NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

## MULTIVARIATE MARKOV CHAINS

When Categorizing Costs You Accuracy: Trade-offs in Applying Multivariate Markov Chains

Guilherme de Almeida Curioso

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade NOVA de Lisboa

# MULTIVARIATE MARKOV CHAINS

by

Guilherme de Almeida Curioso

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

**Supervised by**:

Bruno Damásio, NOVA University Lisbon
Carolina Vasconcelos, NOVA University Lisbon

July, 2025

**STATEMENT OF INTEGRITY**

I, Guilherme Curioso, hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 20250715

**Multivariate Markov Chains**
**When Categorizing Costs You Accuracy: Trade-offs in Applying Multivariate Markov Chains**



This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.3.9) (Lourenço, 2021).

# Acknowledgements

I would like to express my appreciation to my supervisor, Professor Bruno Damásio, for overseeing this work and being part of the process that led to its completion. I am also sincerely thankful to my co-supervisor, Professora Carolina Vasconcelos, whose support arrived at the right moments, especially when it was most needed in the final stages of this work.

Finally, I would like to thank my family and friends for their patience and encouragement during a period marked by exhaustion and frustration. Your presence helped me keep going, even when things felt overwhelmingly unproductive.

# Abstract

Multivariate Markov Chains (MMCs) are increasingly explored as tools for modeling and forecasting categorical time series. However, their application to real-world data typically requires an initial discretization step, as most available data is continuous in nature. This transformation raises an essential trade-off: while discretization allows the use of MMCs and offers potential advantages in terms of model interpretability and robustness, it can also degrade performance by discarding relevant information from the original signal. This thesis investigates the consequences of discretizing continuous variables for the purpose of applying MMCs to forecasting tasks. Through a comprehensive Monte Carlo simulation framework, we assess how different discretization strategies — including Equal Width, Equal Frequency, and K-means — affect both information retention and forecast accuracy. Comparisons are made between discretized traditional linear models (AR and VAR) and their Markov-based counterparts (MC and MTD-Probit MMC), across varying levels of autocorrelation, bin number, and forecast horizons. Forecast performance is evaluated using RMSE, while information loss is quantified through entropy-based metrics including Mutual Information, Information Loss Ratio (ILR), and Information Change Ratio (ICR). The results highlight that while more refined discretizations can reduce information loss, they do not necessarily translate into improved forecasting accuracy. The findings reveal a fundamental limitation: even under optimized discretization conditions, MMCs rarely outperform continuous models in terms of predictive accuracy. Nonetheless, the framework developed here provides a valuable basis for future research on discretization-aware model design and multivariate time series analysis. These insights are particularly relevant for researchers considering MMCs in real-world forecasting applications, as they underscore the importance of carefully balancing model interpretability with the potential loss of predictive performance due to discretization.

**Keywords:** Multivariate Markov Chains, Time Series Forecasting, Discretization, Information Loss

## Sustainable Development Goals (SDG):

# Resumo

As Cadeias de Markov Multivariadas (CMMs) têm sido cada vez mais exploradas como ferramentas para modelar e prever séries temporais discretas. No entanto, a sua aplicação a dados do mundo real geralmente exige uma etapa inicial de discretização, dado que a maioria dos dados disponíveis são de natureza contínua. Essa transformação levanta um *trade-off* essencial: embora a discretização permita o uso de CMMs e ofereça potenciais vantagens em termos de interpretabilidade e robustez do modelo, esta pode também degradar o desempenho ao descartar informações relevantes do sinal original. Esta dissertação investiga as consequências da discretização de variáveis contínuas com o objetivo de aplicar CMMs em tarefas de previsão. Por meio de um abrangente quadro de simulação de Monte Carlo, avaliamos como diferentes estratégias de discretização — incluindo Intervalos de Largura Igual (*Equal Width*), Frequência Igual (*Equal Frequency*) e K-means — afetam tanto a retenção de informação quanto a precisão preditiva. São feitas comparações entre modelos lineares tradicionais (AR e VAR) discretizados e os seus equivalentes baseados em Cadeias de Markov (CM e CMM MTD-Probit), sob diferentes níveis de autocorrelação, número de intervalos e horizontes de previsão. O desempenho preditivo é avaliado com base no RMSE, enquanto a perda de informação é quantificada por meio de métricas baseadas em entropia, incluindo Informação Mútua, Taxa de Perda de Informação e Taxa de Alteração de Informação. Os resultados mostram que, embora discretizações mais refinadas possam reduzir a perda de informação, isso não se traduz necessariamente em maior precisão preditiva. Os resultados revelam uma limitação fundamental: mesmo sob condições de discretização otimizadas, as CMMs raramente superam modelos contínuos em termos de desempenho preditivo. Ainda assim, a estrutura desenvolvida nesta dissertação oferece uma base importante para futuras investigações sobre modelos sensíveis à discretização e análise de séries temporais multivariadas. Estes resultados são particularmente relevantes para investigadores e profissionais envolvidos em tarefas de previsão, pois evidenciam a importância de equilibrar cuidadosamente a interpretabilidade do modelo com a possível perda de desempenho preditivo proveniente da discretização.

**Palavras-chave:** Cadeias de Markov Multivariadas, Previsão de Séries Temporais, Discretização, Perda de Informação

# Contents

**List of Figures** x

**1 Introduction** 1

**2 Theoretical Framework** 3
- 2.1 Preliminary Concepts . . . 3
- 2.2 Related Studies . . . 6
- 2.3 Literature Review . . . 11
- 2.4 Discretization . . . 13
  - 2.4.1 Simple Discretization Methods . . . 13
  - 2.4.2 Symbolic Aggregate Approximation . . . 14
- 2.5 Information Loss . . . 16
  - 2.5.1 Shannon Entropy . . . 16
  - 2.5.2 Kozachenko-Leonenko Estimator . . . 17
  - 2.5.3 Mutual Information Estimation . . . 17
  - 2.5.4 Information Loss Rate . . . 18
  - 2.5.5 Information Change Rate . . . 19

**3 Data and Methods** 20
- 3.1 Methodology . . . 20
- 3.2 Univariate Monte Carlo Simulation Setup . . . 20
- 3.3 Multivariate Monte Carlo Simulation Setup . . . 21
- 3.4 Tools and Implementation Environment . . . 23

**4 Results and Discussion** 25
- 4.1 Monte Carlo Simulation Results . . . 25
  - 4.1.1 Forecast Accuracy . . . 25
  - 4.1.2 Effect of Discretization and Information Loss . . . 32
- 4.2 Key Findings . . . 40
  - 4.2.1 Univariate Results . . . 40

4.2.2 Multivariate Results . . . . . 41

**5 Conclusion** **43**
5.1 Overview . . . . . 43
5.2 Key Findings and Limitations . . . . . 43
5.3 Future Work . . . . . 44

**Bibliography** **46**

**Appendices**

**A Additional Results** **52**
A.1 Forecast RMSE . . . . . 52
A.1.1 Univariate Results . . . . . 52
A.1.2 Multivariate Results . . . . . 57
A.2 Information Loss Results . . . . . 62
A.3 Other Plots . . . . . 63

# List of Figures


2.1 SAX Representation . . . . . 14

4.1 RMSE Comparison of AR(1) for Low Cor and High Cor Scenarios. . . . . 27
4.2 RMSE Comparison of Markov Chain for Low Cor and High Cor Scenarios. 28
4.3 RMSE Comparison of VAR for Low Cor and High Cor Scenarios. . . . . 30
4.4 RMSE Comparison of MMC for Low Cor and High Cor Scenarios. . . . . 31
4.5 Mutual Information estimates under different discretization methods for low (left) and high (right) correlation settings. . . . . 33
4.6 ILR across discretization methods for low (left) and high (right) correlation settings . . . . . 34
4.7 ICR across discretization methods for low (left) and high (right) correlation settings. . . . . 35
4.8 Multivariate Z1: Mutual Information estimates under different discretization methods, comparing low (left) and high (right) correlation settings. 37
4.9 Multivariate Z1: ILR values across discretization methods under low (left) and high (right) correlation . . . . . 38
4.10 Multivariate Z1: ICR values across discretization methods under low (left) and high (right) correlation. . . . . 39

A.1 Entropy of Discretized Variable $H(X_d)$ under Different Correlation Scenarios 62
A.2 AR(1) distributions with 9-bin discretizations using EFI, EWI, and K-means methods. . . . . 63
A.3 MTD-Probit conditional probabilities for each state $z_{1,t}$ (7 state example). 64

# 1

# Introduction

The discretization of continuous data is a fundamental yet often overlooked challenge in time series forecasting. While countless forecasting models have been developed across disciplines, from neural networks in finance to autoregressive processes in econometrics, many assume direct access to continuous, well-behaved inputs. However, some modeling frameworks, such as Markov Chains, require categorical inputs, making discretization an essential preprocessing step. Forecasting remains a cornerstone of data-driven decision-making, supporting activities in Finance and Economics (Elliott & Timmermann, 2016), Meteorology (Ramage, 1993), Public Health (Soyiri & Reidpath, 2013), and beyond. Accurate prediction of future events or values enables organizations to anticipate risk, allocate resources, and adapt to dynamic environments. Each modeling approach comes with its own strengths and limitations, often trading off complexity, interpretability, and data requirements.

Markov Chains (MCs) and Multivariate Markov Chains (MMCs) are particularly well suited for problems where the system evolves through a finite number of states. However, this discrete nature imposes a significant modeling constraint: when working with continuous data, the first step must be discretization. This transformation, while necessary for applying MMCs, raises an important trade-off between simplifying model structure and losing valuable information from the original signal.

Several discretization methods have been developed across fields such as pattern recognition (Baek & Young, 2017), noise reduction (Pal et al., 2022), and time series analysis (Moskovitch & Shahar, 2015). However, the majority of these techniques have been designed with supervised classification tasks in mind, where discretization is guided by the goal of improving predictive accuracy, with respect to a known target variable, and reducing computacional cost. As a result, their effectiveness in unsupervised or forecasting contexts (where no target variable is available) remains less well understood and often underexplored. This is partly because discretization is generally less suitable in regression settings, where the response variable is continuous, and because many discretization techniques assume smooth, well-behaved data distributions — an assumption that often fails in real-world time series, which can be noisy, irregular,

and exhibit abrupt changes. However, regardless of method, converting continuous variables into categorical ones inevitably implies some degree of information loss. This is especially problematic when the objective is forecasting, as reduced signal fidelity can degrade model performance.

Although MMC models have been explored in areas ranging from economics to finance (D'Amico & de Blasis, 2019), relatively few works have systematically addressed the consequences of discretization in these settings. In particular, the trade-off between forecast accuracy and information retention has not been thoroughly quantified. This thesis seeks to fill that gap by directly addressing the research question: "When categorizing costs you accuracy, to what extent do the benefits of applying Multivariate Markov Chains outweigh the information loss introduced by discretization?" Through this lens, we evaluate the practical viability of MMCs in forecasting tasks, balancing their interpretability and robustness against potential decline in predictive performance.

To achieve this, we conduct a series of controlled Monte Carlo simulations. In the univariate setting, we compare traditional autoregressive (AR) models with first-order Markov Chains. In the multivariate case, we assess the performance of Vector Autoregressive (VAR) models and MTD-Probit MMCs. Forecasting accuracy is measured using RMSE across multiple horizons, while information retention is evaluated using entropy, mutual information, and derived metrics like the Information Loss Ratio (ILR) and Information Change Ratio (ICR).

The remainder of this thesis is structured as follows. Chapter 2 provides the theoretical foundation, introducing key concepts in time series modeling, discretization, and information theory, as well as a review of relevant literature on MMCs and its comparison with AR/VAR models. Chapter 3 outlines the simulation design, detailing the methodology used to assess the impact of discretization on forecast performance and information loss in both univariate and multivariate settings. Chapter 4 presents and discusses the empirical results, comparing different models and discretization methods under various configurations. Finally, Chapter 5 summarizes the main findings, discusses their limitations, and suggests potential directions for future research.

The key contribution of this thesis lies in providing an empirical framework to study under what conditions, MMCs offer a viable alternative to continuous models, given the unavoidable cost of discretization.

# 2

# Theoretical Framework

## 2.1 Preliminary Concepts

In order to understand the models used throughout this thesis, we begin by presenting foundational stochastic modeling tools. First, we introduce Markov processes, which provide the theoretical basis for both Markov Chains (MCs) and Autoregressive (AR) models. We then extend the discussion to multivariate formulations: Vector Autoregressive (VAR) models and Multivariate Markov Chains (MMCs), which serve as competing frameworks for multivariate time series forecasting.

A Markov process is a stochastic process that satisfies the Markov property, meaning that the future evolution of the process depends only on its present state, and not on the sequence of states that preceed it.

Formally, a stochastic process $\{X_t\}_{t\geq 0}$ is a Markov process if:

$$P(X_{t+1} \in A \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = P(X_{t+1} \in A \mid X_t = x_t) \tag{2.1}$$

for all measurable sets $A$, times $t \geq 0$, and all states $x_0, x_1, \ldots, x_t$.

Depending on the context, Markov processes can be either:

- Discrete-time or continuous-time, depending on whether the index $t$ takes values in $\mathbb{N}$ or $\mathbb{R}_0^+$ respectively,
- Discrete-state or continuous-state, depending on the nature of the state space.

Markov chains (MC) are a particular case of Markov processes in which time is discrete and the state space is finite or countable.

A discrete-time Markov chain is a sequence of random variables $\{X_t\}_{t\geq 0}$ defined on a state space $S$ such that:

$$P(X_{t+1} = s' \mid X_t = s, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = P(X_{t+1} = s' \mid X_t = s)$$

for all states $s_0, s_1, \ldots, s, s' \in S$ and all times $t \geq 0$. This is known as the *Markov property*.

If $S$ is finite or countable, the transition probabilities can be represented by a transition probability matrix $P$, where each element is given by:

$$P_{ij} = P(X_{t+1} = j \mid X_t = i), \quad \forall i, j \in S$$

In other words, $P_{ij}$ represents the probability of transitioning from state $i$ to state $j$ in one time step.

Assuming a finite-state Markov chain with $n$ possible states labeled from 1 to $n$, the transition probability matrix can be written as:

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix}$$

A Markov chain is fully specified by the transition probability matrix $P$ and its initial distribution $\pi^{(0)}$, where $\pi_i^{(0)} = P(X_0 = i)$.

The use of Markov Chain models has been studied in many areas such as Biology (Berchtold, 2001; Gottschau, 1992; Raftery & Tavaré, 1994), Economics (Mehran, 1989), Finance (Fung & Siu, 2012; Siu et al., 2005), Physics (Boccaletti et al., 2014; Gómez et al., 2010), Sports (Bukiet et al., 1997) and others (Asadabadi, 2017; Horvath et al., 2005; Maskawa, 2003; Sahin & Sen, 2001; Sericola, 2013; Tsiliyannis, 2018; Turchin, 1986).

As researchers explored more complex dependencies, higher-order Markov chains were presented. A Higher-Order Markov Chain (HOMC) is a generalization of a first-order Markov chain, in which the future state depends on multiple previous states, rather than just the immediate past state.

A discrete-time $m$-th order Markov chain is a sequence of random variables $\{X_t\}_{t \geq 0}$ defined on a state space $S$ such that:

$$\begin{aligned} &P(X_{t+1} = s' \mid X_t = s, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) \\ =&P(X_{t+1} = s' \mid X_t = s, X_{t-1} = s_{t-1}, \ldots, X_{t-m+1} = s_{t-m+1}) \end{aligned}$$

for all states $s_0, s_1, \ldots, s, s' \in S$ and for all $t \geq 0$.

The transition probability for an $m$-th order Markov chain is given by:

$$P(X_{t+1} = j \mid X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_{t-m+1} = i_{t-m+1}) = P_{i_{t-m+1}, i_{t-m+2}, \ldots, i_t, j}$$

where $P_{i_{t-m+1}, i_{t-m+2}, \ldots, i_t, j}$ represents the probability of transitioning from the sequence $(i_{t-m+1}, i_{t-m+2}, \ldots, i_t)$ to state $j$.

Finally, suppose we have $s > 1$ interrelated categorical time series, where each series represents a discrete stochastic process. When the future state of a given series depends not only on its own past states (intra-transitions) but also on the past states of other series (inter-transitions), we obtain a Multivariate Markov Chain (MMC).

Formally, let us consider a multivariate discrete-time Markov process $\{(S_{1t}, \ldots, S_{st})\}_{t=1}^{T}$, where each $S_{jt}$ takes values in a countable set $E = \{1, \ldots, m\}$ for $j = 1, \ldots, s$. Without loss of generality, we assume a first-order MMC, meaning that:

$$P(S_{jt} = k \mid \mathcal{F}_{t-1}) = P(S_{jt} = k \mid S_{1,t-1} = i_1, \ldots, S_{s,t-1} = i_s)$$

where $\mathcal{F}_{t-1}$ denotes the information set generated by all past values up to time $t-1$.

This assumption implies that, in order to predict or explain $S_{jt+1}$, past values of the process $S_{jt-i}$ for $i > 0$ are unnecessary, as the current values of all series provide sufficient information for modeling the next state.

In contrast, autoregressive time series models typically assume that the current value of a variable depends on several of its past values. An Autoregressive model of order $p$, denoted as AR($p$), is a univariate model in which the current observation is expressed as a linear combination of its $p$ most recent past values, a constant term, and a stochastic error term:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t \tag{2.2}$$

where:

- $y_t$ is the value of the time series at time $t$,
- $c$ is a constant (intercept),
- $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients,
- $u_t$ is a white noise error term.

A special case of this model is the AR(1), where the current value depends only on its immediately preceding value. The AR(1) model is given by:

$$y_t = c + \phi y_{t-1} + u_t \tag{2.3}$$

In this case:

- $\phi$ captures the influence of the past value $y_{t-1}$ on the current value $y_t$.

If $|\phi| < 1$, the process is stationary, meaning that shocks dissipate over time and the process has a constant long-term variance.

AR(1) models satisfy the Markov property under the assumption that the state space is continuous and that the conditional distribution of $y_t$ depends only on $y_{t-1}$.

A Vector Autoregressive model of order $p$, denoted as VAR($p$), is a multivariate time series model used to capture the linear interdependencies among multiple time series. In a VAR($p$) model, each variable is expressed as a linear function of its own past values and the past values of all other variables in the system, up to $p$ lags.

Let $\mathbf{y}_t \in \mathbb{R}^k$ be a $k \times 1$ vector of variables at time $t$. A VAR($p$) model is defined as:

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \cdots + \Phi_p \mathbf{y}_{t-p} + \mathbf{u}_t \tag{2.4}$$

where:

- $\mathbf{c}$ is a $k \times 1$ vector of intercepts,
- $\Phi_i$ for $i = 1, \ldots, p$ are $k \times k$ coefficient matrices,
- $\mathbf{u}_t$ is a $k \times 1$ vector of white noise error terms, possibly contemporaneously correlated.

A special case is the VAR(1) model with $k = 2$ variables, say $y_{1t}$ and $y_{2t}$. It can be written in matrix form as:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{2.5}$$

This corresponds to the following system of equations:

$$y_{1t} = c_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + u_{1t} \tag{2.6}$$

$$y_{2t} = c_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + u_{2t} \tag{2.7}$$

where:

- $c_1, c_2$ are constants (intercepts),
- $\phi_{ij}$ are the autoregressive coefficients,
- $u_{1t}, u_{2t}$ are white noise error terms, potentially contemporaneously correlated but serially uncorrelated.

In this setting:

- $\phi_{11}$ and $\phi_{22}$ capture the influence of each variable's own past value on its current value,
- $\phi_{12}$ and $\phi_{21}$ represent the effect of one variable's past on the other variable's current value.

