fold.
   - Evaluate the mean squared prediction error using the k-th fold as a function of α.

   For each α, compute the average results and select the α that minimizes the mean error.
4. Return the subtree from step 2 that corresponds to the chosen value of α.

The CART algorithm is known for its simplicity, interpretability, and versatility, making it a widely used tool in statistical modeling. However, due to its piecewise response, it may not adapt well to the smoothness of the relation between the covariates and the response variable. In the next section, we will explore a method that addresses these limitations.

3.3 Probabilistic Regression Trees

Probabilistic Regression Trees are a generalization of decision trees proposed by Alkhoury et al. (2020). Similarly to the Soft trees and STR-Trees methods, PRTrees suggests altering the tree splits to yield smooth decisions and a continuous response based on probability functions that relate each data point to each region of the tree. This method maintains the interpretability of predictions and can be shown to be consistent.

3.3.1 The model

Let $\boldsymbol{X} = (X_1, \cdots, X_p)$ be a p-dimensional random vector in a subspace $\mathcal{X}$ of $\mathbb{R}^p$, and let

$$Y := f(\boldsymbol{X}; \Theta) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where Θ are the parameters associated with f. While Soft trees and STR-Trees replace hard splits with different sigmoid functions, PRTrees replace the indicator function with a smooth function Ψ, leading to the prediction function given by

$$f_{PR}(\boldsymbol{X}; \Theta) = \sum_{m=1}^{M} \gamma_m \Psi(\boldsymbol{X}; R_m, \boldsymbol{\sigma}), \tag{3.3}$$

where the set of parameters $\Theta = (\{R_m\}_{m=1}^M, \boldsymbol{\gamma}, \boldsymbol{\sigma})$ correspond, respectively, to the set of regions, the weights $\boldsymbol{\gamma} \in \mathbb{R}^p$ associated with these regions, and a vector of parameters $\boldsymbol{\sigma} \in \mathbb{R}_+^p$ which captures the potential noise in the input variables. For example when $\boldsymbol{X}$ are measurements done while calibrating machines, the noise corresponds to the measurement errors.

Alkhoury et al. (2020) propose to define the function Ψ in (3.3) through the relation

$$\Psi(\boldsymbol{X}; R_m, \boldsymbol{\sigma}) := \left[\prod_{i=1}^{p} \sigma_i\right]^{-1} \int_{R_m} \phi\left(\frac{u_1 - X_1}{\sigma_1}, \cdots, \frac{u_p - X_p}{\sigma_p}\right) d\boldsymbol{u}, \quad \boldsymbol{X} \in \mathcal{X}, \tag{3.4}$$

where $\phi : \mathbb{R}^p \to \mathbb{R}$ is a sufficiently regular probability density function, that is, ϕ is in L^2, its first derivative is continuous and the support of its Fourier transform is $\mathbb{R}^p$. In practice, choosing which ϕ to use is a problem itself. A priori knowledge about the distribution of the errors can be useful for delimiting potential candidate probability densities to be used in (3.4). The optimal choice for ϕ can then be determined through cross-validation.

Assumption (3.4) implies that Ψ relates the data points to different regions of the decision trees through a probability density function and smooths the predictions made. Note that when

$$\Psi(\boldsymbol{X}; R_m, \boldsymbol{\sigma}) = I(\boldsymbol{X} \in R_m), \quad \forall\, m \in \{1, \cdots, M\},$$

the standard regression tree model is obtained. Figure 3.3 presents a comparison between the output of the standard regression tree and a PRTree, black dots corresponding to a sample $\{(X_i, Y_i)\}_{i=1}^n$, of size $n = 200$, from

$$Y = \cos(X) + \varepsilon, \quad \text{where} \quad X \sim U(0, 10), \quad \varepsilon \sim \mathcal{N}(0, 0.05^2),$$

with X and ε independent from each other. The red and blue lines represent the predicted curves generated by the CART and PRTree algorithms, respectively. For the PRTree method ϕ was taken to be the standard Gaussian probability density function and σ was set to be the sample standard deviation of $\{X_i\}_{i=1}^n$. For the sake of comparison, in both cases, exactly 9 regions were constructed. In this example, it is evident that the curve estimated by PRTree (MSE $= 0.0024$) captures the smooth behavior of the cosine function more effectively than CART (MSE $= 0.027$). Considering a grid of 200 equally spaced values $X \in \{i/20\}_{i=1}^{200}$ and $f(X) = \cos(X) = E[Y|X]$, it is possible to assess how close the curves estimated by these two methods are to the cosine function (the conditional mean of the underlying process). In this case, the PRTree demonstrated even better performance (MSE $=$ 0.00016), while the CART exhibited a similar MSE to that of the training set (MSE $= 0.028$).

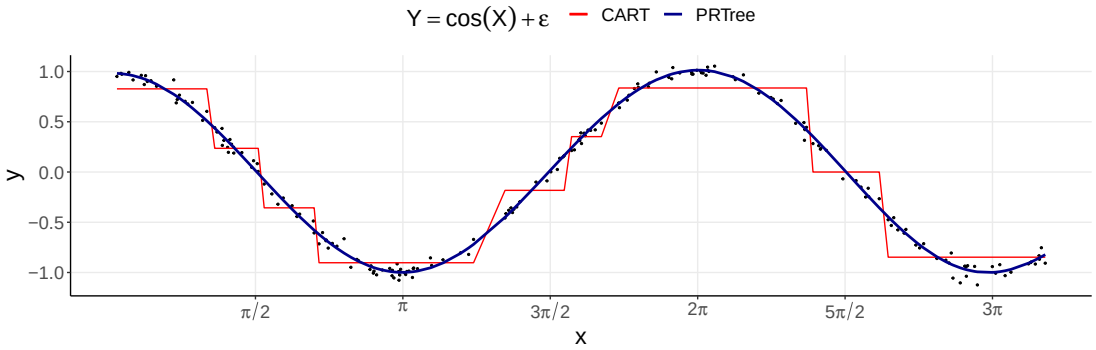

Figure 3.3: Comparison between the estimations of the cosine function using a standard regression tree and a PRTree, both with 9 regions.

3.3.2 Parameter estimation

Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, where $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, be a training data set. Upon considering the quadratic loss function, the algorithm for probabilistic regression tree estimation aims to find the parameters Θ which minimizes

$$L(\Theta) := \sum_{i=1}^{n} \left(Y_i - \sum_{m=1}^{M} \gamma_k P_{im} \right)^2, \quad P_{im} := \Psi(\boldsymbol{X}_i; R_m, \boldsymbol{\sigma}). \tag{3.5}$$

The entries P_{im} of the matrix $P_{n\times M}$ establish the relationship between each observation $\boldsymbol{X}_i$ and each region R_m, adhering to the conditions

$$0 \leq P_{im} \leq 1, \quad i \in \{1, \cdots, n\}, \quad m \in \{1, \cdots, M\},$$

and also satisfying

$$\sum_{m=1}^{M} P_{im} = 1, \quad \text{for all} \quad i \in \{1, \cdots, n\}.$$

Due to their nature, throughout the study the values of P_{im} will be referred to as the probability of observation $\boldsymbol{X}_i$ belonging to region R_m.

The estimation of Θ alternates between region and weight estimates, as in standard regression trees, till a stopping criterion is met. The most commonly used stopping criteria are the number

of observations in each region, the maximum number of regions, a minimum decrease in the mean squared error of estimates, and so forth. Firstly, the optimal splitting point is determined within the existing regions, the region is divided, the matrix P is recalculated, and the weights γ are updated.

3.3.2.1 *Estimating the weights γ*

Given the regions $\{R_m\}_{m=1}^{M_1}$ and $\boldsymbol{\sigma}$, if $P'P$ is nonsingular, minimizing equation (3.5) with respect to γ consists in finding the least squares estimator

$$\hat{\gamma} = (P'P)^{-1}P'\boldsymbol{Y}, \quad \boldsymbol{Y} = (Y_1, \cdots, Y_n)'. \tag{3.6}$$

However, numerical instability may arise when the dimensionality of P is high or when the entries in a column of matrix P have low values.

Assume that $M_1 - 1$ regions have been identified, i.e., the tree already has $M_1 - 1$ leaves. Hence, for each $m \in \{1, \cdots, M_1 - 1\}$, the region R_m can be partitioned into two subregions upon considering the jth variable, for any $j \in \{1, \cdots, p\}$, and a split point s_m^j. Any tested split point results in M_1 new regions and updates P and γ which became, respectively, a $n \times M_1$ matrix and a vector of size M_1. Upon replacing γ in (3.5) by the updated value, the best split for the current region R_m is given by

$$\underset{1 \le j \le p,\, s \in S_m^j}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{M_1} \hat{\gamma}_l P_{il} \right)^2 \right\}, \quad \hat{\gamma} = (P'P)^{-1}P'\boldsymbol{Y},$$

where S_m^j denotes the set of split points for region R_m in covariate $\boldsymbol{X}_j$.

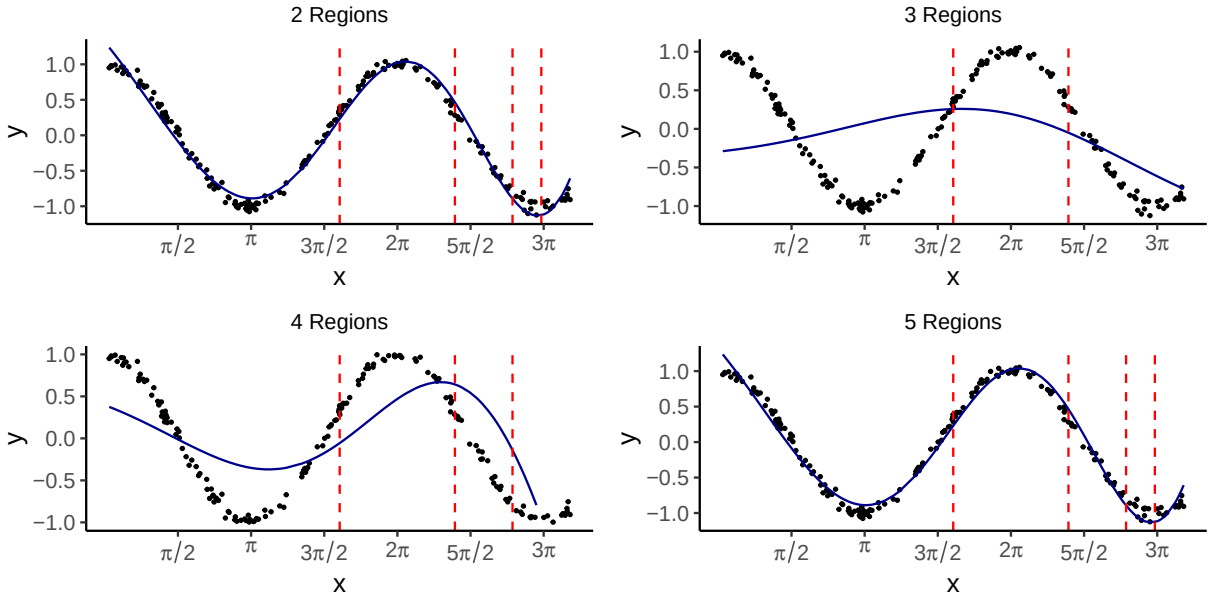

Figure 3.4: Estimates of the cosine function using the PRTree method with increasing number of regions $(2 - 5)$.

The same PRTree configurations as in Figure 3.3 are presented in Figure 3.4, with the only difference being the number of regions, now increasing from 2 to 5. The red dashed lines represent

the split points chosen by the method and the blue lines represent the predicted curves generated by the CART algorithm. In this figure, it becomes evident that the behavior of the cosine function is well reconstructed through the division into 5 regions, although the fit may not be as precise as that in Figure 3.3, it is more parsimonious. It is worth noting that the estimated curve's behavior for m regions resembles a polynomial of degree $m-1$.

3.3.2.2 Estimating the deviation parameter $\boldsymbol{\sigma}$

The best way to determine the vector $\boldsymbol{\sigma}$ is through a priori knowledge. However, as it is not always possible to have prior knowledge of a subject, another viable solution is to apply a grid search.

3.4 PRTree R package

Alkhoury et al. (2020) implemented the PRTrees algorithm, building upon the `Scikit-Learn` implementation of standard regression trees (Varoquaux et al., 2015). The `python` code for this implementation is available at `https://gitlab.com/sami.courie/pr-tree`. An adaptation of the existing code was necessary to incorporate the changes proposed in this work for handling missing data. Given that `R` is an open-source language widely employed by statisticians, the code was written using this programming language.

Initial tests revealed that implementing the PRTree algorithm solely in `R` would result in a relatively high computational cost, potentially limiting its practical use. Motivated by the need for faster processing speed and considering that a significant portion of the algorithm relies on basic operations and loops, an `R` package was developed, incorporating `FORTRAN` and `C` functions. Despite encountering challenges during implementation, such as coordinating communication between multiple programming languages and the author's limited experience with extensive `FORTRAN` coding, the results are promising: the current version of the code runs approximately 30 times faster than the initial `R` version. The code was transformed in the `R` package PRTree (Neimaier and Prass, 2024), which was made available on `CRAN`. This effort provides an important venue for the dissemination of the proposed methodology to the statistical community.

3.4.1 An adaptation to handle missing data

Originally the PRTree algorithm is not able to handle missing data. In what follows we propose an adaptation to the method that can be applied if one or more covariate values are missing. Given M regions $\{R_m\}_{m=1}^{M}$, define Ψ^* through the relation

$$\Psi^*(\boldsymbol{X}; R_m, \boldsymbol{\sigma}) = \begin{cases} \Psi(\boldsymbol{X}; R_m, \boldsymbol{\sigma}), & \text{if there are no missing values in } \boldsymbol{X}, \\ M^{-1}, & \text{otherwise.} \end{cases} \tag{3.7}$$

where Ψ is defined in (3.4). This definition implies that, if for some $1 \leq i \leq n$, one or more coordinates of $\boldsymbol{X}_i$ are missing, (3.7) attributes a uniform probability to all regions so that the ith row of P is constant and equal to M^{-1}. If there are no missing values in the covariates $\{\boldsymbol{X}_i\}_{i=1}^{n}$ the standard PRTree model is obtained.

3.4.2 Computational issues

Testing all possible split points for all variables in all regions is computationally infeasible. Therefore, it is essential to cleverly limit the set of split points to be tested in each iteration of the algorithm. Following the suggestion of Alkhoury et al. (2020), the `PRTree` package considers only the midpoint of the intervals for each variable in each region as potential split points. In cases where the lower limit is $-\infty$ or the upper limit is ∞, the respective observed minimum and maximum values for that variable in that region are considered. Using this criterion, the minimization problem for each region can be formulated as

$$\underset{1\leq j\leq p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{M_1} \hat{\gamma}_l P_{il} \right)^2 \right\}, \quad \hat{\boldsymbol{\gamma}} = (P'P)^{-1} P' \boldsymbol{Y},$$

given that the split points are deterministic.

Another challenge is the potential singularity of the matrix $P'P$. In the PRTree package, the inverse matrix in (3.6) is replaced with the generalized inverse, obtained through the LU factorization of the matrix $P'P$. The computation of the generalized inverse involves the utilization of the `LAPACK` subroutines `DGETRF` and `DGETRI`, which respectively perform LU factorization and compute the inverse matrix based on the results of the factorization.

In principle, a wide range of distribution functions can be used to define Ψ. However, the only one currently implemented in the algorithm is the Gaussian distribution, utilizing the `C` routines pulled from `R`. Moreover, it is assumed that $\sigma_j = \sigma$ for all $1 \leq j \leq p$. In other words, the implemented function in the algorithm is

$$\Psi(\boldsymbol{X}; R_m; \boldsymbol{\sigma}) = \frac{1}{(2\pi\sigma^2)^{p/2}} \int_{R_m} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{p} (u_j - X_j)^2 \right\} d\boldsymbol{u}.$$

It is worth noting that despite this implementation, the assumption of independence among covariates is not valid when the covariates are assumed to be lagged versions of the original time series.

3.4.3 Stopping criteria

Stopping criteria are essential to determine when the tree-building process should stop. Without them, the algorithm could in principle run indefinitely, causing excessive computational cost and potential overfitting. The proposed algorithm employs several usual stopping criteria for decision trees and some novel ones, proposed in this work (`perc_x` and `p_min`), for the context of PRTrees that utilize properties specific to the probability matrix P, as defined in (3.5). These criteria are

- `cp`: the complexity parameter. Any split that does not decrease the mean squared error (MSE) by a factor of `cp` will not be attempted. This prevents overfitting in the case of a simple problem where the MSE gain in an iteration is not worth the complexity added to the model.
- `max_depth`: the maximum depth of the decision tree. The depth is defined as the length of the longest path from the root to a leaf. Controls the number of splits made in a specific region, preventing overfitting in a particular region.

- max_terminal_nodes: the maximum number of regions in the output tree. Controls the number of columns in the matrix P, preventing problems with overfitting and non-invertibility.
- n_min: the minimum number of observations in a final node. Prevents the algorithm from splitting a region with too few observations;
- perc_x and p_min: given a column of P, the value perc_x is the percentage of rows in this column that must have a probability higher than the threshold p_min for a splitting attempt to be made in the corresponding region. This ensures that fewer cuts are tested in regions with lower weight in the PRTree algorithm prediction.

3.4.4 PRTree functions

The functions implemented in the PRTree package are pr_tree and pr_tree.predict. The function pr_tree function is responsible for processing the input data and the output from FORTRAN. It has the following arguments

- y - a numeric vector corresponding to the dependent variable;
- X - a numeric vector, matrix, or dataframe corresponding to the independent variables, with the same number of observations as y;
- sigma_grid - (optional) a numeric vector with candidate values for the parameter sigma, to be used in the grid search algorithm.

Additionally, the stopping criteria described previously can be chosen by the user, otherwise, the default values (max_terminal_nodes $= 15$, cp $= 0.01$, max_depth $= 5$, n_min $= 5$, perc_x $= 0.10$, p_min $= 0.05$) are used.

This function is also responsible for the initial processing of inputs necessary to the call of the FORTRAN subroutine. Tasks include creating variables used by the subroutine, ensuring all variables are in the correct format for interpretation by FORTRAN, and, if sigma_grid is not provided by the user, assigning to this variable a vector with the standard deviations of the columns of X. The function also processes the output of the FORTRAN subroutine to make it more user-friendly in R. This involves removing entries in vectors not used by FORTRAN and organizing the R output into matrices with explanatory column names, returning to the user a list with the following arguments:

- yhat - the estimated values for y;
- P - the matrix of probabilities calculated for the returned tree;
- gamma - the values of the gamma weights for the returned tree;
- MSE - the mean squared error calculated for the returned tree;
- sigma - the parameter for the standard deviation for the returned tree;
- nodes_matrix_info - information related to each node of the returned tree
  - node: node identification number;
  - isTerminal: boolean if the node is a leaf;

  - fatherNode: identification number of the father node of this node;
  - varCut: if this node is not terminal, in which covariate it was split;
  - cutpoints: if this node is not terminal, where in the varCut it was split.

- `region` - information related to each region of the returned tree
  - node: node identification number;
  - var: covariates identification number;
  - inf: lower limit of the interval for each covariate in each node;
  - sup upper limit of the interval for each covariate in each node;
  - isTerminal boolean if the node is a leaf.

Due to its complexity and significance, the main subroutine implemented in `FORTRAN` is presented in pseudocode format (see Algorithm 2). We shall not discuss the other subroutines implemented in `FORTRAN` and available in the package since they are auxiliary functions used to process the input (output) from (to) `R`.

Algorithm 2: PRTree main_calc

do while (*number of terminal nodes* $<$ `max_terminal_node`)

if (*number of nodes to split* $= 0$) go to 999

do i $= 1$, n
  do j $= 1$, *number of terminal nodes*
    P[i,j] $= \Psi^*(X_i, R_j, \sigma)$
  end do
end do

if ((*depths of all candidate nodes* $=$ `max_depth`) or
    (`perc_max` $<$ `p_min` *for all columns*)) go to 999

do i $= 1$, *number of candidate nodes*
  do j $= 1$, p
    s $=$ *midpoint of variable j in node i*
    if (*number of points in any splited region* $<$ `min_split`) go to 999
    if (*new MSE* $<$ *old MSE*) *update the auxiliary tree*
  end do
end do

if ((*no split was made*) or (*reduction in MSE* $<$ cp)) go to 999

end do

999 if (*any split was made*) *update the main tree*

return

The function `predict.pr_tree` works as a standard `prediction function in R` taking the inputs `object` (object of class inheriting from `"prtree"`) and `newdata` (a matrix with new values for the covariates). It uses the estimated values for γ and σ to compute the probability matrix `P` for the observations considering `X` $=$ `newdata` and returns the predicted values `yhat` for this new dataset.

3.5 How to fill missing data with tree-based methods?

The first challenge in handling missing values with tree-based methods is to determine which covariates to utilize when imputing time series, as it is essential to consider the characteristics of the underlying process. Hence, in this work, we aim to present a reasonable approach that takes into account the dependence structure of the time series. Time series exhibit correlation between different time points, with these dependencies typically weakening as the distance between two observations increases. For example, AR(p) and MA(q) processes, which are particular cases of ARMA(p, q) processes (respectively when $q = 0$ and $p = 0$), have a very interesting characteristic regarding the behavior of their autocorrelation (ACF) and partial autocorrelation (PACF) functions. As shown in Table 3.1, the ACF on an AR model behaves as the PACF of an MA model with the same order and vice-versa. In practice, when the time series is complete or enough data is available, this characteristic helps in the model identification step. Taking this into account, in the context of missing data, observations surrounding the point of interest are used as covariates to fill in the missing values.

Table 3.1: Behavior of the ACF and PACF functions for ARMA(p, q) processes, for $p \geq 0$ and $q \geq 0$.

Process	ACF	PACF
AR(p)	Not null for all h. Decays rapidly towards 0	Not null only for $\|h\| \leq p$
MA(q)	Not null only for $\|h\| \leq q$	Not null for all h. Decays rapidly towards 0
ARMA(p, q)	Not null for all h. Decays rapidly towards 0	Not null for all h. Decays rapidly towards 0

The procedure to fill in missing values in a time series using tree-based methods can be described as follows. Let $\{X_t\}_{t=1}^n$ be a time series, and $T \subset \{1, \cdots, n\}$ denote the set of indexes corresponding to missing values. Define $X_t^{\text{miss}} = \texttt{NA} \times I(t \in T) + X_t \times I(t \in T^C)$. The reconstruction procedure for $\{X_t^{\text{miss}}\}_{t=1}^n$ is described in Algorithm 3.

Algorithm 3: Missing data imputation using decision trees

1. Set h_1 and h_2 and define X_t^{miss} as the response variable, for $h_1 < t \leq n - h_2$.
2. Construct the matrix containing the covariates, using the previous h_1 observations and the subsequent h_2 observations associated to X_t^{miss}.
3. Train the decision tree model using only the responses and corresponding covariates for which the indexes satisfy $t \notin T$.
4. For the CART algorithm, prune the tree using the `prune` function, with the `cp` parameter set to the smallest error value calculated from cross-validation via `rpart`. PRTree algorithm skips this step.
5. Utilizing the decision tree obtained in the previous step, predict the values of X_t^{miss} for $t \in T$, resulting in a sequence of predicted values $\{\hat{X}_t\}_{t \in T}$.

Chapter 4

Detrended Variance and Covariance Analysis

In the existing literature, the definitions of DFA and DCCA are commonly presented in a heuristic manner. This section closely follows the definitions and notations outlined in Prass and Pumi (2021), which provides a more rigorous framework. Additionally, the paper presents unique results that are crucial for evaluating the simulation results in the subsequent chapter.

In the sequel, given any sequence $\{Y_t\}_{t=1}^n$, let $\boldsymbol{Y}_j^{\langle i\rangle}$ be defined by

$$\boldsymbol{Y}_j^{\langle i\rangle} = (Y_i, \cdots, Y_j)' \qquad i, j \in \{1, \cdots, n\},\ i \leq j.$$

For any $\ell \times \ell$ matrix A_ℓ, let $A_\ell^{\langle i\rangle}$ be the matrix containing the elements of A_ℓ from row i up to row l. Given a block matrix A, let $[A]^{p,q}$ denote its (p,q)-th block. Also, let 0_n and 1_n denote vectors of zeros and ones in $\mathbb{R}^n$, respectively. Similarly, let $0_{m,n}$ and $1_{m,n}$ denote the $m \times n$ matrices of zeros and ones, respectively, and I_n denote the $n \times n$ identity matrix.

4.1 Integrated process and detrended residual

Let $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ be two stochastic processes and let $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ be two samples of size n obtained from these processes, respectively. Define the integrated signals $\{R_{k,t}\}_{t=1}^n$ by

$$R_{k,t} := \sum_{i=1}^{t} X_{k,j}, \quad k \in \{1,2\}, \quad t \in \{1, \ldots, n\}.$$

Let J_ℓ be the $\ell \times \ell$ matrix whose (r,s)-th element is given by $[J_\ell]_{r,s} = I(1 \leq r \leq s \leq \ell)$, that is, J_ℓ is a lower triangular matrix with all entries equal to 1. It follows that, for $0 < m < n$,

$$\boldsymbol{R}_{k,n}^{\langle 1\rangle} = J_n \boldsymbol{X}_{k,n}^{\langle 1\rangle}, \qquad \boldsymbol{R}_{k,m+i}^{\langle i\rangle} = J_{m+i}^{\langle i\rangle} \boldsymbol{X}_{k,m+i}^{\langle 1\rangle}, \quad i \in \{1, \ldots, n-m\}. \tag{4.1}$$

The set $\{R_{k,m+i}^{\langle i\rangle}\}_{i=1}^{n-m}$, defined through (4.1), is a sequence of $n-m$ overlapping boxes containing $m+1$ values from the integrated signals, starting at i and ending at $m+i$. Upon considering non-overlapping boxes, all definitions, theorems, corollaries, and lemmas that follow can be stated analogously with slight modifications.

Now, for each $k \in \{1,2\}$, and $i \in 1, \cdots, n-m$, let $\tilde{\boldsymbol{R}}_{k,i}$ be the vector with the ordinates $\tilde{R}_{k,t}(i)$, $i \leq t \leq m+i$ of a polynomial least-squares fit associated with the i-th box $\boldsymbol{R}_{k,m+1}^{\langle i\rangle}$, and $\boldsymbol{\mathcal{E}}_{k,i}$ be the

vector with the residual terms $\mathcal{E}_{k,t}(i)$, $i \leq t \leq m+i$. That is,

$$\begin{aligned} \tilde{\boldsymbol{R}}_{k,i} &= P_{m+1}\boldsymbol{R}_{k,m+i}^{\langle i \rangle} = \left(\tilde{\boldsymbol{R}}_{k,i}(i), \ldots, \tilde{\boldsymbol{R}}_{k,m+i}(i)\right)', \\ \boldsymbol{\mathcal{E}}_{k,i} &= \boldsymbol{R}_{k,m+i}^{\langle i \rangle} - \tilde{\boldsymbol{R}}_{k,i} = Q_{m+1}\boldsymbol{R}_{k,m+i}^{\langle i \rangle} = (\boldsymbol{\mathcal{E}}_{k,i},(i), \ldots, \boldsymbol{\mathcal{E}}_{k,m+i}(i))', \end{aligned} \tag{4.2}$$

with

$$D'_{m+1} := \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & m+1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2^{\nu+1} & \cdots & (m+1)^{\nu+1} \end{pmatrix},$$

$$P_{m+1} := D_{m+1}(D'_{m+1}D_{m+1})^{-1}D'_{m+1} \quad \text{and} \quad Q_{m+1} := I_{m+1} - P_{m+1}$$

being, respectively, the design, the projection, and the annihilator matrices of a polynomial regression of degree $\nu+1$, for $\nu \in \mathbb{N}$. Prass and Pumi (2021) point out that, if $\{X_{k,t}\}_{t\in\mathbb{Z}}$ is a stationary process with finite mean, then $\mathbb{E}[\boldsymbol{\mathcal{E}}_{k,t}] = 0_{m+1}$. In lemma 3.1 the authors also prove that, if $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two jointly strictly stationary processes, then so are $\{\boldsymbol{\mathcal{E}}_{1,i}\}_{i=1}^{n-m}$ and $\{\boldsymbol{\mathcal{E}}_{2,i}\}_{i=1}^{n-m}$. Figure 4.1 presents the plot of a stationary time series $\{X_t\}_{t=1}^{11}$, the corresponding integrated signal $\{R_t\}_{t=1}^{11}$ and illustrates the step-by-step process of the polynomial fit with overlapping windows and $m = 9$.

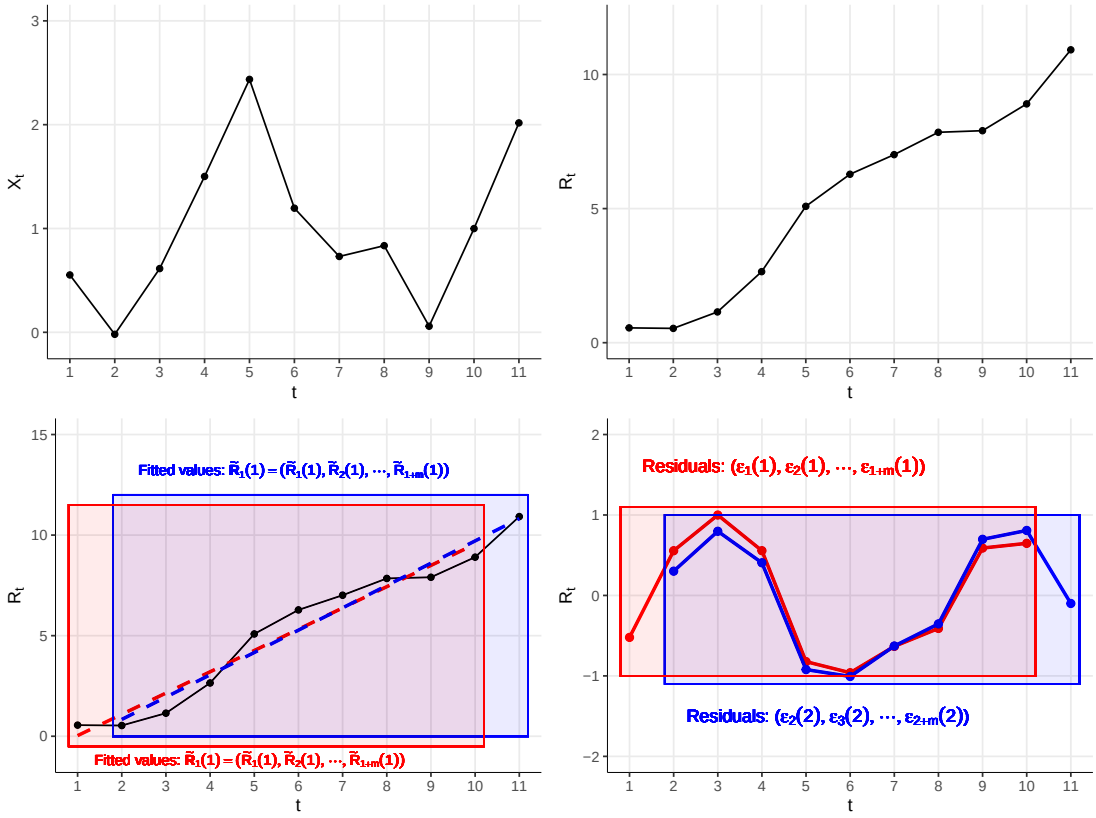

Figure 4.1: A simulated sample $\{X_t\}_{t=1}^{11}$ (top left) from a stationary time series, the corresponding integrated signal $\{R_t\}_{t=1}^{11}$ (top right), the window polynomial fit for the integrated signal (bottom left) and the corresponding residuals (bottom right).

4.2 Detrended variance, covariance and correlation coefficient

For $0 < m < n$ and $i \in \{1, \cdots, n-m\}$, let $f^2_{k,DFA}(m,i)$ be the sample variance of the residuals $\{\mathcal{E}_{k,t}(i)\}_{t=i}^{m+i}$, for $k \in \{1,2\}$ and $f_{DCCA}(m,i)$ be the sample covariance between the residuals $\{\mathcal{E}_{1,t}(i)\}_{t=i}^{m+i}$ and $\{\mathcal{E}_{2,t}(i)\}_{t=i}^{m+i}$, corresponding to the i-th box, that is,

$$f^2_{k,DFA}(m,i) := \frac{1}{m}\boldsymbol{\mathcal{E}}'_{k,i}\boldsymbol{\mathcal{E}}_{k,i} \qquad \text{and} \qquad f_{DCCA}(m,i) := \frac{1}{m}\boldsymbol{\mathcal{E}}'_{1,i}\boldsymbol{\mathcal{E}}_{2,i}.$$

The detrended variance $F^2_{k,DFA}$, $k \in \{1,2\}$, the detrended covariance F_{DCCA} and the detrended correlation coefficient ρ_{DCCA}, are defined respectively by

$$F^2_{k,DFA}(m) = \frac{1}{m-n}\sum_{i=1}^{n-m} f^2_{k,DFA}(m), \quad F_{DCCA}(m) = \frac{1}{m-n}\sum_{i=1}^{n-m} f_{DCCA}(m,i), \tag{4.3}$$

and

$$\rho_{DCCA}(m) = \frac{F_{DCCA}(m)}{\sqrt{F^2_{1,DFA}(m)}\sqrt{F^2_{2,DFA}(m)}}. \tag{4.4}$$

In the literature, the DFA and DCCA are usually defined constructively based on a sample of a given stochastic processes, as in (4.3) and (4.4), and therefore can be seen as an estimator of some quantity. From (4.3) it is easy to see that, $F^2_{k,DFA}(m)$ and $f_{DCCA}(m,i)$ are simply the sample means of $\{f^2_{k,DFA}(m,i)\}_{i=1}^{n-m}$ and $\{f_{DCCA}(m,i)\}_{i=1}^{n-m}$, respectively. Now, if $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two jointly strictly stationary processes, then both processes $\{f^2_{k,DFA}(m,i)\}_{i=1}^{n-m}$ and $\{f_{DCCA}(m,i)\}_{i=1}^{n-m}$ are strictly stationary (Prass and Pumi, 2021, corollary 3.1) and hence

$$\mathbb{E}[F^2_{k,DFA}(m)] = \frac{1}{n-m}\sum_{i=1}^{n-m}\mathbb{E}[f^2_{k,DFA}(m,i)] = \mathbb{E}[f^2_{k,DFA}(m,1)] = \frac{1}{m}\mathbb{E}[\boldsymbol{\mathcal{E}}'_{k,1}\boldsymbol{\mathcal{E}}_{k,1}]$$

and

$$\mathbb{E}[F_{DCCA}(m)] = \frac{1}{n-m}\sum_{i=1}^{n-m}\mathbb{E}[f_{DCCA}(m,i)] = \mathbb{E}[f_{DCCA}(m,1)] = \frac{1}{m}\mathbb{E}[\boldsymbol{\mathcal{E}}'_{k,1}\boldsymbol{\mathcal{E}}_{k,2}].$$

Under this context, the theoretical counterpart of ρ_{DCCA} is given by

$$\rho_{\mathcal{E}}(m) = \frac{\mathbb{E}[F_{DCCA}(m)]}{\sqrt{\mathbb{E}[F^2_{1,DFA}(m)]}\sqrt{\mathbb{E}[F^2_{2,DFA}(m)]}} = \frac{\sum_{t=i}^{m+i}\mathrm{Cov}[\mathcal{E}_{1,t}(i),\mathcal{E}_{2,t}(i)]}{\sqrt{\sum_{t=i}^{m+i}\mathrm{Var}[\mathcal{E}_{1,t}(i)]}\sqrt{\sum_{t=i}^{m+i}\mathrm{Var}[\mathcal{E}_{2,t}(i)]}}, \tag{4.5}$$

$0 < m < n$, $1 \le m \le n-m$, where $\boldsymbol{\mathcal{E}}_{k,i}(i) = (\mathcal{E}_{k,i}(i), \cdots \mathcal{E}_{k,i+m}(i))'$ is defined by equation (4.2), $k \in \{1,2\}$.