If the eigenvalues of the coefficient matrix have absolute value less than 1, the VAR(1) process is stationary. This means that shocks dissipate over time and the process has a constant long-term variance and covariance.

VAR(1) models also satisfy the Markov property under the assumption of a continuous state space, since the current vector $\mathbf{y}_t$ depends only on the previous value $\mathbf{y}_{t-1}$.

## 2.2 Related Studies

This section presents MMCs and related models. It outlines the main developments in the field, including extensions of the Mixed Transition Distribution (MTD) model, the use of MMCs as stochastic covariates, and recent approaches for forecasting and

handling structural breaks. The aim is to give an overview of how MMCs have evolved and how they are used in time series modeling and prediction tasks.

First introduced in 1985 by Raftery, the Mixed Transition Distribution (MTD) model was developed to attempt to estimate transition probabilities in high-order Markov Chains. It is defined as:

$$P(X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-\ell} = i_\ell) = \sum_{g=1}^{\ell} \gamma_g \, P(X_t = i_0 \mid X_{t-g} = i_g) = \sum_{g=1}^{\ell} \gamma_g \, q_{i_g i_0}^{(g)} \quad (2.8)$$

where:

- $i_0, \ldots, i_\ell \in \{1, \ldots, m\}$;
- $Q = [q_{ij}^{(g)}]$ is an $m \times m$ row-stochastic transition matrix (or $\ell$ such matrices if generalized);
- $\gamma = (\gamma_1, \ldots, \gamma_\ell)$ is a vector of lag weights;

Subject to the following constraints:

- $\sum_{g=1}^{\ell} \gamma_g = 1$
- $\gamma_g > 0 \quad$ for all $g = 1, \ldots, \ell$

Since then, multiple authors contributed to the extension of the MTD model: the Multimatrix MTD in Berchtold (1996) and Berchtold (1995), the Infinite-Lag MTD in Le et al. (1996) and Mehran (1989) and the General State Spaces MTD in Adke and Deshmukh (2018), Martin and Raftery (1987), and Wong and Li (2001). These models managed to surpass some limitations of the original MTD model: allowing a different $mm$ transition matrix for each lag, assuming a infinite lag order ($l = \infty$) and allowing more general processes with an arbitrary space state, respectively.

In Ching et al. (2008), the authors addressed the difficulties faced by previous models: in cases involving multiple data sequences, the state space is expanded, treating the MMC as a HOMC. They proposed a model with less parameters, using a distinct $mxm$ transition matrix for each lag, which is capable of handling more than one data sequence.

Siu et al. (2005) proposed a model that, despite its limitation in terms of the number of parameters, was straightforward to implement. Lastly, (Zhu & Ching, 2010) introduced an estimation approach based on minimizing the prediction error under a set of equality and inequality restrictions.

In Nicolau (2014), a new model for multivariate Markov chains was developed, MTD-Probit, a generalization of Raftery's MTD on the MMC model. The model yields more precise transition probability estimates than the standard MTD, and does so

without imposing any parameter constraints. Taking inspiration from the MTD model, the author proposes modelling $P_j(i_0 \mid i_1, \ldots, i_s)$ as follows:

$$P_j(i_0 \mid i_1, \ldots, i_s) = P_j^{\Phi}(i_0 \mid i_1, \ldots, i_s) = \frac{\Phi\left(\eta_{j0} + \eta_{j1} P_{j1}(i_0 \mid i_1) + \cdots + \eta_{js} P_{js}(i_0 \mid i_s)\right)}{\sum_{k=1}^{m} \Phi\left(\eta_{j0} + \eta_{j1} P_{j1}(k \mid i_1) + \cdots + \eta_{js} P_{js}(k \mid i_s)\right)}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, $\eta_{j0}, \eta_{j1}, \ldots, \eta_{js} \in \mathbb{R}$ are parameters estimated by maximum likelihood and the parameters $P_{jk}(i_0 \mid i_1)$, for $k = 1, \ldots, s$, can be estimated in advance through the consistent estimators:

$$\hat{P}_{jk}(i_0 \mid i_1) = \frac{n_{i_1 i_0}}{\sum_{i_0=1}^{n} n_{i_1 i_0}} \tag{2.9}$$

where $n_{i_1 i_0}$ is the number of transitions from $S_{k,t-1} = i_1$ to $S_{jt} = i_0$.

The model comes with several other remarks, in which the following stand out:

1. The numerator utilizes the same logic as the original MTD model, where $\Phi(\cdot)$ is a linear combination of probabilities.

2. As mentioned before, the model is free from constraints since $P_j^{\Phi}(i_0 \mid i_1, \ldots, i_s)$ lies within the interval (0, 1).

3. The likelihood function for the MTD-Probit model is given by:

$$\log L = \sum_{i_1 i_2 \ldots i_s i_0} n_{i_1 i_2 \ldots i_s i_0} \log\left(P_j^{\Phi}(i_0 \mid i_1, \ldots, i_s)\right) \tag{2.10}$$

   and the maximum likelihood estimator is defined by

$$\hat{\boldsymbol{\eta}}_j = \underset{\eta_{j1}, \ldots, \eta_{js}}{\operatorname{argmax}} \log L.$$

4. A constant term $\eta_{j_0}$ was introduced in this model and it generally allows for $P_j^{\Phi}(i_0 \mid i_1, \ldots, i_s)$ to be closer to $P_j(i_0 \mid i_1, \ldots, i_s)$, when comparing to the MTD case. However, it can be set to zero.

A new concept was introduced in Damásio and Nicolau (2014): the use of MMCs as stochastic covariates. This approach leverages the structure of MMCs to improve forecasts of a target variable by modeling its categorical (or discrete) regressors. By capturing the historical interactions between the states of these regressors through an MMC, it can better predict their future states and improve the forecast of the actual dependent variable.

In order to incorporate the multivariate Markov chain (MMC) structure into a regression framework, each possible state of the MMC must be represented numerically. Since the regressors involved in the model are categorical or have been discretized, the set of all possible states of the multivariate process forms a finite state space, which can be mapped into dummy variables suitable for regression analysis.

To form a regression model relating a dependent variable $y_t$ to the categorical (or discretized) variables that form the multivariate Markov chain (MMC), we begin by converting each regressor $S_{jt}$ into a set of dummy variables, as follows:

$$z_{jkt} = \mathbb{I}\{S_{jt} = k\} \tag{2.11}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function: $\mathbb{I}\{S_{jt} = k\} = 1$ if $S_{jt} = k$, and 0 otherwise. Here, $z_{jkt}$ is a binary variable that indicates whether the $j$-th regressor is in state $k$ at time $t$.

The methodology also supports the case where $S_{jt}$ is already a discrete variable with a natural numeric interpretation and state space $\{1, 2, \ldots, m\}$, in which case no dummy variables are needed.

Specifically, the model assumes a linear relationship of the form:

$$y_t = x_t^\top \beta + z_t^\top \delta + u_t \tag{2.12}$$

where $x_t$ is a vector of deterministic or stochastic covariates (e.g. AR(1)), $z_t$ is a vector of dummy variables representing the categorical or discretized states of the MMC at time $t$, $\beta$ and $\delta$ are vectors of unknown parameters, and $u_t$ is a mean-zero error term assumed to be uncorrelated with the regressors.

To forecast the future value $y_{t+h}$, one must account for the fact that the future states $S_{jt+h}$ are unknown at time $t$. Therefore, the forecast is formed by taking the conditional expectation:

$$\mathbb{E}(y_{t+h} \mid \mathcal{F}_t) = \mathbb{E}(x_{t+h}^\top \beta \mid \mathcal{F}_t) + \mathbb{E}(z_{t+h}^\top \delta \mid \mathcal{F}_t) \tag{2.13}$$

where $x_{t+h}$ is a vector of known regressors, $z_{t+h}$ is the vector of dummy variables associated with the MMC states at time $t+h$ and where $\mathcal{F}_t$ represents the sigma-algebra generated by all observed data up to time $t$. Since the error term $u_t$ is assumed to be white noise and mean-independent of the regressors, the forecast simplifies to the expected value of the linear predictor.

The crucial step in evaluating $\mathbb{E}(z_{t+h} \mid \mathcal{F}_t)$ lies in computing the probabilities:

$$\mathbb{E}(z_{jkt+h} \mid \mathcal{F}_t) = \mathbb{P}(S_{jt+h} = k \mid \mathcal{F}_t) \tag{2.14}$$

which represent the $h$-step-ahead predictive probabilities of the MMC components.

In time series analysis, an $h$-step-ahead forecast refers to the prediction of the value of a variable at time $t + h$, based on information available up to time $t$. The objective is to estimate the conditional expectation $\mathbb{E}(y_{t+h} \mid \mathcal{F}_t)$. For $h = 1$, the forecast relies solely on the most recent observation and the model's one-step-ahead dynamics. As $h$ increases, the forecast becomes multi-step, requiring either iterative methods (where one-step-ahead forecasts are used recursively) or direct multi-step prediction models. The complexity and uncertainty of the forecast typically grow with $h$, as errors can accumulate across iterations and the influence of initial conditions tends to decay over time. In iterative forecasting approaches, the model relies on its own previous forecasts

as inputs for future predictions, meaning that each additional step ahead compounds the potential for error.

To this end, Damásio and Nicolau (2014) propose a recursive approach based on the MTD-Probit formulation. The $h$-step-ahead predictive probability is then computed as:

$$\mathbb{P}(S_{jt+h} = i_0 \mid S_{1t} = i_1, \ldots, S_{st} = i_s) = \frac{\Phi\left(\eta_{j0} + \eta_{j1} p_{j1}(i_0 \mid i_1) + \cdots + \eta_{js} p_{js}(i_0 \mid i_s)\right)}{\sum_{k=1}^{m} \Phi\left(\eta_{j0} + \eta_{j1} p_{j1}(k \mid i_1) + \cdots + \eta_{js} p_{js}(k \mid i_s)\right)}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\eta_{j0}, \ldots, \eta_{js}$ are the estimated parameters of the MTD-Probit model for component $j$, and $p_{jl}(i_0 \mid i_l)$ are the one-step transition probabilities from state $i_l$ to $i_0$. This formulation allows the $h$-step-ahead forecast to be computed recursively by propagating the state probabilities forward.

In Damásio and Nicolau (2024), the authors develop a framework for detecting multiple structural breaks in time-inhomogeneous MMCs, particularly when the change points are unknown. The authors model the evolution of categorical multivariate time series using MMCs where the transition probability matrix is allowed to vary over time, thus accommodating structural instability in the underlying process. A key contribution of the paper is a formal testing procedure designed to detect the presence and number of structural breaks in the transition dynamics. The test does not require prior knowledge of the break dates and is consistent against multiple structural changes. This makes it particularly valuable in real world scenarios such as economic regime shifts, political transitions, or market disruptions where breakpoints are not directly observable.

In Vasconcelos and Damásio (2025), the authors propose a novel generalization of Multivariate Markov Chains (MMCs) that extends the traditional autoregressive structure by incorporating exogenous covariates—either deterministic or stochastic—into the modeling framework. While standard MMCs determine future states solely from the past of the multivariate system, this approach allows transition probabilities to depend also on external variables, resulting in a non-homogeneous Markov process. The transition matrix becomes conditional on the covariates and varies over time. The model is formulated using a mixture-style structure and estimated via maximum likelihood with a multinomial logit formulation. A Monte Carlo simulation study confirms its effectiveness in detecting non-homogeneous structures. Additionally, the paper introduces an R package that enables simulation, estimation, and forecasting under this generalized MMC setting, as well as the MTD and MTD-Probit model, which will be used in this study.

## 2.3 Literature Review

This section reviews key studies that compare the forecasting capabilities and modeling assumptions of AR and VAR models with those of Markov Chain-based approaches. While AR and VAR models are widely used for modeling linear dependencies in continuous time series, they can fall short in capturing regime-switching dynamics. In contrast, MC and MMC frameworks provide a flexible structure for modeling categorical dynamics and symbolic state transitions, often proving more suitable in contexts where the process evolves in a non-linear or qualitative manner. However, most of the literature has focused on integrating autoregressive components into Markov-based models, such as in regime-switching AR or mixture models, rather than directly comparing the standalone forecasting performance of AR/VAR versus MC/MMC approaches. As a result, the comparative evaluation of these model classes remains relatively underexplored. This work aims to address that gap by systematically assessing their respective performance under various data generating conditions.

Nonetheless, some studies have conducted partial comparisons between AR/VAR and MC/MMC models, offering valuable insights into their relative strengths. Berchtold and Raftery (2002) provide a thorough evaluation of the MTD model's effectiveness in modeling high-order dependencies compared to traditional Markov chains and autoregressive frameworks. Through empirical examples such as wind direction data and epileptic seizure counts, they show that MTD models consistently achieve better BIC values than first- and higher-order Markov chains, while requiring fewer parameters.

In Damásio and Mendonça (2019), the authors present a novel framework that integrates both MMC and VAR to model the evolution of strategic competition between technological systems, such as insurgent versus incumbent paradigms. The authors argue that while VAR models are effective at capturing short run interdependencies and lagged effects among continuous variables (e.g., economic or technological indicators), they fall short in representing qualitative, regime switching behavior that often characterizes real world strategic dynamics. To address this, the paper proposes using an MMC to model transitions between latent strategic states, such as dominance or competition, allowing for the modeling of symbolic and state dependent dynamics. The MMC captures the discrete and qualitative shifts in system behavior, while VAR accounts for continuous, quantitative dependencies within regimes.

Eberle et al. (2022) investigate the effectiveness of multivariate simulation techniques for offshore weather time series. The authors evaluate several modeling approaches including first and second-order MC, VAR models, and LSTM neural networks on their ability to replicate wind and wave conditions. Their results show that second-order MC and VAR models both achieve strong performance in capturing the distributional and temporal structure of the original time series, while first-order MC are less effective. Notably, the study highlights the complementary strengths of MC and AR models: while MC excel in preserving the transition dynamics and symbolic structure of the

data, VAR models better reproduce smooth linear dependencies.

Wiecek and Kubek (2024) examine the impact of time series characteristics on the forecasting performance of different models in the context of fuel demand prediction. The authors compare ARIMA and SARIMA (AR-based models) and MC models using real-world data from 85 gas stations, assessing their accuracy across a wide range of demand profiles. Their empirical results indicate that while ARIMA and SARIMA models can achieve low forecast errors in specific cases, they are more sensitive to nonstationary behaviors and outliers. In contrast, MC models demonstrate greater robustness across varying series and more consistent error distributions. The study emphasizes the unique strengths of each modeling approach: while autoregressive models are well-suited for capturing smooth and continuous variations in the data, MC are more effective at representing discrete transitions and shifts in demand, particularly in time series that exhibit high volatility or irregular behavior.

Despite the contributions of prior studies, the role of discretization is often treated as a technical pre-processing step rather than a core modeling decision. Most comparisons between Markov-based models and AR/VAR approaches do not explicitly quantify how the choice of discretization method impacts information retention or forecast performance. This leaves a significant gap, especially in contexts where models like MMCs require categorical inputs derived from inherently continuous data.

Rather than treating discretization as a routine step, this study takes a principled approach to evaluate how the choice of unsupervised discretization strategy influences both information retention and predictive performance. Entropy-based metrics are used to quantify information loss, while RMSE is employed to assess forecast accuracy across multiple horizons. Together, these measures offer a dual perspective on the theoretical and practical impact of discretization on model performance.

To contextualize, García et al. (2013) present a detailed taxonomy of discretization techniques and highlight the diversity and complexity involved in choosing an appropriate method. They point out that the effectiveness of any subsequent modeling task can be heavily shaped by characteristics of the discretization process—such as whether it is supervised or unsupervised, global or local, or whether it follows a splitting or merging approach. Crucially, they argue that discretization should not be treated as a simple data transformation, as it involves important trade-offs between interval granularity, data consistency, information retention, and interpretability. However, much of their analysis is restricted to supervised classification tasks. Building on this foundation, this work shifts the focus toward unsupervised discretization methods and their role in time series modeling, a setting that remains relatively overlooked in existing studies. Rather than limiting the scope to classification tasks, it investigates how discretization choices affect both information retention and predictive accuracy in sequential models.

The central contribution of this work is not merely to compare AR/VAR with MC/MMC models, but to demonstrate that the effectiveness of Markov-based models

is heavily dependent on the discretization method used. In particular, some methods lead to significantly lower information loss but do not necessarily translate to better forecasts. This finding suggests that when applying MMCs, discretization is not just a preprocessing step, but a modeling choice with critical implications.

## 2.4 Discretization

All of what was discussed in the previous sections is only possible given that the variables used are discrete or categorical. If one wanted to apply MC or MMC on continuous variables, they would need to transform them into discrete variables.

Discretization or binning is a process in which continuous variables are divided or binned into measurable sections. The role of discretization is one of reducing the computational processing and also of simplifying the values obtained in a way that is easier to apply algorithms or machine learning models. It is used in areas such as Health (Acosta-Mesa et al., 2014; dos Santos Passos et al., 2017), Finance (Lkhagva et al., 2006a), Anomaly and Fraud Detection (Foorthuis, 2020), Pattern extraction (Baek & Kim, 2017; Mörchen et al., 2005), among others (Hranisavljevic et al., 2020; Overlöper et al., 2024).

### 2.4.1 Simple Discretization Methods

The most basic discretization method is Equal Width Discretization (EWD) which divides the variable's range into k equal sized intervals, where k is a parameter input by the user. Another similar method that intended to surpass the previous method's sensitivity to outliers is Equal Frequency Discretization (EFD). It divides a ordered feature into k bins where each bin contains $\frac{m}{k}$ samples (where m is the total number of values). While it solved the outlier problem, it is likely that values with the same label will be put into different bins in order to meet the $\frac{m}{k}$ restriction. It is also known as quantile discretization.

Some other methods, like K-Means, are also used for temporal data discretization, as mentioned in M. Dash et al. (2014). K-Means clustering is an unsupervised machine learning algorithm used to partition a dataset into k distinct clusters based on feature similarity. It starts by randomly initializing k cluster centroids, then iteratively assigns each data point to the nearest centroid and updates the centroids as the mean of the assigned points. This process repeats until centroids no longer change significantly or a maximum number of iterations is reached. The resulting clusters or centroids are used as the states of the discretization process. Although, K-means assigns points to clusters based on feature similarity without considering order, thus losing preservation of temporal order.

Discretization is not without its drawbacks. A significant downside is the loss of information that occurs when continuous values are replaced by a limited set of

categories. This simplification can hide meaningful patterns or subtle distinctions in the data, potentially leading to biased estimates or reduced forecasting accuracy. Additionally, the choice of discretization method and the number of intervals/bins can heavily influence model outcomes. Nonetheless, discretization models are still actively developed to this day, due to their valuable use in dimensionality reduction and simple compatibility with models that use categorical or ordinal variables.

### 2.4.2 Symbolic Aggregate Approximation

Symbolic Aggregate approXimation (SAX) is a symbolic representation of time series that can reduce a time series length to a symbolic string of lower length composed of at least 2 different symbols. The data is firstly transformed into the Piecewise Aggregate Approximation (PAA) representation and then the PAA representation is made into a symbolic discrete string. Piecewise Aggregate Approximation is a dimensionality reduction technique used to convert a time series into a lower dimensional representation while retaining its overall pattern. It works by dividing a time series of length n into w equal sized segments, where each segment represents a portion of the original data. Within each segment, the average value of the data points is calculated, reducing the series to w representative values. The final output is a compressed version of the time series, where the sequence of segment means captures the general trend. After applying PAA to obtain a lower dimensional sequence, SAX discretizes these values using Gaussian breakpoints, which divide the data into bins of equal probability under a normal distribution. Each bin is assigned a unique symbol ('a', 'b', 'c', etc.), transforming the numerical time series into a sequence of discrete symbols. In SAX, the window refers to the number of data points considered in each segment before applying PAA and the alphabet defines the number of symbols used for discretization. Each symbol represents a bin formed by the Gaussian breakpoints, and the alphabet size dictates how many different levels of discretization are available.

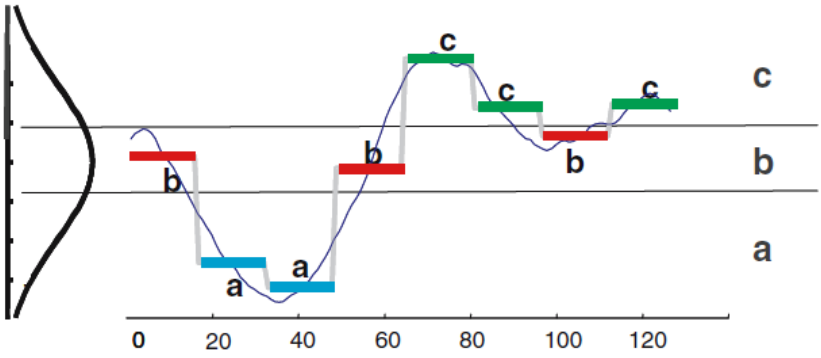



Figure 2.1: A time series is discretized by first obtaining a PAA approximation and then using predetermined breakpoints to map the PAA coefficients into SAX symbols. In the example above, with n = 128, w = 8 and a = 3, the time series is mapped to the word baabccbc (Lin et al., 2007)

This method presents two main benefits:

1. dimensionality reduction - by using PAA, SAX leverages the fact that most data mining and machine learning algorithms scale poorly with dimensionality;

2. creation of lower bounding distance measures - this guarantees that the distance computed after compression will not be smaller than the true distance between the original time series.

Since then, many other SAX-based methods have been developed using the same techniques as the original algorithm but with improved features and with a different scope in mind. Some honorable mentions are:

1. ESAX, Lkhagva et al. (2006b), introduces $min$ and $max$ points, besides the mean value for the data inside a the window, to reduce information loss and computational cost without losing the symbolic nature of the original SAX;

2. GASAX, Muhammad Fuad (2012), uses a genetic algorithm to optimize the breakpoints placement, leading to improved representation quality compared to fixed or equal probability breakpoints;

3. 1dSAX, Malinowski et al. (2013), a well known limitation of SAX is that trends are not taken into account in the symbolization, which this method manages to do by containing information about the average and trend of the series on a segment. The main advantage is the quantity of information needed to represent a time series is the same as the one needed by SAX, for a same number of symbols;

4. MSAX, Anacleto et al. (2020), extends SAX to multivariate time series and takes into account the covariance structure of the data. Overall the method does not surpass the original SAX applied independently to each attribute but is able to be better than some other techniques.

Despite the many advantages of SAX-based algorithms, they are not ideally suited for applications involving Markov Chains and time series forecasting. This is because SAX relies on the assumption that the underlying data follows a normal distribution, whereas the distributions generated by a Markov process typically do not exhibit normality. Moreover, the key strengths of the SAX algorithm can actually work against its effectiveness in the context of Markov Chains. PAA performs best when the underlying data distribution is smooth, as it summarizes each window by averaging its values. However, when the data points are highly dispersed, the average fails to accurately represent the true variability within the window. In Markov processes, such as those generated by AR or VAR models, high variance is common, leading to frequent transitions between distant states, after discretization. When these wide transitions are averaged during the PAA step, they are effectively concealed, giving

the misleading impression that most transitions occur between the middle regions (bins) of the distribution, rather than reflecting the true dynamic range of the process. Furthermore, the primary application of SAX is in time series classification, as its dimensionality reduction process aggregates multiple data points, leading to the loss of specific observations that one might want to predict. For instance, if the goal is to forecast the next 1000 values, using a window size of 4 would yield only 250 predictions, making it difficult to compare with methods that preserve the original resolution. Nevertheless, SAX is regarded as a state-of-the-art technique and is included here due to its relevance, having been considered up to the final stages of this work.

It is also important to note that all discretization methods applied in this study are unsupervised, given the absence of a target variable against which accuracy or precision could be evaluated. In contrast, supervised discretization methods — such as decision tree-based binning, entropy-based discretization or ChiSquare-based discretization (R. Dash et al., 2011) — rely on class labels or a target variable to optimize the discretization process. These approaches are therefore not applicable in the current context.

## 2.5 Information Loss

One cannot talk about discretization without mentioning information loss. As mentioned before, discretizing a variable means grouping unique observations together based on certain rules, which can sometimes miss important patterns, trends, or data points. Even with a high number of bins, turning a continuous variable into an ordinal categorical one, as done in this work, most certainly comes with some loss of detail. For that, some information loss metrics were developed in order to measure and study the trade-off between efficiency and how well the original data is preserved.

### 2.5.1 Shannon Entropy

The concept of entropy was first introduced by Claude Shannon in his seminal 1948 paper (Shannon, 1948). Shannon entropy quantifies the uncertainty or information content of a *discrete* random variable. It is defined as:

$$H(X_d) = -\sum_i p_i \log_2 p_i$$

where $X_d$ is the discretized version of a continuous variable, and $p_i$ is the probability that $X_d$ takes on the value $x_i$. The sum is taken over all possible values that $X_d$ can assume. The base 2 logarithm expresses the result in bits.

In this study, Shannon entropy is used to quantify the information content of the discretized signal. While Shannon entropy and differential entropy are not directly comparable in a strict mathematical sense due to differences in scale and interpretation, they can be used together to approximate relative information loss caused by discretization.

### 2.5.2 Kozachenko-Leonenko Estimator

Estimating the entropy of a continuous variable requires a different approach than for discrete data. The Kozachenko-Leonenko estimator in Kozachenko and Leonenko (1987) provides a non-parametric method to estimate the *differential entropy* of a continuous random variable using $k$-nearest neighbor ($k$-NN) distances. This method does not require any assumptions about the underlying distribution and is particularly effective for univariate time series data, such as AR(1) processes.

In this work, the Kozachenko-Leonenko estimator is applied to the original continuous variables $X$, providing an estimate of its entropy $H(X)$. This value serves as a baseline for evaluating how much information is lost when the signal is discretized.

### 2.5.3 Mutual Information Estimation

Mutual information is a fundamental concept in information theory that quantifies the amount of information shared between two random variables. It measures the reduction in uncertainty about one variable given knowledge of the other. For two discrete random variables $A$ and $B$, mutual information is defined as:

$$I(A;B) = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a,b)\log_2\left(\frac{p(a,b)}{p(a)\,p(b)}\right)$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the sets of possible values taken by $A$ and $B$, respectively. The term $p(a,b)$ represents the joint probability mass function of $A$ and $B$, while $p(a)$ and $p(b)$ are their respective marginal distributions. The logarithm base 2 expresses the result in bits.

In this work, mutual information is applied to evaluate how much information is preserved when a continuous variable $X$ is discretized into $X_d$. However, since mutual information is traditionally defined for pairs of discrete variables, applying it to the continuous–discrete pair $(X, X_d)$ requires specialized estimation techniques. Three different estimation methods are employed:

- Mesner and Shalizi (2020): A $k$-nearest neighbor-based estimator designed to handle mixed type variable pairs (continuous–discrete). It generalizes the Kraskov estimator, which is a widely used non-parametric method for estimating mutual information between continuous variables based on $k$-nearest neighbor statistics, as in Kraskov et al. (2004). While the original Kraskov method is limited to fully continuous data, the Mesner–Shalizi extension adapts it to the mixed variable setting. The estimator computes:

$$\hat{I}_{\mathrm{MS}}(X;X_d) = \psi(k) + \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\left[\psi(n_{x_i}+1) + \psi(n_{x_{d,i}}+1)\right]$$

where $\psi(\cdot)$ is the digamma function, $n_{x_i}$ and $n_{x_{d,i}}$ denote the number of neighbors within a certain distance in the marginal spaces of $X$ and $X_d$, and $n$ is the sample size.

- Huo et al. (2011): A kernel density estimation approach based on Gaussian mixture models. This method assumes that the conditional distribution $p(x \mid x_d)$ follows a Gaussian distribution within each discrete category of $X_d$. The mutual information is computed via:

$$I(X; X_d) = H(X) - \sum_{j=1}^{m} p_j H(X \mid X_d = j)$$

where $p_j$ is the empirical probability of each discrete level $j$, and $H(X \mid X_d = j)$ is the differential entropy of the Gaussian component associated with that bin. This formulation enables an analytical solution using known expressions for the entropy of Gaussian distributions.

- Binned Approximation: The continuous variable $X$ is first discretized into a high number $n$ of bins using either EWD or EFI (usually over 100). This transforms both $X$ and $X_d$ into discrete variables, allowing mutual information to be computed using the standard discrete Shannon formula:

$$I(X; X_d) = \sum_{i,j} p_{i,j} \log_2 \left( \frac{p_{i,j}}{p_i p_j} \right)$$

where $p_{i,j}$ is the joint empirical probability of the $i$-th bin of $X$ and the $j$-th bin of $X_d$, and $p_i$, $p_j$ are the marginal probabilities. Although this approach introduces additional discretization error, it provides a baseline that is computationally simple and interpretable.

### 2.5.4 Information Loss Rate

In Liu et al. (2002), the authors propose the Information Loss Rate (ILR) as a criterion for evaluating discretization quality. It is defined as:

$$\text{ILR} = 1 - \frac{I(X; X_d)}{H(X)}$$

where $I(X; X_d)$ is the mutual information between the original continuous variable and its discretized version. ILR reflects the proportion of the original signal's entropy that is not captured by the discretization. As such, it provides a normalized and interpretable measure of the information degradation caused by binning or quantization.

In this study, ILR is computed separately for all mutual information estimators, in order to understand how estimation assumptions influence the interpretation of information loss.

### 2.5.5 Information Change Rate

To address known limitations in ILR, particularly its tendency to underestimate information loss due to compensatory binning effects Tang et al. (2021) introduces the Information Change Rate (ICR):

$$\text{ICR} = \frac{I(X; X_d)}{H(X_d)}$$

Rather than measuring information lost from the original signal, ICR quantifies the proportion of the discretized signal's entropy that is actually explained by the original variable. This metric is discretization centric and complements ILR by focusing on the effectiveness of the symbolic representation produced.

As with ILR, this thesis reports ICR values computed using three mutual information estimators for a complete picture of discretization quality under different modeling assumptions.

# 3

# Data and Methods

## 3.1 Methodology

In this section, the main objective of this thesis will be tested by evaluating the effectiveness of the proposed methodology: are the benefits of the utilization of the MMC larger or lower than the loss of information caused by the discretization of continuous variables. To investigate this, a simulation framework is developed, beginning with a univariate analysis that compares MCs to traditional AR models. This comparison provides a foundation for later extending the approach to the multivariate setting. To provide statistically reliable results and account for variability, the analysis is conducted through a Monte Carlo simulation framework.

## 3.2 Univariate Monte Carlo Simulation Setup

To evaluate and compare forecasting performance under different discretization strategies, we design a univariate Monte Carlo simulation that contrasts traditional continuous AR models with their discretized and Markov-based counterparts. The procedure assesses both prediction accuracy and information loss, and is repeated under multiple configurations to ensure robustness.

Firstly, we simulate a univariate AR(1) process defined by $X_t = \phi X_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, 1)$, using two values for $\phi$: 0.3 and 0.7. This allows us to mimic both low and high autocorrelation scenarios, providing a broader assessment of model performance under different time series dynamics.

Each time series is generated with a total length of either 5,000 or 10,000 observations and is split into a training set (90%) and a test set (10%).

An AR(1) model is then fitted to the continuous training data, and $h$-step-ahead forecasts are computed for $h \in \{1, 2, 3, 5, 10\}$.

The training set is subsequently discretized into $n \in \{3, \ldots, 10\}$ bins using one of three unsupervised methods: Equal Width Intervalling (EWI), Equal Frequency Intervalling (EFI), or K-means Clustering. The same bin edges are applied to the test

set and to the continuous AR forecasts to ensure consistent comparison.

Forecast performance is measured using the Root Mean Squared Error (RMSE). For the AR(1) model, RMSE is calculated by comparing the discretized forecasts against the discretized test values using bin indices.

In parallel, a first-order Markov Chain is fitted to the discretized training sequence using empirical transition frequencies. Forecasts are generated recursively by selecting the most probable state at each horizon $h$, based on the fitted transition matrix and the most recent observed state. These predicted states are likewise evaluated against the discretized test set using RMSE.

To assess the information loss due to discretization, we compute the entropy of the original series using the Kozachenko–Leonenko estimator and the entropy of the discretized series using Shannon's formula. Mutual Information is estimated using three different methods: (i) Mesner–Shalizi's $k$-nearest neighbor estimator, (ii) Huo et al.'s Gaussian mixture model, and (iii) a binned approximation using fine discretization. Each mutual information estimate is then used to compute the Information Loss Ratio (ILR) and the Information Compression Ratio (ICR), resulting in three variants of each metric.

The entire simulation process is repeated across 100 Monte Carlo replications, with results averaged across runs to ensure statistical reliability.

**In summary, for each 100 replicate of the simulation procedure:**

1. Simulate an AR(1) process defined by $X_t = \phi X_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, 1)$
2. Split the time series into training and test sets
3. Fit an AR(1) model on the continuous training data and generate $h$-step-ahead forecasts
4. Discretize the training set using one of the binning methods (EWI, EFI, or K-means), and apply the same bins to the test set and AR forecasts
5. Fit a first-order Markov Chain to the discretized training data
6. Generate forecasts from both the Markov Chain and discretized AR model for $h \in \{1, 2, 3, 5, 10\}$
7. Calculate RMSE and information loss metrics (Entropy, MI, ILR, ICR)

## 3.3 Multivariate Monte Carlo Simulation Setup

To evaluate forecasting performance in a multivariate setting, we implement a Monte Carlo simulation based on a bivariate VAR(1) process, following a structure that mirrors the univariate testing framework. The objective is to compare continuous VAR forecasts against their discretized counterparts and predictions from a Multivariate

Markov Chain (MMC) model, specifically the MTD-Probit formulation. As in the univariate case, forecast accuracy and information loss are assessed across different discretization methods, bin sizes, and levels of autocorrelation.

The time series data are generated from a bivariate VAR(1) process of the form

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} = A \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu} = (1, 1)$ and $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Two versions of the autoregressive matrix $A$ are considered to simulate different levels of autocorrelation:

$$A_{\text{high}} = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}, \quad A_{\text{low}} = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}.$$

Each simulated series contains either 5,000 or 10,000 observations and is divided into a training set (90%) and a test set (10%). A VAR(1) model is fitted on the continuous training data, and $h$-step-ahead forecasts are generated recursively for $h \in \{1, 2, 3, 5, 10\}$.

Afterward, each variable in the training set is discretized independently into $n \in \{3, \ldots, 10\}$ bins using Equal Width Intervalling (EWI), Equal Frequency Intervalling (EFI), or K-means clustering. The same bin edges are applied to the test set and the continuous VAR forecasts to ensure consistency across evaluations.

To assess performance, VAR forecasts are discretized using the corresponding binning scheme and compared to the discretized test series using RMSE. In parallel, a multivariate MTD-Probit model is estimated using the discretized training data. Forecasts from the MTD-Probit are generated recursively by selecting the most probable state at each horizon $h$, and compared against the discretized test data using RMSE.

To quantify information loss, we compute the entropy of the original continuous series using the Kozachenko–Leonenko estimator and estimate the entropy of the discretized variables using Shannon's formula. Mutual information between the continuous and discretized series is calculated using three methods: (i) the Mesner–Shalizi $k$-nearest neighbor estimator, (ii) Huo et al.'s Gaussian mixture approach, and (iii) a fine-grid binned approximation. From these, ILR and ICR metrics are derived separately for each variable.

The entire simulation process is repeated across 100 Monte Carlo replications. All forecast errors and information-theoretic metrics are averaged to ensure robust conclusions.

**In summary, for each 100 replicate of the simulation procedure:**

1. Simulate a bivariate VAR(1) process with $A = A_{\text{low}}$ or $A_{\text{high}}$
2. Split the time series into training and test sets
3. Fit a continuous VAR(1) model and generate $h$-step-ahead forecasts

4. Discretize each variable in the training set using EWI, EFI, or K-means; apply same bins to test set and VAR forecasts

5. Fit an MTD-Probit model to the discretized training data

6. Forecast both the MTD-Probit and discretized VAR models for $h \in \{1, 2, 3, 5, 10\}$

7. Calculate RMSE and information loss metrics (Entropy, MI, ILR, ICR) for each variable

## 3.4 Tools and Implementation Environment

This subsection provides an overview of the software tools and specific functions used in the implementation of the proposed methodology.

The implementation of both the discretization methods and the MMC models was carried out in R, using RStudio as the integrated development environment (IDE). This choice was made due to the availability of specialized packages with functions well-suited to the needs of this thesis, despite some limitations, which will be discussed later.

Regarding discretization, most of the simpler methods are readily available through core R packages. SAX has also been implemented in R. The $TSclust$ package, as described in Montero and Vilar (2014), is designed to provide a comprehensive suite of well-established, peer-reviewed time series dissimilarity measures. These include methods based on raw data, extracted features, underlying parametric models, complexity measures, and forecast behavior, making it a valuable resource for time series analysis and comparison.

As for the MMC models, Vasconcelos and Damásio (2025) provides implementations of both the MTD and MTD-Probit models through the GenMarkov package. This package was developed in response to the limited availability of tools that support the estimation and application of such models. Indeed, most existing approaches rely on algorithms and software that are either not widely accessible or are applied to very specific use cases.

The main function used, $multi.mtd\_probit()$, implements the MTD-Probit model introduced in Nicolau (2014). It returns the estimated parameters, their standard errors, z-statistics, p-values, and the log-likelihood values for each equation. The main inputs include the matrix $y$ of categorical data sequences, a numerical vector $initial$ containing the starting values for the optimization, and $nummethod$, which specifies the numerical maximization technique to be used. Supported methods include "NR" (Newton-Raphson), "BFGS" (Broyden-Fletcher-Goldfarb-Shanno), "BFGSR" (R's implementation of BFGS), "BHHH" (Berndt-Hall-Hall-Hausman), "SANN" (Simulated Annealing), "CG" (Conjugate Gradients), and "NM" (Nelder-Mead). In this thesis, only

the "BFGS" method was used, as it provided the most valid and stable results across all simulations.

# 4

# Results and Discussion

## 4.1 Monte Carlo Simulation Results

This section presents the results of the Monte Carlo experiments described previously, which evaluate the forecasting performance of discretized models, namely, first-order Markov Chains and Multivariate Markov Chains, in comparison to standard autoregressive and vector autoregressive models. To enable a fair comparison, the continuous forecasts produced by the continuous models were discretized using the same binning schemes applied to the original data. The analysis explores a range of bin sizes, discretization methods, and horizons, and includes information metrics to assess the extent of information loss due to discretization, and, of course, the forecast error itself. The plots shown in the main body are based on simulations with 10,000 observations and were obtained from a Monte Carlo experiment with 100 replications.

### 4.1.1 Forecast Accuracy

To better understand the cost of discretization in forecasting tasks, we report the RMSE of each model directly, instead of using performance ratios. For each simulation setting, we present the RMSE of the continuous baseline model (AR or VAR), and compare it with the RMSE of its discretized counterpart as well as with the RMSE of the corresponding Markov-based model (MC or MMC).

In the univariate setting, we show:

- The RMSE of the continuous AR(1) model in a table.
- The RMSE of the AR(1) model fitted on discretized input.
- The RMSE of the first-order Markov Chain (MC) model fitted on the same discretized data.