As pointed out in Prass and Pumi (2021), the coefficient $\rho_{\varepsilon}(m)$ given in (4.5) can be written as the average covariance divided by the square root of the average variances corresponding to the processes $\{\mathcal{E}_{1,t}(i)\}_{t=i}^{m+i}$ and $\{\mathcal{E}_{2,t}(i)\}_{t=i}^{m+i}$, which are the residuals of a local polynomial fit applied to the i-th window associated to the integrated processes $\{R_{1,t}\}_{t=1}^{n}$ and $\{R_{2,t}\}_{t=1}^{n}$. Hence, (4.5) is clearly not a direct measure of the cross-correlation between the original processes and there is no obvious interpretation for it. Another important point to note is that, since

$$\mathbb{E}[\rho_{DCCA}(m)] = \mathbb{E}\left[\frac{F_{DCCA}(m)}{\sqrt{F^2_{1,DFA}(m)}\sqrt{F^2_{2,DFA}(m)}}\right] \neq \frac{\mathbb{E}[F_{DCCA}(m)]}{\sqrt{\mathbb{E}[F^2_{1,DFA}(m)]}\sqrt{\mathbb{E}[F^2_{2,DFA}(m)]}} = \rho_{\mathcal{E}}(m),$$

the coefficient $\rho_{DCCA}(m)$ is a biased estimator for $\rho_{\mathcal{E}}(m)$, for any fixed n. However, under some reasonable conditions, $\rho_{\mathcal{E}}(m)$ is consistent. Before stating this result formally, we shall introduce some notation. For any $0 < m < n$ and $k_1, k_2 \in \{1, 2\}$, let

$$\Gamma_{k_1,k_2}^{h_1,h_2} := \mathrm{Cov}\left(\boldsymbol{X}_{k_1,m+1+h_1,}^{\langle 1\rangle}, \boldsymbol{X}_{k_2,m+1+h_2}^{\langle 1\rangle}\right), \quad 0 \le h_1, h_2 < n - m.$$

Also, denote by $\kappa_{k_1,k_2}(p, r, q, s)$ the joint cumulant of $(X_{k_1,p}, X_{k_1,r}, X_{k_2,q}, X_{k_2,s})$ and, for any $h \ge 0$, let $\mathcal{K}_{k_1,k_2}(h)$ be the $[(m+1)(m+1+h)] \times [(m+1)(m+1+h)]$ block matrix where the (r, s)-th element in the (p, q)-th block is given by

$$\left[[\mathcal{K}_{k_1,k_2}(h)]^{p,q}\right]_{r,s} := \kappa_{k_1,k_2}(p, r, q, s), \quad 1 \le p, q \le m+1, \quad 1 \le r, s \le m+1+h.$$

For sake of simplicity, for any $h, h_1, h_2 \ge 0$ and $k, k_1, k_2 \in \{1, 2\}$, define

$$\Gamma_k^{h_1,h_2} := \Gamma_{k,k}^{h_1,h_2}, \quad \Gamma_k := \Gamma_{k,k}^{0,0}, \quad \Gamma_{1,2} := \Gamma_{1,2}^{0,0},$$
$$\mathcal{K}_k(h) := \mathcal{K}_{k,k}(h), \quad \mathcal{K}_k := \mathcal{K}_k(0) \quad \text{and} \quad \mathcal{K}_{k_1,k_2} := \mathcal{K}_{k_1,k_2}(0).$$

Finally, let

$$K_{m+1} = K_{m+1}(0) := J'_{m+1} Q_{m+1} J_{m+1}, \quad K_{m+1}^{\otimes} = K_{m+1}^{\otimes}(0) := K_{m+1} \otimes K_{m+1}$$
$$K_{m+1}(h) := [J_{m+1+h}^{\langle h+1\rangle}]' Q_{m+1} J_{m+1+h}^{\langle h+1\rangle} \quad \text{and} \quad K_{m+1}^{\otimes}(h) := K_{m+1} \otimes K_{m+1}(h), \quad h > 0,$$

where $\otimes$ denotes the Kronecker product. Under the assumptions of joint stationarity and finite fourth moment for $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{1,t}\}_{t\in\mathbb{Z}}$, Theorem 4.1 presents closed-form expressions for the expected values, the variances

$$\gamma_{k,DFA}(0) := \mathrm{Var}[f_{k,DFA}^2(m, i)], \ k \in \{1, 2\} \quad \text{and} \quad \gamma_{DCCA}(0) := \mathrm{Var}[f_{DCCA}(m, i)],$$

and the covariance functions

$$\gamma_{k,DFA}(h) := \mathrm{Cov}[f_{k,DFA}^2(m, i), f_{k,DFA}^2(m, i+h)] \quad \text{and}$$
$$\gamma_{DCCA}(h) := \mathrm{Cov}[f_{DCCA}(m, i), f_{DCCA}(m, i+h)], \quad h \ne 0$$

related to the stochastic processes $\{f_{k,DFA}^2(m, i\}_{i=1}^{n-m}$ and $\{f_{DCCA}(m, i\}_{i=1}^{n-m}$. Theorem 4.8 provides sufficient conditions for consistency and almost sure convergence of $\rho_{DCCA}(m)$. Since $\rho_{DCCA}(m)$ is bounded, the asymptotic unbiasedness follows immediately.

Teorema 4.1 (Prass and Pumi (2021), theorem 3.1)**.** *Let $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ be two jointly strictly stationary stochastic processes with $\mathbb{E}[|X_{k,t}|^4] < \infty$, $k \in \{1, 2\}$. Then, for all $0 < m < n$, $1 \le i \le n - m$, $0 \le h < n - m$ and $k \in \{1, 2\}$,*

$$\mathbb{E}[f_{k,DFA}^2(m, i)] = \frac{1}{m}\,\mathrm{tr}(K_{m+1}\Gamma_k),$$
$$\gamma_{k,DFA}(0) = \frac{1}{m^2}\left[tr(K_{m+1}^{\otimes}\mathcal{K}_k + 2K_{m+1}\Gamma_k K_{m+1}\Gamma_k)\right],$$
$$\gamma_{k,DFA}(h) = \frac{1}{m^2}\left[\mathrm{tr}(K_{m+1}^{\otimes}(h)\mathcal{K}_k(h) + 2K_{m+1}\Gamma_k^{0,h} K_{m+1}(h)\Gamma_k^{h,0})\right],$$

and

$$\mathbb{E}[f_{DCCA}(m, i)] = \frac{1}{m}\,\mathrm{tr}(K_{m+1}\Gamma_{1,2}),$$
$$\gamma_{DCCA}(0) = \frac{1}{m^2}[\mathrm{tr}(K_{m+1}^{\otimes}\mathcal{K}_{1,2} + K_{m+1}\Gamma_1 K_{m+1}\Gamma_2 + K_{m+1}\Gamma_{1,2}K_{m+1}\Gamma_{1,2})],$$
$$\gamma_{DCCA}(h) = \frac{1}{m^2}\left[\mathrm{tr}(K_{m+1}^{\otimes}(h)\mathcal{K}_{1,2}(h) + K_{m+1}\Gamma_1^{0,h} K_{m+1}(h)\Gamma_2^{h,0} + K_{m+1}\Gamma_{1,2}^{0,h} K_{m+1}(h)\Gamma_{1,2}^{h,0})\right].$$

Teorema 4.2 (Prass and Pumi (2021), theorem 3.2). *Let $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ be two jointly stationary processes. If $\gamma_{k,DFA}(h)\to 0$ and $\gamma_{DCCA}(h)\to 0$, as $h\to\infty$, then*

$$F^2_{k,DFA}(m) \xrightarrow{P} \mathbb{E}[f^2_{k,DFA}(m,1)] = \frac{1}{m}\operatorname{tr}(K_{m+1}\Gamma_k), \quad \text{as } n\to\infty, \tag{4.6}$$

and

$$F_{DCCA}(m) \xrightarrow{P} \mathbb{E}[f_{DCCA}(m,1)] = \frac{1}{m}\operatorname{tr}(K_{m+1}\Gamma_{1,2}), \quad \text{as } n\to\infty. \tag{4.7}$$

Moreover,

$$\rho_{DCCA}(m) \xrightarrow{P} \frac{\operatorname{tr}(K_{m+1}\Gamma_{1,2})}{\sqrt{\operatorname{tr}(K_{m+1}\Gamma_1)\operatorname{tr}(K_{m+1},\Gamma_2)}} = \rho_{\mathcal{E}}(m), \quad \text{as } n\to\infty, \tag{4.8}$$

Furthermore, if

$$\sum_{h=1}^{\infty}\frac{\gamma_{k,DFA}(h)}{h^{q_k}}<\infty, \qquad \text{and} \qquad \sum_{h=1}^{\infty}\frac{\gamma_{DCCA}(h)}{h^{q_{12}}}<\infty,$$

for some $0\le q_k, q_{12}<1$, then the convergence holds almost surely.

Calculating the limit in probability of $F^2_{k,\text{DFA}}$, F_{DCCA}, and ρ_{DCCA}, as $n\to\infty$, Theorem 4.2 requires that $\gamma_{k,\text{DFA}}(h)\to 0$ and $\gamma_{\text{DCCA}}(h)\to 0$ as $h\to\infty$. Despite being simple assumptions, verifying them can be challenging, therefore, Prass and Pumi (2021) proposes a method to check these conditions using information about the original time series, which are often easier to verify. For Theorem 4.2 to hold, it is sufficient that, $\operatorname{Cov}[X_{k_1,t}, X_{k_2,t+h}]\to 0$ and $\kappa_{k_1,k_2}(p,h+\tau,p,h+q)\to 0$, as $|h|\to\infty$, for $k_1,k_2\in\{1,2\}$ and any fixed $p,q,\tau>0$.

Figure 4.2 displays a heatmap with the values of the matrix K_{m+1} for $m\in\{3,27,81,101\}$. It is noticeable that all matrices exhibit a similar pattern, with a band of positive values around the main diagonal and negative values arranged in entries further away from the main diagonals. Additionally, it is worth noticing that the values in the first row and first column of the matrix are equal to 0, which means that the first row of the matrices $K_{m+1}\Gamma_k$ and $K_{m+1}\Gamma_{1,2}$ always equal to 0. The asymptotic behavior of the quantities $\operatorname{tr}(K_{m+1}\Gamma_k)$ and $\operatorname{tr}(K_{m+1}\Gamma_{12})$ was derived in Prass and Pumi (2021) for the specific case where $\nu=0$ (linear polynomial fit for each window), and the autocovariance functions $\gamma_k(.)$ and the cross-autocovariance $\gamma_{1,2}(.)$ are absolutely summable. The authors show that, in this context,

$$\mathbb{E}[f^2_{k,\text{DFA}}(m,1)]\sim\frac{m}{15}\sum_{h\in\mathbb{Z}}\gamma_k(h), \quad \mathbb{E}[f_{\text{DCCA}}(m,1)]\sim\frac{m}{15}\sum_{h\in\mathbb{Z}}\gamma_{1,2}(h), \quad \text{as } m\to\infty, \tag{4.9}$$

and, as a consequence,

$$\rho_{DCCA}(m) \xrightarrow{P} \frac{\sum_{h\in\mathbb{Z}}\gamma_{1,2}(h)}{\sqrt{\sum_{h\in\mathbb{Z}}\gamma_1(h)}\sqrt{\sum_{h\in\mathbb{Z}}\gamma_2(h)}}, \quad \text{as } m,n\to\infty.$$

Therefore, $F^2_{k,\text{DFA}}(m)$ for $k\in\{1,2\}$ and $F_{\text{DCCA}}(m)$, asymptotically exhibit linear behaviors, while $\rho_{DCCA}(m)$ converges to a constant.

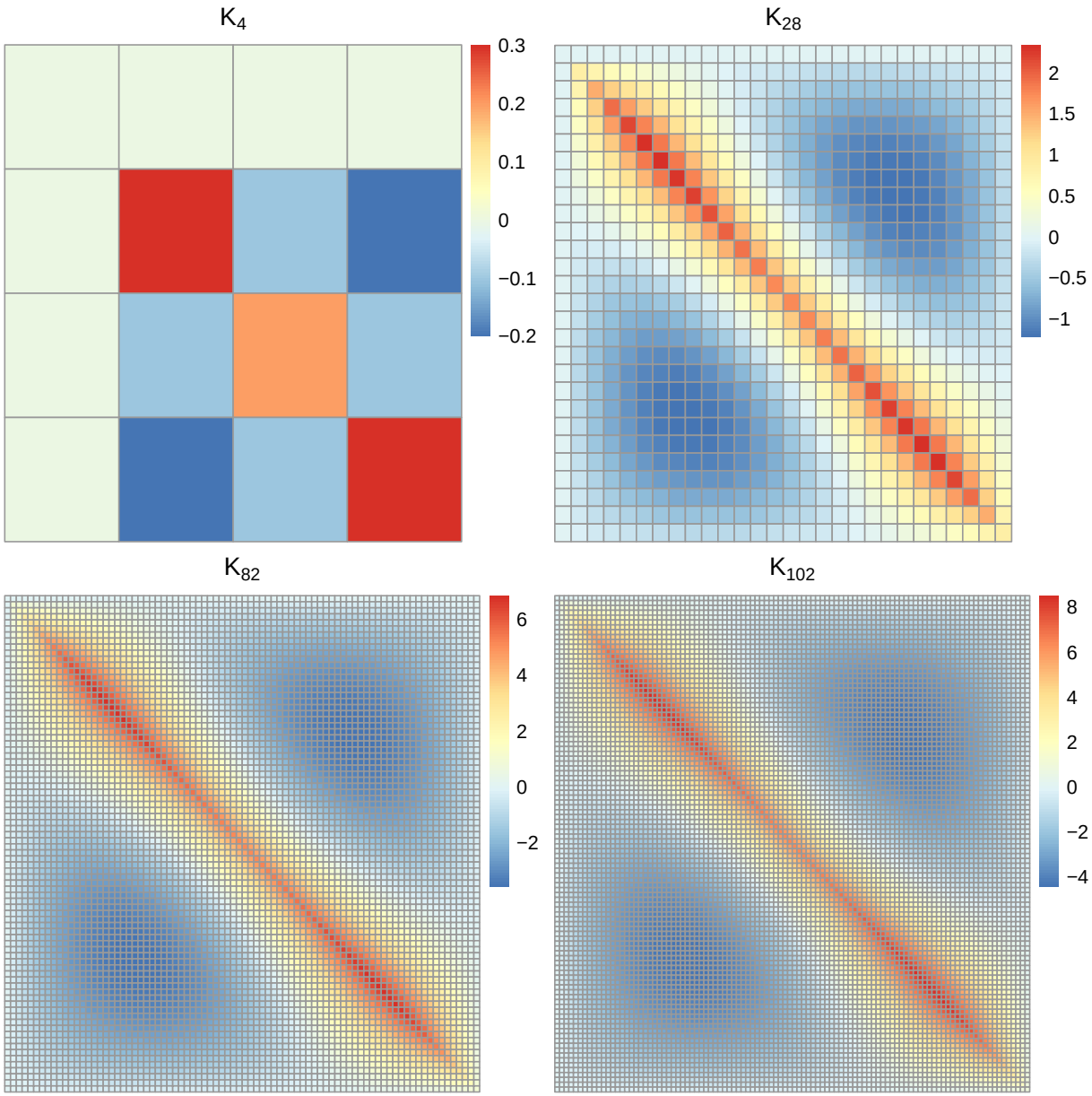

Figure 4.2: Heatmap of the K_{m+1} matrix for $m \in \{3, 27, 81, 101\}$.

4.3 Novel theoretical results for imputed time series

Theoretical results concerning DFA and DCCA in the context of missing data are currently absent from the literature, even in simple cases such as when missing values are filled using the mean. In what follows, asymptotic results for the autocovariance and cross-covariance functions corresponding to a time series filled with the mean are derived, and a case study to investigate its impact on the expected values of the function is conducted.

Let $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ be two jointly weak stationary processes with autocovariance functions $\gamma_{k,k}(\cdot) := \gamma_k(\cdot)$, $k \in \{1, 2\}$, and cross-covariance function $\gamma_{1,2}(\cdot)$ satisfying

$$\sum_{h\in\mathbb{Z}} |\gamma_{k_1,k_2}(h)| < \infty, \quad \text{for any } k_1, k_2 \in \{1, 2\}.$$

Define the set $T = \{1 \cdots, n\}$ and let ρ be the proportion of missing values. Also, let $T_k := \{t_{k,1}, \cdots, t_{k,\lfloor n\rho \rfloor}\} \subset T$ be the set of indexes of missing data and $T_k^c = T - T_k$ the set of indexes of observed values, for $k \in \{1, 2\}$. The time series with missing data can be written as

$$X_{k,t}^{\text{miss}} = \texttt{NA} \times I(t \in T_k) + X_{k,t} \times I(t \in T_k^C), \quad t \in T, \quad k \in \{1, 2\}.$$

Denote by $\gamma_{k,k}^{\text{miss}}(\cdot, \cdot)$, $k \in \{1, 2\}$, and $\gamma_{1,2}^{\text{miss}}(\cdot, \cdot)$, respectively, the autocovariance functions and the cross-covariance function corresponding to the time series obtained after applying some imputation method to reconstruct $\{X_{1,t}^{\text{miss}}\}_{t=1}^n$ and $\{X_{2,t}^{\text{miss}}\}_{t=1}^n$. Then, for any $k_1, k_2 \in \{1, 2\}$,

$$\gamma_{k_1,k_2}^{\text{miss}}(t_1, t_2) = \begin{cases} \text{Cov}[X_{k_1,t_1}, X_{k_2,t_2}], & \text{if } t_1 \in T_{k_1}^C \text{ and } t_2 \in T_{k_2}^C, \\ \text{Cov}[X_{k_1,t_1}, \hat{X}_{k_2,t_2}], & \text{if } t_1 \in T_{k_1}^C \text{ and } t_2 \in T_{k_2}, \\ \text{Cov}[\hat{X}_{k_1,t_1}, X_{k_2,t_2}], & \text{if } t_1 \in T_{k_1} \text{ and } t_2 \in T_{k_2}^C, \\ \text{Cov}[\hat{X}_{k_1,t_1}, \hat{X}_{k_2,t_2}], & \text{if } t_1 \in T_{k_1} \text{ and } t_2 \in T_{k_2}, \end{cases} \tag{4.10}$$

where $\hat{X}_{k_1,t_1}$ and $\hat{X}_{k_2,t_2}$ are the imputed values. In this work, we consider a simple case where missing values are imputed with the mean computed from the non-missing values, that is,

$$\hat{X}_{k,t} = \hat{X}_k = \frac{1}{\#(T_k^c)} \sum_{i \in T_k^c} X_{k,i}, \quad t \in T_k, \tag{4.11}$$

where $\#(T_k^c) = n - \lfloor n\rho \rfloor$, for $k \in \{1, 2\}$. Hence, the estimator $\hat{X}_k$ is the arithmetic mean of the observed values. Considering the mean imputation, expression (4.10) can be simplified for large n. As the estimated values in (4.11) depend on n, it is possible to show that asymptotically, the terms of (4.10) that depend on the imputed values tend to 0.

To see why this is the case, if $\{X_{k,t}\}_{t \in \mathbb{Z}}$ is a weakly stationary process satisfying $\gamma_{k,k}(h) \to 0$, as $h \to \infty$, then for any sample $\{X_{k,t}\}_{t=1}^n$ of this process, $\text{Var}[\bar{X}] \to 0$, as $n \to \infty$. Therefore, it is trivial to note that under the same assumptions, $\text{Var}[\hat{X}_k] \to 0$, as $n \to \infty$, given that the observed values $\{X_{k,t}\}_{t \in T_k^c}$ of the process form a sample (with size $n - \lfloor n\rho \rfloor \sim n(\rho - 1)$) of the complete process, for $k \in \{1, 2\}$. Consequently, for any $k_1, k_2 \in \{1, 2\}$,

$$|\text{Cov}[\hat{X}_{k_1}, \hat{X}_{k_2}]| \leq \sqrt{\text{Var}[\hat{X}_{k_1}]}\sqrt{\text{Var}[\hat{X}_{k_2}]} \longrightarrow 0, \text{ as } n \to \infty,$$

and

$$|\text{Cov}[X_{k_1,t}, \hat{X}_{k_2}]| \leq \sqrt{\text{Var}[X_{k_1}]}\sqrt{\text{Var}[\hat{X}_{k_2}]} \longrightarrow 0, \text{ as } n \to \infty.$$

Therefore, the covariance functions and cross-correlation function corresponding to the imputed series satisfy, respectively

$$\gamma_{k,k}^{\text{miss}}(t_1, t_2) \longrightarrow \begin{cases} \text{Cov}[X_{k,t_1}, X_{k,t_2}], & \text{if } t_1, t_2 \in T_k^C, \\ 0, & \text{otherwise,} \end{cases} \quad k \in \{1, 2\}, \quad n \to \infty, \tag{4.12}$$

and

$$\gamma_{1,2}^{\text{miss}}(t_1, t_2) \longrightarrow \begin{cases} \text{Cov}[X_{1,t_1}, X_{2,t_2}], & \text{if } t_1 \in T_1^C \text{ and } t_2 \in T_2^C, \\ 0, & \text{otherwise,} \end{cases} \quad n \to \infty. \tag{4.13}$$

By letting Γ_k^{miss} and $\Gamma_{1,2}^{\text{miss}}$ be the matrices whose (i,j)-th entries are, respectively, $\gamma_{k,k}^{\text{miss}}(i,j)$ and $\gamma_{1,2}^{\text{miss}}(i,j)$, it can be observed that asymptotically, the values of the $n\rho$ columns and $n\rho$ rows corresponding to the missing observations in the autocovariance matrix Γ_k^{miss} will approach 0. Therefore, $n^2(1-\rho)^2$ elements of the covariance matrix of the time series imputed with the mean are potentially different from zero. It is worth noting that considering that the indexes of missing values in the series may be different, the cross-covariance matrix $\Gamma_{1,2}^{\text{miss}}$ can have up to twice the proportion of missing values of the covariance matrices Γ_1^{miss} and Γ_2^{miss}. Figure 4.3 illustrates the sparsity of the autocovariance and cross-covariance matrices for various proportions of missing data ρ. The blue points represent cases with $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and the red dashed line represents the maximum proportion of missing data for the cross-covariance matrix.

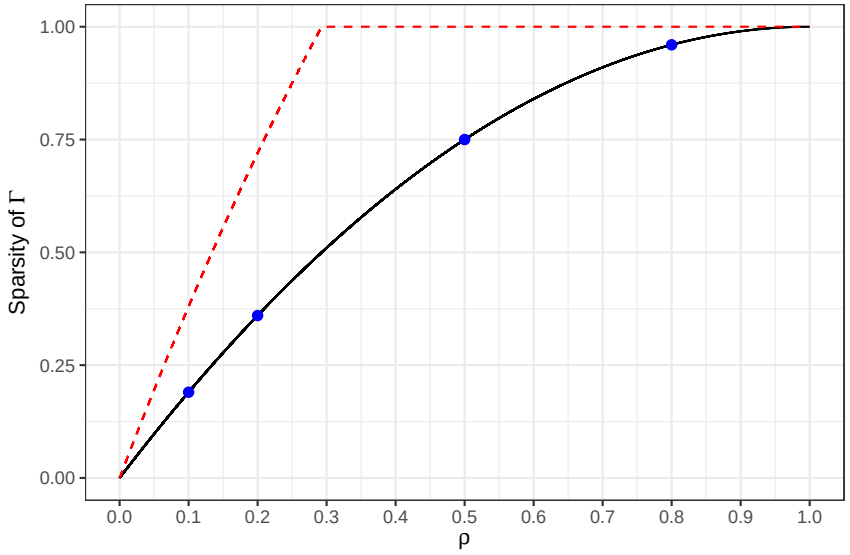

Figure 4.3: Line plot illustrating the sparsity of the autocovariance (solid line) and cross-covariance (dashed line) matrices at varying proportions of missing values ρ. Blue points indicate specific values of $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and the red dashed line represents the maximum proportion of missing data for the cross-covariance matrix.

With the asymptotic expressions derived in (4.12) and (4.13), respectively, for the autocovariance and cross-covariance matrices in the case of missing data, it becomes possible to study the behavior of the limiting matrices

$$K_{m+1}\left(\lim_{n\to\infty} \Gamma_k^{\text{miss}}\right) \quad \text{and} \quad K_{m+1}\left(\lim_{n\to\infty} \Gamma_{1,2}^{\text{miss}}\right),$$

which correspond to the theoretical matrices that should be used in the calculation of the asymptotic expected value of $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ as per (4.6) and (4.7) under the presence of missing values. In what follows, a study showcasing the expected behavior of $K_{m+1}(\lim_{n\to\infty} \Gamma_k^{\text{miss}})$ for an AR(1) and an MA(1) process is presented.

4.3.1 Case study

The objective of this case study is to investigate the behavior of the limiting matrix $K_{m+1}(\lim_{n\to\infty} \Gamma_k^{\text{miss}})$ which replaces $K_{m+1}\Gamma_k$ in the presence of missing data. Since only marginal behaviors are studied,

for sake of simplicity, in this section the following notation shall be adopted

$$\Gamma = \Gamma_k, \quad \Gamma^{\text{miss}} := \Gamma_k^{\text{miss}} \quad \text{and} \quad \Gamma_\infty^{\text{miss}} = \lim_{n\to\infty} \Gamma^{\text{miss}}.$$

For each case, the matrix $K_{m+1}\Gamma$, associated to the time series without missing values will be presented, along with simulated cases for the matrix Γ^{miss} considering $m = 27$ and proportions of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The calculated values for the main diagonal of the matrices for $m = 101$ and proportions of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ will be displayed. Finally, the trace values of the matrices for $m \in \{27, 81, 101\}$ and $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ will be depicted.

4.3.1.1 Autocovariance of AR(1)

In this analysis the time series $\{X_t\}_{t=1}^n$ is a sample from an AR(1) process whose definition and corresponding (i, j)-th term of autocovariance matrix are given by

$$X_t = 0.6X_{t-1} + \varepsilon_t,\ t \in \mathbb{Z}, \qquad \text{and} \qquad [\Gamma]_{i,j} = \frac{0.6^{|i-j|}}{0.64}, \tag{4.14}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, is a sequence of i.i.d. $\mathcal{N}(0, 1)$ random variables.

The heatmap in Figure 4.4 illustrates that the structure of the theoretical $K_{28}\Gamma$ matrix in this example behaves similarly to the K_{28} matrix shown in Figure 4.2, with a band of positive values around the main diagonal with the values in the center of the matrix are not as high as those at the beginning and end of the diagonal. It is interesting to notice that the range of values in $K_{28}\Gamma$ is greater than that of the K_{28} matrix in this example.

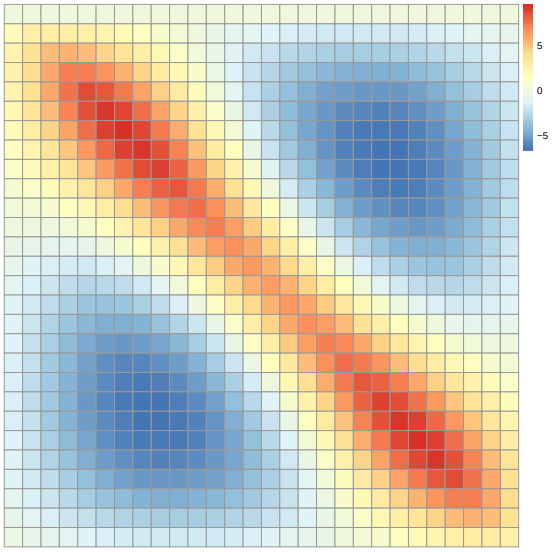

Figure 4.4: Heatmap of the $K_{28}\Gamma$ matrix considering an AR(1) process.

Figure 4.5 displays examples of the calculated values for the $K_{28}\Gamma_\infty^{\text{miss}}$ matrix with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from top to bottom and evenly distributed across all, start, middle, and end columns of the matrix, respectively, from left to right. The non-zero entries are similar to those in the matrix without missing values, both in structure and value range.

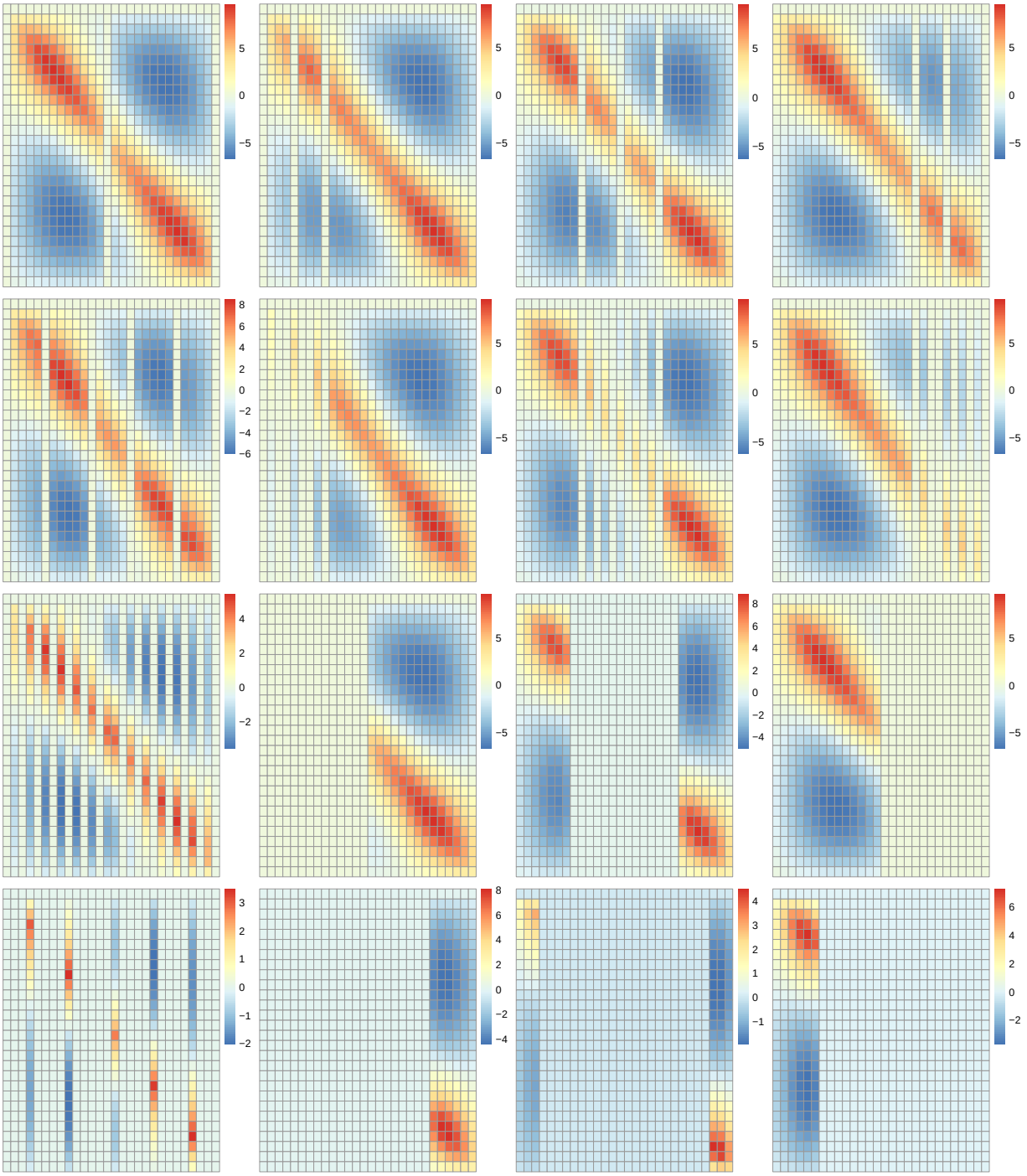

Figure 4.5: Examples of heatmaps of the $K_{28}\Gamma_{\infty}^{\text{miss}}$ matrix considering an AR(1) process with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from top to bottom. Missing data are evenly distributed across all, start, middle, and end columns of the matrix, respectively, from left to right.

Figure 4.6 presents box-plots with the calculated diagonal values of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an AR(1) process defined by (4.14). The graphs are based on 1000 replications, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$, from left to right and top to bottom. The red line represents the calculated values for the diagonal of $K_{102}\Gamma$. It is noticeable that the calculated values for the diagonal of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix are generally lower than those calculated for the complete case. The higher the proportion of missing values ρ, the lower the median values,

but the variability of the calculated values is independent of ρ. The values maintain a "wing" shape, consistent with the behavior shown in Figure 4.4. Cases where the matrix entry was 0 were omitted for better visibility and a more realistic depiction of variability in cases where the entry came from an observed value.

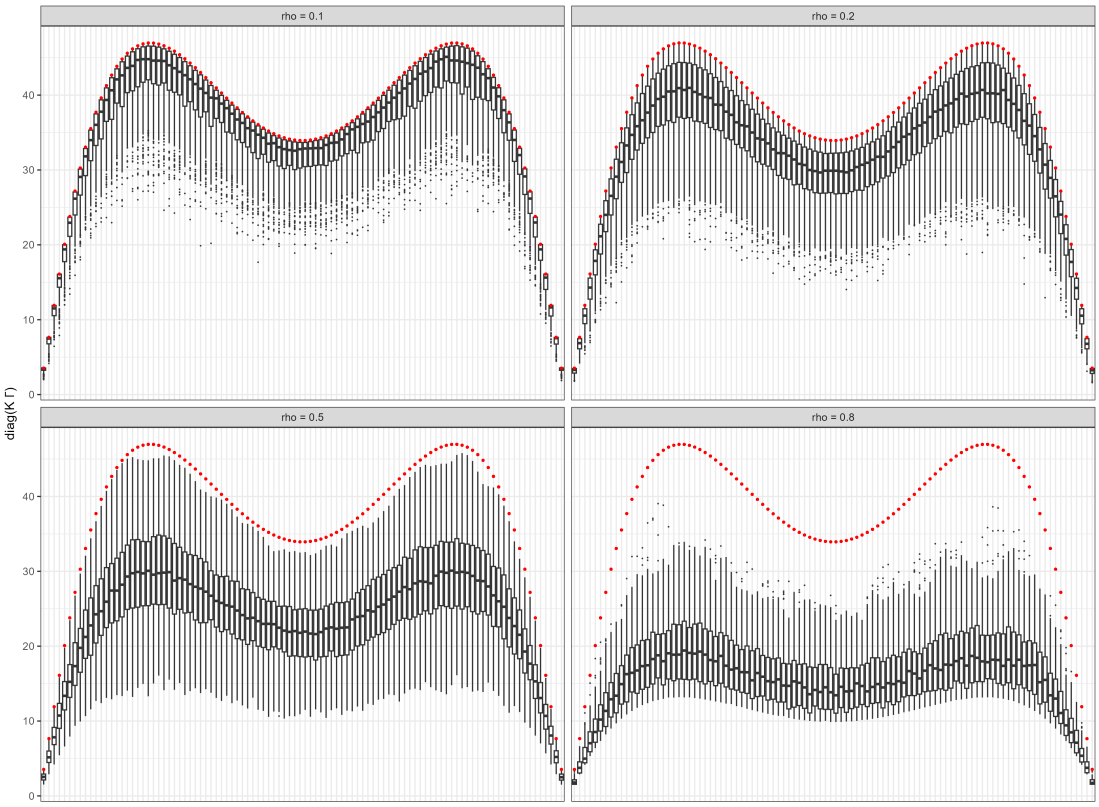

Figure 4.6: Boxplots illustrating the calculated values for 1000 replications of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an AR(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from left to right and top to bottom. The red line represents the calculated values for the diagonal of $K_{102}\Gamma$ without missing values.