In the multivariate setting, we evaluate performance using a single variable (e.g., $Z_1$) for consistency. This is justified by the fact that both $Z_1$ and $Z_2$ were generated

using the same VAR(1) process structure with symmetric parameters. Therefore, their behavior and discretization patterns are statistically equivalent.

We report:

- The RMSE of the continuous VAR(1) model on $Z_1$ in a table.
- The RMSE of the VAR(1) model fitted to discretized data (on $Z_1$).
- The RMSE of the MTD-Probit model fitted on the discretized series.

These RMSE comparisons enable a direct assessment of how predictive performance changes under discretization and alternative modeling frameworks.

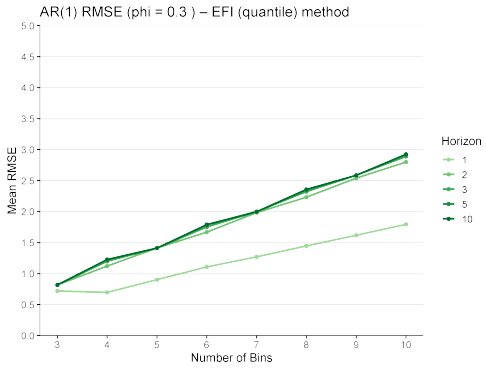
#### 4.1.1.1 Univariate Results

This subsection presents the forecasting results for univariate simulations under different levels of autocorrelation. Specifically, AR(1) processes and their corresponding Markov chain representations are simulated using two autoregressive coefficients: low ($\phi = 0.3$) and high ($\phi = 0.7$). Forecast accuracy is evaluated using the Root Mean Squared Error (RMSE) across various forecast horizons and discretization methods: Equal Frequency Intervalling (EFI), Equal Width Intervalling (EWI), and K-means clustering.

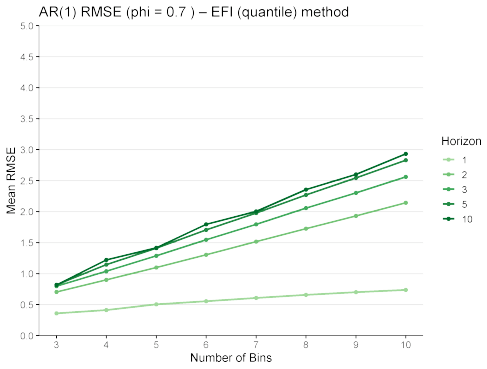
Table 4.1 provides a summary of RMSE values for the continuous AR(1) process across forecast horizons and correlation levels. In Figure 4.1, the green-shaded areas depict the RMSE values of the discretized AR model across different numbers of bins (ranging from 3 to 10) and forecast horizons (1, 2, 3, 5, and 10), under three discretization methods: EWI, EFI, and K-means. Results are shown for both low and high autocorrelation scenarios. Figure 4.2 presents a similar analysis for the Markov Chain model, with the results highlighted in shades of blue.

| Horizon | RMSE Low Correlation | RMSE High Correlation |
|---|---|---|
| 1 | 0.739571 | 0.424894 |
| 2 | 1.029175 | 1.049484 |
| 3 | 1.050768 | 1.243650 |
| 5 | 1.052951 | 1.366312 |
| 10 | 1.052877 | 1.403574 |

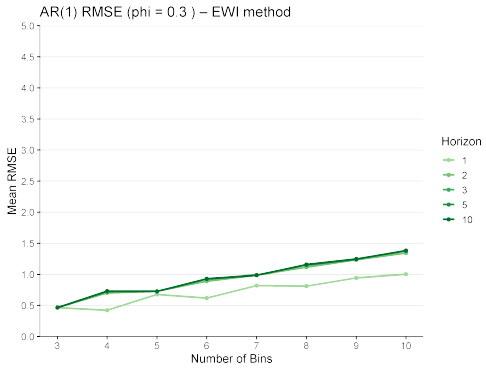
Table 4.1: RMSE $h$-step Ahead Forecast Comparison Between AR(1) Processes with Low and High Autocorrelation Levels



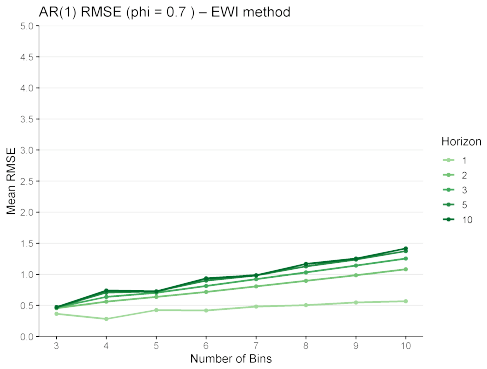

(a) EFI (quantile) Low Cor



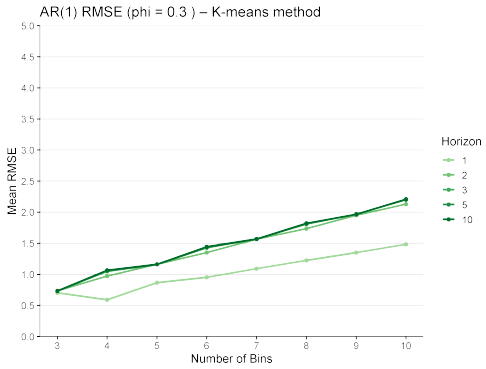

(b) EFI (quantile) High Cor



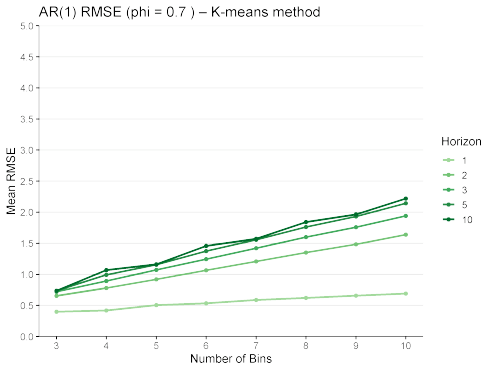

(c) EWI Low Cor




(d) EWI High Cor

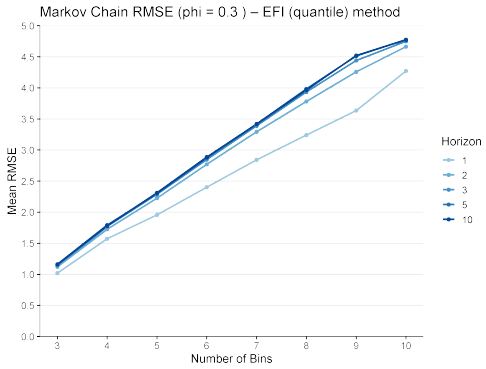



(e) K-means Low Cor

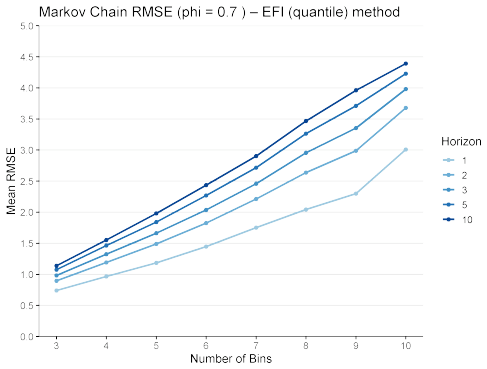



(f) K-means High Cor

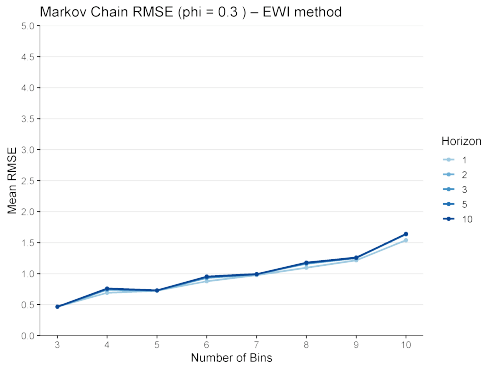
Figure 4.1: RMSE Comparison of AR(1) for Low Cor and High Cor Scenarios.



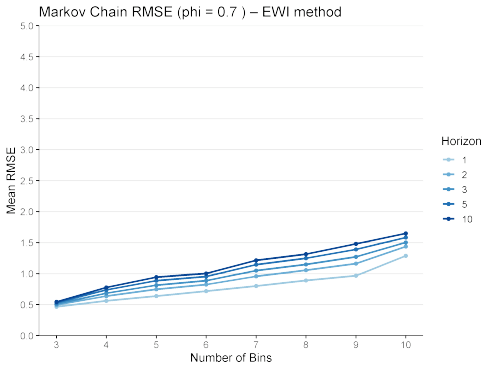

(a) EFI (quantile) Low Cor



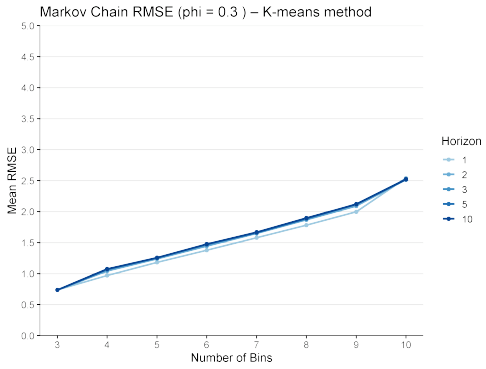

(b) EFI (quantile) High Cor



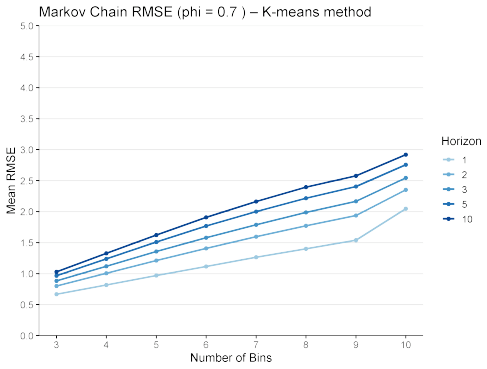

(c) EWI Low Cor




(d) EWI High Cor




(e) K-means Low Cor




(f) K-means High Cor

Figure 4.2: RMSE Comparison of Markov Chain for Low Cor and High Cor Scenarios.

#### 4.1.1.2 Multivariate Results

This subsection presents the forecasting results for multivariate simulations based on VAR(1) and MMC models. Two autoregressive matrix specifications are used to reflect weaker and stronger dependence regimes:

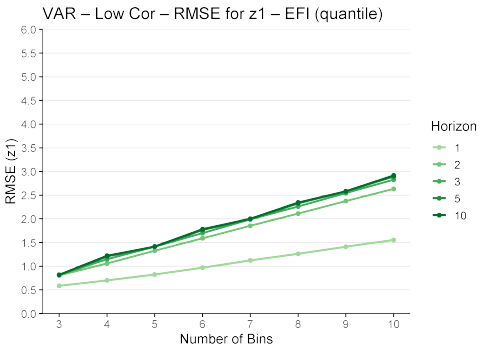
$$A_{\text{low}} = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}, \quad A_{\text{high}} = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}.$$

As in the univariate case, forecast performance is evaluated using RMSE across several horizons and discretization methods (EFI, EWI, and K-means).

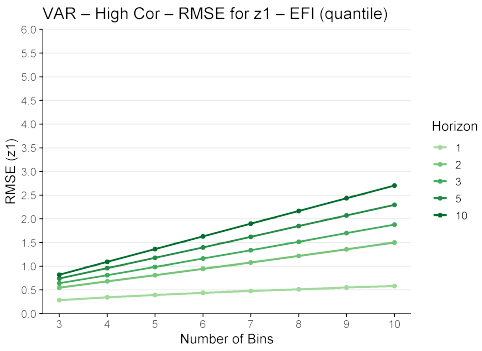
Table 4.2 summarizes the RMSE values for the continuous VAR(1) process (using the first variable $Z_1$). Figures 4.3 and 4.4 show the RMSE performance of the discretized VAR model and its corresponding MMC-based forecasts, respectively. As before, the analysis spans bin sizes from 3 to 10 and forecast horizons of 1, 2, 3, 5, and 10, using the same three discretization methods—EWI, EFI, and K-means—under both low and high autocorrelation scenarios.

| Horizon | RMSE Low Correlation | RMSE High Correlation |
|---|---|---|
| 1 | 0.673654 | 0.351068 |
| 2 | 1.030374 | 1.025149 |
| 3 | 1.095888 | 1.301911 |
| 5 | 1.115780 | 1.610722 |
| 10 | 1.117169 | 1.924371 |

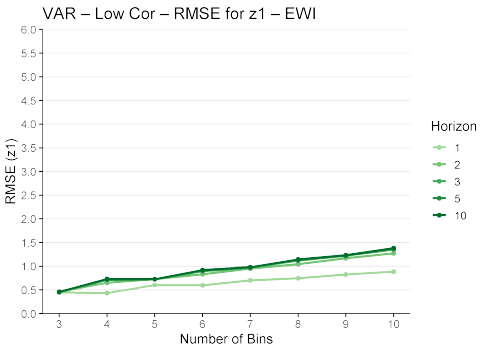
Table 4.2: RMSE (z1) $h$-step Ahead Forecast Comparison Between VAR(1) Processes with Low and High Autocorrelation Scenarios



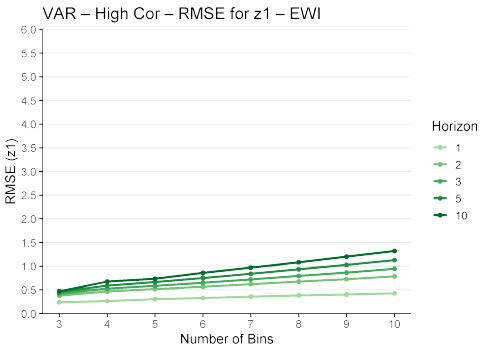

(a) EFI (quantile) Low Cor



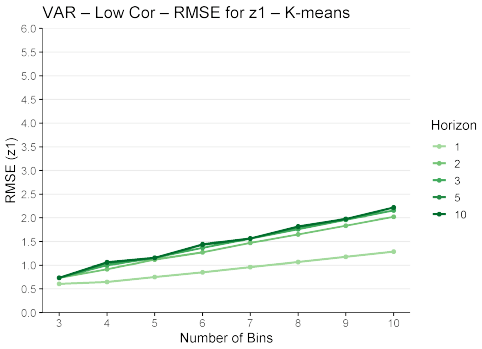

(b) EFI (quantile) High Cor



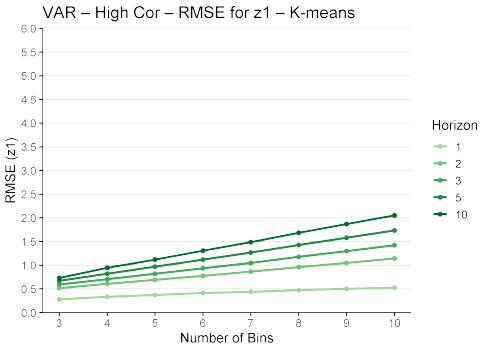

(c) EWI Low Cor




(d) EWI High Cor

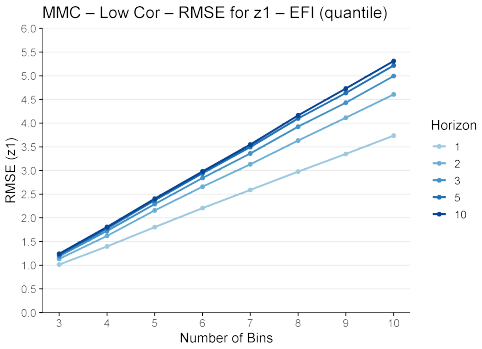



(e) K-means Low Cor

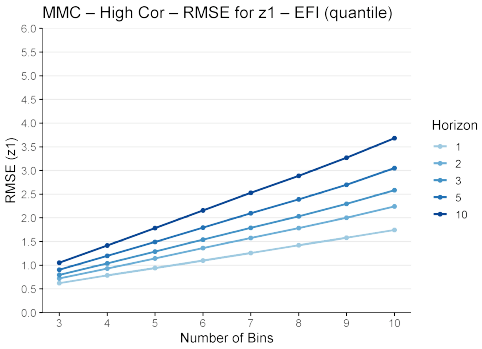



(f) K-means High Cor

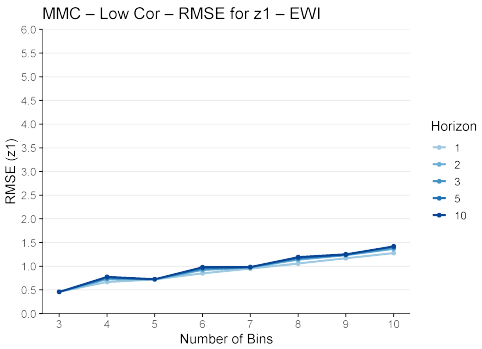
Figure 4.3: RMSE Comparison of VAR for Low Cor and High Cor Scenarios.



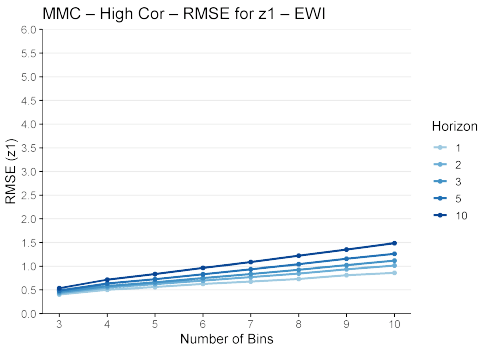

(a) EFI (quantile) Low Cor



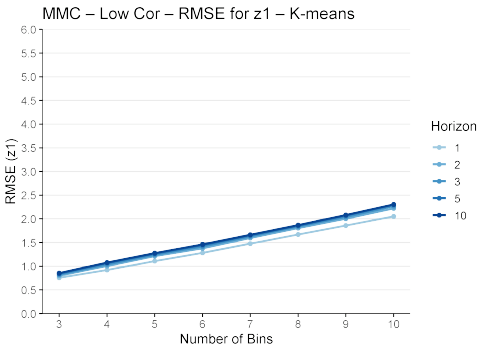

(b) EFI (quantile) High Cor



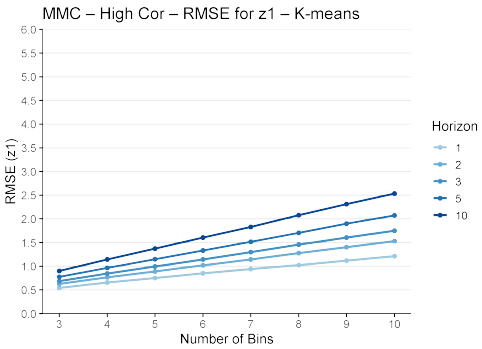

(c) EWI Low Cor




(d) EWI High Cor




(e) K-means Low Cor




(f) K-means High Cor

Figure 4.4: RMSE Comparison of MMC for Low Cor and High Cor Scenarios.

### 4.1.2 Effect of Discretization and Information Loss

To evaluate how discretization impacts the preservation of information in time series data, we analyze several information-theoretic metrics. These include the entropy of the discretized signal, mutual information (MI) between the original and discretized series, and derived ratios such as Mutual Information, ILR and ICR.

We perform this analysis under different discretization strategies: EFI (quantile discretization), EWI and K-means clustering and compare results across low and high correlation regimes. Like mentioned before, in the multivariate scenario, all metrics are evaluated for the first variable ($Z_1$) of the VAR(1) system, as both $Z_1$ and $Z_2$ are symmetrically generated and thus present equal statistical behavior.

The visualizations in this section serve to scope out issues related to information loss, which cannot be fully captured by forecasting RMSE alone. While some discretization algorithms may yield lower prediction errors, they may also discard a substantial amount of information from the original signal. By analyzing entropy, mutual information, and related ratios, we gain a deeper understanding of the trade-offs each method imposes, helping to contextualize forecasting results beyond accuracy.