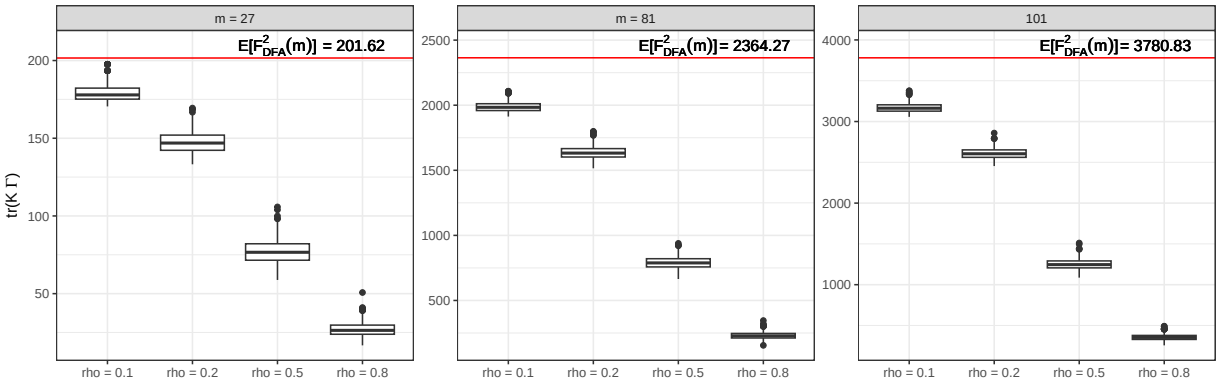

Figure 4.7: Boxplots illustrating the calculated values for 1000 replications of the trace of $K_{102}\Gamma_{\infty}^{\text{miss}}$ corresponding to an AR(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and different window sizes $m \in \{27, 81, 101\}$ from left to right. The red line represents the calculated values for the trace of $K_{m+1}\Gamma$.

Figure 4.7 shows the calculated values of the trace of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an AR(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (top to bottom) and different window sizes $m \in \{27, 81, 101\}$ (left to right), for 1000 replications. It can be observed that the trace of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ was always lower than the value of $K_{102}\Gamma$, with the median decreasing as the proportion of missing values increases, the decay's behavior is similar for different values of m.

4.3.1.2 Autocovariance of an MA(1)

In this analysis the time series $\{X_t\}_{t=1}^{n}$ is a sample from an MA(1) process whose definition and corresponding (i, j)-th term of autocovariance matrix are given by

$$X_t = \varepsilon_t + 0.6\varepsilon_{t-1},\ t \in \mathbb{Z}, \qquad \text{and} \qquad [\Gamma_2]_{i,j} = 1.36I(i = j) + 0.6I(|i - j| = 1), \tag{4.15}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, is a sequence of i.i.d. $\mathcal{N}(0, 1)$ random variables.

The heatmap in Figure 4.8 illustrates that the structure of the theoretical $K_{28}\Gamma_{\infty}^{\text{miss}}$ matrix in this example behaves similarly to the K_{28} matrix shown in Figure 4.2, with a band of positive values around the main diagonal with the values in the center of the matrix are not as high as those at the beginning and end of the diagonal. It is interesting to notice that, in this example, the range of values in $K_{28}\Gamma$ is greater than those of the K_{28} matrix, but lower than those in Figure 4.4.

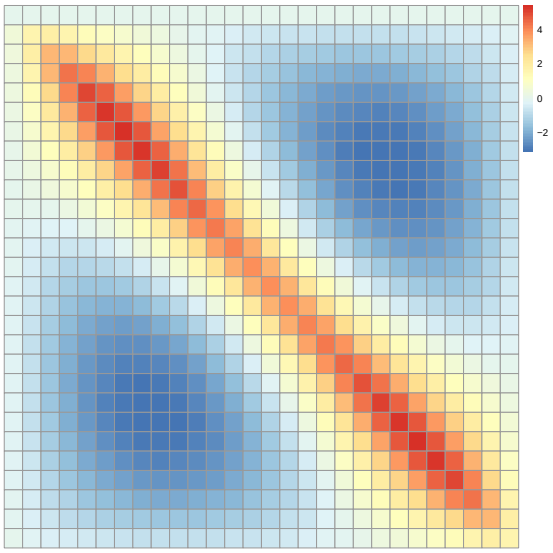

Figure 4.8: Heatmap of the $K_{28}\Gamma$ matrix considering an MA(1) process.

Figure 4.9 displays examples of the calculated values for the $K_{28}\Gamma_{\infty}^{\text{miss}}$ matrix with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from top to bottom and evenly distributed across all, start, middle, and end columns of the matrix, respectively, from left to right. The non-zero entries are similar to those in the matrix without missing values, both in structure and value range.

Figure 4.10 presents box-plots with the calculated diagonal values of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an MA(1) process defined by (4.15). The graphs are based on 1000 replications, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$, from left to right and top to bottom.

The red line represents the calculated values for the diagonal of $K_{102}\Gamma_{\infty}^{\text{miss}}$. Since the only non-zero entries in the matrix Γ are those within 1 entry of distance from the main diagonal, there are fewer possible values for the main diagonal of $K_{102}\Gamma$. It can be observed that for $\rho = 0.1$, the values of the matrix diagonal are concentrated at the same point as the values calculated with the complete matrix, and the median value decreases as the proportion of missing values increases. The values maintain a "wing" shape, consistent with the behavior shown in Figure 4.4. Cases where the matrix entry was 0 were omitted for better visibility and a more realistic depiction of variability in cases where the entry came from an observed value.

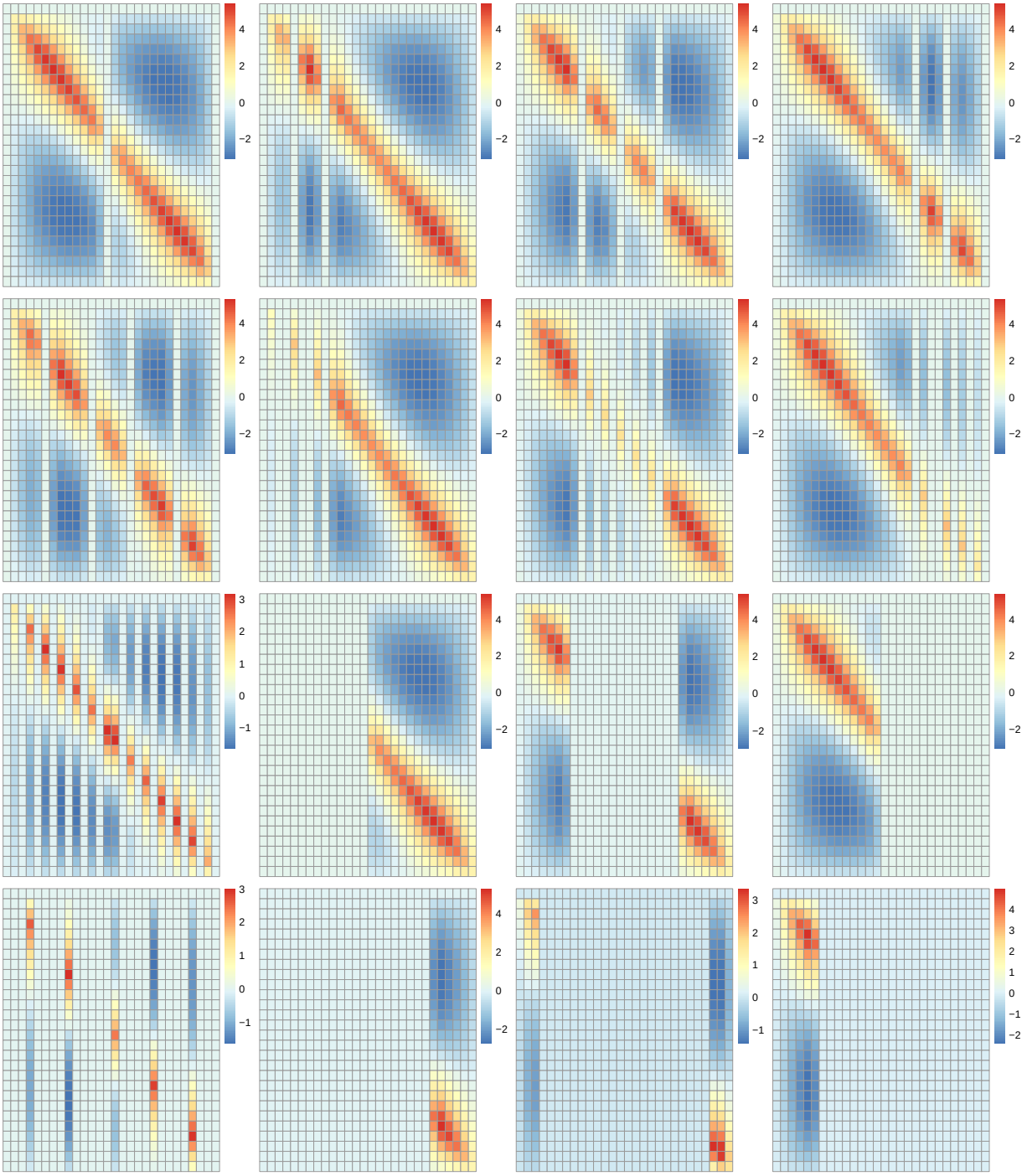

Figure 4.9: Examples of heatmaps of the $K_{28}\Gamma_{\infty}^{\text{miss}}$ matrix considering an MA(1) process with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from top to bottom. Missing data are evenly distributed across all, start, middle, and end columns of the matrix, respectively, from left to right.

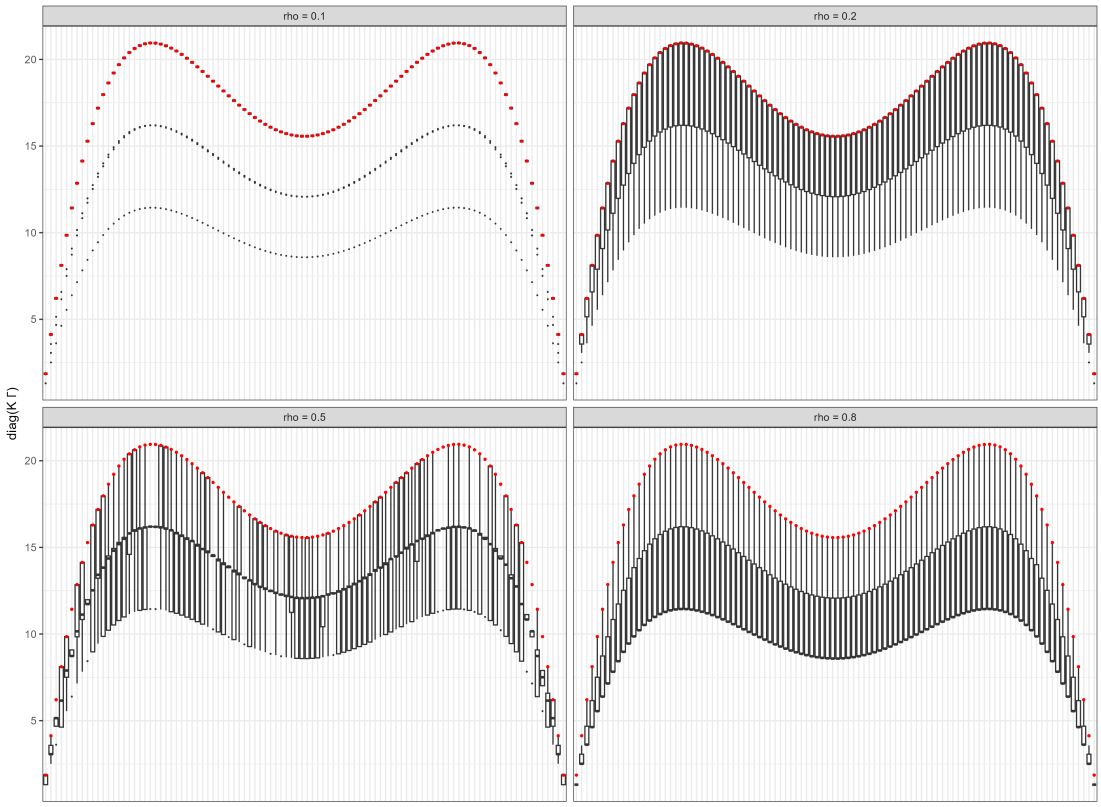

Figure 4.10: Boxplots illustrating the calculated values for 1000 replications of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an MA(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ from left to right and top to bottom. The red line represents the calculated values for the diagonal of $K_{102}\Gamma$ without missing values.

Figure 4.11 shows the calculated values of the trace of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ matrix corresponding to an MA(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ (top to bottom) and different window sizes $m \in \{27, 81, 101\}$ (left to right), for 1000 replications. It can be observed that the trace of the $K_{102}\Gamma_{\infty}^{\text{miss}}$ was always lower than the value of $K_{102}\Gamma$, with the median decreasing as the proportion of missing values increases, the decay's behavior is similar for different values of m.

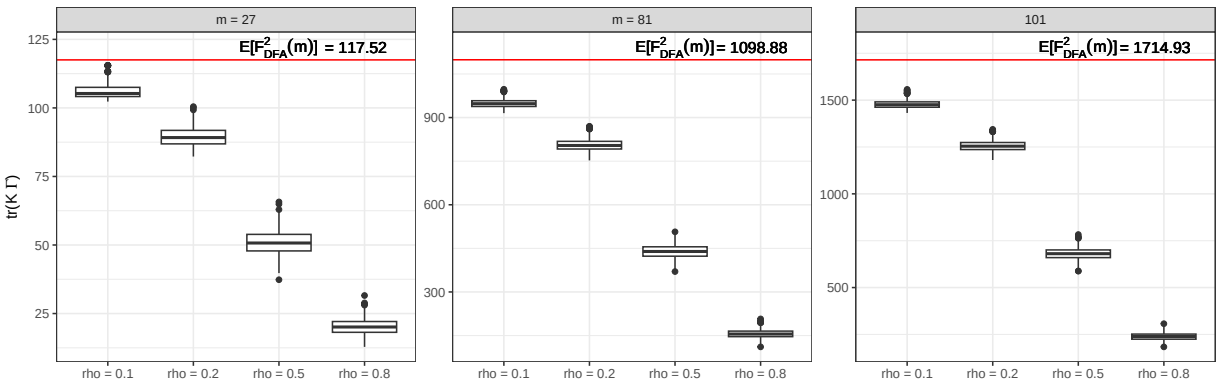

Figure 4.11: Boxplots illustrating the calculated values for 1000 replications of the trace of $K_{102}\Gamma_{\infty}^{\text{miss}}$ considering an MA(1) process, with varying proportions of missing data $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and different window sizes $m \in \{27, 81, 101\}$ from left to right. The red line represents the calculated values for the trace of $K_{m+1}\Gamma$ without missing values.

Chapter 5

Monte Carlo Simulations

Theoretical results concerning DFA and DCCA in the context of missing data are currently absent from the literature, even in the simplest case where missing observations are imputed with the mean. Deriving the theoretical asymptotic behavior of these measures is a challenging task. For this reason, in this work, we focus on Monte Carlo simulation studies. The goal is to investigate the behavior of the DFA, DCCA, and the coefficient ρ_{DCCA} when missing values are imputed considering traditional methods and decision tree algorithms. The questions we aim to answer with the current simulation study are the following.

Q1: Which imputation method presents the best performance in terms of mean square error?

Q2: Are the DFA, DCCA and/or ρ_{DCCA} values calculated using recomposed time series different from those based on the complete time series? If yes, which imputation method leads to estimated values of DFA, DCCA, and ρ_{DCCA} closer to their expected values?

Q3: Do the best-performing imputation methods in Q1 and Q2 coincide?

Q4: Does the dependence structure of the data affect the results?

Q5: Does the proportion of missing values have any influence on the conclusions?

In what follows, the simulation study conducted to address these questions is described in detail.

5.1 Data generating process

In this study, we consider the same scenarios as those simulated in Prass and Pumi (2021), namely,

1. **Uncorrelated processes.** In this context, $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two stationary uncorrelated processes,

$$\Gamma_k = \text{Cov}\left[\boldsymbol{X}^{\langle 1\rangle}_{k,m+1}, \boldsymbol{X}^{\langle 1\rangle}_{k,m+1}\right], \quad k \in \{1,2\}, \quad \text{and} \quad \Gamma_{1,2} = 0_{m+1,m+1}, \quad \forall\, m > 0.$$

   In this case, $\mathbb{E}[F_{\text{DCCA}}(m)] = 0$ and $\rho_{\text{DCCA}}(m) \xrightarrow{P} 0$, as $n \longrightarrow \infty$, for all $m > 0$. The expression for $\mathbb{E}[F^2_{k,\text{DFA}}(m)]$ depends on the scenario considered. For this structure of cross-correlation, two scenarios are presented. Scenario 1 considers two i.i.d. standard Gaussian sequences independent

from each other. Scenario 2 considers an AR(1) and an MA(1) process generated from two i.i.d. standard Gaussian sequences, that is, two processes with autocorrelation but mutually independent.

2. **Bivariate white noise process.** In this context, $\{(X_{1,t}, X_{2,t})\}_{t\in\mathbb{Z}}$ is a bivariate white noise, with $\mathbb{E}[X_{k,t}] = \mu_k$, $\mathrm{Var}[X_{k,t}] = \sigma_k^2$, $k \in \{1,2\}$ and $\mathrm{Cov}[X_{1,t}, X_{2,t}] = \sigma_{12}$, so that

$$\Gamma_k = \sigma_k^2 I_{m+1}, \quad k \in \{1,2\}, \quad \text{and} \quad \Gamma_{1,2} = \sigma_{12} I_{m+1}, \quad \forall\, m > 0.$$

Hence, independent of the scenario considered,

$$\mathbb{E}[F^2_{k,\mathsf{DFA}}(m)] = \left[\frac{1}{15}m + \frac{2}{15} - \frac{1}{5m}\right]\sigma_k^2 \sim \frac{\sigma_k^2}{15}m, \quad \text{as } m \to \infty,$$

$$\mathbb{E}[F_{\mathsf{DCCA}}(m)] = \left[\frac{1}{15}m + \frac{2}{15} - \frac{1}{5m}\right]\sigma_{12} \sim \frac{\sigma_{12}}{15}m, \quad \text{as } m \to \infty,$$

and

$$\rho_{DCCA}(m) \xrightarrow{P} \frac{\sigma_{12}}{\sigma_1\sigma_2} = \mathrm{Corr}[X_{1,t}, X_{2,t}], \quad \text{as} \quad n \longrightarrow \infty, \quad \forall\ m > 0.$$

For this structure, four scenarios are considered. Scenarios 3.1 and 3.2 consider bivariate Gaussian processes with intermediate and high correlation, respectively. Scenarios 4.1 and 4.2 consider signal plus noise processes with low and high variance, respectively. Under these four scenarios, the two processes are i.i.d. sequences with cross-correlation only at lag $h = 0$.

3. **Short-memory cross-correlated processes.** In this context $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two jointly stationary processes that can be written as

$$X_{k,t} = \sum_{j\in\mathbb{Z}} \psi_{k,j}\eta_{k,t-j}, \quad t \in \mathbb{Z}, \quad \text{with} \sum_{j\in\mathbb{Z}} |\psi_{k,j}| < \infty, \quad k \in \{1,2\},$$

where $\{\eta_{k,t}\}_{t\in\mathbb{Z}}$, is a white noise process with zero mean, $\mathrm{Var}[\eta_{k,t}] = \tau_k^2$, and $\mathrm{Cov}[\eta_{1,r}, \eta_{2,s}] = \tau_{1,2} I(r = s)$. In this case, closed expressions for $\mathbb{E}[F^2_{k,\mathsf{DFA}}(m)]$ and $\mathbb{E}[F_{\mathsf{DCCA}}(m)]$ might be hard to derive, but it is known that

$$\rho_{\mathsf{DCCA}}(m) \xrightarrow{P} \mathsf{sign}(\Psi_{1,2})\frac{\tau_{1,2}}{\tau_1\tau_2} = \mathsf{sign}(\Psi_{1,2})\,\mathrm{Corr}[\eta_{1,t}, \eta_{2,t}], \qquad \text{as} \quad n, m \to \infty,$$

where $\Psi_{1,2} = \sum_{j\in\mathbb{Z}} \psi_{1,j} \sum_{\ell\in\mathbb{Z}} \psi_{2,\ell}$. For this structure, four scenarios are considered. Scenarios 5.1 and 5.2 consider couples of processes where the cross-correlation structure is driven by a moving average and an autoregressive structure, respectively. Scenarios 6.1 and 6.2 consider couples of processes sharing the same white noise sequence. Under these four scenarios, the cross-covariance structure is determined by the covariance structure of the underlying noise processes.

In any given scenario, to generate a time series with missing values we proceed as described in the sequel.

In all cases, the complete time series are generated with sample size $n = 2000$. Given a complete time series $\{X_t\}_{t=1}^n$ and a target proportion of missing data ρ, we select a set T_1 with $\lfloor n\rho \rfloor$ elements using a simple random sample without replacement from T. Subsequently, the corresponding observations of the original time series are transformed into missing values through the relation

$$X_t^{\mathsf{miss}} = \mathtt{NA} \times I(t \in T_1) + X_t \times I(t \in T_1^C), \quad t \in \{1, \cdots, n\}.$$

The proportions of missing values considered in the study are $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. For each scenario and proportion of missing data, 1000 replications of the procedure are performed.

5.2 Imputation and estimation

The imputation methods and corresponding abbreviations used in figures and tables are Last Observation Carried Forward (LOCF), Linear Interpolation (LI), Kalman Smoothing (KS), Exponential Moving Average (EMA), Mean (M), Classification and Regression Trees (rpart), and Probabilistic Regression Trees (prtree). The description of these methods is presented in Chapters 2 and 3. The methods available in the `imputeTS` package are employed with default arguments, as described in Moritz and Bartz-Beielstein (2017). The set up used for the decision tree methods are as follows.

- For the CART method, the `rpart` function is employed with the following arguments: `minsplit = 6` (minimum observations on a leaf), `cp = 0.01` (complexity parameter), `usesurrogate = 0` (how surrogates are used), `maxdepth = 30` (maximum depth). It is worth mentioning that with `usesurrogate = 0` then if an observation has a missing value for the primary splitting rule, it is not considered in subsequent splits. The remaining parameters are set to their default values in `rpart`.

- For pruning the tree obtained from the CART method, the `prune` function is used with the `cp` parameter set to the smallest error value calculated from the cross-validation method conducted by `rpart`.

- In the case of probabilistic decision trees, the `prtree` function is employed with the default parameters, as described in Section 3.4.4.

- The covariates considered are X_{t-1}^{miss} and X_{t+1}^{miss}. This selection is based on the preliminary simulation presented in Section 5.2.1.

Calculating the sample and expected detrended variances, cross-covariance and cross-correlation is a computationally intensive task. Therefore, for this purpose, the `R` package `DCCA` (Prass and Pumi, 2020) will be utilized. This package implements the results presented in Prass and Pumi (2021) and the main functions are written in `FORTRAN` for efficiency. For all scenarios, the DFA, DCCA and ρ_{DCCA} are calculated considering overlapping windows with size $m \in \{3, 5, \cdots, 99, 101\}$ and $\nu = 0$, corresponding to a linear fit in each window.

5.2.1 Preliminary study on the decision trees algorithms covariates

In Neimaier and Prass (2023) a method considering standard decision trees was proposed to fill in missing data in ARMA time series and a Monte Carlo simulation study was performed to analyze the performance of the method. In that study, the lagged values $X_{t-h_1}^{\text{miss}}, \cdots, X_{t-1}^{\text{miss}}, X_{t+1}^{\text{miss}}, \cdots X_{t+h_2}^{\text{miss}}$ were used as covariates, with $\boldsymbol{h} = (h_1, h_2) = \{(1,0), (1,1), (2,0), (2,2), (5,0), (5,5)\}$. The results indicated that incorporating information both before and after the observation of interest yields more accurate results and that there is no significant difference between cases where $\boldsymbol{h} = (1,1)$ and $\boldsymbol{h} = (5,5)$. However, the study did not investigate if all lags are indeed important when $\boldsymbol{h} = (5,5)$. Also, Neimaier and Prass (2023) did not consider the fact that using $\boldsymbol{h} = (5,5)$ potentially decreases the size of the training sample since the study was carried out without the use of surrogate variables, which implies that any row with missing data in the covariates was ignored by the `rpart` algorithm.

The upcoming simulation study aims to assess the variable's importance in one of the scenarios considered in Neimaier and Prass (2023). We shall assume that the same pattern follows for the remaining scenarios so that the results of this simulation will be used as a reference for the remaining cases. The scenario selected is the ARMA model, which was the one that presented the worst results (globally) for all imputation methods considered.

For this simulation study, 1000 replications of time series $\{X_t\}_{t=1}^{n}$, with size $n = 2000$, are generated from an ARMA$(1,1)$ process with parameters $\phi = 0.7$ and $\theta = 0.4$. For each replication, time series $\{X_t^{\text{miss}}\}_{t=1}^{n}$ with missing data are created, varying the proportion of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The lagged values $X_{t-5}^{\text{miss}}, \cdots, X_{t-1}^{\text{miss}}, X_{t+1}^{\text{miss}}, \cdots X_{t+5}^{\text{miss}}$ are used as covariates, and the reconstruction is carried out using only the CART method and the procedure described in Section 3.5. The relative importance of each lag as a predictor variable is examined using the "variable importance" metric provided by the `rpart` implementation of CART in `R`.

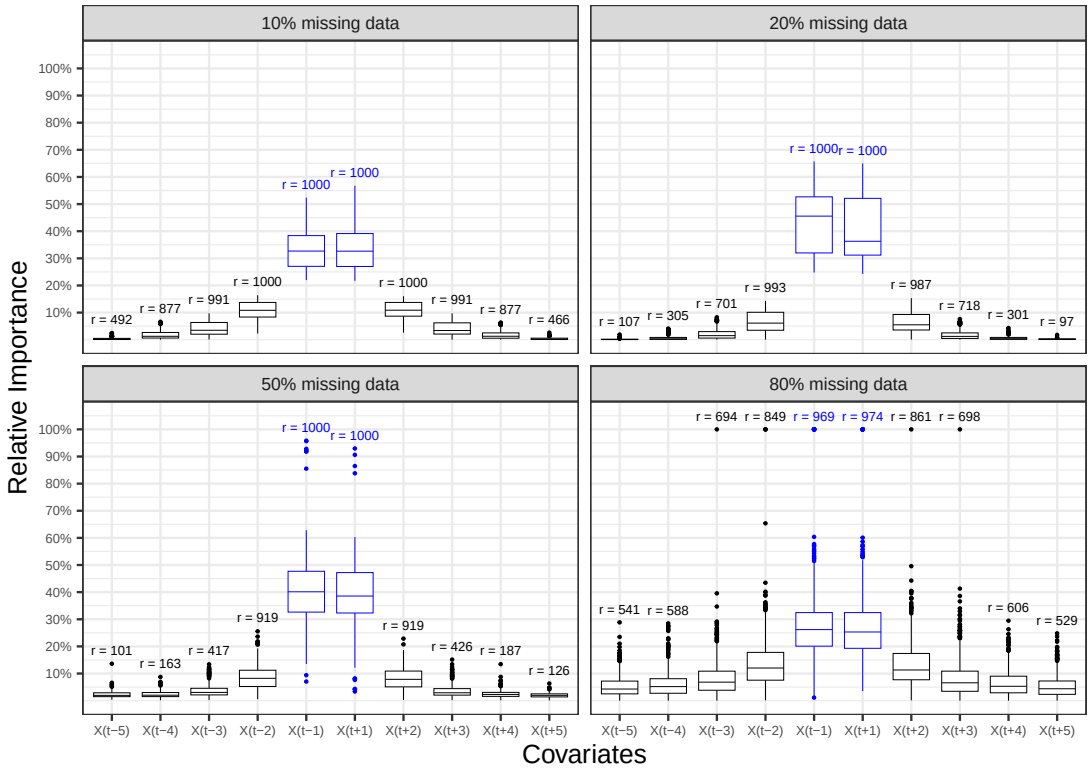

Figure 5.1: Boxplots of the relative importance of the covariates $X_{t-5}^{\text{miss}}, \cdots, X_{t-1}^{\text{miss}}, X_{t+1}^{\text{miss}}, \cdots X_{t+5}^{\text{miss}}$, based on 1000 Monte Carlo replicas, where time series with a proportion ρ of missing values, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$, were reconstructed using the CART decision tree algorithm.

Figure 5.1 presents boxplots illustrating the relative importance of the covariates $X_{t-5}^{\text{miss}}, \cdots, X_{t-1}^{\text{miss}}$, $X_{t+1}^{\text{miss}}, \cdots X_{t+5}^{\text{miss}}$, based on 1000 replications, for each $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The values r above the boxplots indicate the number of replications in which the corresponding variable was used in the main or surrogate tree. The covariates X_{t-1}^{miss} and X_{t+1}^{miss} are highlighted in blue. The first lag consistently stands out as the most important, and the relative importance decreases as the temporal distance between the covariate and the variable of interest increases. Based on these findings, we have decided to consider only one lag from the past and one from the future as predictors for subsequent analysis.

5.3 Results presentation

For each scenario under consideration, a Monte Carlo simulation study is conducted to assess the quality of missing data reconstruction and the finite sample performance of the $F^2_{k,DFA}(m)$, $k \in \{1,2\}$, $F_{DCCA}(m)$ and $\rho_{DCCA}(m)$ in the presence of missing data. Different proportions of missing values and imputation methods are considered and, for all scenarios, the results are summarized following the same order of presentation, as described in the sequel.

1. First, the simulation results considering the complete time series are reported. The results are presented in a figure showing the boxplots of the estimated values of $F^2_{k,\text{DFA}}(m)$, $k \in \{1,2\}$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$, for $m \in \{3,5,\cdots,99,101\}$, and a horizontal line corresponding to the theoretical expected values $\mathbb{E}(F^2_{k,DFA}(m))$ and $\mathbb{E}(F_{DCCA}(m))$, and the limit $\rho_{\mathcal{E}}(m)$ for which the coefficient $\rho_{DCCA}(m)$ converges to, as $n \longrightarrow \infty$. In these graphs, the color blue is used to emphasize the values of m that will be considered in the graphs corresponding to the imputed time series. These graphs are equivalent to those in Prass and Pumi (2021). They are useful for analyzing the bias, variability, and the decay of $F^2_{k,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$, as functions of m.

2. Second, for each time series $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ the simulation results regarding the imputation of missing values are reported. The figure that summarizes the simulation results consists of boxplots of the mean square prediction error (MSE), defined by

$$MSE = \frac{1}{\#(T_1^C)} \sum_{t \in T_1^c} (X_{k,t} - \hat{X}_{k,t})^2,$$

   for each one of the seven imputation methods applied, and each proportion of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. This part of the simulation aims to answer question Q1 and partially answer question Q5, for each scenario considered. Moreover, the comparison among different scenarios might provide a partial answer to question Q5.

3. Third, the simulation results regarding the DFA under the presence of missing data are reported. The figure presented shows boxplots of the $F^2_{k,\text{DFA}}(m)$ values obtained from the imputed time series and a horizontal line corresponding to the theoretical expected values $\mathbb{E}(F^2_{k,DFA}(m))$, under no missing data. For simplicity, the figure only shows the quantities corresponding to $m \in \{3, 27, 81, 101\}$, for each proportion of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and each one of the seven imputation methods. This part of the simulation aims to partially answer question Q5 and the comparison with the first reported results might give a partial answer to questions Q2 and Q3, for each scenario considered. Moreover, the comparison among different scenarios partially answers question Q4.

4. Fourth, the simulation results regarding the DCCA under the presence of missing data are reported. Analogously to the DFA, the results are summarized in a figure showing boxplots of the $F_{\text{DCCA}}(m)$ values obtained from the imputed time series and a horizontal line corresponding to the theoretical expected values $\mathbb{E}(F_{DCCA}(m))$, under no missing data. For simplicity, the figure only shows the quantities corresponding to $m \in \{3, 27, 81, 101\}$, for each proportion of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and each one of the seven imputation methods. These results complement the previous analysis.

5. Finally, the simulation results regarding the coefficient ρ_{DCCA} under the presence of missing data are reported. The figure presented shows the boxplots of the $\rho_{DCCA}(m)$ values obtained from the imputed time series and a horizontal line corresponding to the theoretical limit $\rho_{\mathcal{E}}(m)$ for which the coefficient $\rho_{DCCA}(m)$ converges to, under no missing data, as $n \longrightarrow \infty$. For simplicity, the figure only shows the quantities corresponding to $m \in \{3, 27, 81, 101\}$, for each proportion of missing values $\rho \in \{0.1, 0.2, 0.5, 0.8\}$ and each one of the seven imputation methods. Combining the analysis of these results with the previous ones should answer the five questions raised.

As the scenarios are sequentially represented, any of the marginal analyses that have already been presented in a previous scenario will be omitted, and the location with its interpretations will be referenced. Following the guideline just provided, in the sequel we outline the findings for each distinct scenario. A final discussion gathering the results, and aiming to clarify questions Q4 and Q5, is presented at the end of this chapter.

5.3.1 Scenario 1: uncorrelated i.i.d. processes

In this scenario, $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are i.i.d. $\mathcal{N}(0,1)$ sequences, independent from each other. The corresponding autocovariance and cross-covariance matrices are given by

$$\Gamma_1 = \Gamma_2 = I_{m+1}, \qquad \text{and} \qquad \Gamma_{1,2} = 0_{m+1,m+1}.$$

and

$$\mathbb{E}[F^2_{k,\text{DFA}}(m)] = \frac{1}{m}\operatorname{tr}(K_{m+1}) = \frac{m^2+2m-3}{15m} \sim \frac{m}{15}, \quad \text{as } m \to \infty, \quad k \in \{1,2\}.$$

Since the two processes are identically distributed, in what follows, the marginal results shall be presented only for $\{X_{1,t}\}_{t=1}^{n}$.

Figure 5.2 shows that the estimates of $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ are close to their expected values, especially when m is small. The values of $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ are all close to zero which, in this scenario, is the value of the theoretical counterpart $\rho_{\mathcal{E}}(m)$, for all m. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

From Figure 5.3 it can be observed that the average-based methods (M, rpart, and prtree) performed better than the other methods in filling the missing values of $\{X_{1,t}\}_{t=1}^{n}$. This outcome is reasonable from a theoretical perspective, given that this process is a sequence of i.i.d. random variables with a standard Gaussian distribution and, in terms of mean squared error, the sample mean is the best linear predictor. It is also noticeable that for these methods, the larger the ρ, the lower the variability of the MSE. Although it may seem counter-intuitive, this occurs because the predictions of missing values are calculated using $\lfloor n(1-\rho) \rfloor$ observations, while the MSE is calculated using $\lfloor n\rho \rfloor$ observations. As the proportion of missing values ρ rises, the size of the sample available to predict the missing values decreases without compromising the imputation quality due to the simplicity of the underlying process. Simultaneously, for each replication, the size of the sample used to calculate the MSE values increases, resulting in MSE estimates with less variability.

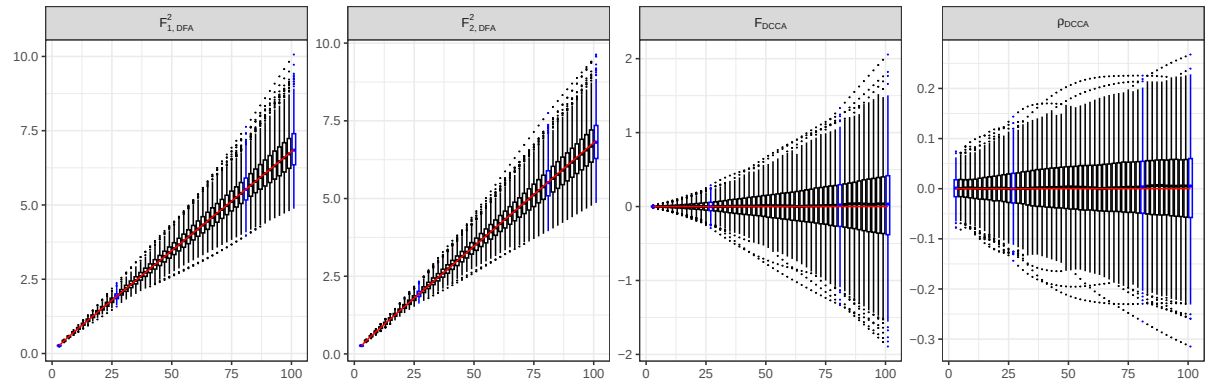

Figure 5.2: Scenario 1: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

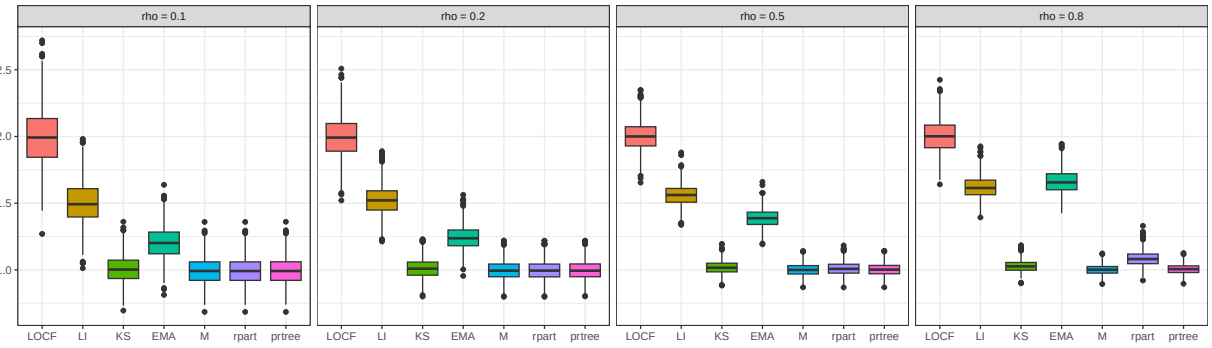

Figure 5.3: Scenario 1: Boxplots of the imputation MSE values for $\{X_{1,t}\}_{t=1}^{2000}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

Figure 5.4 illustrates that when $m = 3$, $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$ is always underestimated, with a significant reduction in the performance of all methods as ρ increases. Also, there is almost no difference among the methods but LOCF, followed by the average-based methods (except EMA) and by the KS method, resulted in $F^2_{1,\text{DFA}}(m)$ values slightly closer to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$. For $m \in \{27, 81, 101\}$, the LOCF, LI, and EMA methods always overestimate $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$, while the other methods underestimate this value. This becomes more evident as the proportion of missing values ρ increases. For $\rho \in \{0.1, 0.2\}$, the estimates seem reasonable, but for higher proportions of missing values either the estimated values are much higher than the theoretical ones or they concentrate very close to zero. Despite this, the KS and the average-based methods consistently produced values closer to the theoretical expected value. Figures 5.5 and 5.6 show that all imputation methods had similar performance for all values of m and ρ, except for $\rho \in \{0.5, 0.8\}$, in which cases the LOCF, LI and EMA methods produced values of $F_{\text{DCCA}}(m)$ with higher variability. In all cases, $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ are close to 0, which is the theoretical target.

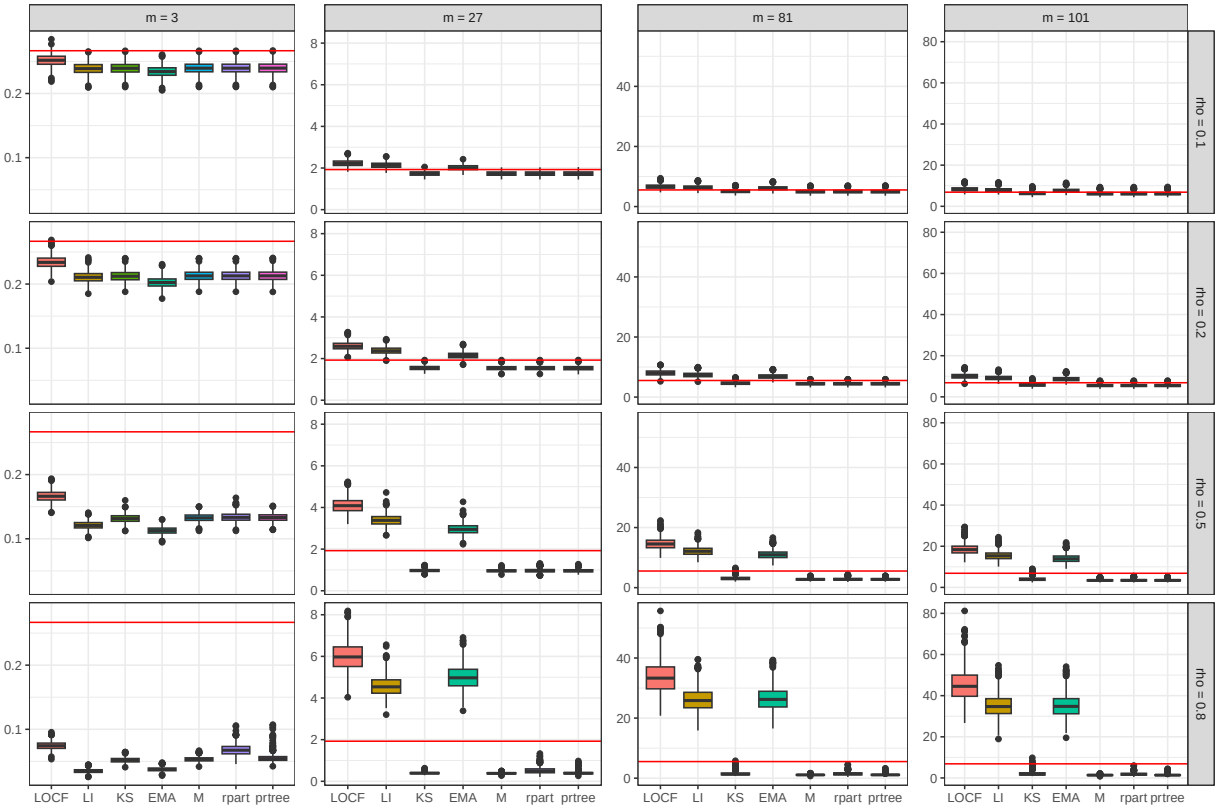

Figure 5.4: Scenario 1: Boxplots of $F^2_{1,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

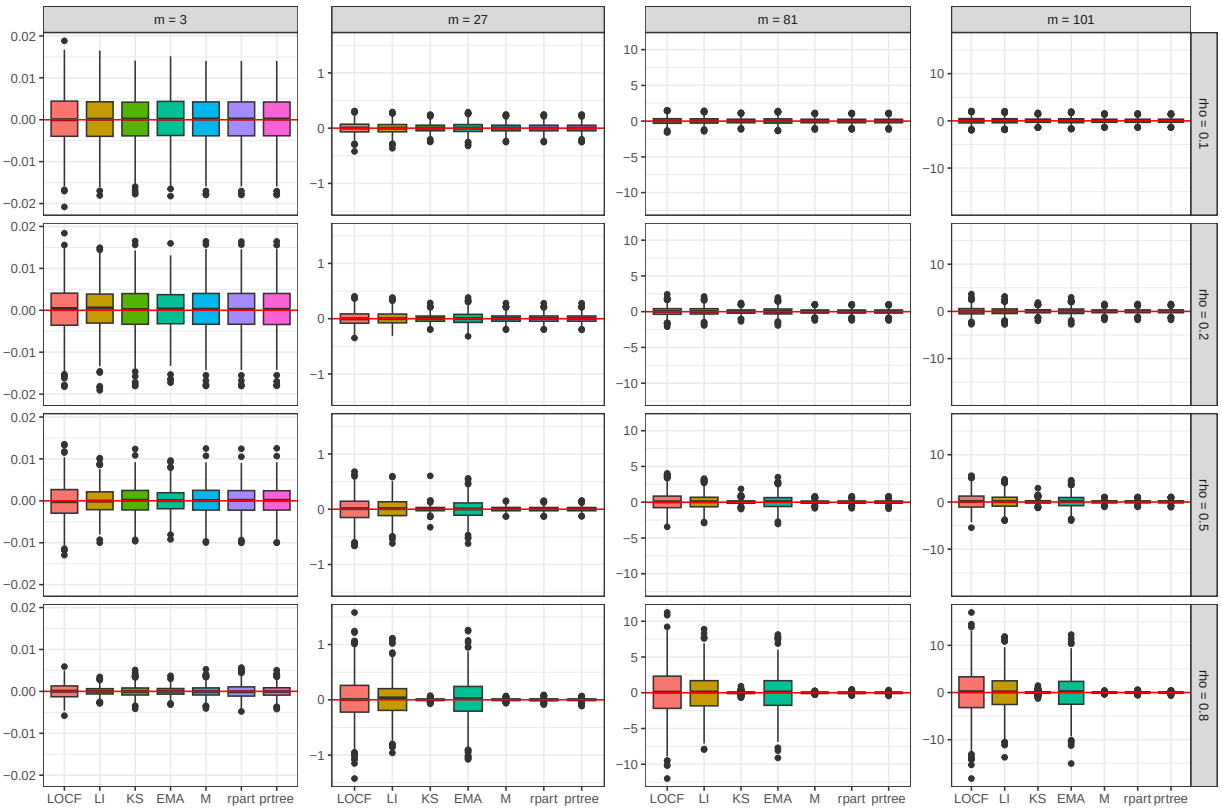

Figure 5.5: Scenario 1: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

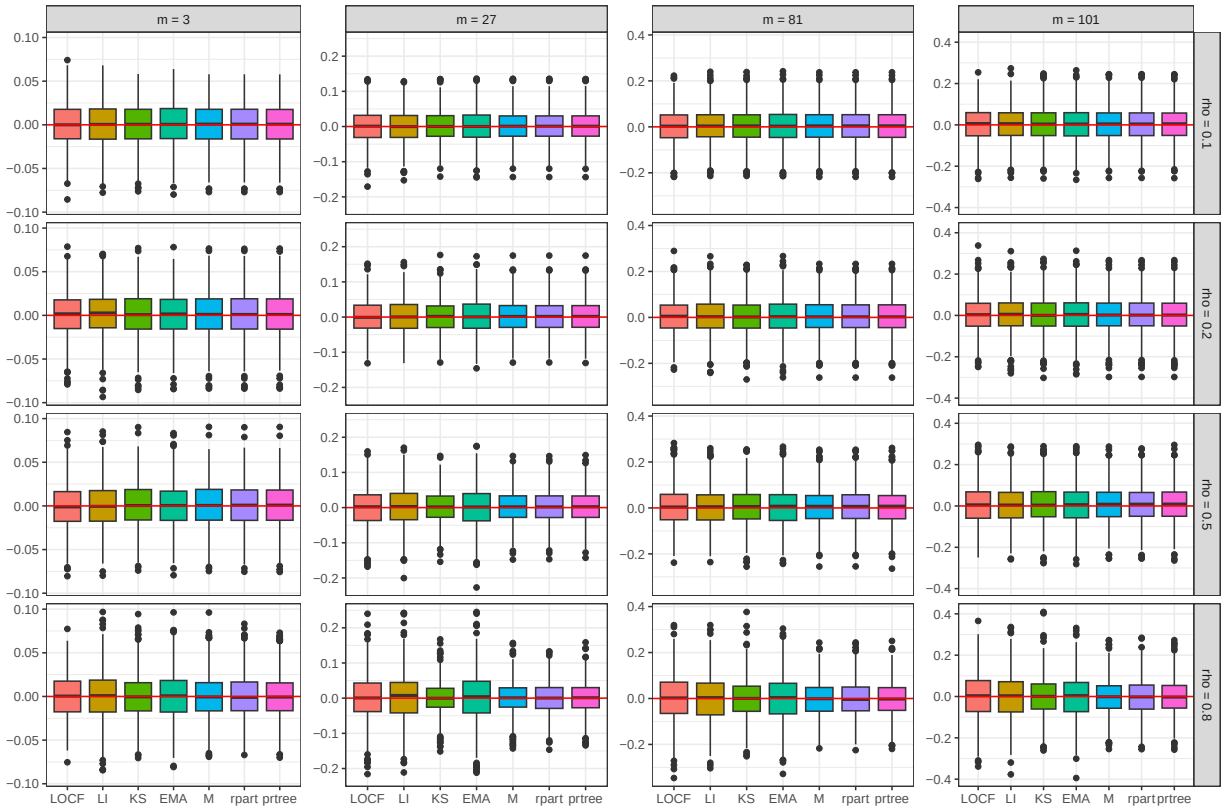

Figure 5.6: Scenario 1: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

From Figures 5.4 - 5.6 it can also be observed that the same pattern shown in Figure 5.2 is presented when estimates of $F^2_{k,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ are obtained after imputation, that is, the variability increases with m. Moreover, as shown in more detail in Figure 5.7, for the imputed time series, the values of $F^2_{k,\text{DFA}}(m)$ also increase linearly with m, independently of ρ. The results reported in Figure 5.7 correspond to the estimated values of $F^2_{k,\text{DFA}}(m)$ obtained when the missing values were imputed using the LOCF method. The behavior observed when other imputation methods were used is analogous.

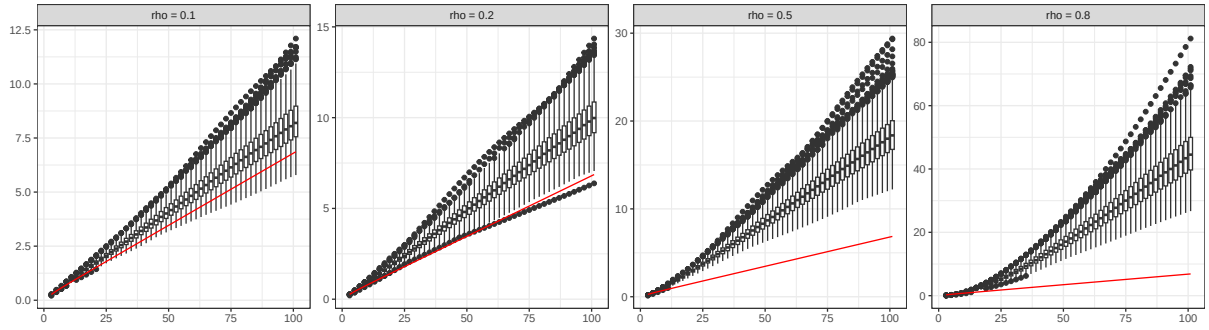

Figure 5.7: Scenario 1: Boxplots of $F^2_{1,\text{DFA}}(m)$, $m \in \{3, 5, \cdots, 99, 101\}$, based on $r = 1000$ replications, considering the LOCF imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

For the current scenario, the estimations of the DFA and DCCA functions with complete time series had results closer to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods. Also, the estimates of $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ were closer to the corresponding expected values when the time series were reconstructed using those methods. The time series reconstructed using the average-based methods were the ones that led to the best results for the estimates of $F_{\text{DCCA}}(m)$ across different values of m and ρ. Regarding $\rho_{\text{DCCA}}(m)$, almost no differences were observed among the imputation methods. Therefore, these observations suggest that for uncorrelated processes, the methods that excel in missing data imputation also provide more accurate estimates for the DFA and DCCA functions.

5.3.2 Scenario 2: uncorrelated processes with autocorrelation

In this scenario the time series $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = 0.6X_{t-1} + \varepsilon_{1,t}, \qquad \text{and} \qquad X_{2,t} = \varepsilon_{2,t} + 0.6\varepsilon_{2,t-1}, \quad t \in \mathbb{Z}, \tag{5.1}$$

where $\{\varepsilon_{k,t}\}_{t\in\mathbb{Z}}$, $k \in \{1,2\}$, are sequences of i.i.d. $\mathcal{N}(0,1)$ random variables and $\varepsilon_{1,r}$ and $\varepsilon_{2,s}$ are independent, for all $r, s \in \mathbb{Z}$. All time series are generated considering the recurrence (5.1) with burn-in size equal to 10. Hence $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ correspond to time series from an AR(1) and an MA(1) process, respectively, which are independent from each other. It follows that the (i,j)-th term in the corresponding autocovariance matrices and the cross-covariance matrix are given by

$$[\Gamma_1]_{i,j} = \frac{0.6^{|i-j|}}{0.64}, \quad [\Gamma_2]_{i,j} = 1.36I(i=j) + 0.6I(|i-j|=1) \quad \text{and} \quad \Gamma_{1,2} = 0_{m+1,m+1}.$$

Moreover, from Prass and Pumi (2021),

$$\mathbb{E}[F^2_{1,\text{DFA}}(m)] = \frac{m^3 + O(m^2)}{2.4(m^2+3m+2)} \sim \frac{5m}{12} \quad \text{and} \quad \mathbb{E}[F^2_{2,\text{DFA}}(m)] = \frac{2.56m^2 + O(m)}{15m} \sim \frac{64m}{375}.$$

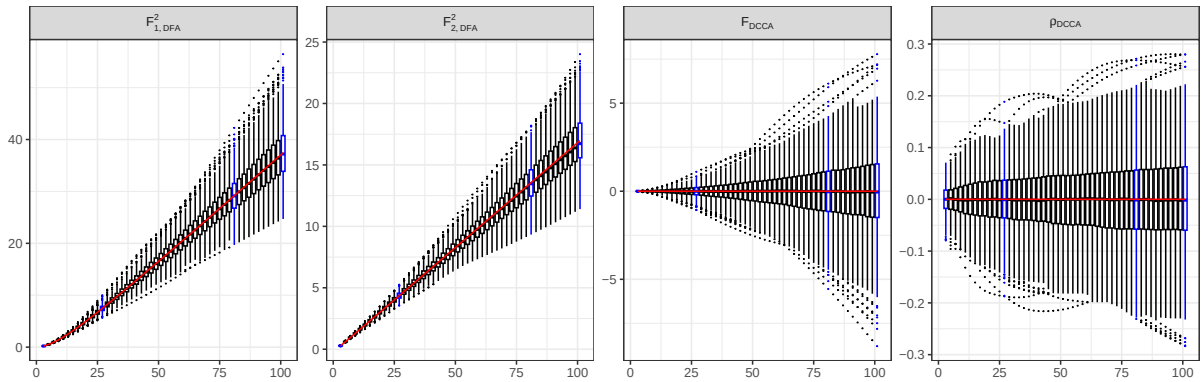

Figure 5.8: Scenario 2: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

Figure 5.8 shows that the estimates of $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ are close to their expected values, especially when m is small. The values of $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ are all close to zero which, in this scenario, is the value of the theoretical counterpart $\rho_{\mathcal{E}}(m)$, for all m. In all cases, the variability increases with m and it is higher than in scenario 1. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

From Figure 5.9 (top row) it can be observed that, for $\rho \in \{0.1, 0.2, 0.5\}$, the methods that best filled the missing values of $\{X_{1,t}\}_{t=1}^n$ were LI, KS, and EMA. It is coherent that these methods performed better under these circumstances, given that, in the current scenario, $\{X_{1,t}\}_{t=1}^n$ is a sample from a stationary Gaussian AR(1) process. As shown in Table 3.1, in this context, the surrounding observations are the ones that contribute the most in predicting X_t, which explains the good performance of LI and EMA. Also, since KS is a likelihood-based method and the underlying distribution is correctly specified, it is expected that this method will be among the best ones. Notably, for $\rho = 0.8$, the methods with the best performance were rpart (smallest median), KS, M, and prtree. Since, in this case, the number of missing observations is very high, the non-missing observations are usually far apart (in terms of time). This could explain why LI and EMA are no longer among the imputation methods with the best performance. Since LOCF only uses the last known observation, its performance is expected to decrease as the proportion of missing values increases. While other methods had a decrease in performance, M maintains about the same median MSE value, regardless of ρ, with a decrease in the variability as ρ increases. Given that rpart and prtree are conditional mean-based methods, it makes sense that their performance with a high proportion of missing values is similar and usually slightly better than the global mean.

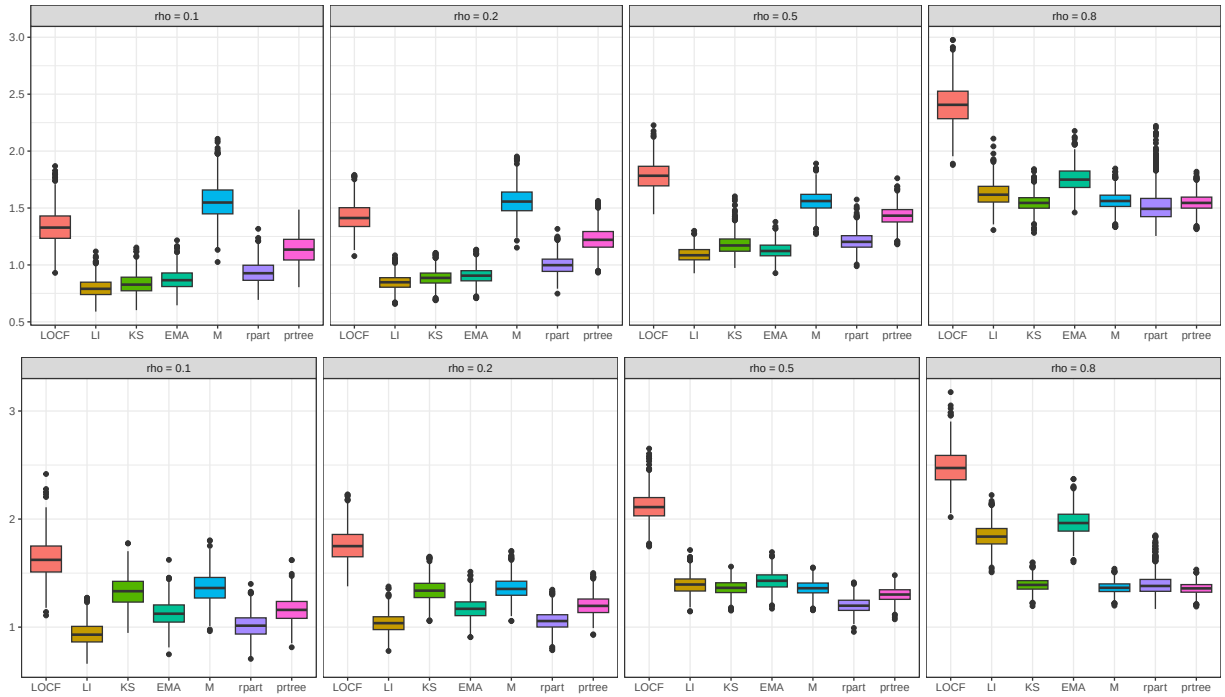

Figure 5.9: Scenario 2: Boxplots of the imputation MSE value for $\{X_{1,t}\}_{t=1}^{2000}$ (top row) and $\{X_{2,t}\}_{t=1}^{2000}$ (bottom row), based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

From Figure 5.9 (bottom row), it is observed that for the methods KS, M, rpart and prtree the MSE values are always less than 1.5, except for a few outliers. Also, these are the methods with smaller variability, for all values of ρ. For $\rho \in \{0.1, 0.2\}$, the methods that performed best in filling the missing values of $\{X_{2,t}\}_{t=1}^{n}$ were the LI and rpart. EMA and prtree performed similarly when compared to each other, but their MSE values were generally higher than those corresponding to the best-performing methods. For $\rho \in \{0.5, 0.8\}$, the best methods were rpart, prtree, M and KS. It makes sense that the average-based methods excels in the imputation of an MA(1) process, given the MA(1) dependence structure as shown in Table 3.1. Therefore, the relevance of the more temporally distant observations decays rapidly, making the local mean a much more useful information for the prediction of missing values. This is even more noticeable for large proportions of missing data.

For all values of ρ, LOCF was the method with the worst performance. Upon comparing the top and bottom rows in Figure 5.9 one observes that, for small values of ρ, the best-performing methods for the AR(1) and MA(1) processes are not the same. While for the AR(1) process the autocorrelation function is non-zero for all $h > 0$, for the MA(1) process the only non-zero lags are 0 and 1. Hence, it is expected that methods that consider a small neighborhood of X_t to predict the missing value will perform better in the AR case given that, in the absence of X_{t-1} and/or X_{t+1} the other values in the neighborhood still have relatively high impact in the prediction.

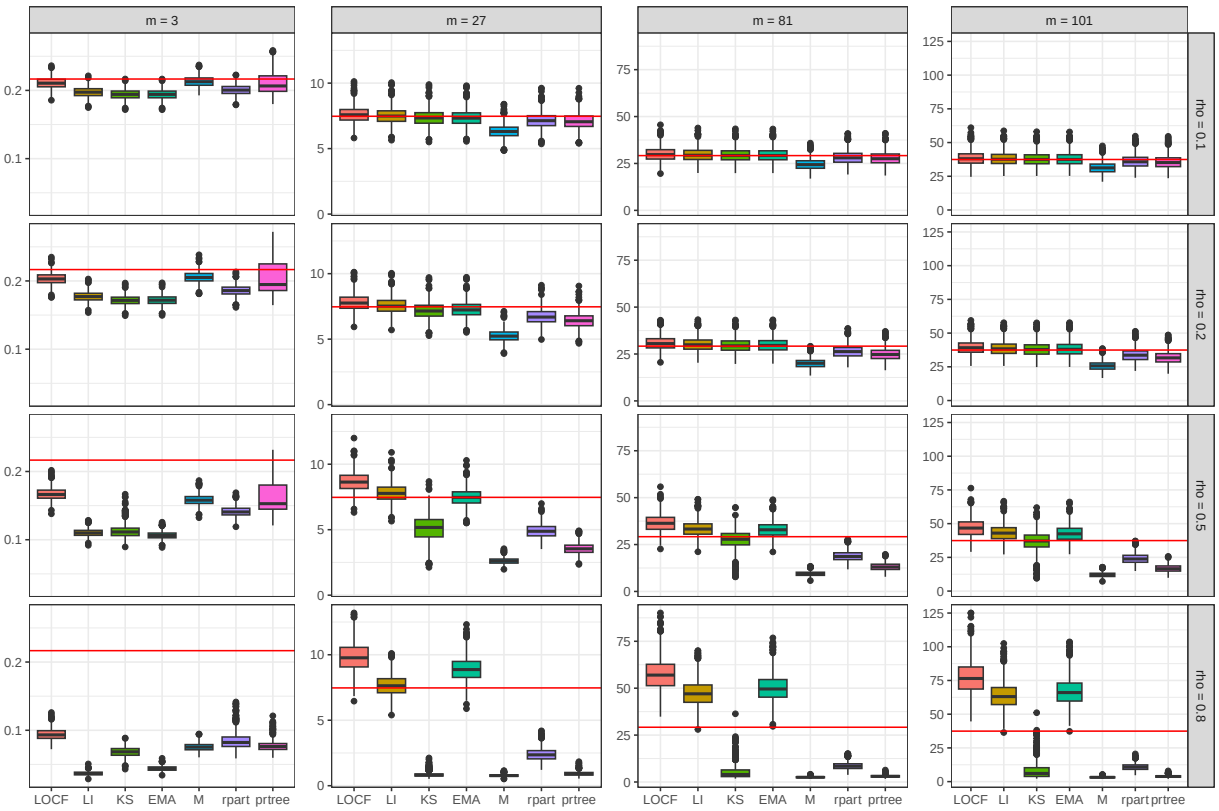

Figure 5.10: Scenario 2: Boxplots of $F^2_{1,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

In Figure 5.10, it can be observed that, for $\rho \in \{0.1, 0.2\}$ all imputation methods yielded satisfactory results in the sense that the calculated values $F^2_{1,\text{DFA}}(m)$ are always close to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$, with very low variability. The average-based methods (excluding M) and LOCF exhibiting the best performance, despite being the worst performing methods in terms of imputation for these values of ρ. For $m = 3$ and $\rho \in \{0.5, 0.8\}$, all methods underestimated $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$ and the closest values were obtained when LOCF and M were used to imput the missing values. For $m \in \{27, 81, 101\}$ and $\rho \in \{0.5, 0.8\}$, the methods LOCF, LI, and EMA overestimated, while the methods KS, M, rpart, and prtree underestimated $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

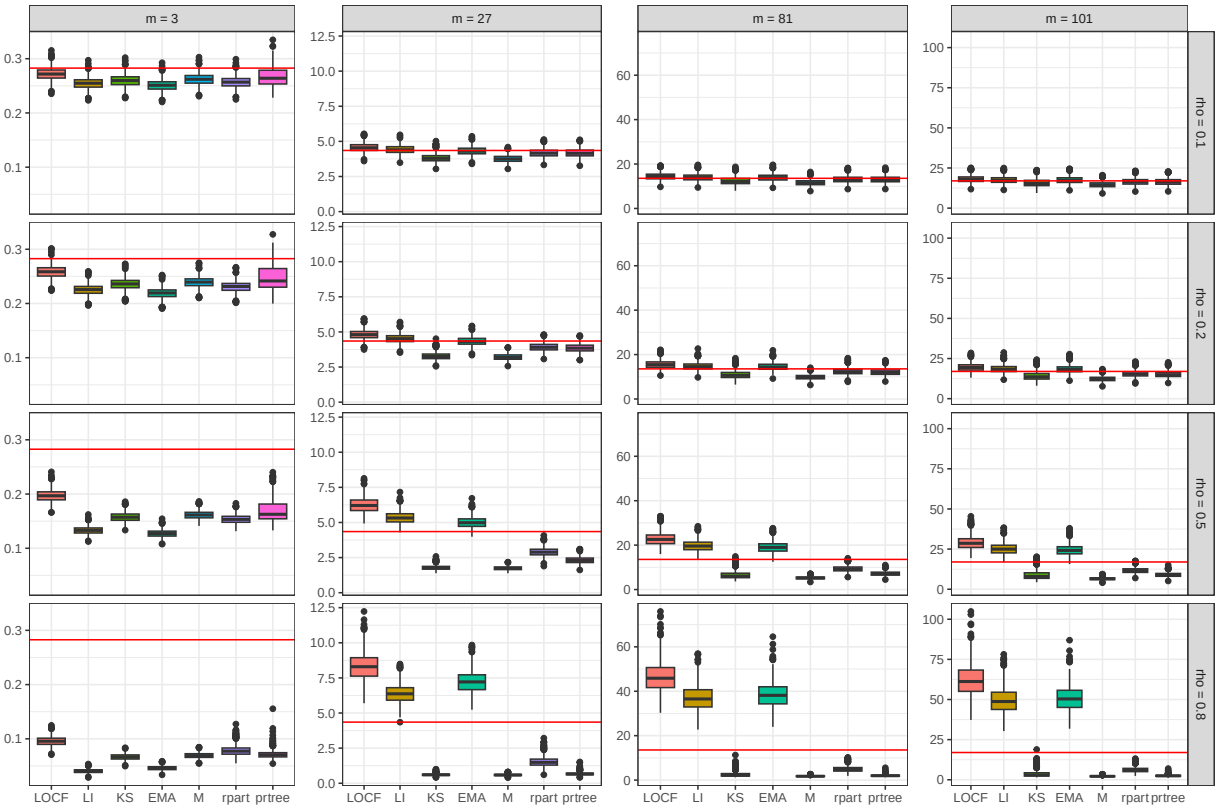

Figure 5.11: Scenario 2: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

Figure 5.11 illustrates that, regarding $F^2_{2,\text{DFA}}(m)$ for $\rho \in \{0.1, 0.2\}$, all methods had satisfactory results. For $m = 3$ and $\rho \in \{0.5, 0.8\}$, all methods underestimated $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$, although LOCF had the best performances. For $m \in \{27, 81, 101\}$ and $\rho \in \{0.5, 0.8\}$, the methods LOCF, LI, and EMA overestimated and the methods KS, M, rpart, and prtree underestimated $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$. Figures 5.12 and 5.13 show that the methods have similar performance (in terms of median MSE) for all values of m and ρ for the estimates of $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$. The estimates using the methods KS, M, rpart, and prtree methods exhibited less variation.