#### 4.1.2.1 Univariate Results

In the univariate setting, we analyze information retention in AR(1) processes simulated under two autocorrelation levels: $\phi = 0.3$ (low) and $\phi = 0.7$ (high). The evaluation is conducted using three discretization strategies—Equal Frequency Intervalling (EFI), Equal Width Intervalling (EWI), and K-means clustering—across multiple estimators of mutual information. Each plot displays the metric values (e.g., MI, ILR, ICR) as a function of the number of bins, ranging from 3 to 10. The lines in each figure represent the different discretization methods, allowing for a direct comparison of how each strategy preserves information across levels of resolution and correlation strength. Negative values are omitted due to numerical instability in the estimation of mutual information and entropies, which resulted in negative or invalid ILR values.

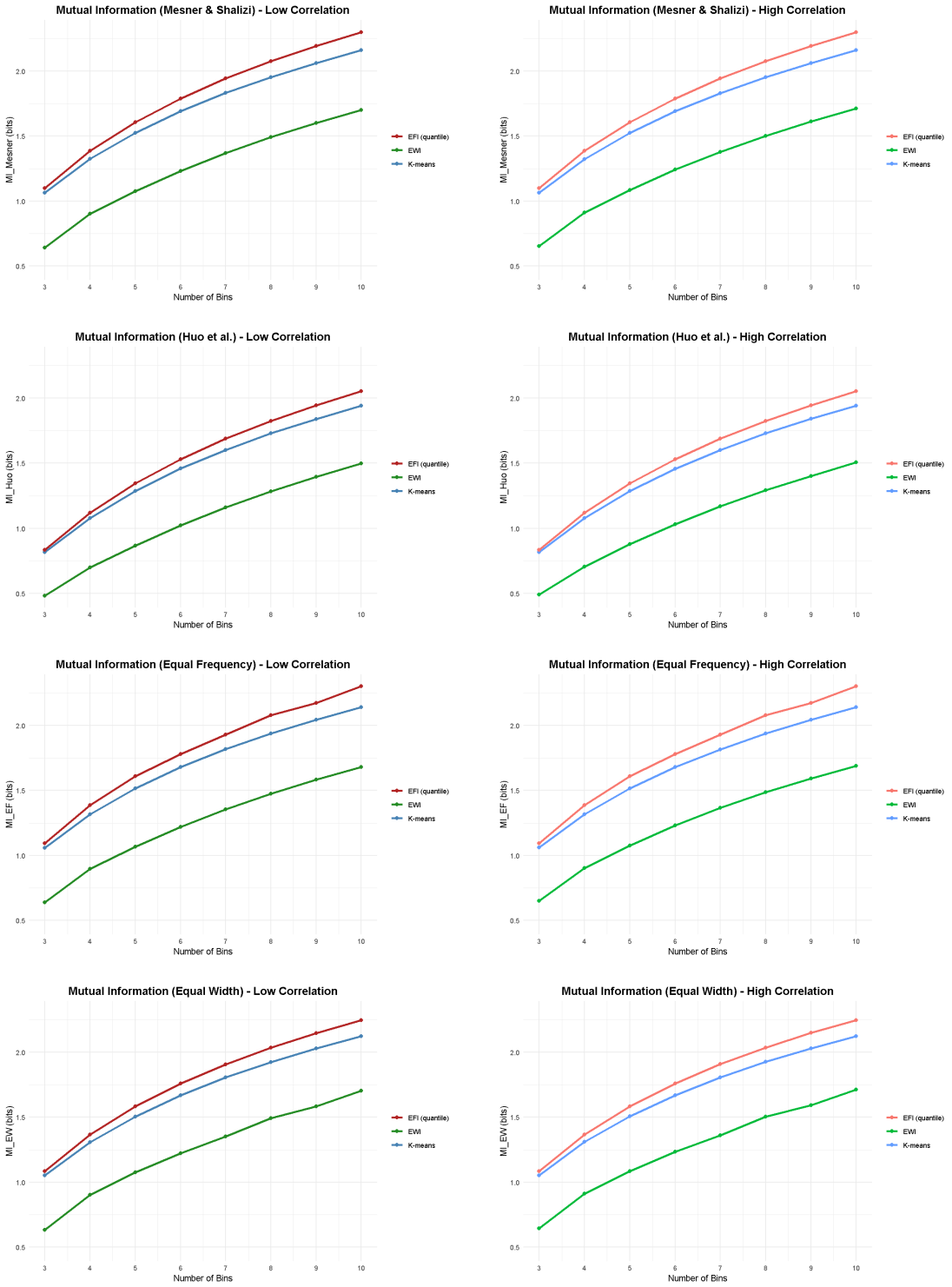



Figure 4.5: Mutual Information estimates under different discretization methods for low (left) and high (right) correlation settings.

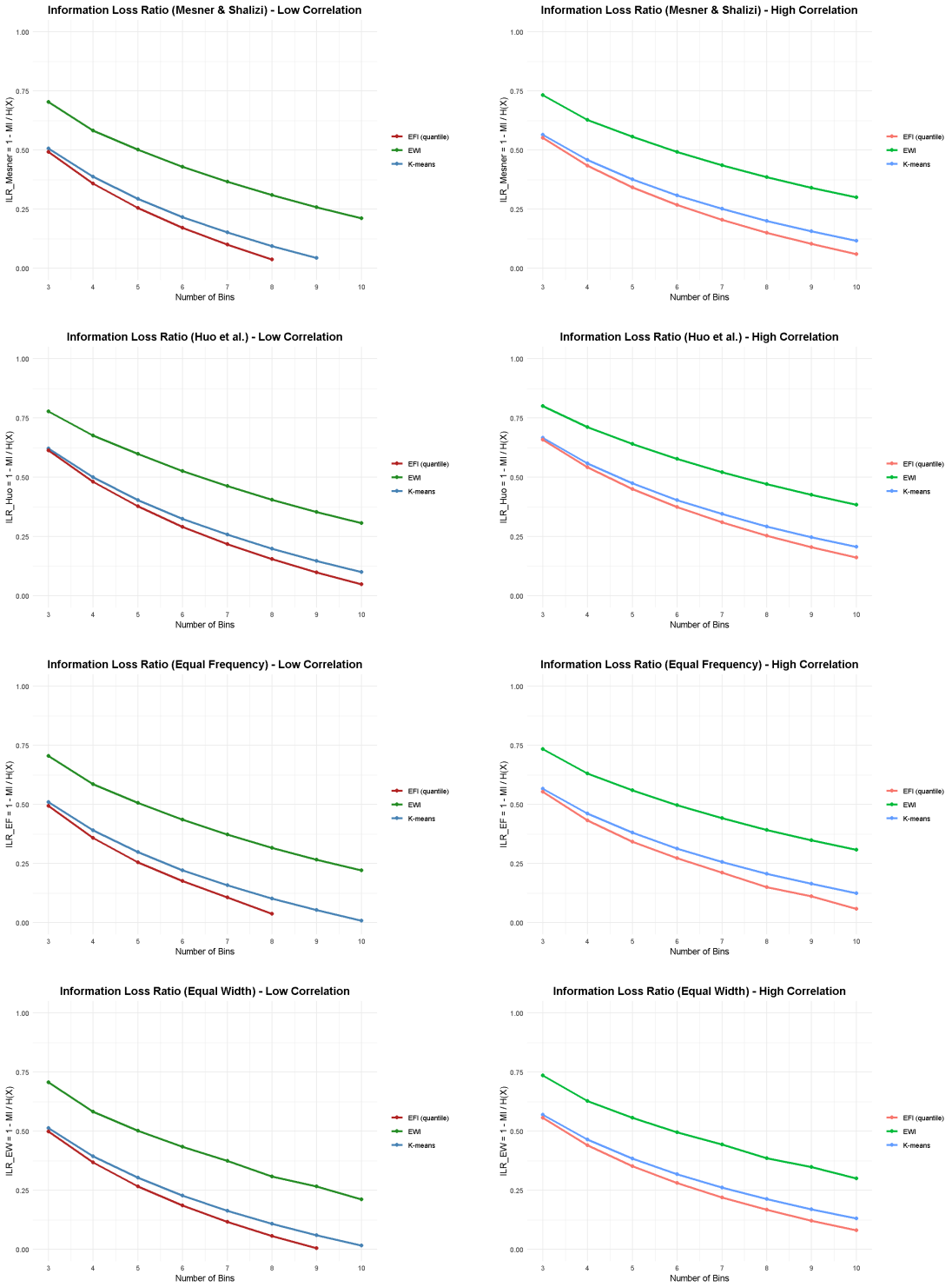



Figure 4.6: ILR across discretization methods for low (left) and high (right) correlation settings

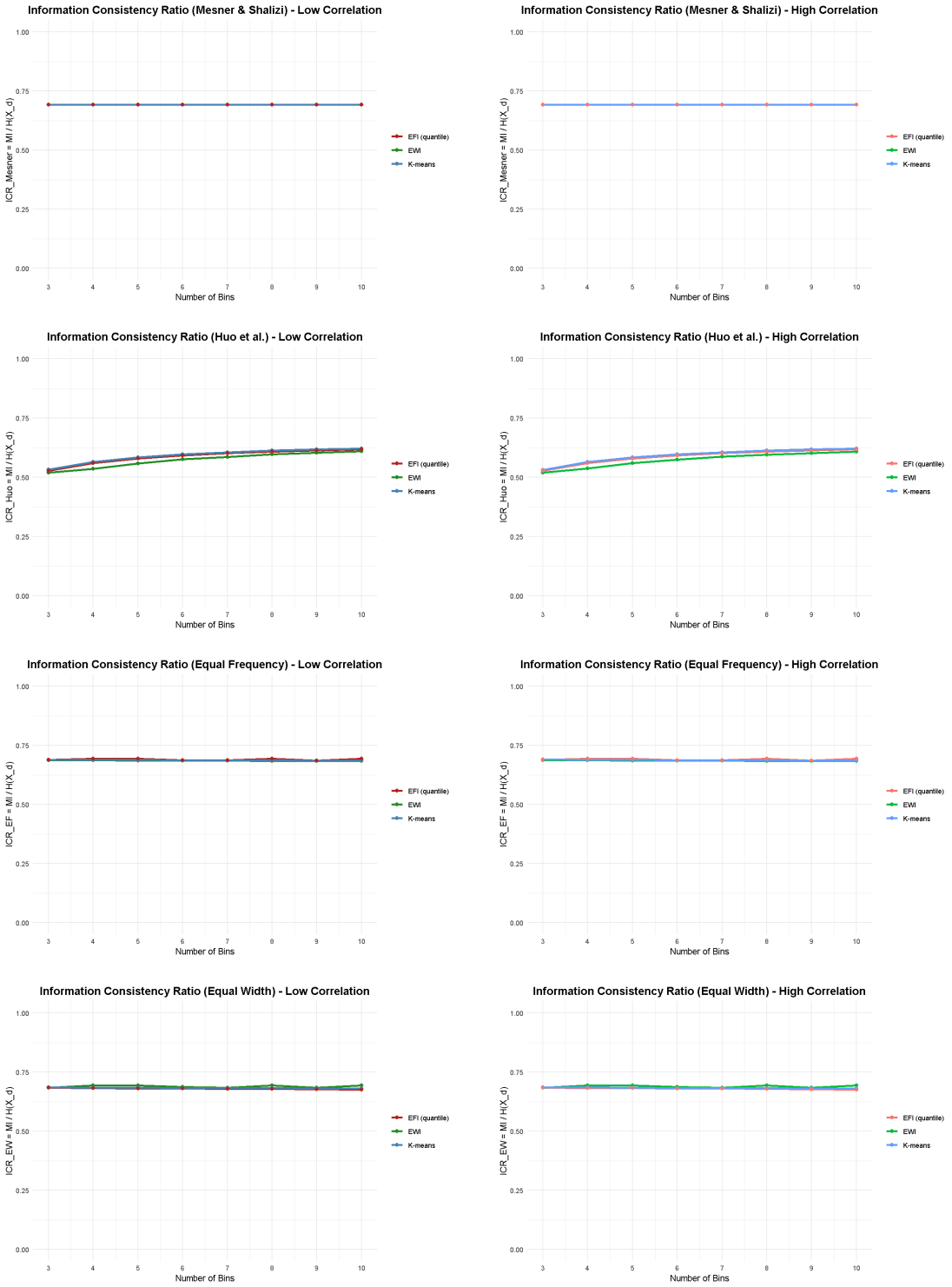



Figure 4.7: ICR across discretization methods for low (left) and high (right) correlation settings.

#### 4.1.2.2 Multivariate Results

In the multivariate setting, we assess the same information-theoretic metrics on bivariate VAR(1) processes under two dependence regimes, defined by the following autoregressive matrices:

$$A_{\text{low}} = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}, \quad A_{\text{high}} = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{bmatrix}.$$

As in the univariate case, we report metrics for the first component ($Z_1$) only. The figures plot each metric against the number of bins (3–10), with separate lines corresponding to each discretization method. Like before, negative values are omitted due to numerical instability in the estimation of mutual information and entropies, which resulted in negative or invalid ILR values.

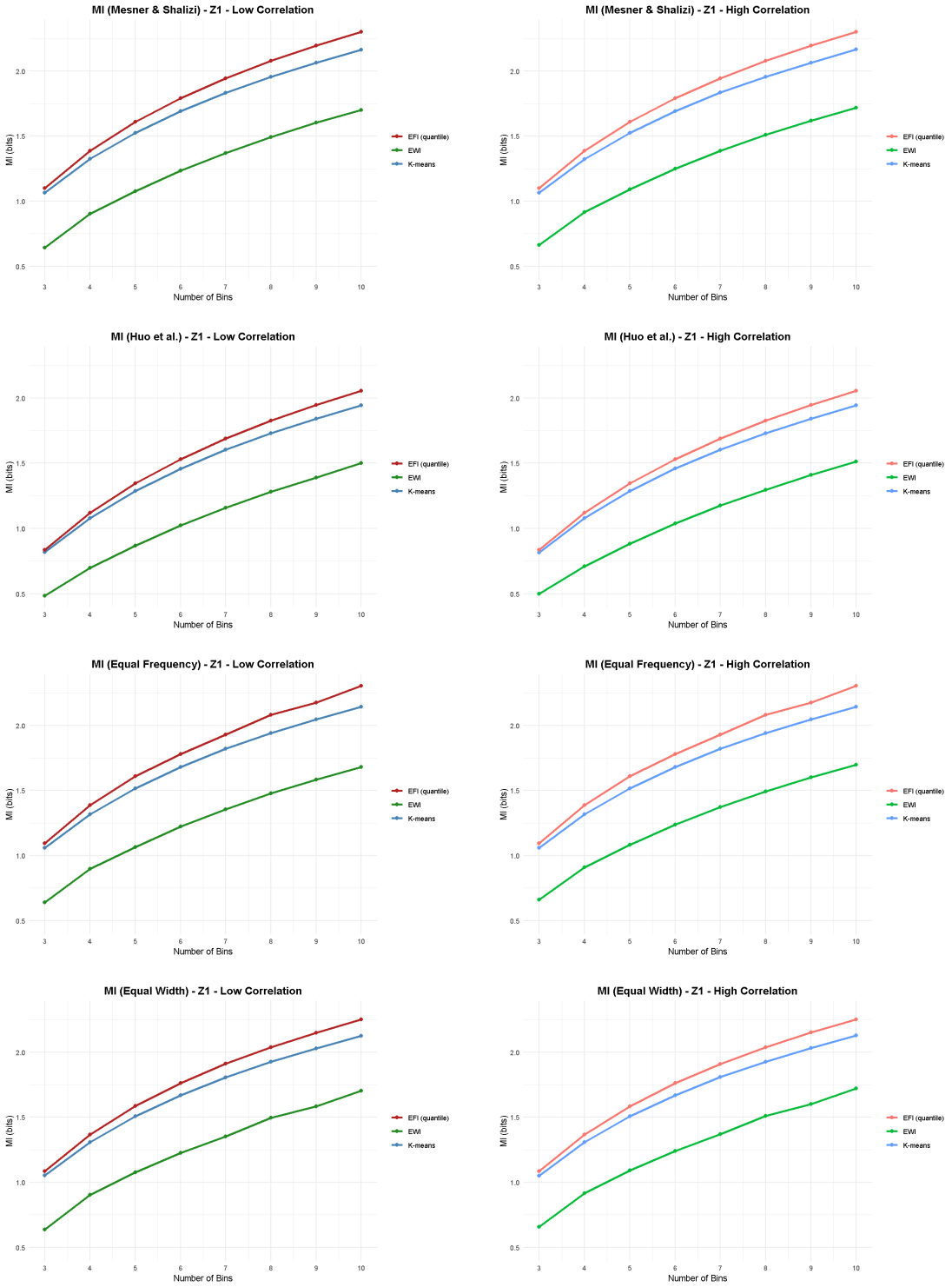



Figure 4.8: Multivariate Z1: Mutual Information estimates under different discretization methods, comparing low (left) and high (right) correlation settings.