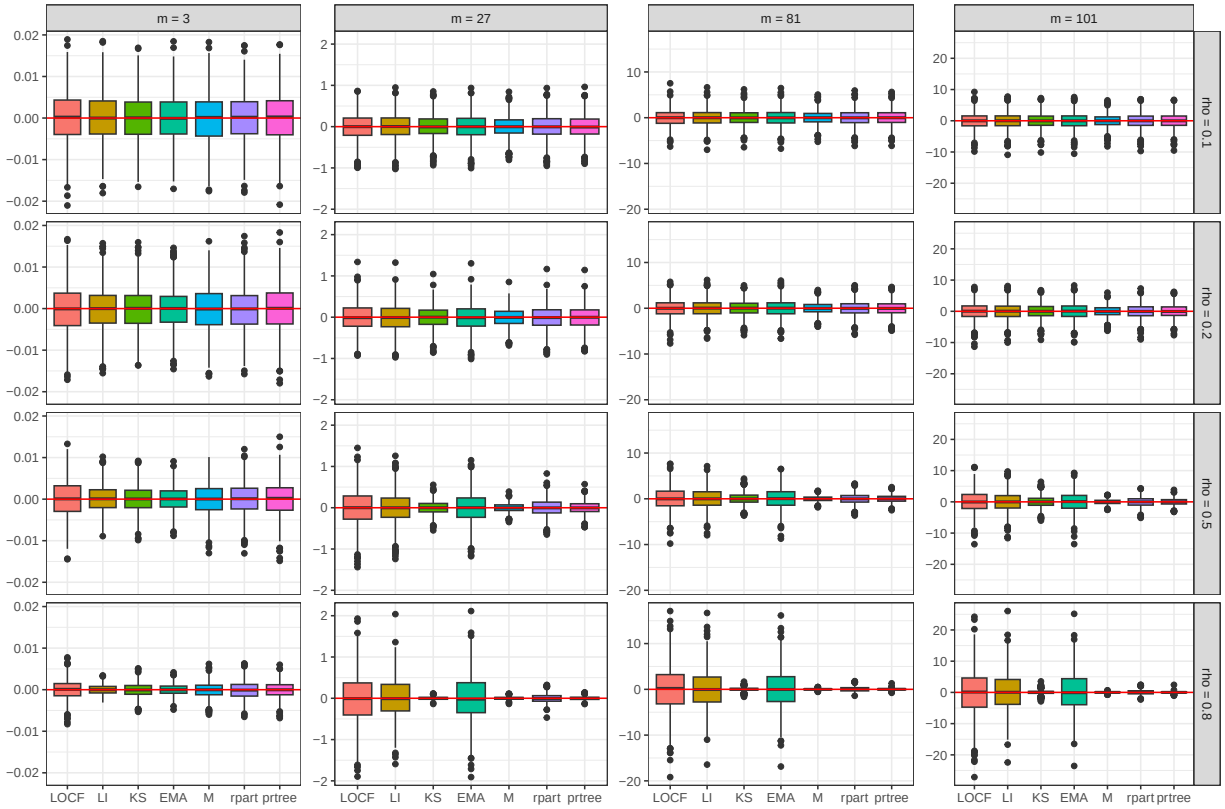

Figure 5.12: Scenario 2: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

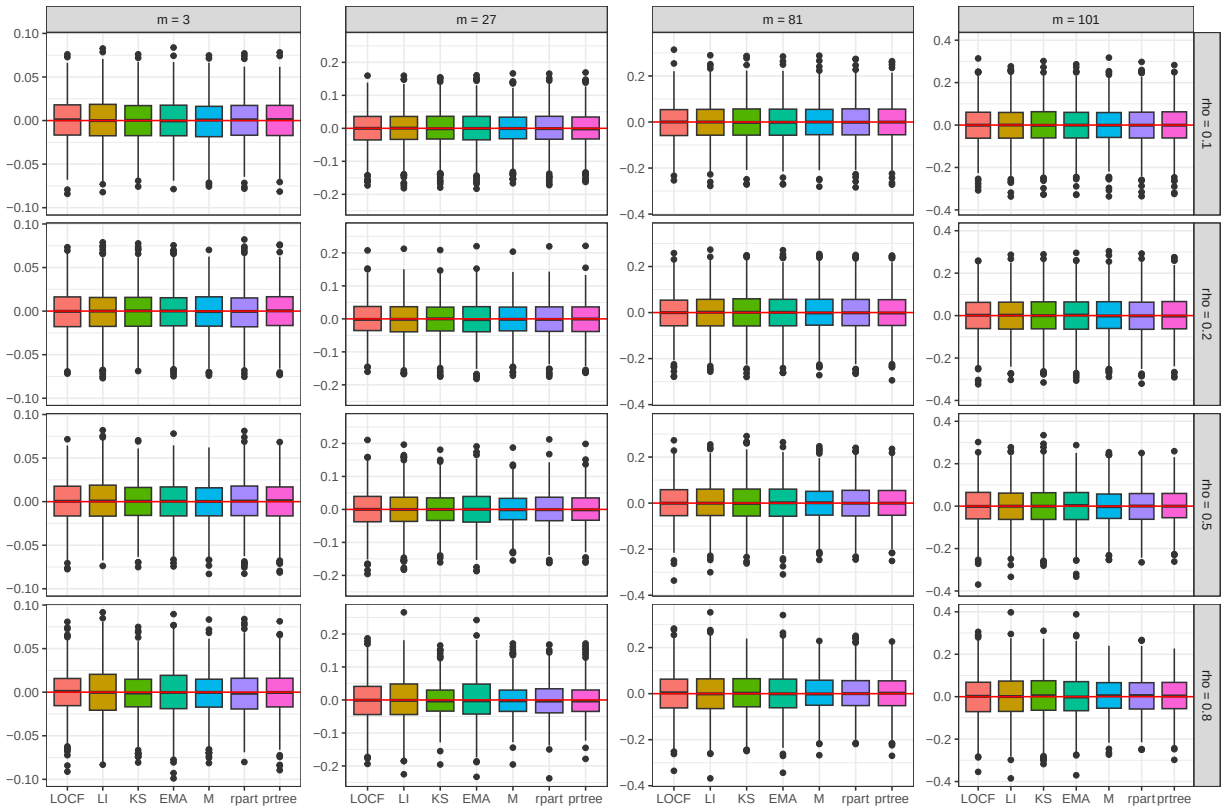

Figure 5.13: Scenario 2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

For this scenario, the estimates of the DFA and DCCA functions with complete time series had results closer to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, for lower proportions of missing data, LI, KS and EMA had superior performances for the AR(1) process, while for the MA(1) process, the best methods were LI and rpart. For higher proportions of missing data, the methods with the best performances were rpart, prtree, M and KS. Also, the estimates of $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$ were closer to the corresponding expected values when the time series were reconstructed using rpart, prtree, EMA and LOCF for $\rho \in \{0.1, 0.2\}$. However, for $\rho \in \{0.5, 0.8\}$, no method stood out for the quality of predictions. For $m = 3$, all methods underestimated $\mathbb{E}[F^2_{k,\text{DFA}}(m)]$ and for $m \in \{27, 81, 101\}$, the methods LOCF, LI and EMA overestimated and KS, M, rpart e prtree underestimated $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$. For $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$, all methods exhibited good performances in terms of median and the time series reconstructed using KS, rpart, prtree and M led to estimates with less variability. Therefore, these observations suggest that for processes with autocorrelation but mutually uncorrelated, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

5.3.3 Scenario 3.1: bivariate Gaussian white noise process with 0.5 correlation

In this scenario the time series $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ are samples from a bivariate Gaussian process $\{(X_{1,t}, X_{2,t})\}_{t\in\mathbb{Z}}$ where $\mathbb{E}[X_{k,t}] = 0$, $k \in \{1, 2\}$, and

$$\text{Cov}[X_{k_1,t}, X_{k_2,t}] = I(k_1 = k_2) + 0.5I(k_1 \neq k_2), \quad k_1, k_2 \in \{1, 2\}.$$

The corresponding autocovariance matrices and cross-covariance matrix are given by

$$\Gamma_1 = \Gamma_2 = I_{m+1}, \qquad \text{and} \qquad \Gamma_{1,2} = 0.5I_{m+1},$$

and $\rho_{\text{DCCA}}(m) \xrightarrow{P} 0.5$, as $n \to \infty$, for all $m > 0$. Since the marginal processes $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ have a standard Gaussian distribution, the quality of their imputation and the estimates of $F^2_{k,\text{DFA}}$, $k \in \{1, 2\}$, have already been discussed in Scenario 1 (average-based methods provided the best results) and thus will be omitted. Therefore, only results related to $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ will be presented.

Figure 5.14 shows that the estimates of $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ are close to their expected values, especially when m is small. The values of $\rho_{\text{DCCA}}(m)$ are all close to 0.5 which, in this scenario, is the value of its theoretical counterpart, for all m. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9). In Figure 5.15, it is possible to notice that $\mathbb{E}[F_{\text{DCCA}}(m)]$ was consistently underestimated across all scenarios. For $m = 3$, the methods M, KS, rpart, and prtree had superior performance, while for $m \in \{27, 81, 101\}$, LOCF, LI, and EMA methods performed significantly better, especially for $\rho \in \{0.5, 0.8\}$. In 5.16, it is possible to notice that regarding $\rho_{\text{DCCA}}(m)$, the methods KS, M, rpart had a very similar MSE distribution and consistently outperformed the others for the estimates of $\rho_{\mathcal{E}}(m)$. As the proportion of missing values increased, all methods uniformly degraded.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased.

Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods for filling missing values for $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ and estimating the functions $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$. The time series reconstructed using LOCF, LI, and EMA had the best results for the estimates of $F_{\text{DCCA}}(m)$ and the average-based methods outperformed other methods for the estimates of $\rho_{\text{DCCA}}(m)$ across different values of m and ρ. Therefore, these observations suggest that for bivariate Gaussian processes, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

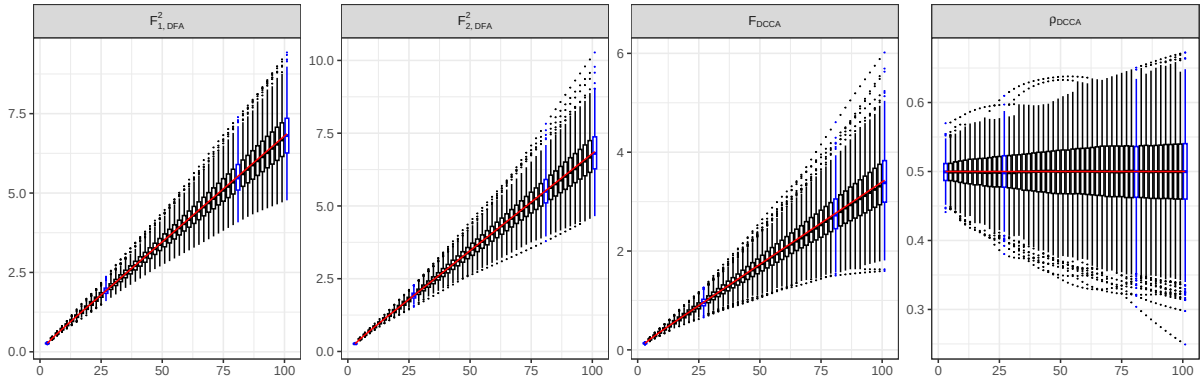

Figure 5.14: Scenario 3.1: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

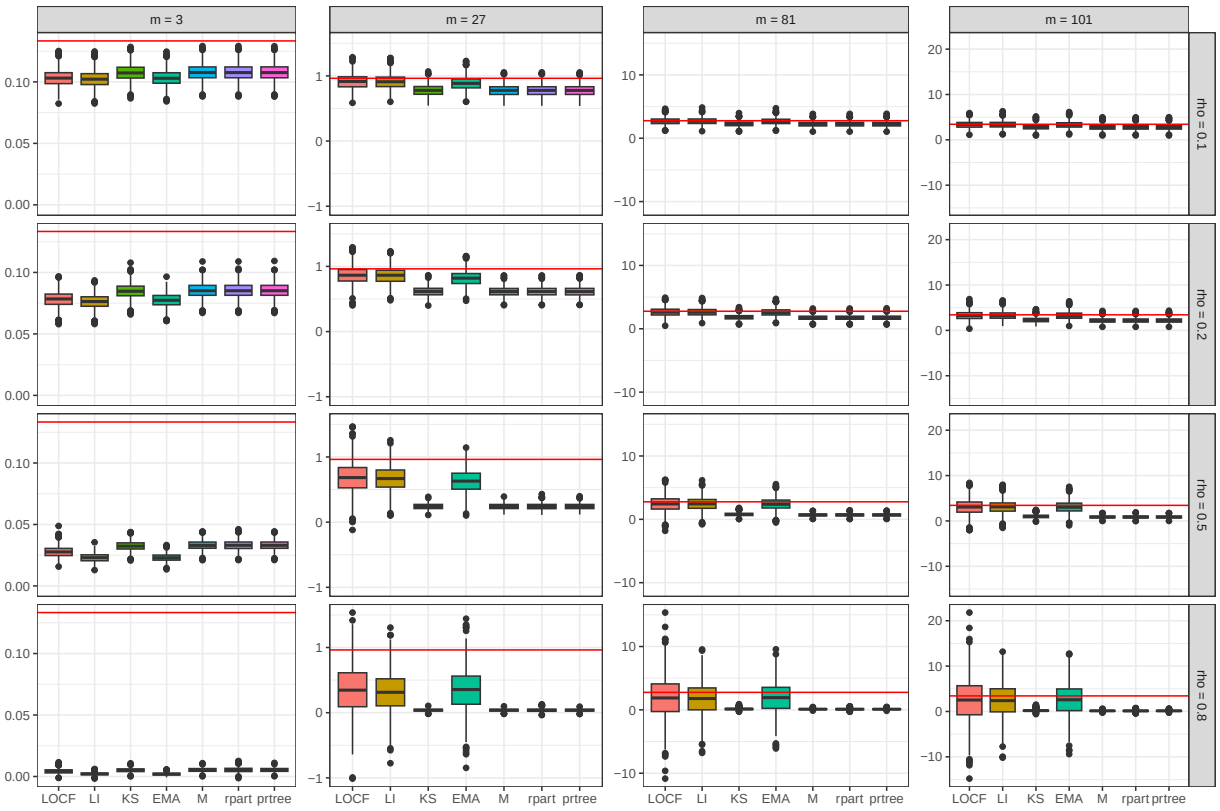

Figure 5.15: Scenario 3.1: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

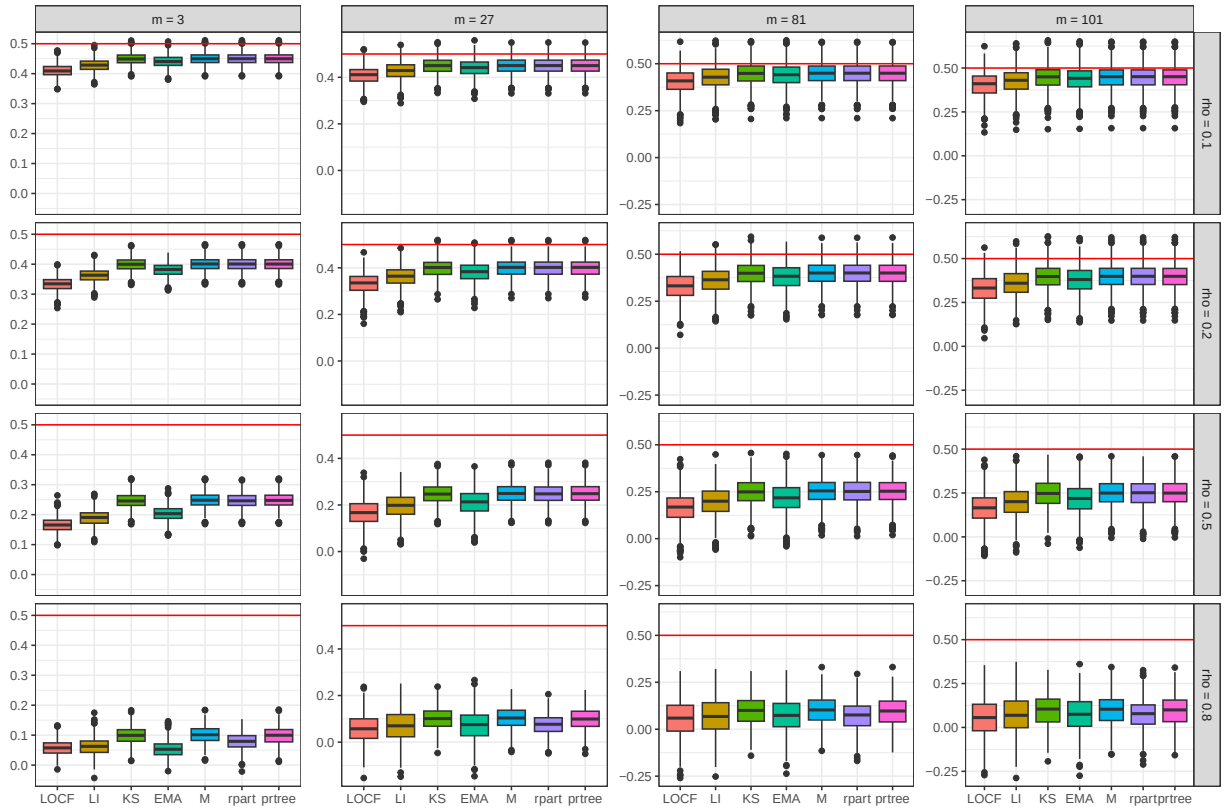

Figure 5.16: Scenario 3.1: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

5.3.4 Scenario 3.2: bivariate Gaussian white noise process with 0.8 correlation

In this scenario the time series $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are samples from a bivariate Gaussian process $\{(X_{1,t}, X_{2,t})\}_{t\in\mathbb{Z}}$ where $\mathbb{E}[X_{k,t}] = 0$, $k \in \{1, 2\}$, and

$$\text{Cov}[X_{k_1,t}, X_{k_2,t}] = I(k_1 = k_2) + 0.8I(k_1 \neq k_2), \quad k_1, k_2 \in \{1, 2\}.$$

This scenario is analogous to Scenario 3.1, only with a higher correlation. The corresponding autocovariance matrices and cross-covariance matrix are given by

$$\Gamma_1 = \Gamma_2 = I_{m+1}, \qquad \text{and} \qquad \Gamma_{1,2} = 0.8I_{m+1},$$

and $\rho_{\text{DCCA}}(m) \xrightarrow{P} 0.8$, as $n \to \infty$, for all $m > 0$. Since the marginal processes $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ have a standard Gaussian distribution, the quality of their imputation and the estimates of $F^2_{k,\text{DFA}}$ $k \in \{1, 2\}$ have already been discussed in Scenario 1 (average-based methods had the best results) and thus will be omitted. Therefore, only results related to $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ will be presented.

Figure 5.17 shows that the estimates of $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ are close to their expected values, especially when m is small. The values of $\rho_{\text{DCCA}}(m)$ are all close to 0.8 which, in this scenario, is the value of its theoretical counterpart, for all m. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

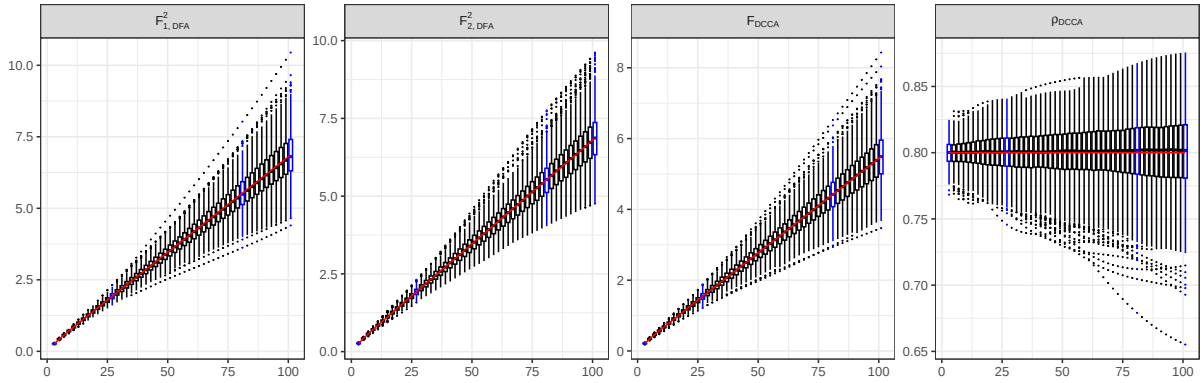

Figure 5.17: Scenario 3.2: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

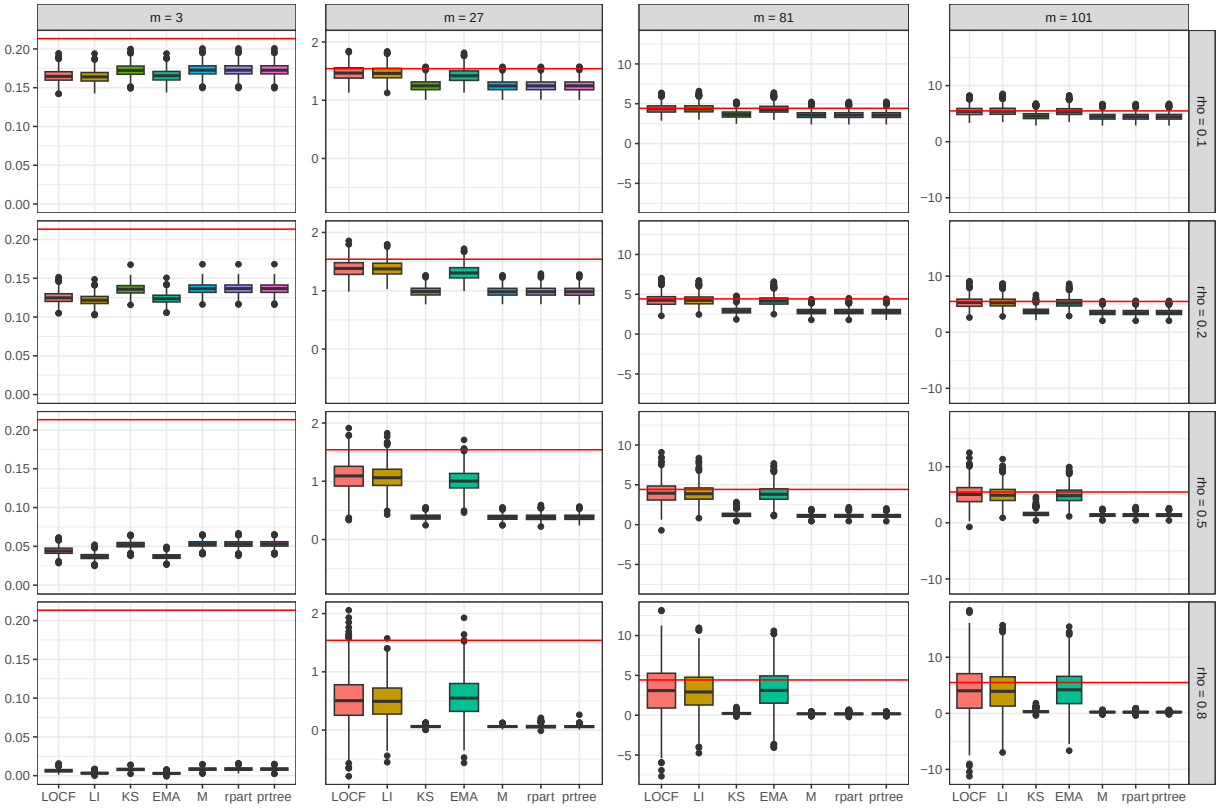

Figure 5.18: Scenario 3.2: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

In Figure 5.18, it is possible to notice that $\mathbb{E}[F_{\text{DCCA}}(m)]$ was consistently underestimated across all scenarios. For $m = 3$, the methods M, KS, rpart, and prtree had superior performance, while for $m \in \{27, 81, 101\}$, LOCF, LI, and EMA methods performed significantly better, especially for $\rho \in \{0.5, 0.8\}$. From Figure 5.19, it is possible to notice that regarding the $\rho_{\text{DCCA}}(m)$ function, the methods KS, M, rpart had a very similar MSE distribution and consistently outperformed the others

for the estimates of $\rho_{\mathcal{E}}(m)$. As the proportion of missing values increased, all methods uniformly degraded.

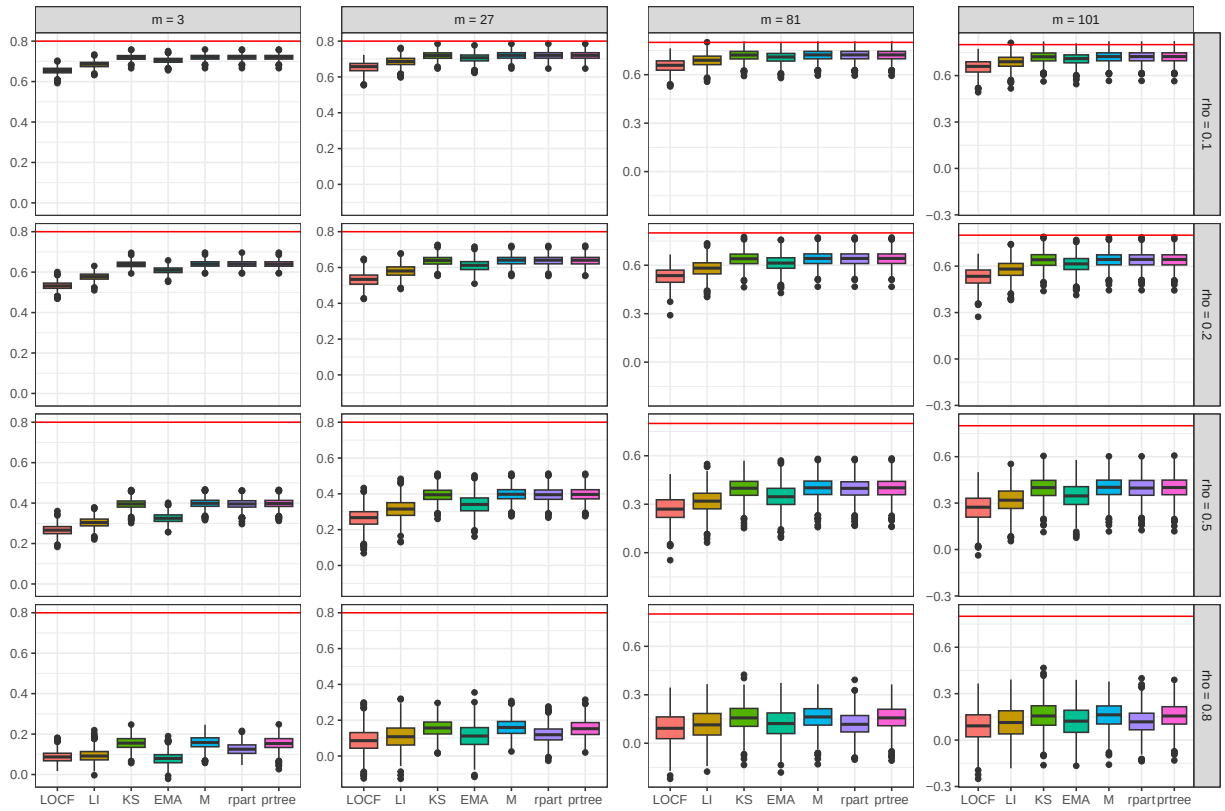

Figure 5.19: Scenario 3.2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods for filling missing values for $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ and estimating the functions $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$. The time series reconstructed using LOCF, LI, and EMA had the best results in estimating $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ across different values of m and ρ. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

5.3.5 Scenario 4.1: bivariate white noise with a signal plus noise (low variance) structure

In this scenario the time series $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = \varepsilon_{1,t}, \qquad \text{and} \qquad X_{2,t} = 3 + 2X_{1,t} + \varepsilon_{2,t}, \quad t \in \mathbb{Z}, \tag{5.2}$$

where $\{\varepsilon_{k,t}\}_{t\in\mathbb{Z}}$, $k \in \{1, 2\}$, are sequences of i.i.d. $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$ random variables, respectively,

and $\varepsilon_{1,r}$ and $\varepsilon_{2,s}$ are independent, for all $r, s \in \mathbb{Z}$. All time series are generated considering the recurrence (5.2). Hence $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ correspond to time series from a Gaussian white noise with variance 1 and a Gaussian signal plus noise process (which can be proven to be a white noise) with variance 8. It follows, that this scenario is similar to Scenarios 3.1 and 3.2 in the sense that here $\{(X_{1,t}, X_{2,t})\}_{t=1}^{n}$ is also a sample from a bivariate gaussian white noise. The corresponding autocovariance matrices and the cross-covariance matrix are given by

$$\Gamma_1 = I_{m+1}, \qquad \Gamma_2 = 8I_{m+1}, \qquad \text{and} \qquad \Gamma_{1,2} = 2I_{m+1},$$

and $\rho_{\text{DCCA}}(m) \xrightarrow{P} \sqrt{2}/2$, as $n \to \infty$, for all $m > 0$. In what follows, marginal results shall be presented only for $\{X_{2,t}\}_{t=1}^{n}$, as results for $\{X_{1,t}\}_{t=1}^{n}$ were presented in Scenario 1 (see Figures 5.3 and 5.4).

Figure 5.20 shows that the estimates of $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ are close to their expected values, especially when m is small. The values of $\rho_{\text{DCCA}}(m)$ are all close to $\sqrt{2}/2 \simeq 0.71$ which, in this scenario, is the value of its theoretical counterpart, for all m. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From Figure 5.20, one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, reflecting the theoretical result stated in (4.9).

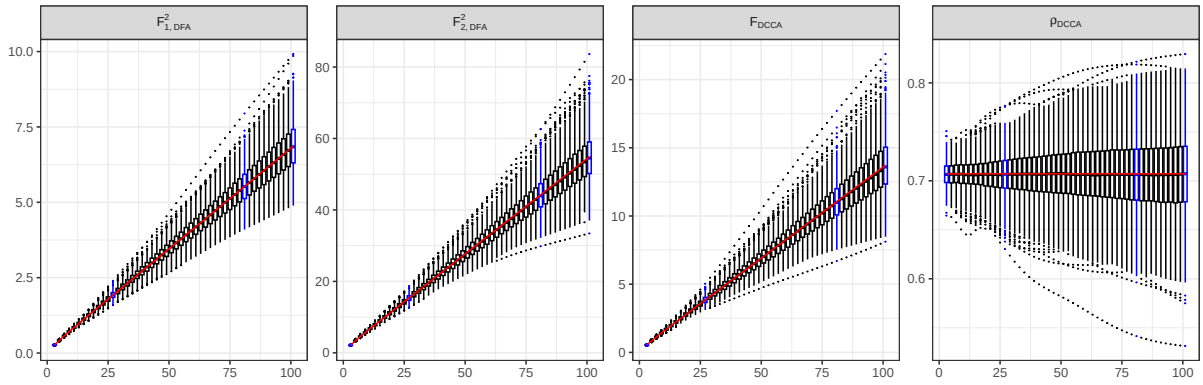

Figure 5.20: Scenario 4.1: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the limit obtained by letting $n \to \infty$.

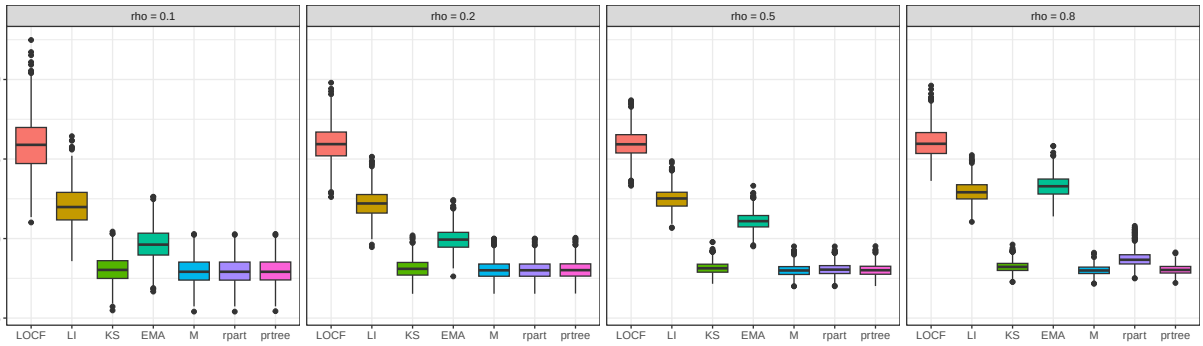

Figure 5.21: Scenario 4.1: Boxplots of the imputation MSE values for $\{X_{2,t}\}_{t=1}^{2000}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

As seen on Figures 5.21 and 5.22, the outcomes of $\{X_{2,t}\}_{t=1}^{n}$ and $F^2_{2,\text{DFA}}(m)$ closely resemble those of $\{X_{1,t}\}_{t=1}^{n}$ and $F^2_{1,\text{DFA}}(m)$ (see Figures 5.3 and 5.4), respectively, with the average-based methods (M, rpart and prtree) outperforming the other methods, only with a higher MSE. This result is expected, given that this process is also an i.i.d. Gaussian process, but with higher variance.

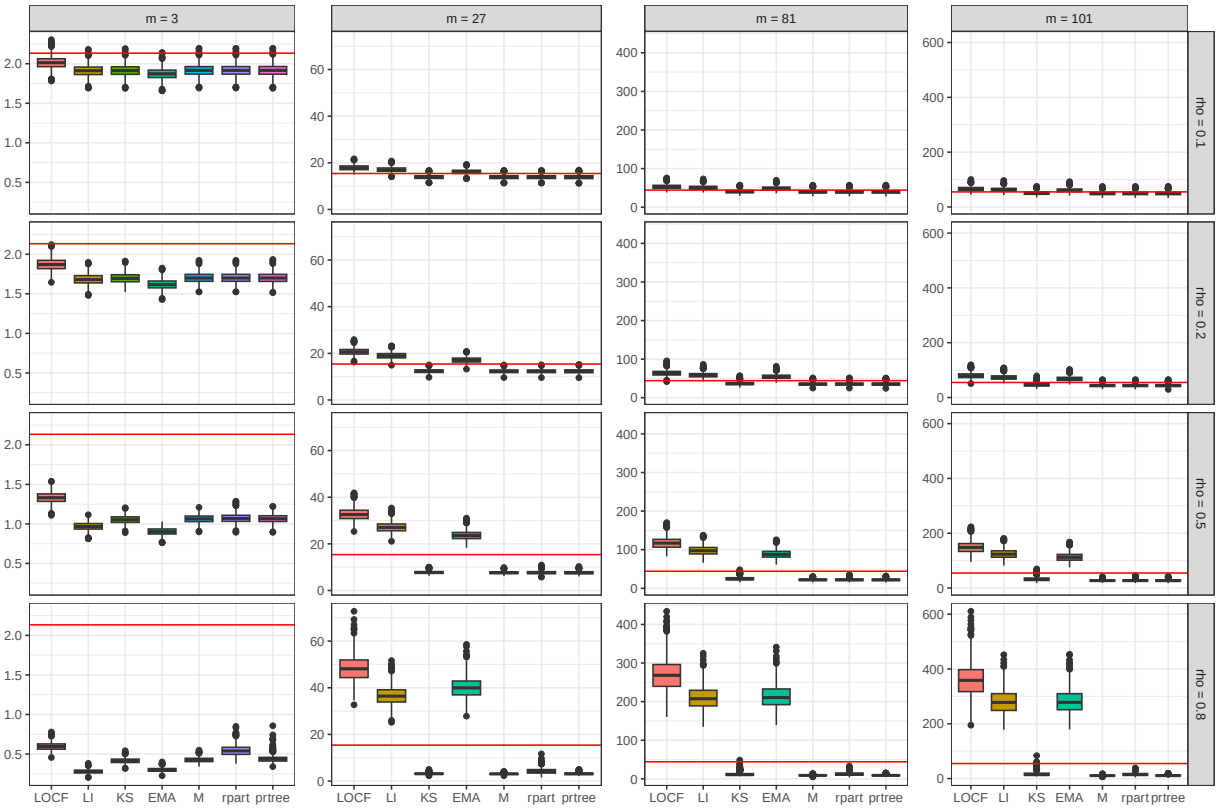

Figure 5.22: Scenario 4.1: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

From Figure 5.23 one observes that $\mathbb{E}[F_{\text{DCCA}}(m)]$ was consistently underestimated, across all scenarios. For $m = 3$, the methods M, KS, rpart, and prtree demonstrated superior performance, while for $m \in \{27, 81, 101\}$, the LOCF, LI, and EMA methods exhibited significantly better performance, especially for $\rho \in \{0.5, 0.8\}$. In Figure 5.24, it is possible to notice that regarding $\rho_{\text{DCCA}}(m)$, the methods KS, M, rpart had very similar MSE values and consistently outperformed the other methods, providing better results in terms of the estimation of $\rho_{\mathcal{E}}(m)$. As the proportion of missing values increased, all methods uniformly degraded.

The estimations of the DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods for filling missing values and also in the context of the estimates $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$. The time series reconstructed using LOCF, LI, and EMA led to the best results when used to calculate the estimates $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$, across different values of m and ρ. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

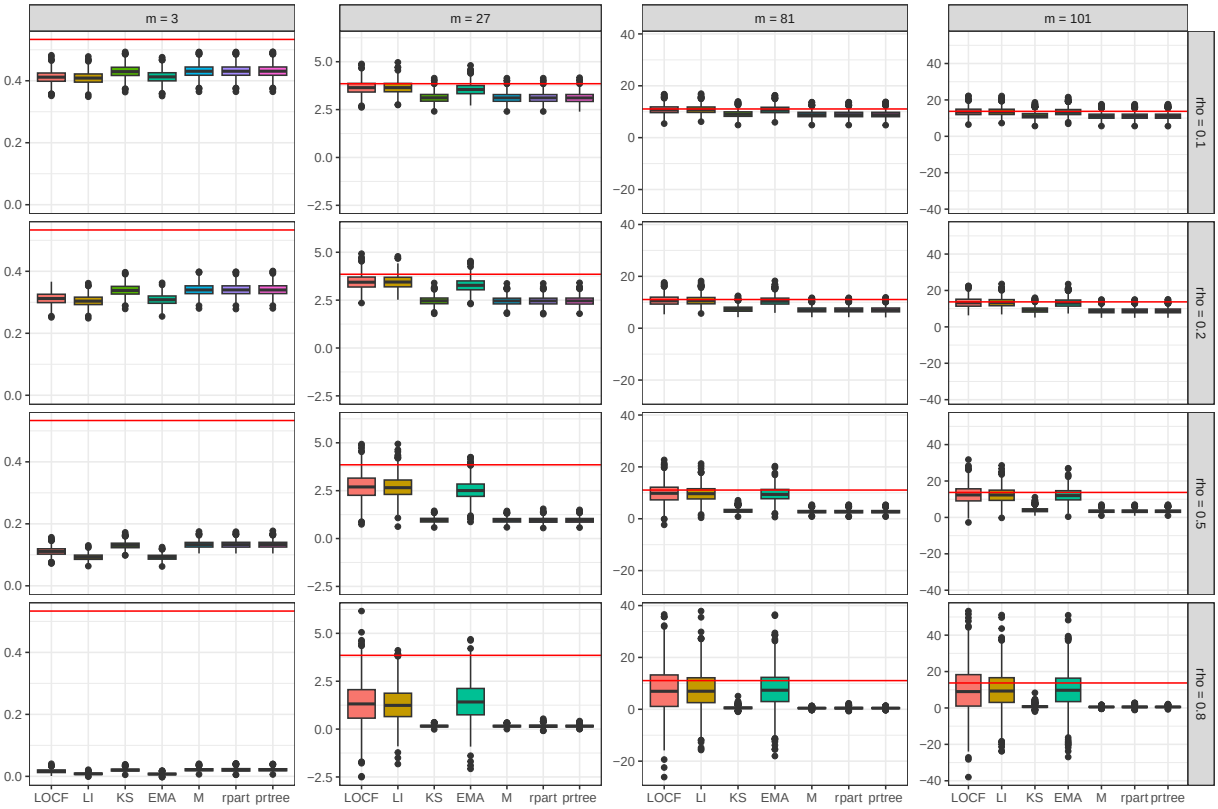

Figure 5.23: Scenario 4.1: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

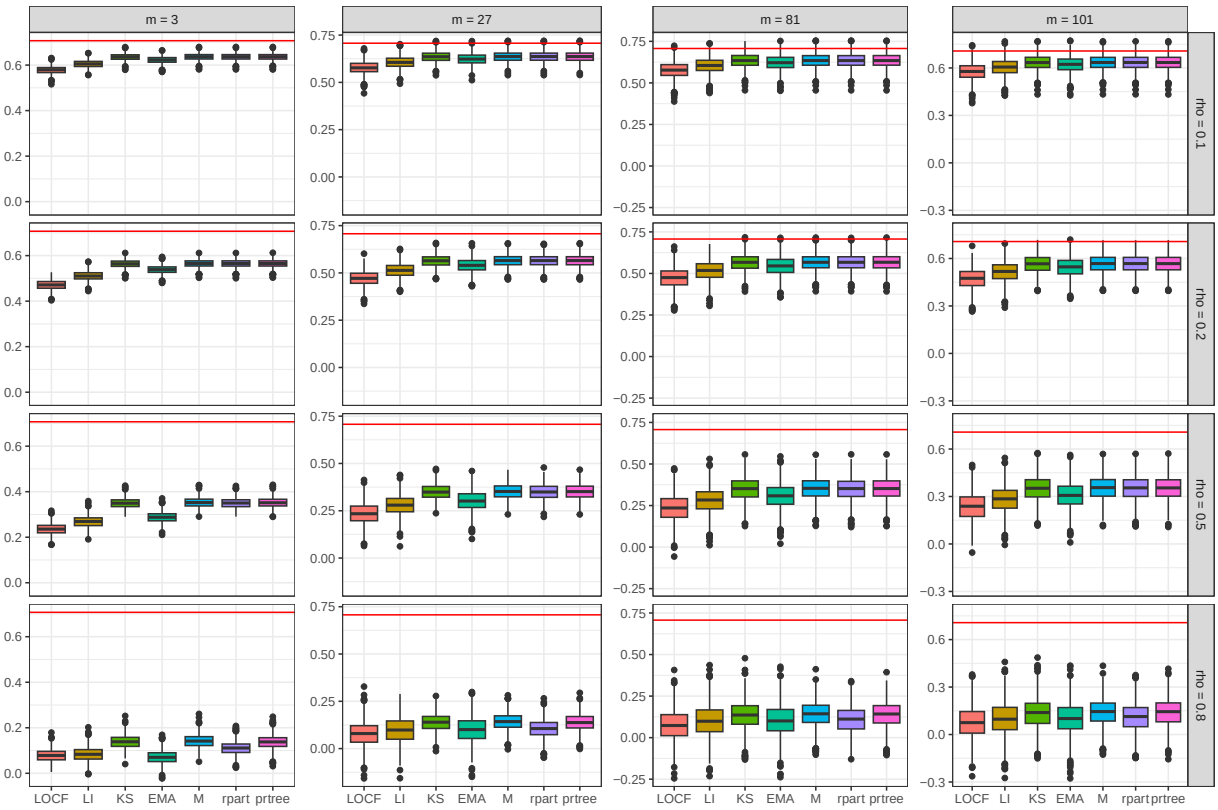

Figure 5.24: Scenario 4.1: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

5.3.6 Scenario 4.2: bivariate white noise with a signal plus noise (high variance) structure

In this scenario the time series $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = \varepsilon_{1,t}, \qquad \text{and} \qquad X_{2,t} = 3 + 2X_{1,t} + \varepsilon_{2,t}, \quad t \in \mathbb{Z}, \tag{5.3}$$

where $\{\varepsilon_{k,t}\}_{t\in\mathbb{Z}}$, $k \in \{1,2\}$, are sequences of i.i.d. $\mathcal{N}(0,1)$ and $\mathcal{N}(0,64)$ random variables, respectively, and $\varepsilon_{1,r}$ and $\varepsilon_{2,s}$ are independent, for all $r, s \in \mathbb{Z}$. Hence $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ correspond to time series from a Gaussian white noise with variance 1 and Gaussian signal plus noise process (which can be proven to be a white noise) with variance 68. The process $\{X_{2,t}\}_{t\in\mathbb{Z}}$ defined in (5.3) shares the same structure as the process $\{X_{2,t}\}_{t\in\mathbb{Z}}$ in (5.2). Consequently, the outcomes of this scenario are expected to resemble those of scenario 4.1. The corresponding autocovariance matrices and the cross-covariance matrix are given by

$$\Gamma_1 = I_{m+1}, \qquad \Gamma_2 = 68I_{m+1}, \qquad \text{and} \qquad \Gamma_{1,2} = 8I_{m+1},$$

and $\rho_{\text{DCCA}}(m) \xrightarrow{P} \sqrt{17}/17$, as $n \to \infty$, for all $m > 0$. In what follows, the marginal results shall be presented only for $\{X_{2,t}\}_{t=1}^{n}$, as results related to the processes $\{X_{1,t}\}_{t=1}^{n}$ were already presented in Scenario 1 (see Figures 5.3 and 5.4).

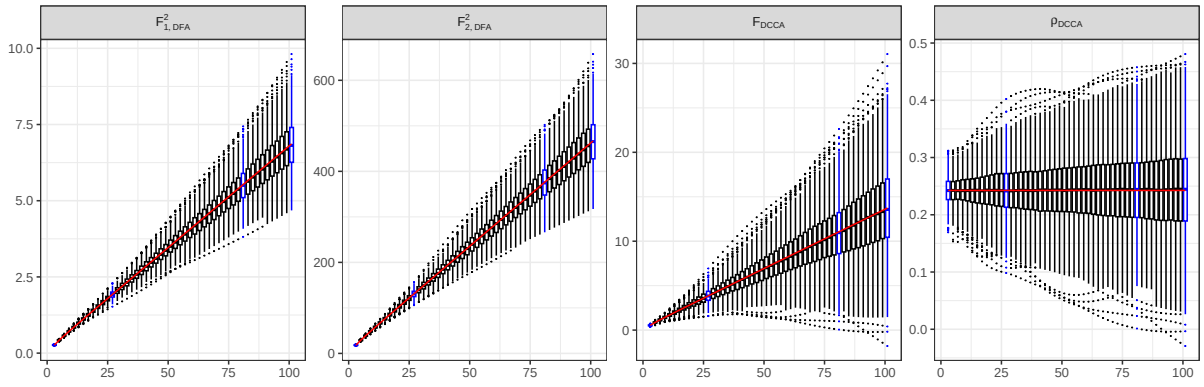

Figure 5.25: Scenario 4.2: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

Figure 5.25 shows that the estimates of $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ are close to their expected values, especially when m is small. The values of $\rho_{\text{DCCA}}(m)$ are all close to $\sqrt{17}/17 \simeq 0.24$, which in this scenario, is the value of its theoretical counterpart for all m. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From Figure 5.25 one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

As seen on Figures 5.26 and 5.27, the outcomes of $\{X_{2,t}\}_{t=1}^{n}$ and $F^2_{2,\text{DFA}}(m)$ closely resemble those of $\{X_{1,t}\}_{t=1}^{n}$ and $F^2_{1,\text{DFA}}(m)$ (see Figures 5.3 and 5.4), respectively, with the average-based methods (M, rpart and prtree) outperforming the other methods, only with a higher MSE. This result is expected, given that this process is also an i.i.d. Gaussian process, but with higher variance.

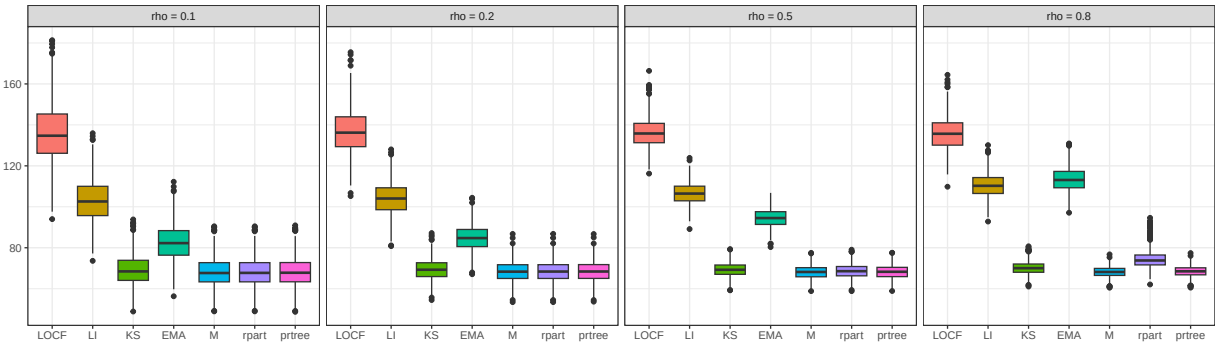

Figure 5.26: Scenario 4.2: Boxplots of the imputation MSE values for $\{X_{2,t}\}_{t=1}^{2000}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

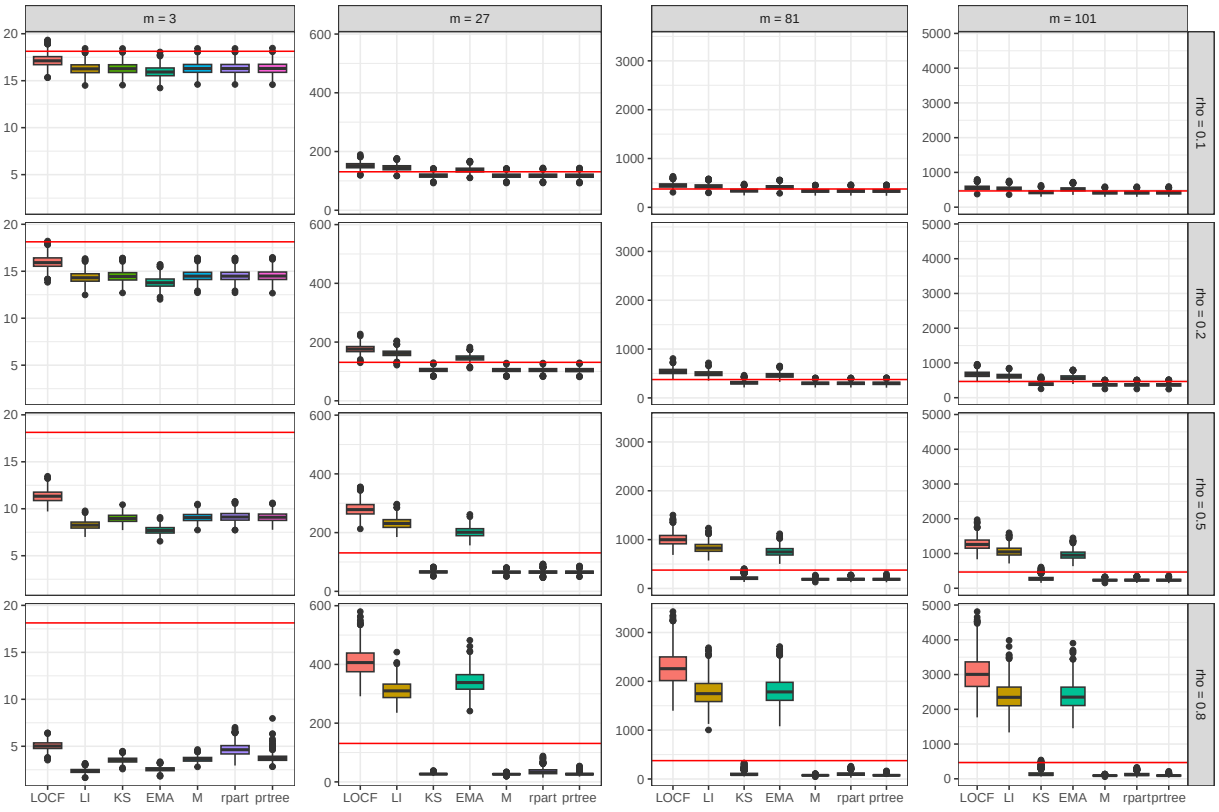

Figure 5.27: Scenario 4.2: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

From Figure 5.28 one observes that $\mathbb{E}[F_{\text{DCCA}}(m)]$ was consistently underestimated, across all scenarios. For $m = 3$, the methods M, KS, rpart, and prtree demonstrated superior performance, while for $m \in \{27, 81, 101\}$, the LOCF, LI, and EMA methods exhibited significantly better performance, especially for $\rho \in \{0.5, 0.8\}$. In Figure 5.29, it is possible to notice that regarding $\rho_{\text{DCCA}}(m)$, the methods KS, M, rpart had very similar MSE values and consistently outperformed the other methods, providing better results in terms of the estimation of $\rho_{\mathcal{E}}(m)$. As the proportion of missing values increased, all methods uniformly degraded.

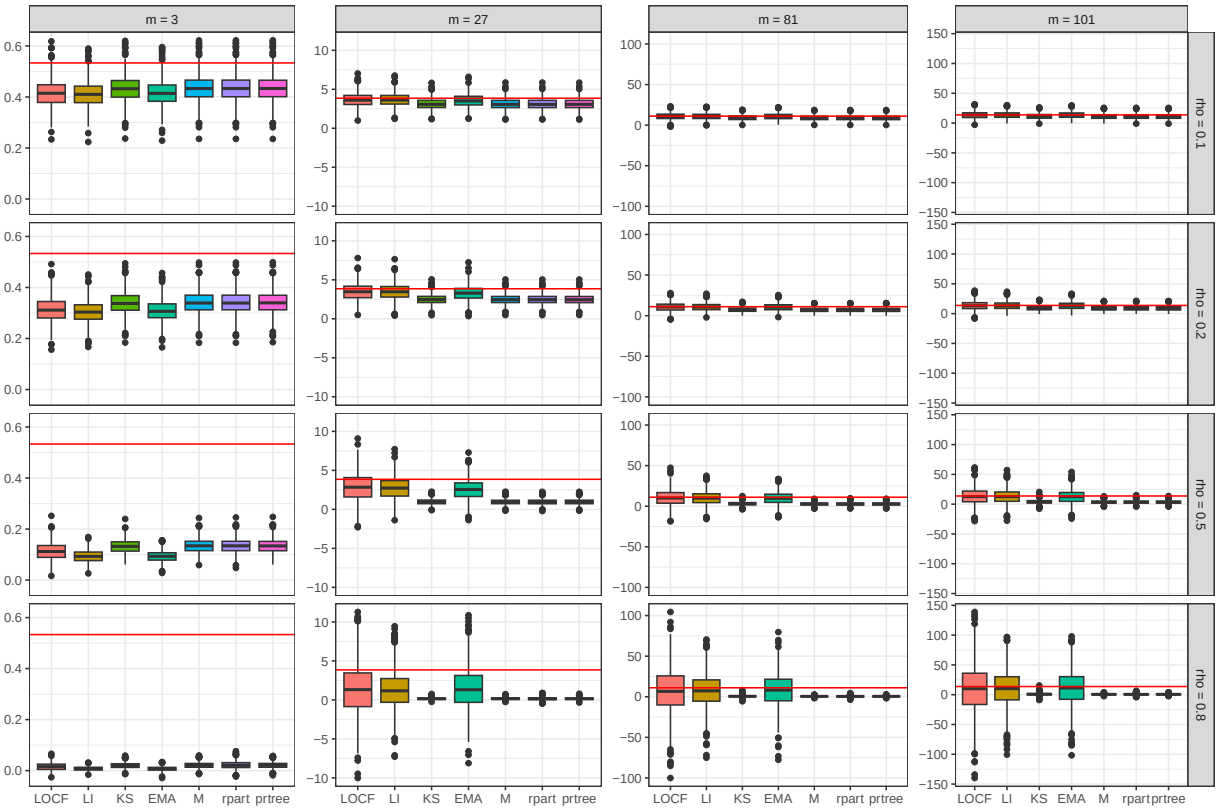

Figure 5.28: Scenario 4.2: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

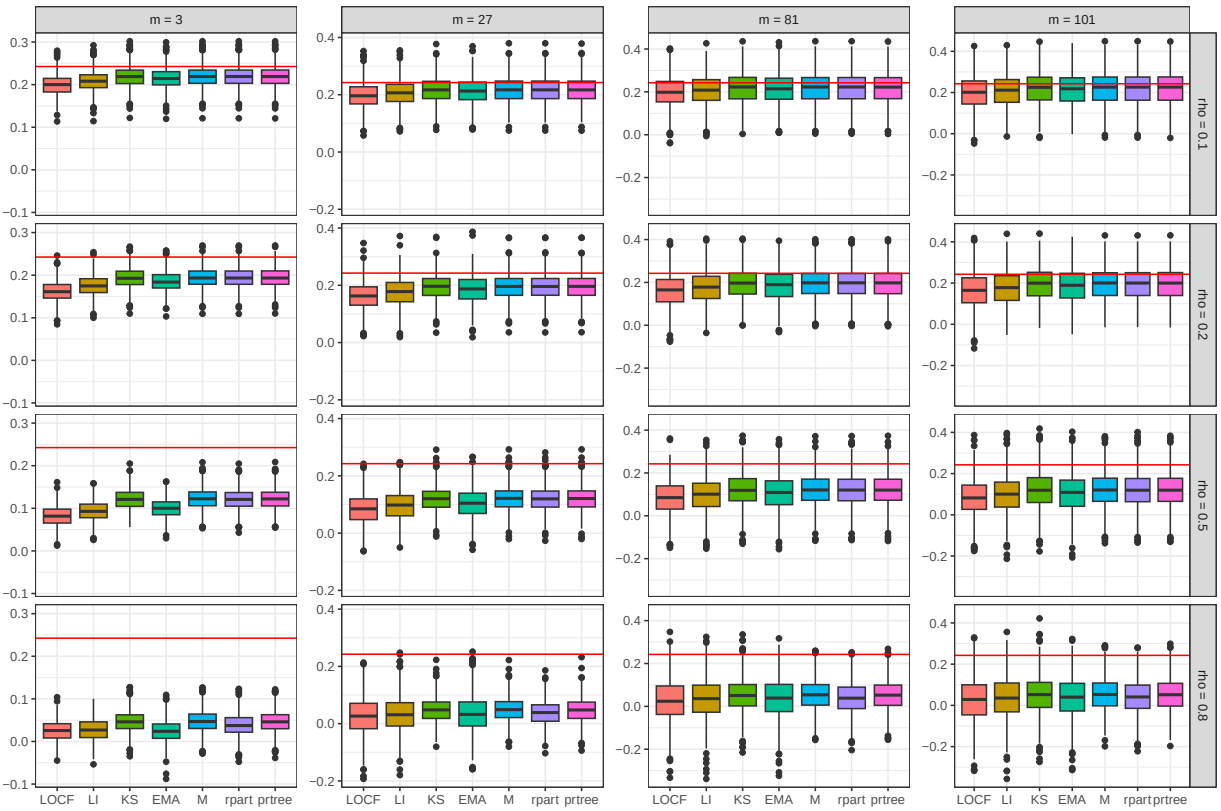

Figure 5.29: Scenario 4.2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

The estimations of the DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods for filling missing values and also in the context of the estimates $F^2_{1,\text{DFA}}(m)$ and $F^2_{2,\text{DFA}}(m)$. The time series reconstructed using LOCF, LI, and EMA led to the best results when used to calculate the estimates $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$, across different values of m and ρ. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

5.3.7 Scenario 5.1: correlated process with dependence driven by an MA structure

In this scenario the time series $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = \varepsilon_t \quad \text{and} \quad X_{2,t} = \varepsilon_t + \sum_{k=1}^{20} \frac{21-k}{10}\varepsilon_{t-k}, \quad t \in \mathbb{Z}, \tag{5.4}$$

where $\{\varepsilon_1\}_{t\in\mathbb{Z}}$, is sequences of i.i.d. $\mathcal{N}(0,1)$ random variables. All time series are generated considering the recurrence (5.4). with burn-in size equal to 20. Hence $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ correspond to time series from a Gaussian white noise and an MA(20) process, respectively, which are cross-correlated. It follows that the (i,j)-th term in the corresponding autocovariance and cross-covariance matrices are given by

$$[\Gamma_1]_{i,j} = I(i=j), \quad [\Gamma_2]_{i,j} = \frac{297}{10}I(i=j) + \left(\frac{|h|^3}{600} - \frac{1321|h|}{600} + \frac{154}{5}\right)I(i\neq j),$$

and

$$[\Gamma_{1,2}]_{i,j} = I(i=j) + \frac{21+j-i}{10}I(0\leq i\leq j\leq 20).$$

The exact expressions for $\mathbb{E}[F^2_{k,\text{DFA}}(m)]$, $\mathbb{E}[F_{\text{DCCA}}(m)]$ and, consequently, for $\rho_{\mathcal{E}}(m)$, can be found in the supplementary material provided by Prass and Pumi (2021). In summary, the following holds

$$\mathbb{E}[F^2_{1,\text{DFA}}(m)] = \frac{m}{15} + O(1) \sim \frac{1}{15}m, \quad \mathbb{E}[F^2_{2,\text{DFA}}(m)] = \frac{484}{15}m + O(1) \sim \frac{22^2}{15}m.$$

$$\mathbb{E}[F_{\text{DCCA}}(m)] = \frac{22}{15}m + O(1) \sim \frac{22}{15}m \quad \text{and} \quad \rho_{\text{DCCA}}(m) \sim 1, \quad \text{as} \quad m\to\infty.$$

In what follows, the marginal results shall be presented only for $\{X_{2,t}\}_{t=1}^n$, as results related to the processes $\{X_{1,t}\}_{t=1}^n$ have already been presented in Scenario 1(see Figures 5.3 and 5.4).

Figure 5.30 shows that the estimates for the functions are close to their expected values, especially when m is small. Also, $\mathbb{E}[\rho_{\text{DCCA}}(m)]$ in a logarithmic behavior. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

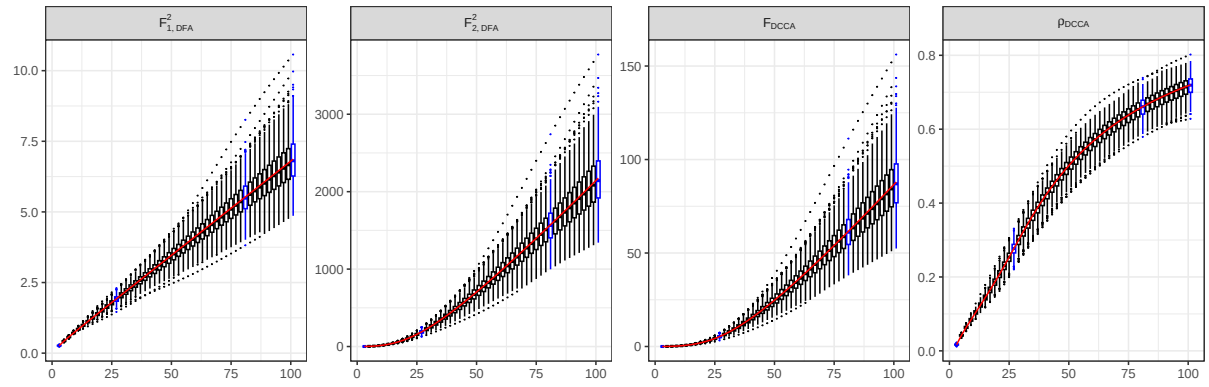

Figure 5.30: Scenario 5.1: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

As seen in Figure 5.31, the methods that had better results filling the missing data for $\{X_{2,t}\}_{t=1}^{n}$ were the LI, KS, and EMA for $\rho \in \{0.1, 0.2, 0.5\}$. For $\rho = 0.8$, the rpart, prtree, M and KS outperformed the other methods. In all cases, LOCF was among the worst performances. Considering that the ACF of a MA(20) process is non-zero for all $|h| \leq 20$, as illustrated in Table 3.1, it is consistent that LI, KS, and EMA methods yielded the best results for $\rho \in \{0.1, 0.2, 0.5\}$, while average-based methods performed better for $\rho = 0.8$. As the proportion of missing values increases, the observations available for imputation using these methods become more temporally distant, and consequently, the relevance of these observations diminishes, making the information from conditional means more significant for the prediction of the missing values. A similar behavior was observed for the AR(1) and MA(1) models in Scenario 2.

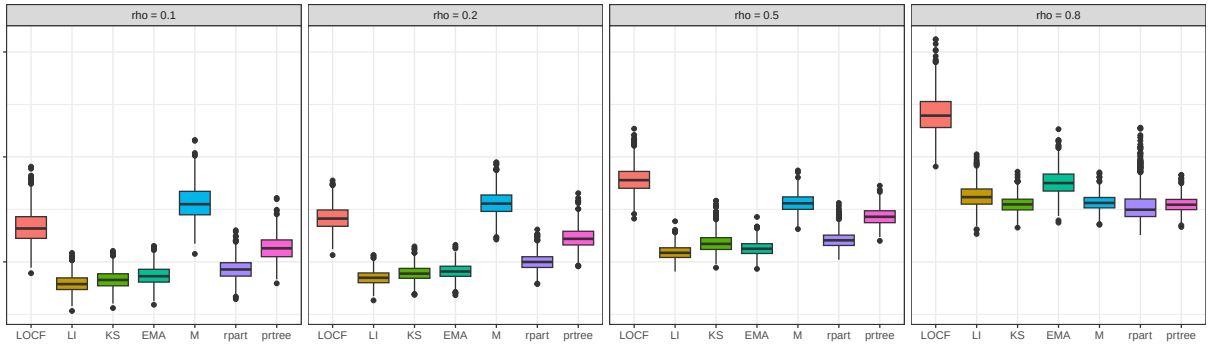

Figure 5.31: Scenario 5.1: Boxplots of the imputation MSE values for $\{X_{2,t}\}_{t=1}^{2000}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

From Figure 5.32, every method estimated $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$ reasonably for $\rho = \{0.1, 0.2\}$, with LOCF and M being the best methods for $m = 3$ and LI, KS and EMA yielding the best results for $m \in \{27, 81, 101\}$. For $\rho \in \{0.5, 0.8\}$, all methods significantly underestimated $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$ when $m = 3$ and for $m \in \{27, 81, 101\}$ the methods LOCF, LI, and EMA overestimated and the methods KS, M, rpart, and prtree underestimated $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

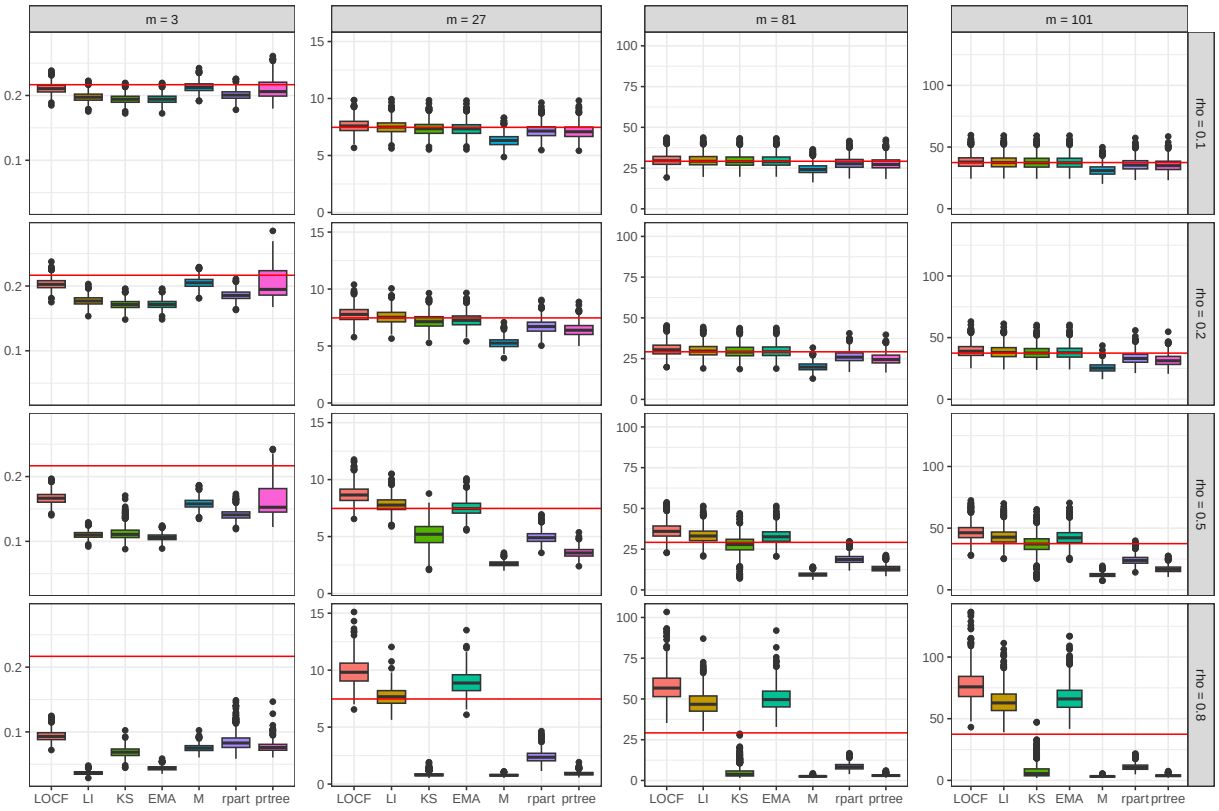

Figure 5.32: Scenario 5.1: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

From Figure 5.33, for $m = 3$, time series reconstructed with all methods had reasonable results for the estimates of $\mathbb{E}[F_{\text{DCCA}}(m)]$, with LI and EMA having the least variability and decrease in quality as ρ increases. For $m \in \{27, 81, 101\}$, LOCF, LI, and EMA methods had the best results, regardless of ρ, while the other methods significantly underestimated $\mathbb{E}[F_{\text{DCCA}}(m)]$. Figure 5.34 illustrates that, for $m = 3$, the expected value of ρ_{DCCA} was reasonably well estimated regardless of method and ρ, with LOCF and KS having the best results with the smaller variability. For $m \in \{27, 81, 101\}$, all methods underestimated $\rho_{\mathcal{E}}(m)$, with EMA and KS having the overall best results. This becomes more evident as the proportion of missing values increases.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, rpart and prtree) outperformed other methods for filling missing values for $\{X_{1,t}\}_{t=1}^n$ and LI, KS and EMA had the best results for $\{X_{2,t}\}_{t=1}^n$. The time series reconstructed using KS and M had the best results for the estimates of $F^2_{1,\text{DFA}}(m)$ while LOCF, LI, KS and EMA yielded the best results for $F^2_{2,\text{DFA}}(m)$, with LOCF standing out as the superior method when $m = 3$. LOCF, LI and EMA estimated values closer to the expected for $F_{\text{DCCA}}(m)$ while LOCF and KS had the best results for $\rho_{\text{DCCA}}(m)$ for $m = 3$ and LOCF, LI, and EMA for $m \in \{27, 81, 101\}$. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation also tend to provide more accurate estimates for the DFA functions, but not necessarily for the DCCA functions.

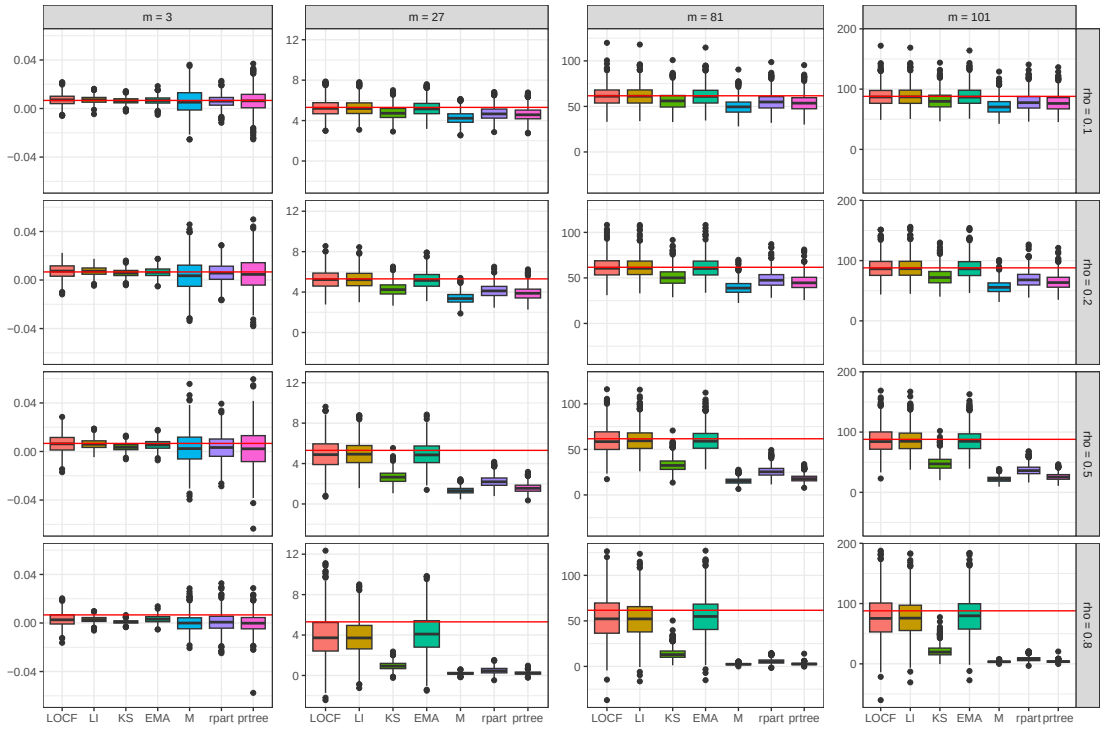

Figure 5.33: Scenario 5.1: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

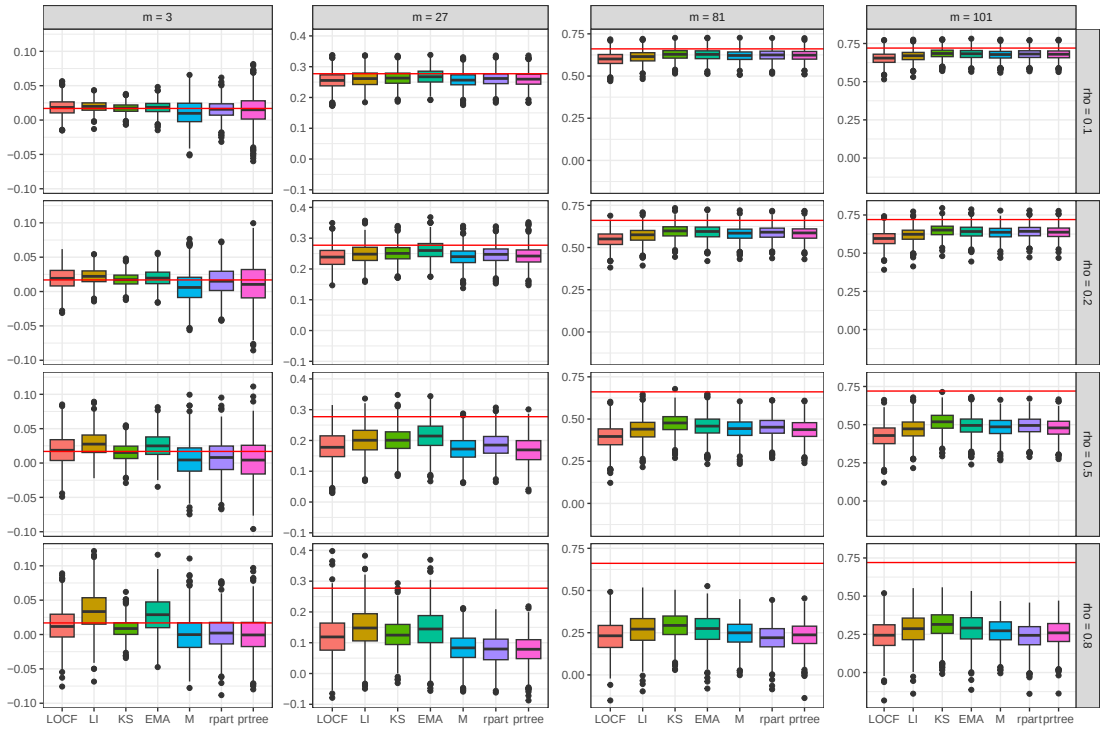

Figure 5.34: Scenario 5.1: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

5.3.8 Scenario 5.2: correlated process with dependence driven by an AR structure

In this scenario the time series $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = \varepsilon_t \quad \text{and} \quad X_{2,t} = 0.6X_{2,t-1} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{5.5}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is sequences of i.i.d. $\mathcal{N}(0,1)$ random variables. All time series are generated considering the recurrence (5.5). with burn-in size equal to 10. Hence $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ correspond to time series from an Gaussian white noise and an AR(1) process, respectively, which are cross-correlated. It follows that the (i,j)-th term in the corresponding autocovariance matrices and cross-covariance matrix are given by

$$[\Gamma_1]_{i,j} = I(i=j), \quad [\Gamma_2]_{i,j} = \frac{0.6^{|i-j|}}{0.64} \quad \text{and} \quad [\Gamma_{1,2}]_{i,j} = 0.6^{j-i}I(i \leq j)$$

From Prass and Pumi (2021),

$$\mathbb{E}[F^2_{1,\text{DFA}}(m)] = \frac{m}{15} + O(1) \sim \frac{m}{15}, \quad \mathbb{E}[F^2_{2,\text{DFA}}(m)] = \frac{m^3 + O(m^2)}{2.4(m^2+3m+2)}, \sim \frac{5m}{12}$$

$$\mathbb{E}[F_{\text{DCCA}}(m)] = \frac{m^3 + O(m^2)}{6(m^2+3m+2)} \sim \frac{m}{6} \quad \text{and} \quad \rho_{\text{DCCA}}(m) \sim 1, \quad \text{as} \quad m \to \infty.$$

Marginal results related to the processes $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ shall be omitted given that they have already been presented in Scenario 1 (see Figures 5.3 and 5.4) and Scenario 2 (see Figures 5.9 and 5.10), respectively. Therefore, only results related to $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ will be presented.