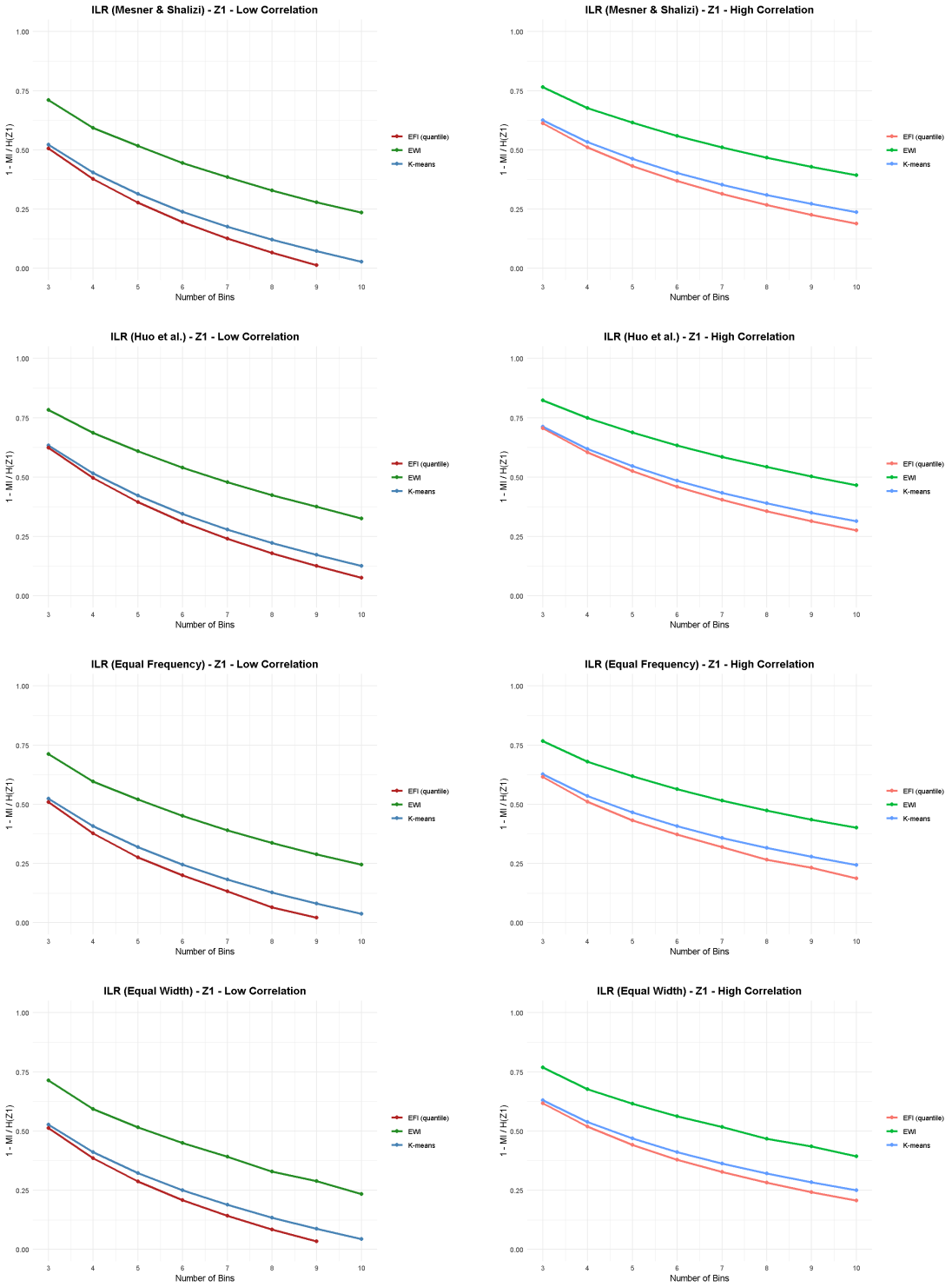



Figure 4.9: Multivariate Z1: ILR values across discretization methods under low (left) and high (right) correlation

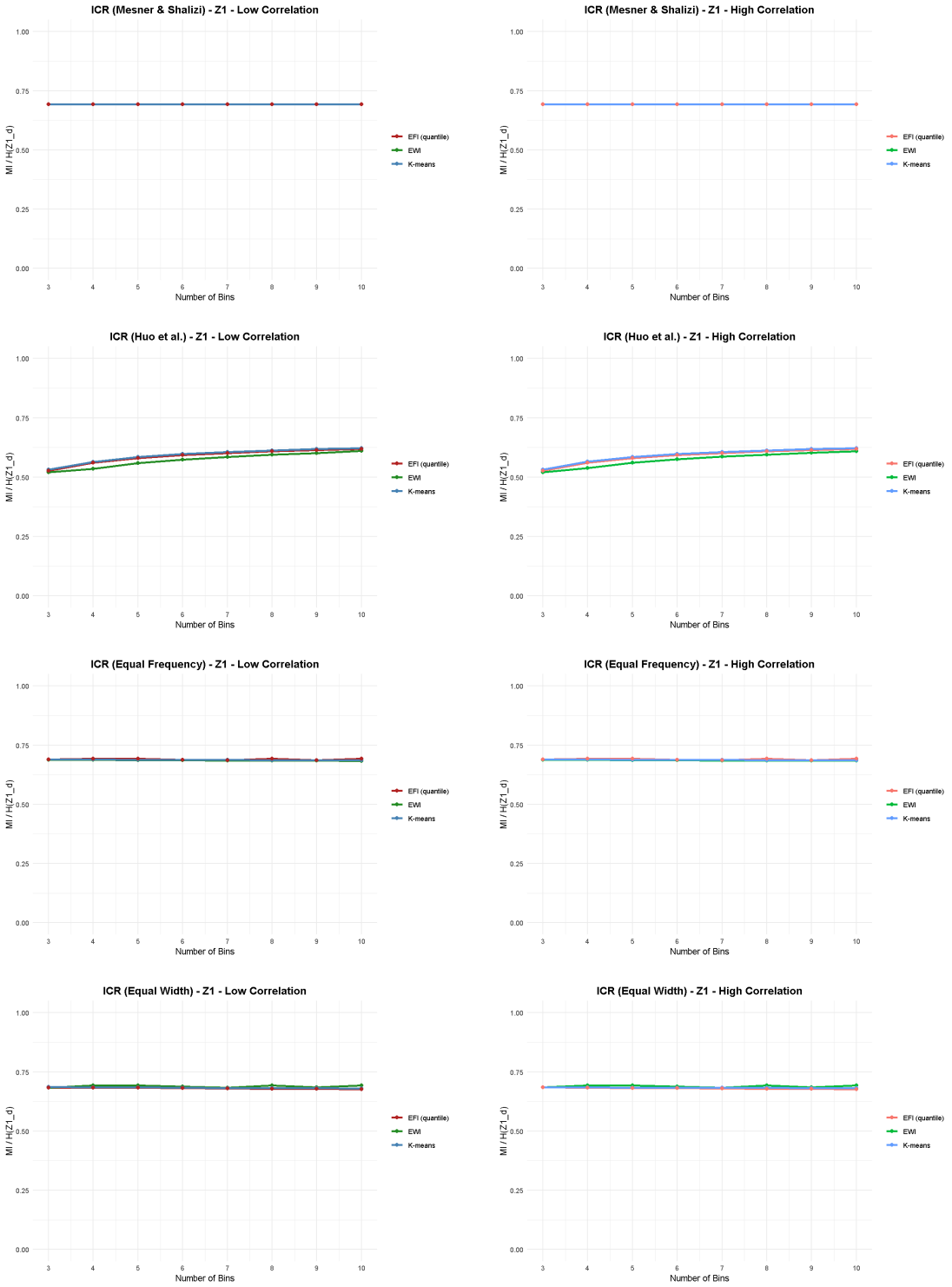



Figure 4.10: Multivariate Z1: ICR values across discretization methods under low (left) and high (right) correlation.

## 4.2 Key Findings

### 4.2.1 Univariate Results

The RMSE plots in Figures 4.1 and 4.2 reveal a clear distinction between low and high autocorrelation scenarios. When the autoregressive coefficient is high ($\phi = 0.7$), both the discretized AR(1) model and the first-order Markov Chain (MC) exhibit significantly better predictive performance compared to the low-correlation case. Among the discretization methods, EWI consistently achieves the lowest RMSE across all bin settings, especially for one-step-ahead forecasts. This result may be partially explained by the distribution of values: most observations lie near the center of the support, so even as the number of bins increases, the predictions tend to fall into the central bins where the model is most accurate.

When compared to the continuous AR(1) model (see Table 4.1), the discretized versions sometimes show lower RMSE values, especially for short horizons and under high autocorrelation. However, this does not necessarily indicate better predictive performance. The RMSE is computed on different scales: the continuous model forecasts values typically ranging from around -6 to 6, while the discretized models predict bin indices (e.g., from 1 to 3 or 1 to 7). Because these values are on a smaller numeric scale, the resulting RMSE can appear artificially lower. Therefore, direct RMSE comparisons between continuous and discretized models should be interpreted with caution. Table 4.1 was included primarily to give a sense of how RMSE progresses across increasing forecast horizons.

When comparing the overall performance of discretized AR versus MC, the discretized AR model generally outperforms the Markov Chain approach. This is particularly evident under the EFI method, where MC yields relatively poor results. The likely cause is the non-uniformity of bin widths in EFI: bins near the center of the distribution are very narrow, while bins at the tails are considerably wider. This non-uniformity creates problems for the Markov model because it relies on transitions between bins to make predictions. In EFI, the bins near the center are very narrow, so even small changes in the data can cause the model to jump between states, leading to noisy transitions. On the other hand, the bins at the edges are much wider and may aggregate together very different values, which can hide important patterns. As a result, the transition matrix becomes harder to estimate accurately, and the MC model struggles to capture the true behavior of the process. In contrast, the discretized AR model still keeps some of the structure of the original process and is less affected by uneven bin sizes.

The information-theoretic metrics confirm several expected patterns. As the number of bins increases, the mutual information between the continuous and discretized variable also increases (see Figure 4.5), reflecting improved resolution of the original signal. At the same time, ILR, which measures the proportion of entropy not captured by

the discretized variable, decreases (see Figure 4.6). This inverse relationship highlights the trade-off between granularity and information retention. Across discretization methods, EWI consistently results in greater information loss compared to EFI and K-means. This is evident from ILR curves, where EWI shows higher ILR values across all bin sizes, for all methods. This suggests that EWI is less efficient in preserving the information structure of the original continuous variable, especially when the data is unevenly distributed. EFI and K-means, which adapt more closely to the distribution of the data, retain more information overall.

Interestingly, the ICR remains relatively stable across bin sizes, hovering around a value of 0.70 (see Figure 4.7). This indicates that approximately 70% of the entropy in the discretized signal $H(X_d)$ is actually informative about the original continuous variable $X$, with the remaining 30% representing noise or redundant structure introduced by the discretization. The stability of ICR suggests that, while mutual information and entropy both grow with the number of bins, they do so proportionally—implying a consistent signal-to-noise ratio in the discretized representation across different granularities.

While the overall trends are consistent across correlation settings, some differences do emerge. In the low correlation scenario, a few ILR values become negative or undefined. This is likely due to numerical instability in the mutual information and entropy estimators, which can occasionally produce invalid results, particularly when the discretized variable captures little structure from a weakly autocorrelated series. These cases were omitted from the plots to maintain interpretability.

### 4.2.2 Multivariate Results

The multivariate results follow a similar pattern to the univariate setting. Forecast errors were generally lower across all models, likely due to the inclusion of a second variable, which provides additional insights for the model to learn. The VAR model, even when applied to discretized data, consistently outperformed the MTD-Probit (MMC) model across horizons and discretization methods (see Figures 4.3 and 4.4). This suggests that the linear dependency structure in the continuous data is better preserved through the VAR framework than through the first-order MMC dynamics.

Among discretization methods, EWI once again emerged as the most reliable option, particularly for short-term forecasts. Its performance remained stable across different numbers of bins, likely benefiting from the concentration of data near the center of the distribution. As with the univariate case, the MMC model struggled the most with EFI. Overall, the findings reinforce that while discretization can yield competitive results under favorable conditions, traditional models like VAR tend to be more robust.

In the multivariate setting, the behavior of the information-theoretic metrics closely mirrors the univariate case. Mutual information increases with the number of bins, while ILR decreases (see Figures 4.8 and 4.9). EWI continues to show the highest ILR values across methods, confirming that it retains less information compared to EFI and

K-means. As before, the correlation structure has minimal impact on the overall trends. However, under low correlation, some ILR values become negative or undefined due to numerical issues in the entropy and mutual information estimators. These instances were excluded from the plots for clarity. ICR remains stable across bin sizes, with approximately 70% of the entropy in the discretized variables being informative about the original continuous series (see Figure 4.10).

# 5

# Conclusion

## 5.1 Overview

This work set out to evaluate the viability of using Multivariate Markov Chain models on continuous time series data through discretization, with a particular focus on their forecasting performance compared to autoregressive models. Beyond this central question, the study has also made significant contributions to the understanding of discretization as a modeling tool. Through a simulation framework, we assessed various discretization algorithms including Equal Frequency Intervalling, Equal Width Intervalling, and K-means clustering across multiple bin sizes and autocorrelation settings.

The results revealed important insights about the trade-offs between information retention and predictive performance, and shed light on which discretization strategies are best suited for MC/MMC applications. Specifically, we demonstrated that while increasing the number of bins tends to reduce information loss, it can also introduce numerical instability in probabilistic models. Furthermore, we identified consistent performance patterns across methods, with EFI and K-means typically outperforming EWI in mutual information retention.

## 5.2 Key Findings and Limitations

The results of this thesis show a clear pattern: both univariate Markov Chains and multivariate variants (MMC via the MTD-Probit model) consistently underperform when compared to their discretized autoregressive counterparts (AR/VAR) across all evaluated horizons, bin sizes, and discretization strategies. While it was hypothesized that MMCs might capture dependencies across variables more effectively, the empirical evidence suggests that their forecasting power remains limited when compared to simpler AR-based approaches, even after controlling for discretization losses.

Moreover, these performance limitations are compounded by significant computational drawbacks. In particular, the estimation procedure for the MMC model, based

on the MTD-Probit formulation in R, is highly sensitive to initialization and prone to non-convergence. To mitigate this, a robust estimation pipeline had to be implemented, involving multiple initializations within a for-loop, ultimately selecting the specification with the highest log-likelihood. Despite this workaround, the computational cost remained substantial: running 100 Monte Carlo replications for a single MMC configuration took on average 48 hours, making large-scale experimentation much slower than a VAR simulation pipeline (with a average runtime of 2 hours for 100 replications). This makes MMC models harder to use in practice, especially when faster and more accurate alternatives are available.

One limitation of this study stems from the reliance on R, primarily due to the availability of the MTD-Probit model implementation. However, the optimization routines in R are not particularly efficient for large-scale simulations. The MTD-Probit model, in particular, suffers from long run times and unstable convergence, which significantly slows down the process. A second and more fundamental drawback lies in the discretization itself. Once a continuous variable is discretized, the original value is lost — all future operations rely only on the bin assignment. This makes it difficult to interpret or reuse the data in later stages of analysis. Additionally, when using methods like quantile-based or k-means discretization, the bin widths can vary considerably, which further complicates the interpretation and mapping back to real-world values.

Another limitation lies in the restricted class of data-generating processes considered in the simulation framework. All synthetic time series were generated from linear AR models. While this choice provides a controlled environment for evaluating discretization and forecasting, it may have inadvertently favored the performance of AR-based models over Markov-based approaches. As a result, the findings may not fully capture the comparative strengths of MC or MMC models in more complex, nonlinear settings.

## 5.3 Future Work

There are several directions for future work that could build upon this study. The Monte Carlo simulation framework can be further expanded to explore a wider range of configurations. This includes testing additional levels of autocorrelation, incorporating smaller sample sizes to assess robustness in limited data settings, and evaluating a broader range of bin sizes beyond those considered here. Moreover, additional error metrics could provide a more nuanced view of forecasting performance, while alternative information-theoretic measures could enrich the analysis of discretization quality.

A natural next step would be to apply the methodology to real-world time series data, to test its practical applicability and limitations. On the discretization side, future work should consider methods that explicitly aim to preserve the correlation structure between variables, as the current approach discretized each series independently.

Investigating multivariate discretization techniques could be particularly important in this context. Additionally, extending the analysis to multivariate settings with more than two variables would provide a more realistic and challenging test of the MTD-Probit model's scalability and forecasting capabilities.

Another important direction for future work involves exploring a more diverse set of data-generating processes. In particular, incorporating nonlinear time series models such as GARCH (Generalized Autoregressive Conditional Heteroskedasticity), threshold autoregressive (TAR), or regime-switching models. These models exhibit features like conditional heteroskedasticity, nonlinearity, or abrupt state transitions—dynamics that linear AR models fail to capture, but which may be better suited to the strengths of Markov and multivariate Markov chain frameworks.

# Bibliography


Acosta-Mesa, H.-G., Rechy-Ramírez, F., Mezura-Montes, E., Cruz-Ramírez, N., & Jiménez, R. H. (2014). Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions. *Journal of Biomedical Informatics*, *49*, 73–83. https://doi.org/10.1016/j.jbi.2014.03.004 (cit. on p. 13).

Adke, S. R., & Deshmukh, S. R. (2018). Limit distribution of a high order markov chain. *Journal of the Royal Statistical Society: Series B (Methodological)*, *50*(1), 105–108. https://doi.org/10.1111/j.2517-6161.1988.tb01715.x (cit. on p. 7).

Anacleto, M., Vinga, S., & Carvalho, A. (2020). *Msax: Multivariate symbolic aggregate approximation for time series classification* [Master's thesis, IST]. https://doi.org/10.1007/978-3-030-63061-4_9 (cit. on p. 15).

Asadabadi, M. R. (2017). A customer based supplier selection process that combines quality function deployment, the analytic network process and a markov chain. *European Journal of Operational Research*, *263*(3), 1049–1062. https://doi.org/10.1016/j.ejor.2017.06.006 (cit. on p. 4).

Baek, S., & Kim, D. Y. (2017). Empirical sensitivity analysis of discretization parameters for fault pattern extraction from multivariate time series data. *IEEE Transactions on Cybernetics*, *47*(5), 1198–1209. https://doi.org/10.1109/TCYB.2016.2540657 (cit. on p. 13).

Baek, S., & Young, D. (2017). Empirical sensitivity analysis of discretization parameters for fault pattern extraction from multivariate time series data. *IEEE TRANSACTIONS ON CYBERNETICS*, *47*(5), 1198–1209. https://doi.org/10.1109/TCYB.2016.2540657 (cit. on p. 1).

Berchtold, A. (1996). Modélisation autorégressive des chaînes de Markov : Utilisation d'une matrice différente pour chaque retard. *Revue de Statistique Appliquée*, *44*(3), 5–25. https://www.numdam.org/item/RSA_1996__44_3_5_0/ (cit. on p. 7).

Berchtold, A. (2001). Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, *22*(4), 379–397. https://doi.org/10.1111/1467-9892.00231 (cit. on p. 4).

Berchtold, A. (1995). Autoregressive modelling of markov chains. In G. U. H. Seeber, B. J. Francis, R. Hatzinger, & G. Steckel-Berger (Eds.), *Statistical modelling* (pp. 19–26). Springer New York. https://doi.org/https://doi.org/10.1007/978-1-4612-0789-4_3 (cit. on p. 7).

Berchtold, A., & Raftery, A. E. (2002). The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science*, *17*(3), 328–356. https://doi.org/10.1214/ss/1042727943 (cit. on p. 11).

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., & Romance, M. e. a. (2014). The structure and dynamics of multilayer networks. *Phys Rep*, *544*(1), 1–122. https://doi.org/10.48550/arXiv.1407.0742 (cit. on p. 4).

Bukiet, B., Harold, E., & Palacios, J. (1997). A markov chain approach to baseball. *Operations Research*, *45*, 14–23. https://doi.org/10.1287/opre.45.1.14 (cit. on p. 4).

Ching, W.-K., Ng, M. K., & Fung, E. S. (2008). Higher-order multivariate markov chains and their applications [Special Issue devoted to the Second International Conference on Structured Matrices]. *Linear Algebra and its Applications*, *428*(2), 492–507. https://doi.org/10.1016/j.laa.2007.05.021 (cit. on p. 7).

Damásio, B., & Mendonça, S. (2019). Modelling insurgent-incumbent dynamics: Vector autoregressions, multivariate markov chains, and the nature of technological competition. *Applied Economics Letters*, *26*(10), 843–849. https://doi.org/10.1080/13504851.2018.1502863 (cit. on p. 11).

Damásio, B., & Nicolau, J. (2014). Combining a regression model with a multivariate markov chain in a forecasting problem. *Statistics & Probability Letters*, *90*, 108–113. https://doi.org/10.1016/j.spl.2014.03.026 (cit. on pp. 8, 10).

Damásio, B., & Nicolau, J. (2024). Time inhomogeneous multivariate markov chains: Detecting and testing multiple structural breaks occurring at unknown dates. *Chaos, Solitons & Fractals*, *180*, 114478. https://doi.org/10.1016/j.chaos.2024.114478 (cit. on p. 10).

D'Amico, G., & de Blasis, R. (2019). A multivariate markov chain stock model. *SCANDINAVIAN ACTUARIAL JOURNAL*. https://doi.org/10.1080/03461238.2019.1661280 (cit. on p. 2).

Dash, M., Liu, H., & Motoda, H. (2014). Discretization of temporal data: A survey. *arXiv preprint arXiv:1402.4283* (cit. on p. 13).

Dash, R., Paramguru, R., & Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *Int. J. Adv. Sci. Technol.*, *2* (cit. on p. 16).

dos Santos Passos, H., Teodoro, F. G. S., Duru, B. M., de Oliveira, E. L., Peres, S. M., & Lima, C. A. M. (2017). Symbolic representations of time series applied to biometric recognition based on ecg signals. *2017 International Joint Conference on Neural Networks (IJCNN)*, 3199–3207. https://doi.org/10.1109/IJCNN.2017.7966255 (cit. on p. 13).

Eberle, S., Cevasco, D., Schwarzkopf, M.-A., Hollm, M., & Seifried, R. (2022). Multivariate simulation of offshore weather time series: A comparison between markov chain, autoregressive, and long short-term memory models. *Wind*, *2*(2), 394–414. https://doi.org/10.3390/wind2020021 (cit. on p. 11).