Figure 5.35 shows that the estimates for the functions are close to their expected values, especially when m is small. Also, $\mathbb{E}[\rho_{\text{DCCA}}(m)]$ decreases from $m=3$ to $m=7$ and increases thereafter in a logarithmic behavior. In all cases, the variability increases with m. Given how these quantities are defined, this behavior is to be expected. From this figure one also observes that $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increase linearly with m, which reflects the theoretical result stated in (4.9).

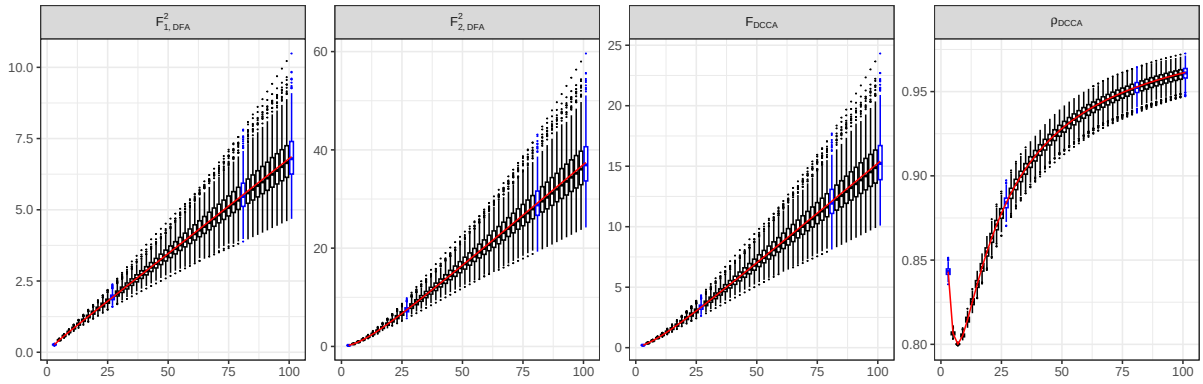

Figure 5.35: Scenario 5.2: Boxplots considering 1000 replications of the complete time series and $m \in \{3,5,\cdots,99,101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

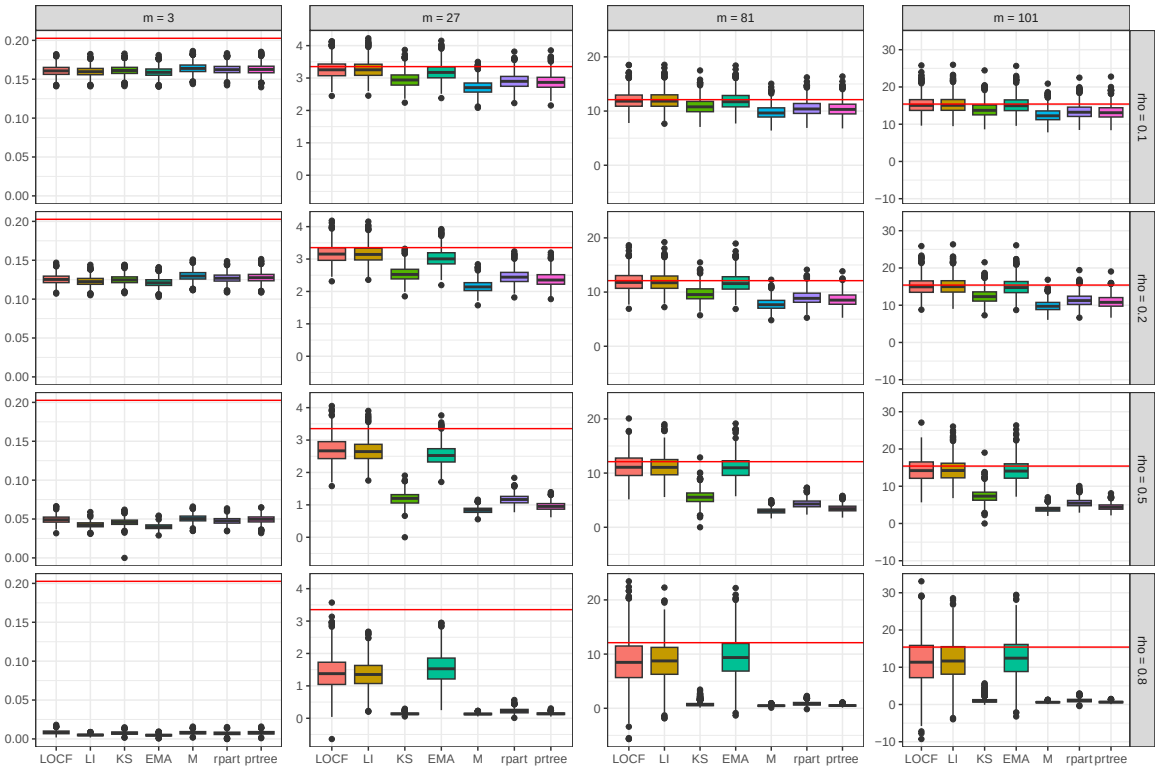

Figure 5.36: Scenario 5.2: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

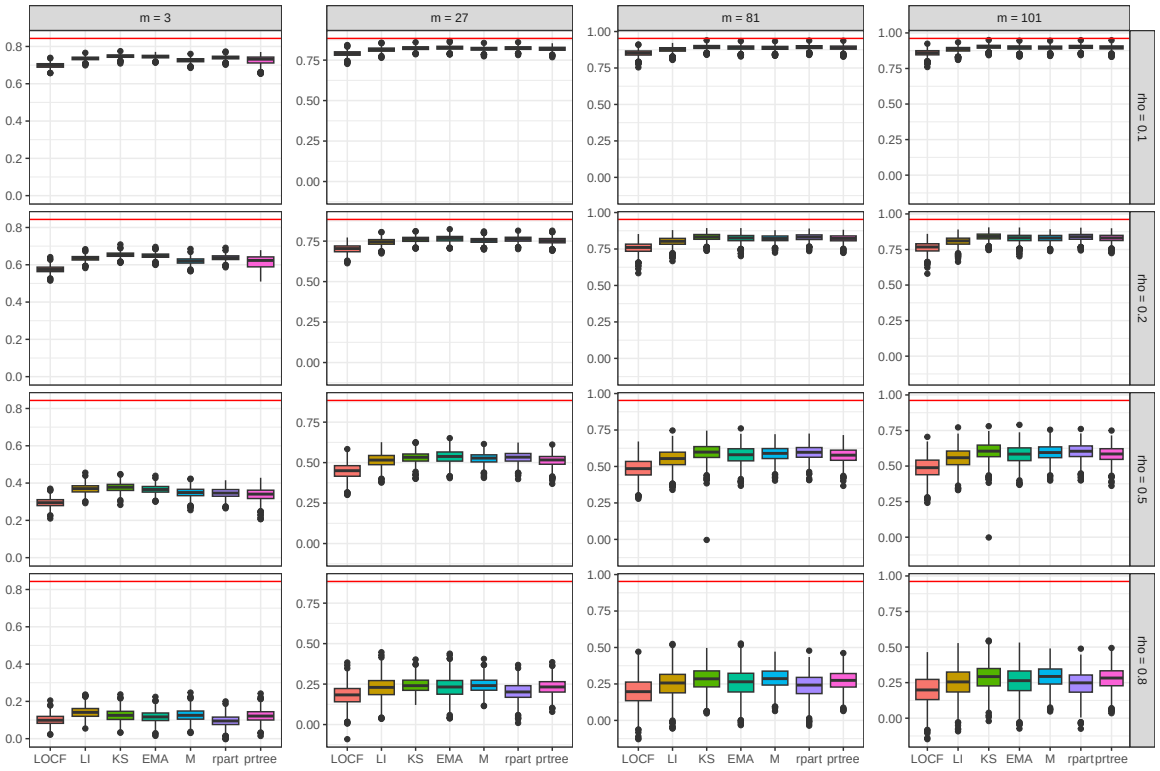

Figure 5.37: Scenario 5.2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

In Figure 5.36, for $m = 3$, all methods underestimated $\mathbb{E}[F_{\text{DCCA}}(m)]$, with M, rpart, and prtree being the closest to $\mathbb{E}[F_{\text{DCCA}}(m)]$. For $m \in \{27, 81, 101\}$, the LOCF, LI, and EMA methods estimated values closer to $\mathbb{E}[F_{\text{DCCA}}(m)]$ for all ρ values, albeit with higher variability than the other methods. The KS, M, rpart, and prtree methods estimated values near zero for $\rho = 0.8$. Concerning $\rho_{\text{DCCA}}(m)$, on Figure 5.37 it is evident that for all values of ρ and m, the expected value is underestimated for all methods. The estimates are reasonable for $\rho = 0.1$ but degrade rapidly as the proportion of missing data is increased, with the time series reconstructed using KS, EMA, and rpart yielding the best results. Figure 5.38 zooms in on the first row of Figure 5.37, providing a closer look to facilitate the visualization of the methods that had superior performance for $\rho = 0.1$.

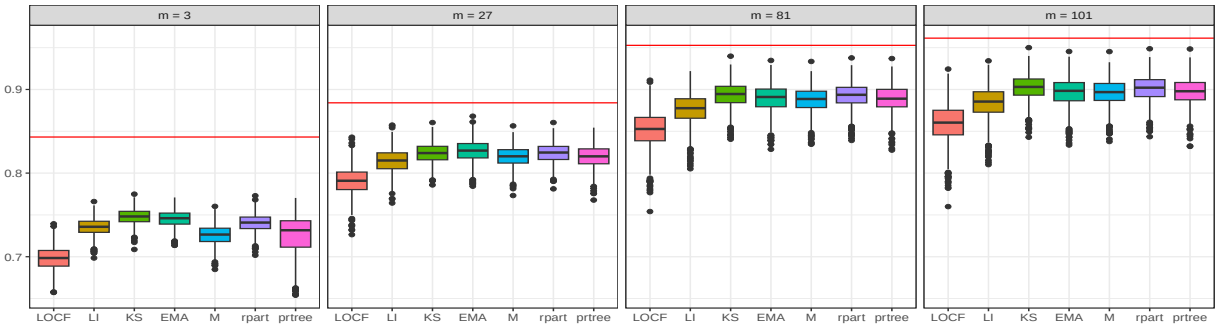

Figure 5.38: Scenario 5.2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho = 0.1$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, average-based methods (M, part, and prtree) outperformed other methods for $\{X_{1,t}\}_{t=1}^{n}$. For $\{X_{2,t}\}_{t=1}^{n}$ M and LOCF had the best results for $m = 3$, while LI, KS, and EMA performed better for $m \in \{27, 81, 101\}$. The time series reconstructed using M, rpart, prtree and KS had the best results for the estimates of $F^2_{1,\text{DFA}}(m)$, while for $F^2_{2,\text{DFA}}(m)$, the methods depended on the window size m. For $m = 3$, the best methods were LOCF and M, while for $m \in \{27, 81, 101\}$ the methods that yielded the best results were LOCF, LI and EMA. Series reconstructed by LOCF and LI estimated values closer to the expected for $F_{\text{DCCA}}(m)$, while $\rho_{\text{DCCA}}(m)$ was better estimated with time series filled by the KS, EMA or rpart. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation also tend to provide more accurate estimates for the DFA functions, but not necessarily for the DCCA function.

5.3.9 Scenario 6.1: couple of AR(2) processes with the same error

In this scenario the time series $\{X_{1,t}\}_{t=1}^{n}$ and $\{X_{2,t}\}_{t=1}^{n}$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = 0.2X_{1,t-1} + \varepsilon_t \quad \text{and} \quad X_{2,t} = 0.6X_{2,t-1} + \varepsilon_t \quad \text{with} \quad \varepsilon_t = 0.7\varepsilon_{t-1} + \eta_t, \quad t \in \mathbb{Z}, \tag{5.6}$$

where $\{\eta_t\}_{t\in\mathbb{Z}}$, is a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. Upon rewriting (5.6) as

$$(1 - 0.2L)(1 - 0.7L)X_{1,t} = \eta_t \quad \text{and} \quad (1 - 0.6L)(1 - 0.7L)X_{2,t} = \eta_t, \quad t \in \mathbb{Z}, \tag{5.7}$$

where L is the back-shift operator, one observes that $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two AR(2) processes with the same residuals. All time series are generated considering the recurrence equations that follow from (5.7), that is,

$$X_{1,t} = 0.9X_{1,t-1} - 0.14X_{1,t-2} + \eta_t \quad \text{and} \quad X_{2,t} = 1.3X_{1,t-1} - 0.42X_{1,t-2} + \eta_t, \quad t \in \mathbb{Z},$$

with burn-in size equal to 60.

From (5.7), one concludes that the causal representation $\{X_{k,t}\}_{t\in\mathbb{Z}}$ is given by

$$X_{k,t} = \sum_{j=0}^{\infty} \psi_{k,j}\eta_{t-j}, \quad \psi_{k,j} = \frac{0.7^{j+1} - \alpha_k^{j+1}}{0.7 - \alpha_k}, \quad \alpha_k = \begin{cases} 0.2, & k = 1, \\ 0.6, & k = 2, \end{cases} \quad k \in \{1, 2\}.$$

It follows that the (i, j)-th term in the corresponding autocovariance and cross-covariance matrices are given by

$$[\Gamma_k]_{i,j} = \frac{1}{(0.7 - \alpha_k)^2}\left(\frac{0.7^{2+|i-j|}}{1 - 0.49} - \frac{0.7^{1+|i-j|}\alpha_k}{1 - 0.7\alpha_k} - \frac{0.7\alpha_k^{1+|i-j|}}{1 - 0.7\alpha_k} + \frac{\alpha_k^{2+|i-j|}}{1 - \alpha_k^2}\right), \quad k \in \{1, 2\},$$

and, by letting $a_{ij} = \alpha_1 I(i \le j) + \alpha_2 I(i > j)$ and $b_{ij} = \alpha_2 I(i \le j) + \alpha_1 I(i > j)$,

$$[\Gamma_{1,2}]_{i,j} = \frac{1}{(0.7 - \alpha_1)(0.7 - \alpha_2)}\left(\frac{0.7^{2+|i-j|}}{1 - 0.49} - \frac{0.7^{1+|i-j|}a_{ij}}{1 - 0.7a_{ij}} - \frac{0.7b_{ij}^{1+|i-j|}}{1 - 0.7b_{ij}} + \frac{a_{ij}b_{ij}^{1+|i-j|}}{1 - \alpha_1\alpha_2}\right).$$

Moreover, from Prass and Pumi (2021),

$$\mathbb{E}[F^2_{k,\text{DFA}}(m)] = \frac{m^3(m^2 + 3m + 2)^{-1}}{1.35(1 - \alpha_k)^2} + O(1) \sim \frac{m}{1.35(1 - \alpha_k)^2}, \quad k \in \{1, 2\},$$

$$\mathbb{E}[F_{DCCA}(m)] = \frac{m^3(m^2 + 3m + 2)^{-1}}{1.35 \times 0.32} + O(1) \sim \frac{m}{1.35 \times 0.32},$$

and $\rho_{\text{DCCA}}(m) \sim 1$, as $m \to \infty$

Figure 5.39 shows that the estimated functions using the complete series closely approximate the expected values for all values of m. The expected value of the function $\rho_{\text{DCCA}}(m)$ increases logarithmically. The values of $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increases linearly as the window size m increases, and the variability of the estimates also increases with m, which reflects the theoretical result stated in (4.9)

As seen on Figures 5.40, the methods that performed the best in filling missing values were LI, KS, and EMA. Notably, all methods, except for M, exhibit a worsening in terms of MSE as the values of ρ increase. Since both these processes are samples from AR(2) processes, the surrounding observations are the ones that contribute the most to predicting X_t (see Table 3.1), which explains the good performance of LI and EMA. Also, since KS is a likelihood-based method and the underlying distribution is correctly specified, it is expected that this method will be among the best ones.

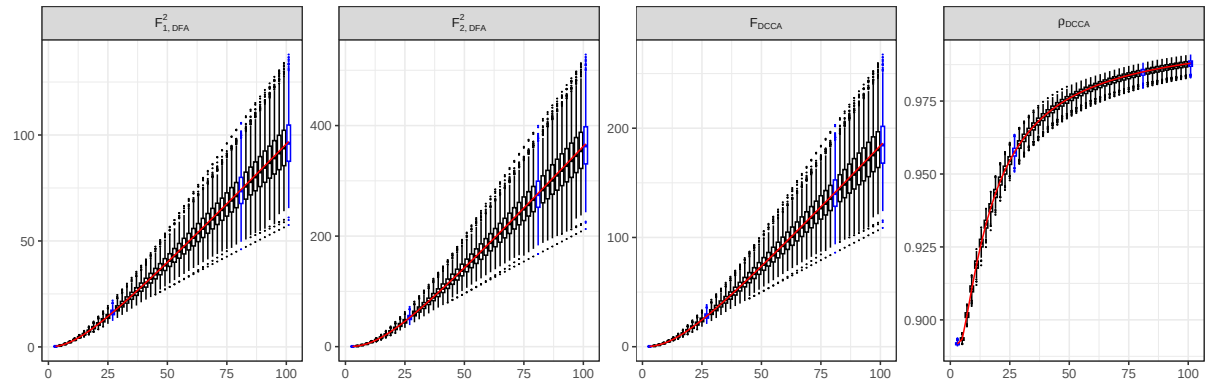

Figure 5.39: Scenario 6.1: Boxplots considering 1000 replications of the complete time series and $m \in \{3, 5, \cdots, 99, 101\}$. From left to right $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and $\rho_{\text{DCCA}}(m)$. In all cases, the red line represents the theoretical limit obtained by letting $n \to \infty$.

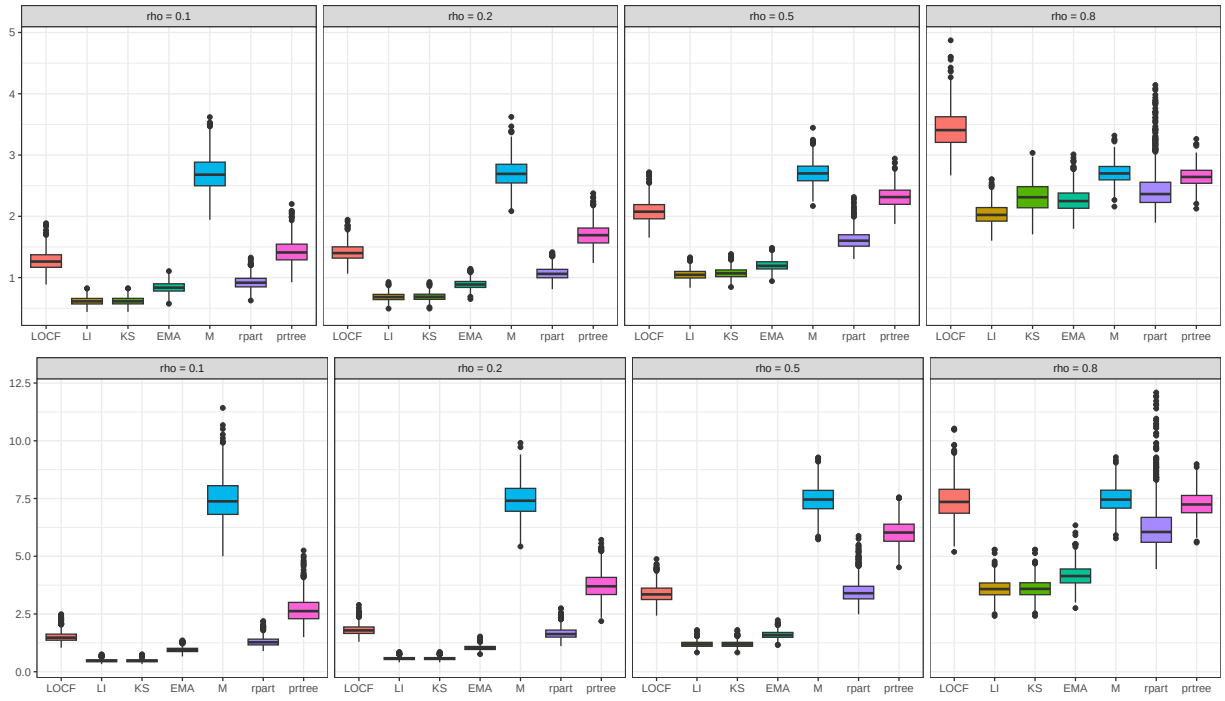

Figure 5.40: Scenario 6.1: Boxplots of the imputation MSE value for $\{X_{1,t}\}_{t=1}^{2000}$ (top row) and $\{X_{2,t}\}_{t=1}^{2000}$ (bottom row), based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

As seen in Figures 5.41 and 5.42, for $m = 3$, the best estimates overall for $F^2_{k,\text{DFA}}(m)$ were achieved in time series filled by LOCF and by the average-based methods. For $m \in \{27, 81, 101\}$, LOCF, LI, KS, and EMA methods all showed similar and reasonable results, regardless of the value of ρ. Also for this values of m, for $\rho \in \{0.5, 0.8\}$, the average based methods significantly underestimated $\mathbb{E}[F^2_{k,\text{DFA}}(m)]$. It is coherent that the methods that provided the best estimates for $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$ are the same as for $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$, as despite having different specifications, both processes share a similar autocovariance structure.

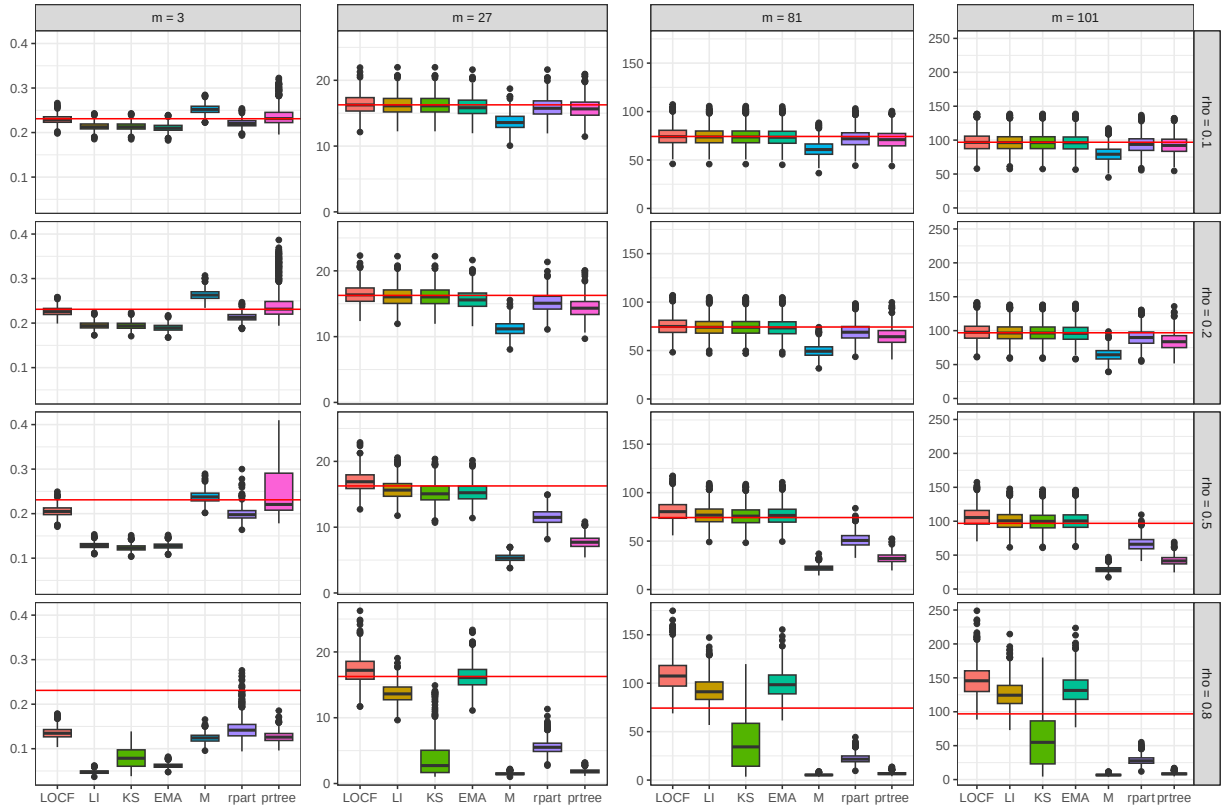

Figure 5.41: Scenario 6.1: Boxplots of $F^2_{1,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

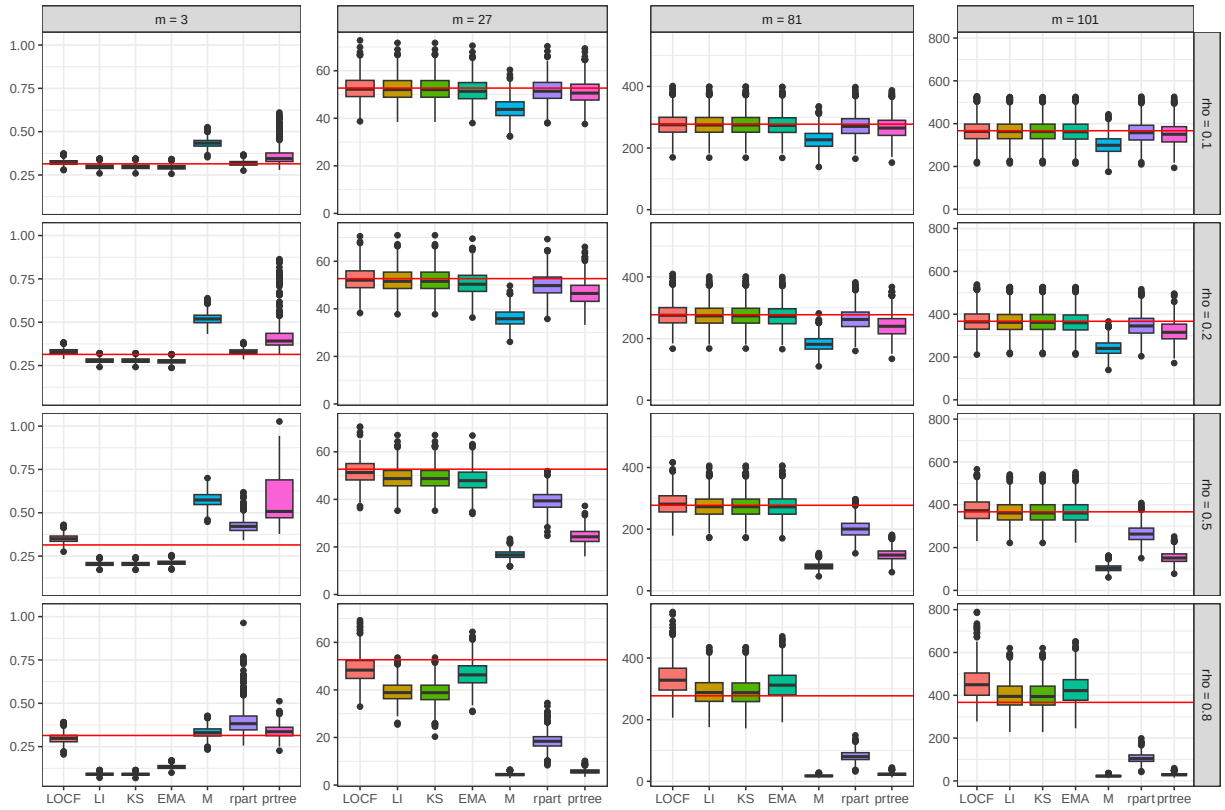

Figure 5.42: Scenario 6.1: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

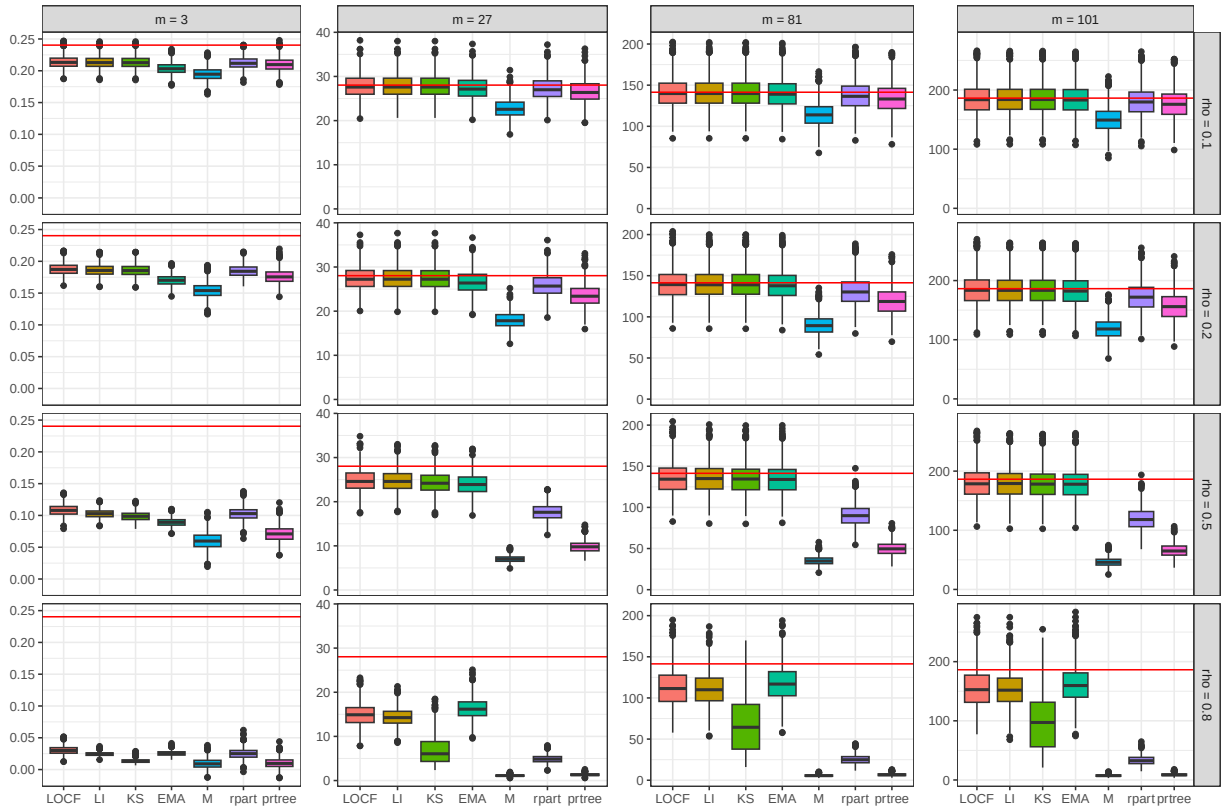

Figure 5.43: Scenario 6.1: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

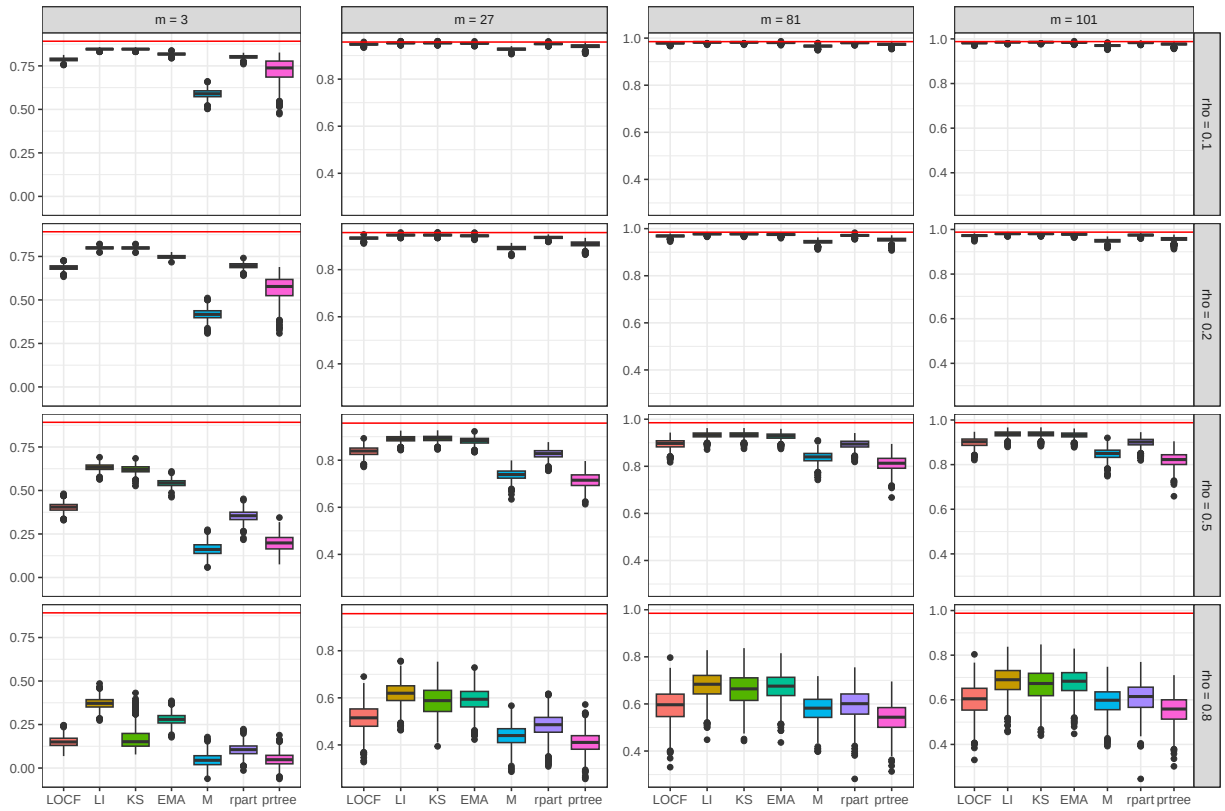

Figure 5.44: Scenario 6.1: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

In Figure 5.43, it is evident that for $m = 3$, all methods underestimated $\mathbb{E}[F_{\text{DCCA}}(m)]$, independently of ρ. LOCF and LI had the best results. For $m \in \{27, 81, 101\}$, LOCF, LI, and EMA all showed similar and reasonable results, falling only well below $\mathbb{E}[F_{\text{DCCA}}(m)]$ for $\rho = 0.8$. Concerning $\rho_{\text{DCCA}}(m)$, it is evident on Figure 5.44 that for $\rho \in \{0.1, 0.2\}$, all methods performed well, providing estimates close to the expected, especially for the time series reconstructed by KS and LI. For $\rho \in \{0.5, 0.8\}$, all methods substantially underestimated the expected values of the function, with LOCF, LI, KS and EMA performing better than the average-based methods.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, LOCF and the average-based methods outperformed other methods for $m = 3$ and for $m \in \{27, 81101\}$ the best methods were LOCF, LI, KS and EMA. The time series reconstructed using LOCF, LI, and EMA had the best estimates for $F^2_{k,\text{DFA}}$, $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ across different values of m and ρ. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation provide more accurate estimates for the DFA and DCCA functions.