Elliott, G., & Timmermann, A. (2016). Forecasting in economics and finance. In P. Aghion & H. Rey (Eds.), *Annual review of economics, vol 8* (pp. 81–110, Vol. 8). https://doi.org/10.1146/annurev-economics-080315-015346 (cit. on p. 1).

Foorthuis, R. (2020). The impact of discretization method on the detection of six types of anomalies in datasets. https://arxiv.org/abs/2008.12330 (cit. on p. 13).

Fung, E. S., & Siu, T. K. (2012). A flexible markov chain approach for multivariate credit ratings. *Computational Economics*, *39*(2), 135–143. https://doi.org/10.1007/s10614-011-9258-y (cit. on p. 4).

García, S., Luengo, J., Sáez, J. A., López, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, *25*. https://doi.org/10.1109/TKDE.2012.35 (cit. on p. 12).

Gómez, S., Arenas, A., Borge-Holthoefer, J., Meloni, S., & Moreno, Y. (2010). Discrete-time markov chain approach to contact-based disease spreading in complex networks. *Europhys Lett*, *89*(3), 38009. https://doi.org/10.48550/arXiv.0907.1313 (cit. on p. 4).

Gottschau, A. (1992). Exchangeability in multivariate markov chain models. *Biometrics*, 751–763. https://doi.org/10.2307/2532342 (cit. on p. 4).

Horvath, P. A., Autry, C. W., & Wilcox, W. E. (2005). Liquidity implications of reverse logistics for retailers: A markov chain approach. *Journal of Retailing*, *81*(3), 191–203. https://doi.org/10.1016/j.jretai.2005.07.003 (cit. on p. 4).

Hranisavljevic, N., Maier, A., & Niggemann, O. (2020). Discretization of hybrid cpps data into timed automaton using restricted boltzmann machines. *Engineering Applications of Artificial Intelligence*, *95*, 103826. https://doi.org/10.1016/j.engappai.2020.103826 (cit. on p. 13).

Huo, L., Lv, C., Fei, S., & Zhou, D. (2011). The correlation analysis of discrete variables and continuous variables based on mutual information. *Applied Mechanics and Materials*, *121-126*, 4203–4207. https://doi.org/10.4028/www.scientific.net/AMM.121-126.4203 (cit. on p. 18).

Kozachenko, L. F., & Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, *23*(2), 95–101 (cit. on p. 17).

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *69*. https://doi.org/10.1103/PhysRevE.69.066138 (cit. on p. 17).

Le, N. D., Martin, R. D., & Raftery, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of*

*the American Statistical Association*, *91*(436), 1504–1515. https://doi.org/10.1080/01621459.1996.10476718 (cit. on p. 7).

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing sax: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, *15*, 107–144. https://doi.org/10.1007/s10618-007-0064-z (cit. on p. 14).

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*(4), 393–423. https://doi.org/10.1023/A:1016304305535 (cit. on p. 18).

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006a). New time series data representation esax for financial applications. *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, x115–x115. https://doi.org/10.1109/ICDEW.2006.99 (cit. on p. 13).

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006b). Extended sax: Extension of symbolic aggregate approximation for financial time series data representation. *Proceeding of IEICE the 17th Data Engineering Workshop* (cit. on p. 15).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. ii).

Malinowski, S., Guyet, T., René, Q., & Tavenard, R. (2013). 1d-sax: A novel symbolic representation for time series. https://doi.org/10.1007/978-3-642-41398-8_24 (cit. on p. 15).

Martin, R. D., & Raftery, A. E. (1987). Non-gaussian state-space modeling of nonstationary time series: Comment: Robustness, computation, and non-euclidean models. *Journal of the American Statistical Association*, *82*(400), 1044–1050. https://doi.org/10.2307/2289377 (cit. on p. 7).

Maskawa, J.-i. (2003). Multivariate markov chain modeling for stock markets [Proceedings of the International Econophysics Conference]. *Physica A: Statistical Mechanics and its Applications*, *324*(1), 317–322. https://doi.org/10.1016/S0378-4371(02)01868-X (cit. on p. 4).

Mehran, F. (1989). Longitudinal analysis of employment and unemployment based on matched rotation samples. *Labour*, *3*(1), 3–20. https://doi.org/10.1111/j.1467-9914.1989.tb00146.x (cit. on pp. 4, 7).

Mesner, O., & Shalizi, C. (2020). Conditional mutual information estimation for mixed, discrete and continuous, data. *IEEE Transactions on Information Theory*. https://doi.org/10.1109/TIT.2020.3024886 (cit. on p. 17).

Montero, P., & Vilar, J. A. (2014). Tsclust: An r package for time series clustering. *Journal of Statistical Software*, *62*(1), 1–43. https://doi.org/10.18637/jss.v062.i01 (cit. on p. 23).

Mörchen, F., Ultsch, A., & Hoos, O. (2005). Extracting interpretable muscle activation patterns with time series knowledge mining. *KES Journal*, *9*, 197–208. https://doi.org/10.3233/KES-2005-9304 (cit. on p. 13).

Moskovitch, R., & Shahar, Y. (2015). Classification-driven temporal discretization of multivariate time series. *DATA MINING AND KNOWLEDGE DISCOVERY*, *29*(4, SI), 871–913. https://doi.org/10.1007/s10618-014-0380-z (cit. on p. 1).

Muhammad Fuad, M. M. (2012). Genetic algorithms-based symbolic aggregate approximation. In A. Cuzzocrea & U. Dayal (Eds.), *Data warehousing and knowledge discovery* (pp. 105–116). Springer Berlin Heidelberg. (Cit. on p. 15).

Nicolau, J. (2014). A new model for multivariate markov chains. *Scandinavian Journal of Statistics*, *41*(4), 1124–1135. https://doi.org/10.1111/sjos.12087 (cit. on pp. 7, 23).

Overlöper, P. J., Moddemann, L., Hranisavljevic, N., Windmann, A., & Niggemann, O. (2024). Discretization of cps time series with neural networks. *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 1–8. https://doi.org/10.1109/ETFA61755.2024.10710901 (cit. on p. 13).

Pal, S., Ghosh, S., Biswas, H., & Patwari, M. (2022). Discretization using combination of heuristics for high accuracy with huge noise reduction. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *34*(4), 1710–1722. https://doi.org/10.1109/TKDE.2020.2997719 (cit. on p. 1).

Raftery, A., & Tavaré, S. (1994). Estimation and modelling repeated patterns in high order markov chains with the mixture transition distribution model. *Applied Statistics*, 179–199. https://doi.org/10.2307/2986120 (cit. on p. 4).

Ramage, C. (1993). Forecasting in meteorology. *BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY*, *74*(10), 1863–1871. https://doi.org/10.1175/1520-0477(1993)074<1863:FIM>2.0.CO;2 (cit. on p. 1).

Sahin, A. D., & Sen, Z. (2001). First-order markov chain approach to wind speed modelling [10th International Conference on Wind Engineering]. *Journal of Wind Engineering and Industrial Aerodynamics*, *89*(3), 263–269. https://doi.org/10.1016/S0167-6105(00)00081-7 (cit. on p. 4).

Sericola, B. (2013). *Markov chains: Theory and applications*. John Wiley & Sons. (Cit. on p. 4).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 16).

Siu, T. K., Ching, W. K., Fung, S. E., & Ng, M. K. (2005). On a multivariate markov chain model for credit risk measurement. *Quantitative Finance*, *5*(6), 543–556. https://doi.org/10.1080/14697680500383714 (cit. on pp. 4, 7).

Soyiri, I. N., & Reidpath, D. D. (2013). An overview of health forecasting. *ENVIRONMENTAL HEALTH AND PREVENTIVE MEDICINE*, *18*(1), 1–9. https://doi.org/10.1007/s12199-012-0294-6 (cit. on p. 1).

Tang, X., Machimura, T., Liu, W., Li, J., & Hong, H. (2021). A novel index to evaluate discretization methods: A case study of flood susceptibility assessment based

on random forest. *Geoscience Frontiers*, *12*(6), 101253. https://doi.org/10.1016/j.gsf.2021.101253 (cit. on p. 19).

Tsiliyannis, C. A. (2018). Markov chain modeling and forecasting of product returns in remanufacturing based on stock mean-age. *European Journal of Operational Research*, *271*(2), 474–489. https://doi.org/10.1016/j.ejor.2018.05.026 (cit. on p. 4).

Turchin, P. B. (1986). Modelling the effect of host patch size on mexican bean beetle emigration. *Ecology*, *67*(1), 124–132. https://doi.org/10.2307/1938510 (cit. on p. 4).

Vasconcelos, C., & Damásio, B. (2025). Genmarkov: Modeling generalized multivariate markov chains in r. *The R Journal*, *16*, 96–113. https://doi.org/10.32614/RJ-2024-006 (cit. on pp. 10, 23).

Wiecek, P., & Kubek, D. (2024). The impact of time series selected characteristics on the fuel demand forecasting effectiveness based on autoregressive models and markov chains. *Energies*, *17*(16), 4163. https://doi.org/10.3390/en17164163 (cit. on p. 12).

Wong, C. S., & Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, *96*(455), 982–995. https://doi.org/10.1198/016214501753208645 (cit. on p. 7).

Zhu, D.-M., & Ching, w. K. (2010). A new estimation method for multivariate markov chain model with application in demand predictions. *Proceedings - 3rd International Conference on Business Intelligence and Financial Engineering, BIFE 2010*, 126–130. https://doi.org/10.1109/BIFE.2010.39 (cit. on p. 7).

# A

# Additional Results

## A.1 Forecast RMSE

### A.1.1 Univariate Results

Table A.1: Mean RMSE of AR(1) forecasts under high correlation ($\phi = 0.7$). Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.36 | 0.36 | 0.40 |
| 3 | 2 | 0.70 | 0.46 | 0.65 |
| 3 | 3 | 0.79 | 0.47 | 0.72 |
| 3 | 5 | 0.82 | 0.47 | 0.74 |
| 3 | 10 | 0.82 | 0.47 | 0.74 |
| 4 | 1 | 0.41 | 0.28 | 0.42 |
| 4 | 2 | 0.90 | 0.56 | 0.78 |
| 4 | 3 | 1.04 | 0.64 | 0.89 |
| 4 | 5 | 1.15 | 0.71 | 0.99 |
| 4 | 10 | 1.22 | 0.74 | 1.07 |
| 5 | 1 | 0.50 | 0.43 | 0.51 |
| 5 | 2 | 1.10 | 0.64 | 0.92 |
| 5 | 3 | 1.29 | 0.70 | 1.07 |
| 5 | 5 | 1.41 | 0.73 | 1.16 |
| 5 | 10 | 1.42 | 0.73 | 1.16 |
| 6 | 1 | 0.55 | 0.42 | 0.53 |
| 6 | 2 | 1.30 | 0.72 | 1.06 |
| 6 | 3 | 1.54 | 0.81 | 1.24 |
| 6 | 5 | 1.70 | 0.90 | 1.37 |
| 6 | 10 | 1.79 | 0.94 | 1.46 |
| 7 | 1 | 0.61 | 0.48 | 0.59 |

*Continued on next page*

Table A.1 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 7 | 2 | 1.51 | 0.81 | 1.21 |
| 7 | 3 | 1.79 | 0.92 | 1.42 |
| 7 | 5 | 1.98 | 0.98 | 1.55 |
| 7 | 10 | 2.00 | 0.98 | 1.57 |
| 8 | 1 | 0.66 | 0.50 | 0.62 |
| 8 | 2 | 1.72 | 0.90 | 1.35 |
| 8 | 3 | 2.06 | 1.03 | 1.60 |
| 8 | 5 | 2.27 | 1.13 | 1.76 |
| 8 | 10 | 2.36 | 1.17 | 1.84 |
| 9 | 1 | 0.70 | 0.55 | 0.66 |
| 9 | 2 | 1.93 | 0.99 | 1.48 |
| 9 | 3 | 2.30 | 1.14 | 1.76 |
| 9 | 5 | 2.55 | 1.24 | 1.93 |
| 9 | 10 | 2.60 | 1.25 | 1.96 |
| 10 | 1 | 0.74 | 0.57 | 0.69 |
| 10 | 2 | 2.14 | 1.08 | 1.64 |
| 10 | 3 | 2.56 | 1.25 | 1.94 |
| 10 | 5 | 2.83 | 1.37 | 2.14 |
| 10 | 10 | 2.93 | 1.42 | 2.22 |

Table A.2: Mean RMSE of AR(1) forecasts under low correlation ($\phi = 0.3$). Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.72 | 0.46 | 0.70 |
| 3 | 2 | 0.82 | 0.46 | 0.73 |
| 3 | 3 | 0.82 | 0.46 | 0.73 |
| 3 | 5 | 0.82 | 0.46 | 0.73 |
| 3 | 10 | 0.82 | 0.46 | 0.73 |
| 4 | 1 | 0.70 | 0.42 | 0.59 |
| 4 | 2 | 1.12 | 0.70 | 0.97 |
| 4 | 3 | 1.20 | 0.73 | 1.05 |
| 4 | 5 | 1.23 | 0.73 | 1.06 |
| 4 | 10 | 1.23 | 0.73 | 1.07 |
| 5 | 1 | 0.90 | 0.68 | 0.86 |
| 5 | 2 | 1.41 | 0.73 | 1.16 |
| 5 | 3 | 1.41 | 0.73 | 1.16 |
| 5 | 5 | 1.41 | 0.73 | 1.16 |

*Continued on next page*

Table A.2 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 5 | 10 | 1.41 | 0.73 | 1.16 |
| 6 | 1 | 1.11 | 0.62 | 0.95 |
| 6 | 2 | 1.67 | 0.89 | 1.35 |
| 6 | 3 | 1.75 | 0.92 | 1.43 |
| 6 | 5 | 1.79 | 0.93 | 1.44 |
| 6 | 10 | 1.79 | 0.93 | 1.44 |
| 7 | 1 | 1.27 | 0.82 | 1.09 |
| 7 | 2 | 1.98 | 0.99 | 1.57 |
| 7 | 3 | 2.00 | 0.99 | 1.57 |
| 7 | 5 | 2.00 | 0.99 | 1.57 |
| 7 | 10 | 2.00 | 0.99 | 1.57 |
| 8 | 1 | 1.45 | 0.81 | 1.22 |
| 8 | 2 | 2.23 | 1.11 | 1.73 |
| 8 | 3 | 2.32 | 1.15 | 1.81 |
| 8 | 5 | 2.36 | 1.16 | 1.82 |
| 8 | 10 | 2.36 | 1.16 | 1.82 |
| 9 | 1 | 1.62 | 0.94 | 1.35 |
| 9 | 2 | 2.54 | 1.23 | 1.95 |
| 9 | 3 | 2.59 | 1.25 | 1.97 |
| 9 | 5 | 2.59 | 1.25 | 1.97 |
| 9 | 10 | 2.59 | 1.25 | 1.97 |
| 10 | 1 | 1.79 | 1.00 | 1.48 |
| 10 | 2 | 2.80 | 1.34 | 2.13 |
| 10 | 3 | 2.89 | 1.38 | 2.20 |
| 10 | 5 | 2.92 | 1.38 | 2.21 |
| 10 | 10 | 2.93 | 1.38 | 2.21 |

Table A.3: Mean RMSE of MC forecasts under high correlation ($\phi = 0.7$). Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.74 | 0.46 | 0.67 |
| 3 | 2 | 0.90 | 0.49 | 0.80 |
| 3 | 3 | 0.98 | 0.51 | 0.88 |
| 3 | 5 | 1.07 | 0.53 | 0.97 |
| 3 | 10 | 1.14 | 0.54 | 1.03 |
| 4 | 1 | 0.97 | 0.56 | 0.82 |
| 4 | 2 | 1.19 | 0.64 | 1.00 |

*Continued on next page*

Table A.3 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 4 | 3 | 1.32 | 0.68 | 1.12 |
| 4 | 5 | 1.46 | 0.74 | 1.24 |
| 4 | 10 | 1.55 | 0.77 | 1.33 |
| 5 | 1 | 1.19 | 0.64 | 0.97 |
| 5 | 2 | 1.49 | 0.75 | 1.21 |
| 5 | 3 | 1.66 | 0.81 | 1.36 |
| 5 | 5 | 1.84 | 0.89 | 1.51 |
| 5 | 10 | 1.98 | 0.94 | 1.62 |
| 6 | 1 | 1.44 | 0.72 | 1.11 |
| 6 | 2 | 1.83 | 0.82 | 1.41 |
| 6 | 3 | 2.04 | 0.88 | 1.58 |
| 6 | 5 | 2.27 | 0.95 | 1.77 |
| 6 | 10 | 2.43 | 1.00 | 1.91 |
| 7 | 1 | 1.75 | 0.80 | 1.26 |
| 7 | 2 | 2.21 | 0.95 | 1.60 |
| 7 | 3 | 2.46 | 1.05 | 1.79 |
| 7 | 5 | 2.72 | 1.14 | 2.00 |
| 7 | 10 | 2.90 | 1.21 | 2.16 |
| 8 | 1 | 2.04 | 0.89 | 1.40 |
| 8 | 2 | 2.64 | 1.06 | 1.77 |
| 8 | 3 | 2.95 | 1.15 | 1.99 |
| 8 | 5 | 3.26 | 1.25 | 2.21 |
| 8 | 10 | 3.47 | 1.31 | 2.40 |
| 9 | 1 | 2.30 | 0.97 | 1.54 |
| 9 | 2 | 2.99 | 1.16 | 1.94 |
| 9 | 3 | 3.35 | 1.27 | 2.17 |
| 9 | 5 | 3.71 | 1.39 | 2.40 |
| 9 | 10 | 3.96 | 1.48 | 2.58 |
| 10 | 1 | 3.01 | 1.29 | 2.05 |
| 10 | 2 | 3.68 | 1.44 | 2.35 |
| 10 | 3 | 3.98 | 1.50 | 2.54 |
| 10 | 5 | 4.23 | 1.58 | 2.76 |
| 10 | 10 | 4.39 | 1.65 | 2.92 |

Table A.4: Mean RMSE of MC forecasts under low correlation ($\phi = 0.3$). Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 1.02 | 0.46 | 0.74 |
| 3 | 2 | 1.12 | 0.46 | 0.74 |
| 3 | 3 | 1.15 | 0.46 | 0.74 |
| 3 | 5 | 1.16 | 0.46 | 0.74 |
| 3 | 10 | 1.16 | 0.46 | 0.74 |
| 4 | 1 | 1.57 | 0.69 | 0.97 |
| 4 | 2 | 1.72 | 0.74 | 1.04 |
| 4 | 3 | 1.77 | 0.75 | 1.06 |
| 4 | 5 | 1.79 | 0.76 | 1.07 |
| 4 | 10 | 1.78 | 0.76 | 1.08 |
| 5 | 1 | 1.96 | 0.72 | 1.18 |
| 5 | 2 | 2.23 | 0.73 | 1.24 |
| 5 | 3 | 2.29 | 0.73 | 1.25 |
| 5 | 5 | 2.30 | 0.73 | 1.26 |
| 5 | 10 | 2.31 | 0.73 | 1.26 |
| 6 | 1 | 2.40 | 0.88 | 1.38 |
| 6 | 2 | 2.77 | 0.93 | 1.44 |
| 6 | 3 | 2.84 | 0.94 | 1.47 |
| 6 | 5 | 2.88 | 0.95 | 1.48 |
| 6 | 10 | 2.88 | 0.95 | 1.48 |
| 7 | 1 | 2.84 | 0.98 | 1.58 |
| 7 | 2 | 3.29 | 0.99 | 1.65 |
| 7 | 3 | 3.39 | 0.99 | 1.66 |
| 7 | 5 | 3.42 | 0.99 | 1.67 |
| 7 | 10 | 3.42 | 0.99 | 1.67 |
| 8 | 1 | 3.24 | 1.10 | 1.78 |
| 8 | 2 | 3.78 | 1.16 | 1.87 |
| 8 | 3 | 3.93 | 1.17 | 1.88 |
| 8 | 5 | 3.98 | 1.18 | 1.89 |
| 8 | 10 | 3.97 | 1.18 | 1.90 |
| 9 | 1 | 3.63 | 1.21 | 2.00 |
| 9 | 2 | 4.26 | 1.25 | 2.09 |
| 9 | 3 | 4.44 | 1.26 | 2.12 |
| 9 | 5 | 4.51 | 1.26 | 2.12 |
| 9 | 10 | 4.52 | 1.26 | 2.12 |
| 10 | 1 | 4.27 | 1.54 | 2.54 |