5.3.10 Scenario 6.2: couple of ARMA(1,1) with the same error

In this scenario the time series $\{X_{1,t}\}_{t=1}^n$ and $\{X_{2,t}\}_{t=1}^n$ are samples from the stochastic processes defined, respectively, by

$$X_{1,t} = 0.2\varepsilon_{t-1} + \varepsilon_t, \quad X_{2,t} = 0.6\varepsilon_{t-1} + \varepsilon_t \quad \text{and} \quad \varepsilon_t = 0.7\varepsilon_{t-1} + \eta_t, \quad t \in \mathbb{Z}, \tag{5.8}$$

where $\{\eta_t\}_{t\in\mathbb{Z}}$, is a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. Upon rewriting (5.8) as

$$(1 - 0.7L)X_{1,t} = (1 + 0.2L)\eta_t \quad \text{and} \quad (1 - 0.7L)X_{2,t} = (1 + 0.6L)\eta_t, \quad t \in \mathbb{Z}, \tag{5.9}$$

where L is the back-shift operator, one observes that $\{X_{1,t}\}_{t\in\mathbb{Z}}$ and $\{X_{2,t}\}_{t\in\mathbb{Z}}$ are two ARMA(1,1) processes with the same residuals. All time series are generated considering the recurrence equations that follow from (5.9), that is,

$$X_{1,t} = 0.7X_{1,t-1} + 0.2\eta_{t-1} + \eta_t \quad \text{and} \quad X_{2,t} = 0.7X_{1,t-1} + 0.6\eta_{t-1} + \eta_t, \quad t \in \mathbb{Z},$$

with burn-in size equal to 60.

From (5.9), one concludes that the causal representation $\{X_{k,t}\}_{t\in\mathbb{Z}}$, $k \in \{1, 2\}$, is given by

$$X_{k,t} = \sum_{j=0}^{\infty} \psi_{k,j}\eta_{t-j}, \quad \psi_{k,j} = I(j = 0) + 0.7^{j-1}(0.7 + \alpha_k)I(j > 0), \quad \alpha_k = \begin{cases} 0.2, & k = 1, \\ 0.6, & k = 2. \end{cases}$$

It follows that the (i, j)-th term in the corresponding autocovariance and cross-covariance matrices are given by

$$[\Gamma_k]_{i,j} = \psi_{k,|i-j|} + \frac{0.7^{|i-j|}}{1 - 0.49}(0.7 + \alpha_k)^2, \quad k \in \{1, 2\},$$

and

$$[\Gamma_{1,2}]_{i,j} = \psi_{2,j-i}I(i \le j) + \psi_{1,i-j}I(i > j) + \frac{0.7^{|j-i|}}{1 - 0.49}(0.7 + \alpha_1)(0.7 + \alpha_2).$$

Moreover, from Prass and Pumi (2021),

$$\mathbb{E}[F^2_{k,\text{DFA}}(m)] = \frac{(0.7+\alpha_k)^2 m^3}{15(0.21)^2(m^2+3m+2)} + O(1) \sim \frac{(0.7+\alpha_k)^2 m}{15(0.21)^2}, \quad k \in \{1,2\},$$

$$\mathbb{E}[F_{DCCA}(m)] = \frac{1.17m^3}{15(0.21)^2(m^2+3m+2)} + O(1) \sim \frac{1.17m}{15(0.21)^2},$$

and $\rho_{\text{DCCA}}(m) \sim 1$, as $m \to \infty$.

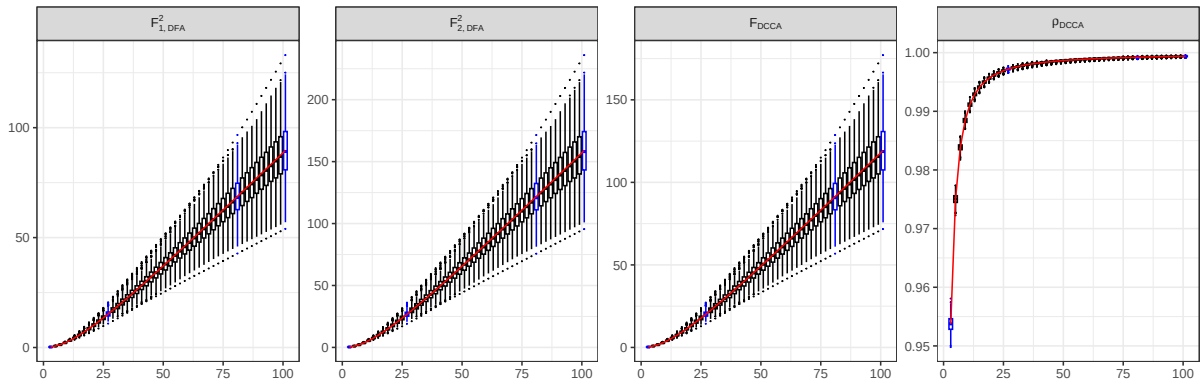

Figure 5.45: Scenario 6.2: Boxplots of the imputation MSE values, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

Figure 5.45 shows that the estimated functions using the complete series closely approximate the expected values for all values of m. The values of $F^2_{1,\text{DFA}}(m)$, $F^2_{2,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ increases linearly as the window size m increases, and the variability of the estimates also increases with m, which reflects the theoretical result stated in (4.9). The expected value of the function $\rho_{\text{DCCA}}(m)$ increases logarithmically.

As seen on Figures 5.46, the methods that performed the best in filling missing values were LI, KS and EMA. Notably, all methods, except for M, exhibit a worsening in terms of MSE as the values of ρ increase. As both these time series are samples from ARMA(1,1) processes, the surrounding observations are the ones that contribute the most to predicting X_t (see Table 3.1), which explains the good performance of LI and EMA. Also, since KS is a likelihood-based method and the underlying distribution is correctly specified, it is expected that this method will be among the best ones.

As observed in Figures 5.47 and 5.48, for $m = 3$, the best estimates for $F^2_{2,\text{DFA}}(m)$ were achieved in time series filled by the LOCF and rpart methods, while the worst-performing methods were KS, LI, and EMA. For $m \in \{27, 81, 101\}$ and $\rho \in \{0.1, 0.2, 0.5\}$, LOCF, LI, KS, and EMA had the best results, However, for $\rho = 0.8$, the methods closest to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$ and $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$ were LI and EMA. It is coherent that the methods which provided the best estimates for $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$ are the same as for $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$, as despite having different specifications, both processes share a similar autocovariance structure. In Figure 5.49, for $m = 3$, all methods underestimated $\mathbb{E}[F_{\text{DCCA}}(m)]$, which especially noticeable for $\rho \in \{0.5, 0.8\}$. For $m \in \{27, 81, 101\}$, LOCF, LI, and EMA methods showed reasonable results, falling well below the expected value of the function only for $\rho = 0.8$. Concerning $\rho_{\text{DCCA}}(m)$, it is evident on 5.50 that for $\rho \in \{0.1, 0.2\}$, all methods performed well, providing estimates close to the expected. However, for $\rho \in \{0.5, 0.8\}$, all methods substantially underestimated $\rho_{\mathcal{E}}(m)$.

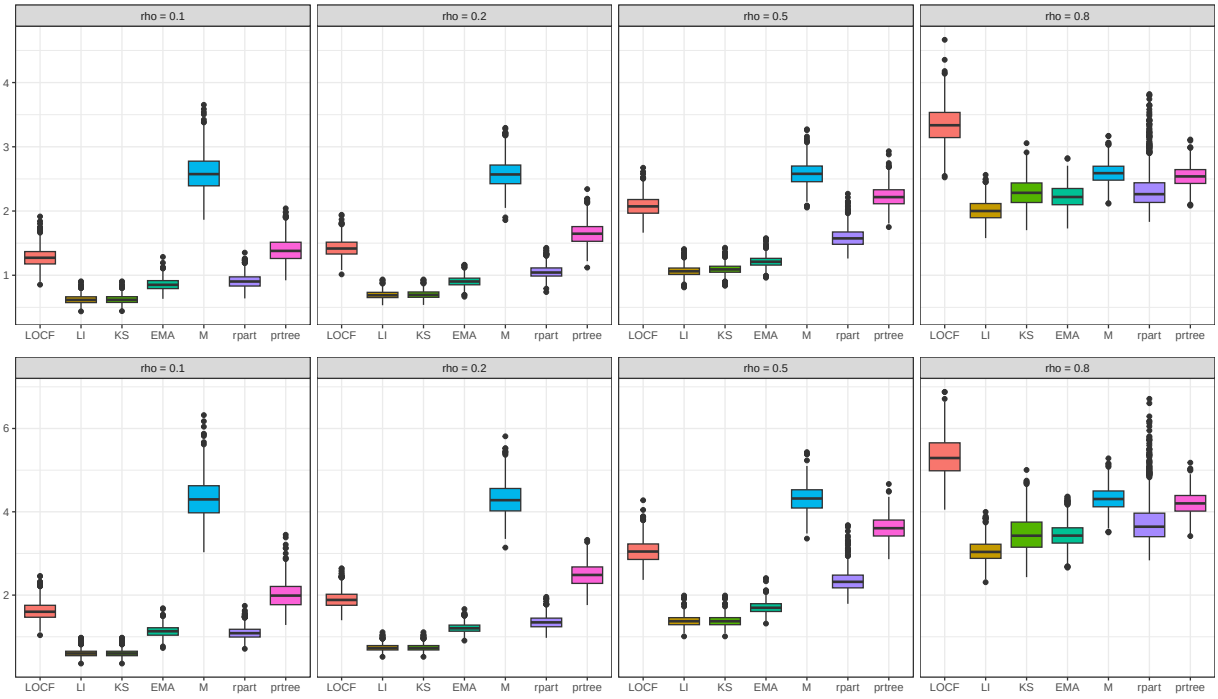

Figure 5.46: Scenario 6.2: Boxplots of the imputation MSE value for $\{X_{1,t}\}_{t=1}^{2000}$ (top row) and $\{X_{2,t}\}_{t=1}^{2000}$ (bottom row), based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$.

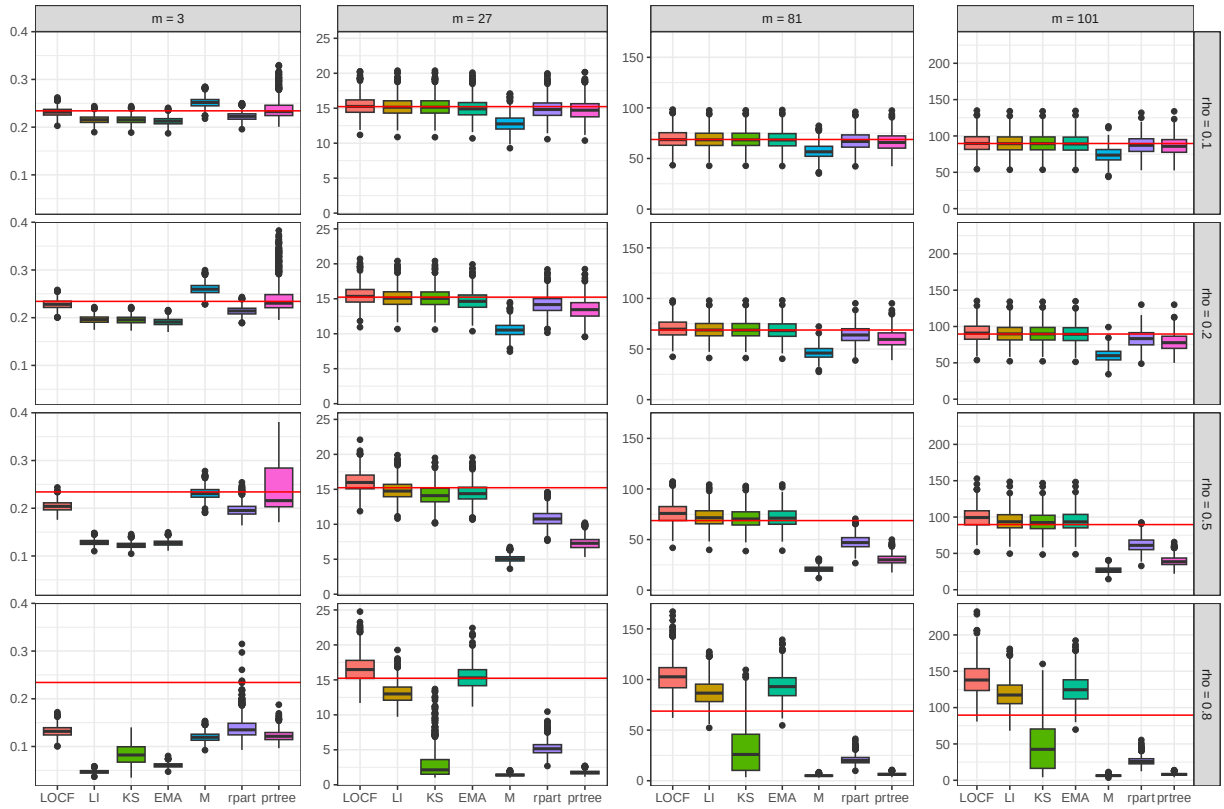

Figure 5.47: Scenario 6.2: Boxplots of $F^2_{1,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{1,\text{DFA}}(m)]$.

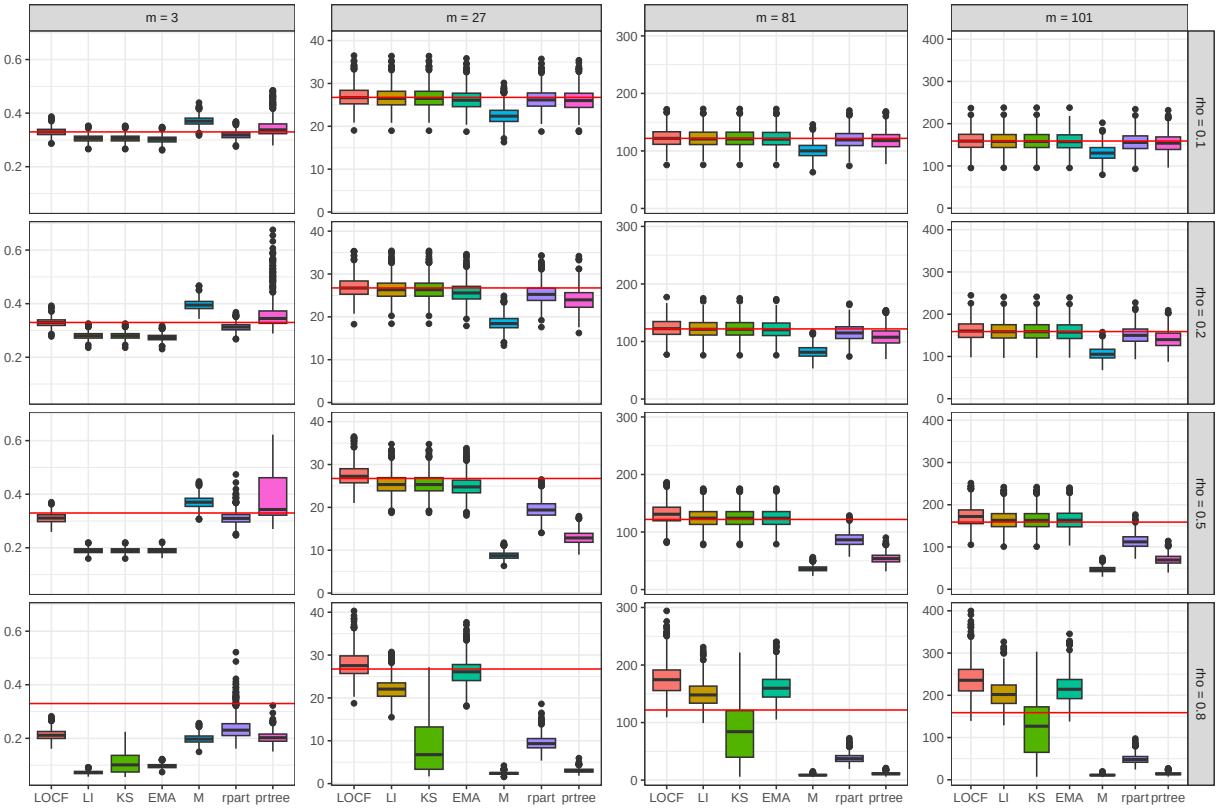

Figure 5.48: Scenario 6.2: Boxplots of $F^2_{2,\text{DFA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F^2_{2,\text{DFA}}(m)]$.

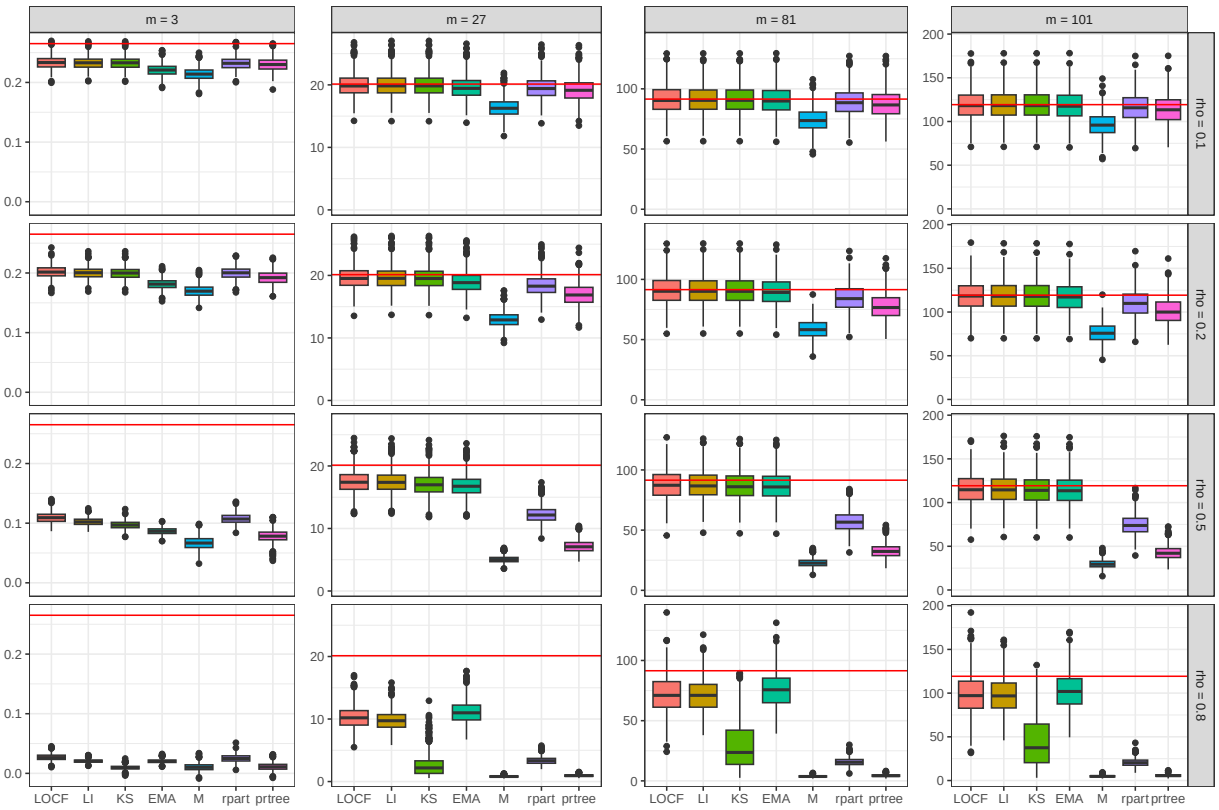

Figure 5.49: Scenario 6.2: Boxplots of $F_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\mathbb{E}[F_{\text{DCCA}}(m)]$.

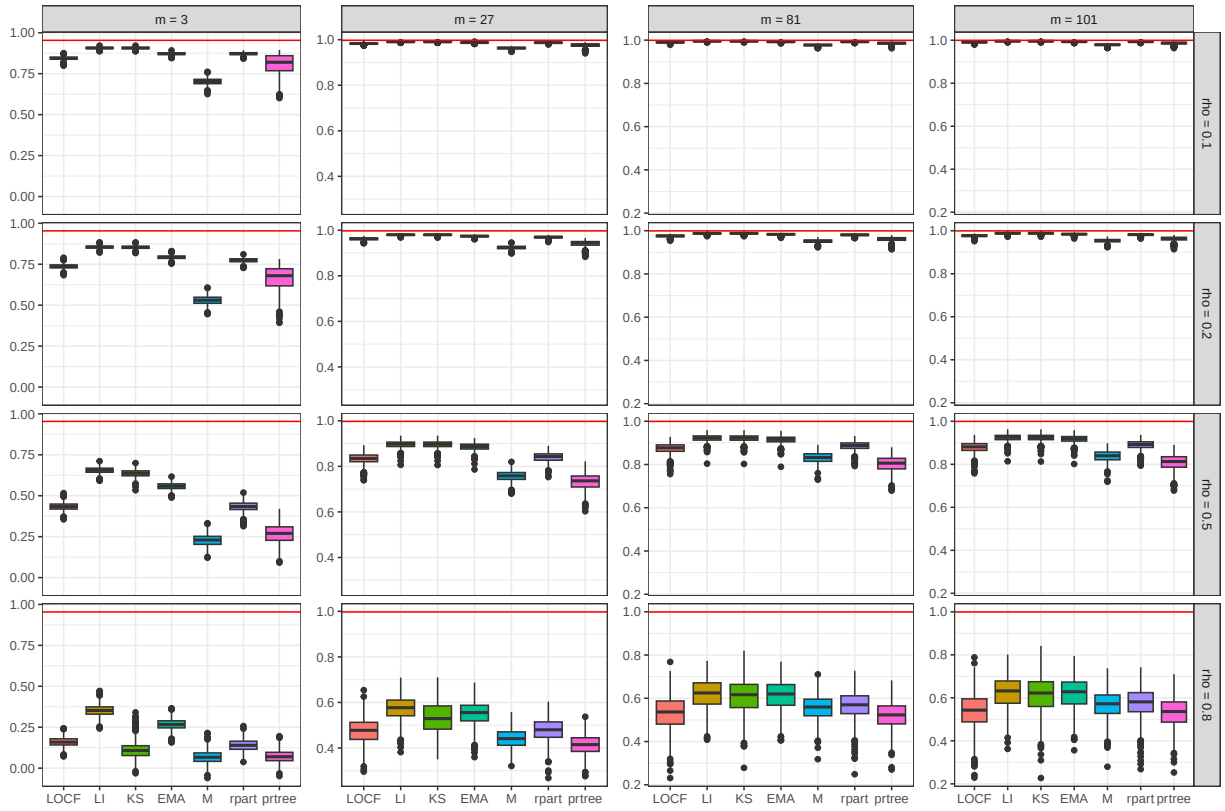

Figure 5.50: Scenario 6.2: Boxplots of $\rho_{\text{DCCA}}(m)$, $m \in \{3, 27, 81, 101\}$, based on $r = 1000$ replications, considering 7 imputation methods, for $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. The red line correspond to $\rho_{\mathcal{E}}(m)$.

The estimates of DFA and DCCA functions with complete time series had results close to the expected values in terms of median, with an increase in variability as the window size m increased. Regarding missing data imputation, LI and KS outperformed other methods.The time series reconstructed using LOCF, LI, and EMA had the best estimates for $F^2_{k,\text{DFA}}$, $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ across different values of m and ρ. Therefore, these observations suggest that for this scenario, the methods that excel in missing data imputation might not provide more accurate estimates for the DFA and DCCA functions.

5.4 *Discussion of the simulation results*

Regarding the simulation results with complete time series, the sample estimators of $F^2_{k,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$ are both very close to their expected values, especially when m is small. Consequently, ρ_{DCCA} estimator behaves closely to its theoretical counterpart. These analyses have been previously presented in Prass and Pumi (2021), and it is coherent that the simulations in this study have yielded similar results, that is, for all values of $m \in \{3, 5, \cdots, 101\}$, the median estimate of ρ_{DCCA} is always very close to the expected values. As m increases, the variances of $F^2_{k,\text{DFA}}(m)$, $F_{\text{DCCA}}(m)$, and ρ_{DCCA} increase. This is expected since $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ are averages calculated from certain quantities obtained by considering boxes of size $m+1$. Since m determines the size of the boxes, the higher the m, the smaller the number of boxes available, hence, the smaller the number of terms used in calculating $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$ leading to an increase in variance.

Addressing Q1 and Q4, the optimal imputation depended both on the underlying process and the proportion of missing data. The average-based methods excel when the underlying processes are sequences of uncorrelated variables or when the autocovariance is non-zero only for lag $h = 1$. For scenarios where the surrounding observations are the ones that contribute the most to predicting X_t, the methods LI, KS and EMA had superior results. Interestingly, in response to Q5, non-average-based methods exhibit a significant decline in performance with higher proportions of missing values, for $\rho = 0.8$, average-based methods yield comparable outcomes in scenarios dominated by LI, KS, and EMA.

When examining the results related to the reconstructed time series, a noticeable difference emerges in comparison to the values obtained from the complete time series. This difference becomes more pronounced with increasing proportions of missing values (Q2, Q5). However, even in the scenarios where the estimates are not close to the expected values, the linear decay os $F^2_{k,\text{DFA}}(m)$ and $F_{\text{DCCA}}(m)$, as $m \to \infty$, still holds. Concerning the $F^2_{k,\text{DFA}}$ functions, the imputation methods that achieved results closest to those calculated with complete time series follow a logic close to that observed in the missing imputation. Average-based methods perform better when the autocovariance is zero for all lags $h > 1$ and methods such as LI, KS, and EMA excel in scenarios with non-zero autocovariance. Notably, LOCF stands out among the top-performing methods for most scenarios when $m \in \{27, 81, 101\}$. There is no evident pattern for the $F_{\text{DCCA}}(m)$ and $\rho_{\text{DCCA}}(m)$ results given the time series reconstruction method used. LOCF, LI, and EMA consistently deliver good results across all scenarios, yet the optimal methods vary depending on the specific values of ρ and m in each scenario. Average-based methods are most effective when the expected value is close to or equals zero. Consequently, the best imputation method does not necessarily yield results closer to the expected values for the DFA and DCCA functions. While this statement holds true for most scenarios regarding $F^2_{k,\text{DFA}}$, it does not extend well to F_{DCCA} and ρ_{DCCA}.

Finally, which method should be used for time series imputation to calculate the DFA and DCCA functions? It depends. In a quick response, LI emerges as a favorable choice due to its simplicity, consistent performance even in scenarios where it may not be the top-performing method, and reliability in both imputation and the estimates of the DFA and DCCA functions. However, in certain scenarios, average-based methods may prove to be superior choices in terms of both median performance and variability of estimators. Therefore, the selection of an imputation method should be guided by a priori knowledge of the correlation structure, the proportion of missing values, and the window size of interest.

Chapter 6

Conclusions and Future Work

This work discussed the generating mechanisms of missing values, including a didactic and humorous example illustrating the difference among these probabilistic models and described various methods for handling missing values. During this master's thesis, the missing data were considered to be MCAR and the methods used to adress this problem were imputation methods.

This work described fundamental concepts about the decision tree structure and provided a review on usual several decision trees algorithms and some algorithms that alter the traditional structure of decision trees by introducing "soft" decisions at internal nodes. CART and an implementation of a modified version of the Probabilistic Regression Tree algorithm modified to handle missing data were used as imputation methods. The PRTree algorithm worked efficiently and had very promising results for the fitting of smooth functions. It is currently avaiable for download at https://cran.r-project.org/package=PRTree

The evolution of the DFA and DCCA functions is traced, supplemented with various application examples and theoretical results, including works that proposed methods to handle missing data in the context of time series with long-range dependence. This work closely followed the definitions and notations outlined in Prass and Pumi (2021), which provided a more theoretical perspective and unique results that were crucial for evaluating the results found. It has contributed with an asymptotic result for the covariance and cross-covariance functions corresponding to processes with missing values imputed using the mean, along with a comprehensive study of Monte Carlo simulations exploring the behavior of $F^2_{k,\text{DFA}}$, F_{DCCA}, and ρ_{DCCA} in the context of processes with short-term dependence and a varying the proportion of missing data from 10% to 80% with different imputation methods. The initial case study indicated that the matrix $K_{m+1}\Gamma_k$ exhibited the same structure in both complete and missing data cases. However, the values on the main diagonal of the matrix were generally smaller in cases with missing data, resulting in a reduced trace of the matrix (utilized to compute the expected values of DFA and DCCA functions) in the context of processes with missing values. Monte Carlo simulations provided evidence that the optimal methods for reconstructing time series, with respect to the estimates of DFA and DCCA, depend on the correlation structure of the underlying processes, the proportion of missing values, and the window size used to calculate these quantities. Average-based methods performed well when covariances were non-zero only in lags ≤ 1, while LI, KS, and EMA methods excelled when covariances extended beyond lag 1. LI demonstrated the most consistent overall performance, proving to be a simple and effective method for both imputation and estimating $F^2_{k,\text{DFA}}$, F_{DCCA}, and ρ_{DCCA}

In Alkhoury et al. (2020), the authors consider that the function ψ depends on a parameter

vector $\boldsymbol{\sigma}$ associated with the input residuals, without taking into account the potential correlation between the independent variables. Future plans involve generalizing the existing model by replacing the vector $\boldsymbol{\sigma}$ with a covariance matrix Σ and considering joint distribution functions, such as the multivariate Gaussian. Another potential improvements to the algorithm include adopting different strategies to select split points, incorporating additional native methods for handling missing data, and developing a way to include categorical variables in the algorithm using an appropriate distribution. Furthermore, the asymptotic result on the autocovariance and cross-covariance matrices in the case of mean imputation is a novelty in the literature but can be extended, both to other missing data imputation methods and to asymptotic results of DFA and DCCA. Finally, Monte Carlo simulations can be conducted considering different sample sizes and other scenarios of missingness, such as only one of the time series having missing values and both time series having the same indexes of missing data, which would lead to the same number of observations used for the calculation of the DFA and DCCA functions.

Bibliography

Alkhoury, S., Devijver, E., Clausel, M., Tami, M., Gaussier, E., Oppenheim, g., 2020. Smooth and consistent probabilistic regression trees, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 11345–11355. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/8289889263db4a40463e3f358bb7c7a1-Paper.pdf.

Bardet, J.M., Kammoun, I., 2008. Asymptotic properties of the detrended fluctuation analysis of long range dependence processes. IEEE Transactions on Information Theory, Institute of Electrical and Electronics Engineers 54, 2041–2052.

Batista, G., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17, 519–533.

Blythe, D.A.J., 2013. A rigorous and efficient asymptotic test for power-law cross-correlation. ArXiv:1309.4073.

Blythe, D.A.J., Nikulin, V.V., Müller, K.R., 2016. Robust statistical detection of power-law cross-correlation. Scientific Reports 6, 27089.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Greiner, R., Grove, A., Kogan, A., 1997. Knowing what doesn't matter: exploiting the omission of irrelevant data. Artificial Intelligence 97, 345–380.

Grewal, M.S., 2011. Kalman Filtering. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics, Springer New York.

Josse, J., Prost, N., Scornet, E., Varoquaux, G., 2019. On the consistency of supervised learning with missing values. arXiv:1902.06931 .

Kantelhardt, J.W., Koscielny-Bunde, E., Rego, H.H., Havlin, S., Bunde, A., 2001. Detecting long-range correlations with detrended fluctuation analysis. Physica A: Statistical Mechanics and its Applications 295, 441–454.

Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 20, 119–127.

Leo Breiman, Jerome Friedman, C.J.S., Olshen, R., 1984. Classification and Regression Trees. Chapman and Hall/CRC.

Linero, A.R., Yang, Y., 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. arXiv:1707.09461.

Little, R.J.A., 1995. Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90, 1112–1121.

Løvsletten, O., 2017. Consistency of detrended fluctuation analysis. Physical Review E 96.

Luukkonen, R., Saikkonen, P., Terasvirta, T., 1988. Testing linearity against smooth transition autoregressive models. Biometrika 75, 491–499. arXiv:https://academic.oup.com/biomet/article-pdf/75/3/491/641068/75-3-491.pdf.

Marinho, E., Sousa, A., Andrade, R., 2013. Using Detrended Cross-Correlation Analysis in geophysical data. Physica A: Statistical Mechanics and its Applications 392, 2195–2201.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., Verbeke, G., 2020. Handbook of Missing Data Methodology. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Taylor & Francis Group.

Moritz, S., Bartz-Beielstein, T., 2017. imputeTS: Time Series Missing Value Imputation in R. The R Journal 9.

Nakagawa, S., Freckleton, R.P., 2008. Missing inaction: the dangers of ignoring missing data. Trends in ecology & evolution 23, 592–596.

Neimaier, A.S., Prass, T.S., 2023. Missing values imputation in time series using decision trees. Submitted .

Neimaier, A.S., Prass, T.S., 2024. PRTree: Probabilistic Regression Trees. URL: https://CRAN.R-project.org/package=PRTree. r package version 0.1.0.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L., 1994. Mosaic organization of dna nucleotides. Physical Review E 49, 1685–1689.

Podobnik, B., Stanley, H.E., 2008. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. Physical Review Letters 100.

Prass, T.S., Pumi, G., 2020. DCCA: Detrended Fluctuation and Detrended Cross-Correlation Analysis. URL: https://CRAN.R-project.org/package=DCCA. r package version 0.1.1.

Prass, T.S., Pumi, G., 2021. On the behavior of the DFA and DCCA in trend-stationary processes. Journal of Multivariate Analysis 182, 104703.

Pratama, I., Permanasari, A., Ardiyanto, I., Indrayani, R., 2016. A review of missing values handling methods on time-series data, in: 2016 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 1–6.

Quinlan, J.R., 1986. Induction of Decision Trees. Machine Learning 1, 81–106.

Correa da Rosa, J., Veiga, A., Medeiros, M., 2008. Tree-structured smooth transition regression models. Computational Statistics & Data Analysis 52, 2469–2488.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M., 2016. Missing Data. Springer International Publishing, Cham.

Therneau, T., Atkinson, B., 2019. rpart: Recursive Partitioning and Regression Trees. URL: https://CRAN.R-project.org/package=rpart. r package version 4.1-15.

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A., 2015. Scikit-learn: Machine learning without learning the machinery. GetMobile: Mobile Computing and Communications 19.

Wilson, P.S., Tomsett, A.C., Toumi, R., 2003. Long-memory analysis of time series with missing values. Physical Review E 68, 017103.

Zebende, G., 2011. DCCA cross-correlation coefficient: Quantifying level of cross-correlation. Physica A: Statistical Mechanics and its Applications 390, 614–618.

Zebende, G., Brito, A., Castro, A., 2020. DCCA cross-correlation analysis in time-series with removed parts. Physica A: Statistical Mechanics and its Applications 545.

İrsoy, O., Yıldız, O.T., Alpaydın, E., 2012. Soft decision trees, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1819–1822.