*Continued on next page*

Table A.4 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 10 | 2 | 4.66 | 1.63 | 2.51 |
| 10 | 3 | 4.75 | 1.64 | 2.51 |
| 10 | 5 | 4.77 | 1.64 | 2.52 |
| 10 | 10 | 4.78 | 1.64 | 2.52 |

### A.1.2 Multivariate Results

Table A.5: Mean RMSE of discretized VAR forecasts under low correlation. Results are reported across discretization methods (EFI, EWI, K-means) and bin sizes.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.59 | 0.45 | 0.61 |
| 3 | 2 | 0.81 | 0.46 | 0.73 |
| 3 | 3 | 0.82 | 0.46 | 0.73 |
| 3 | 5 | 0.82 | 0.46 | 0.73 |
| 3 | 10 | 0.82 | 0.46 | 0.73 |
| 4 | 1 | 0.70 | 0.43 | 0.64 |
| 4 | 2 | 1.06 | 0.65 | 0.91 |
| 4 | 3 | 1.15 | 0.71 | 1.00 |
| 4 | 5 | 1.21 | 0.73 | 1.05 |
| 4 | 10 | 1.22 | 0.73 | 1.06 |
| 5 | 1 | 0.83 | 0.60 | 0.75 |
| 5 | 2 | 1.33 | 0.72 | 1.12 |
| 5 | 3 | 1.41 | 0.73 | 1.16 |
| 5 | 5 | 1.41 | 0.73 | 1.16 |
| 5 | 10 | 1.41 | 0.73 | 1.16 |
| 6 | 1 | 0.97 | 0.60 | 0.85 |
| 6 | 2 | 1.59 | 0.83 | 1.27 |
| 6 | 3 | 1.70 | 0.90 | 1.36 |
| 6 | 5 | 1.76 | 0.92 | 1.43 |
| 6 | 10 | 1.78 | 0.92 | 1.44 |
| 7 | 1 | 1.12 | 0.70 | 0.96 |
| 7 | 2 | 1.85 | 0.95 | 1.47 |
| 7 | 3 | 1.99 | 0.98 | 1.56 |
| 7 | 5 | 2.00 | 0.98 | 1.57 |
| 7 | 10 | 2.00 | 0.98 | 1.57 |
| 8 | 1 | 1.26 | 0.75 | 1.07 |
| 8 | 2 | 2.11 | 1.04 | 1.65 |

*Continued on next page*

Table A.5 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 8 | 3 | 2.26 | 1.12 | 1.76 |
| 8 | 5 | 2.33 | 1.14 | 1.81 |
| 8 | 10 | 2.35 | 1.14 | 1.82 |
| 9 | 1 | 1.41 | 0.82 | 1.18 |
| 9 | 2 | 2.38 | 1.17 | 1.83 |
| 9 | 3 | 2.55 | 1.22 | 1.96 |
| 9 | 5 | 2.58 | 1.23 | 1.98 |
| 9 | 10 | 2.59 | 1.23 | 1.98 |
| 10 | 1 | 1.56 | 0.89 | 1.29 |
| 10 | 2 | 2.63 | 1.27 | 2.02 |
| 10 | 3 | 2.83 | 1.35 | 2.16 |
| 10 | 5 | 2.89 | 1.38 | 2.21 |
| 10 | 10 | 2.92 | 1.38 | 2.22 |

Table A.6: Mean RMSE of VAR discretized forecasts under high correlation. Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.28 | 0.24 | 0.28 |
| 3 | 2 | 0.55 | 0.38 | 0.51 |
| 3 | 3 | 0.64 | 0.42 | 0.59 |
| 3 | 5 | 0.74 | 0.46 | 0.67 |
| 3 | 10 | 0.82 | 0.47 | 0.73 |
| 4 | 1 | 0.35 | 0.27 | 0.33 |
| 4 | 2 | 0.68 | 0.47 | 0.61 |
| 4 | 3 | 0.81 | 0.53 | 0.71 |
| 4 | 5 | 0.96 | 0.59 | 0.82 |
| 4 | 10 | 1.09 | 0.68 | 0.94 |
| 5 | 1 | 0.39 | 0.30 | 0.37 |
| 5 | 2 | 0.81 | 0.52 | 0.69 |
| 5 | 3 | 0.98 | 0.59 | 0.82 |
| 5 | 5 | 1.18 | 0.66 | 0.97 |
| 5 | 10 | 1.36 | 0.74 | 1.12 |
| 6 | 1 | 0.44 | 0.33 | 0.41 |
| 6 | 2 | 0.94 | 0.57 | 0.77 |
| 6 | 3 | 1.16 | 0.65 | 0.93 |
| 6 | 5 | 1.40 | 0.75 | 1.12 |
| 6 | 10 | 1.63 | 0.86 | 1.30 |

*Continued on next page*

Table A.6 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 7 | 1 | 0.48 | 0.36 | 0.44 |
| 7 | 2 | 1.08 | 0.62 | 0.86 |
| 7 | 3 | 1.34 | 0.72 | 1.05 |
| 7 | 5 | 1.62 | 0.84 | 1.27 |
| 7 | 10 | 1.90 | 0.97 | 1.49 |
| 8 | 1 | 0.51 | 0.38 | 0.47 |
| 8 | 2 | 1.22 | 0.68 | 0.96 |
| 8 | 3 | 1.52 | 0.79 | 1.18 |
| 8 | 5 | 1.85 | 0.94 | 1.43 |
| 8 | 10 | 2.16 | 1.08 | 1.68 |
| 9 | 1 | 0.55 | 0.40 | 0.50 |
| 9 | 2 | 1.36 | 0.73 | 1.05 |
| 9 | 3 | 1.70 | 0.87 | 1.30 |
| 9 | 5 | 2.07 | 1.03 | 1.58 |
| 9 | 10 | 2.43 | 1.20 | 1.87 |
| 10 | 1 | 0.58 | 0.43 | 0.53 |
| 10 | 2 | 1.50 | 0.78 | 1.14 |
| 10 | 3 | 1.88 | 0.94 | 1.42 |
| 10 | 5 | 2.30 | 1.13 | 1.74 |
| 10 | 10 | 2.70 | 1.32 | 2.05 |

Table A.7: Mean RMSE of MMC (MTD-Probit) forecasts under low correlation. Results are grouped by discretization method and bin size.

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 1.01 | 0.46 | 0.75 |
| 3 | 2 | 1.13 | 0.46 | 0.80 |
| 3 | 3 | 1.19 | 0.46 | 0.82 |
| 3 | 5 | 1.22 | 0.46 | 0.85 |
| 3 | 10 | 1.24 | 0.46 | 0.85 |
| 4 | 1 | 1.40 | 0.67 | 0.92 |
| 4 | 2 | 1.62 | 0.73 | 1.00 |
| 4 | 3 | 1.72 | 0.75 | 1.04 |
| 4 | 5 | 1.78 | 0.77 | 1.07 |
| 4 | 10 | 1.81 | 0.78 | 1.08 |
| 5 | 1 | 1.80 | 0.72 | 1.11 |
| 5 | 2 | 2.16 | 0.73 | 1.21 |
| 5 | 3 | 2.29 | 0.73 | 1.24 |

*Continued on next page*

Table A.7 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 5 | 5 | 2.37 | 0.73 | 1.27 |
| 5 | 10 | 2.41 | 0.73 | 1.28 |
| 6 | 1 | 2.21 | 0.85 | 1.28 |
| 6 | 2 | 2.66 | 0.92 | 1.38 |
| 6 | 3 | 2.84 | 0.95 | 1.42 |
| 6 | 5 | 2.94 | 0.97 | 1.45 |
| 6 | 10 | 2.98 | 0.98 | 1.46 |
| 7 | 1 | 2.59 | 0.95 | 1.48 |
| 7 | 2 | 3.13 | 0.98 | 1.59 |
| 7 | 3 | 3.36 | 0.98 | 1.63 |
| 7 | 5 | 3.49 | 0.98 | 1.65 |
| 7 | 10 | 3.55 | 0.98 | 1.66 |
| 8 | 1 | 2.98 | 1.06 | 1.67 |
| 8 | 2 | 3.63 | 1.14 | 1.80 |
| 8 | 3 | 3.93 | 1.16 | 1.84 |
| 8 | 5 | 4.10 | 1.18 | 1.86 |
| 8 | 10 | 4.16 | 1.19 | 1.87 |
| 9 | 1 | 3.35 | 1.17 | 1.86 |
| 9 | 2 | 4.11 | 1.23 | 2.00 |
| 9 | 3 | 4.43 | 1.24 | 2.04 |
| 9 | 5 | 4.64 | 1.25 | 2.07 |
| 9 | 10 | 4.73 | 1.25 | 2.08 |
| 10 | 1 | 3.74 | 1.28 | 2.05 |
| 10 | 2 | 4.61 | 1.37 | 2.22 |
| 10 | 3 | 4.99 | 1.40 | 2.27 |
| 10 | 5 | 5.22 | 1.41 | 2.30 |
| 10 | 10 | 5.31 | 1.42 | 2.31 |

Table A.8: Mean RMSE of MTD-Probit (MMC) forecasts under high correlation. Results are shown side-by-side across discretization methods (EFI, EWI, K-means) and bins.

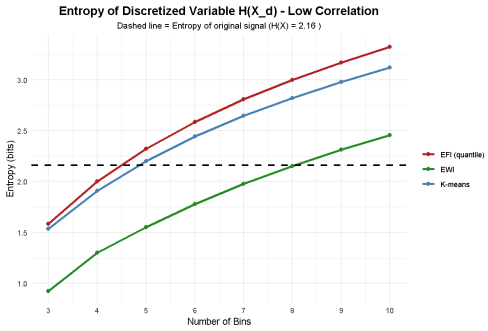
| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 3 | 1 | 0.62 | 0.40 | 0.54 |
| 3 | 2 | 0.72 | 0.43 | 0.63 |
| 3 | 3 | 0.80 | 0.46 | 0.69 |
| 3 | 5 | 0.90 | 0.49 | 0.77 |
| 3 | 10 | 1.05 | 0.54 | 0.90 |
| 4 | 1 | 0.78 | 0.50 | 0.66 |

*Continued on next page*

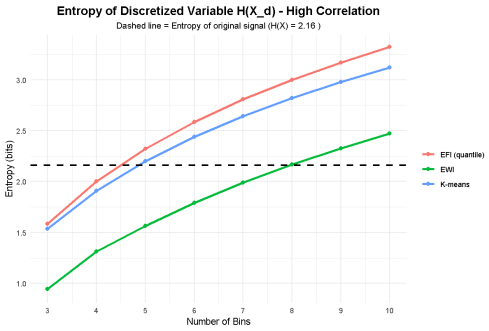
Table A.8 – *continued from previous page*

| Bins | Horizon | EFI (quantile) | EWI | K-means |
|---|---|---|---|---|
| 4 | 2 | 0.93 | 0.55 | 0.77 |
| 4 | 3 | 1.04 | 0.58 | 0.84 |
| 4 | 5 | 1.20 | 0.63 | 0.97 |
| 4 | 10 | 1.42 | 0.71 | 1.14 |
| 5 | 1 | 0.94 | 0.56 | 0.75 |
| 5 | 2 | 1.14 | 0.62 | 0.89 |
| 5 | 3 | 1.29 | 0.66 | 0.99 |
| 5 | 5 | 1.49 | 0.73 | 1.15 |
| 5 | 10 | 1.79 | 0.83 | 1.37 |
| 6 | 1 | 1.10 | 0.63 | 0.85 |
| 6 | 2 | 1.36 | 0.69 | 1.02 |
| 6 | 3 | 1.54 | 0.75 | 1.15 |
| 6 | 5 | 1.79 | 0.83 | 1.33 |
| 6 | 10 | 2.16 | 0.96 | 1.60 |
| 7 | 1 | 1.26 | 0.68 | 0.94 |
| 7 | 2 | 1.57 | 0.77 | 1.14 |
| 7 | 3 | 1.79 | 0.83 | 1.29 |
| 7 | 5 | 2.09 | 0.93 | 1.51 |
| 7 | 10 | 2.53 | 1.09 | 1.83 |
| 8 | 1 | 1.42 | 0.73 | 1.03 |
| 8 | 2 | 1.78 | 0.85 | 1.28 |
| 8 | 3 | 2.03 | 0.93 | 1.45 |
| 8 | 5 | 2.39 | 1.04 | 1.71 |
| 8 | 10 | 2.89 | 1.22 | 2.07 |
| 9 | 1 | 1.58 | 0.81 | 1.12 |
| 9 | 2 | 2.00 | 0.93 | 1.40 |
| 9 | 3 | 2.30 | 1.02 | 1.60 |
| 9 | 5 | 2.70 | 1.16 | 1.90 |
| 9 | 10 | 3.27 | 1.35 | 2.31 |
| 10 | 1 | 1.74 | 0.86 | 1.21 |
| 10 | 2 | 2.24 | 1.01 | 1.53 |
| 10 | 3 | 2.58 | 1.12 | 1.75 |
| 10 | 5 | 3.05 | 1.26 | 2.07 |
| 10 | 10 | 3.68 | 1.48 | 2.53 |

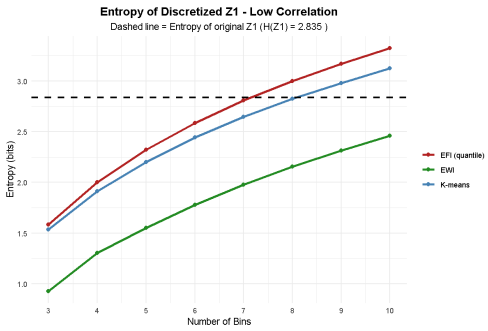
## A.2 Information Loss Results



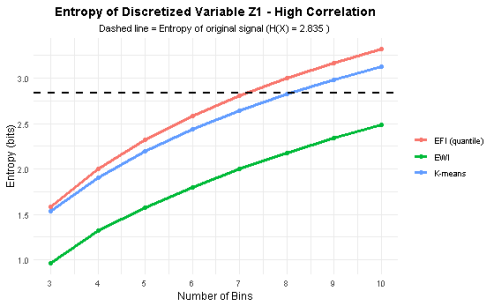

(a) Univariate - Low Correlation




(b) Univariate - High Correlation




(c) Multivariate - Low Correlation




(d) Multivariate - High Correlation

Figure A.1: Entropy of Discretized Variable $H(X_d)$ under Different Correlation Scenarios

## A.3 Other Plots

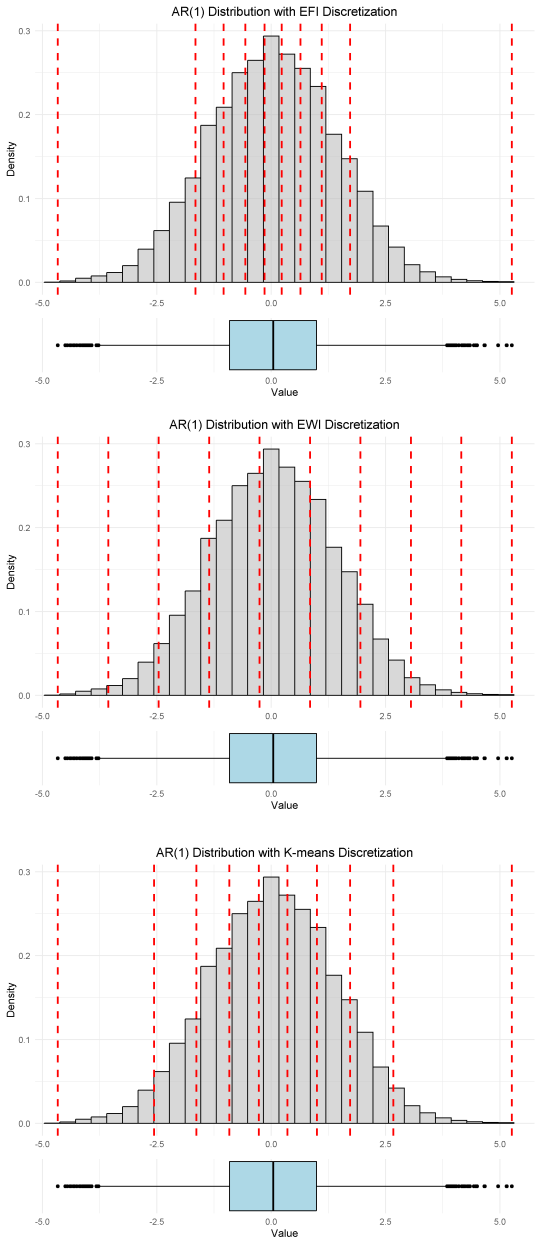
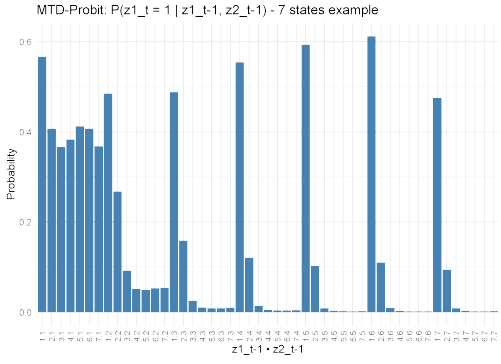
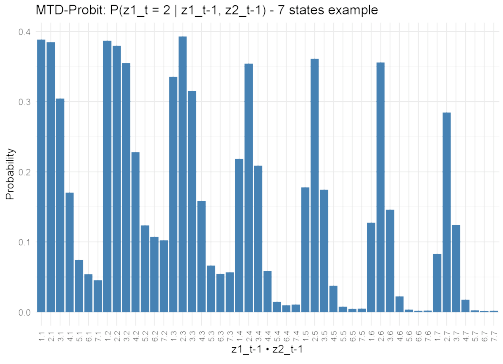


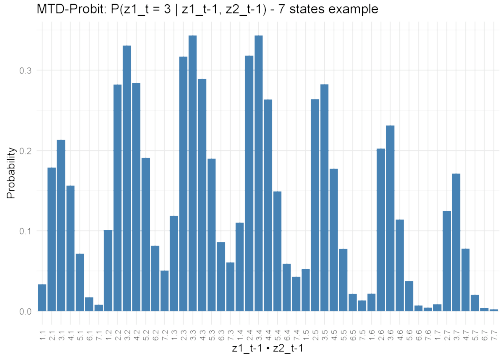

Figure A.2: AR(1) distributions with 9-bin discretizations using EFI, EWI, and K-means methods.



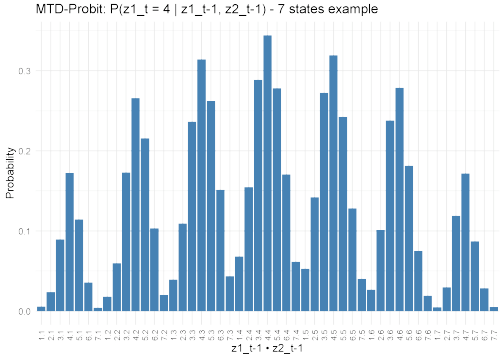

(a) $z_{1,t} = 1$



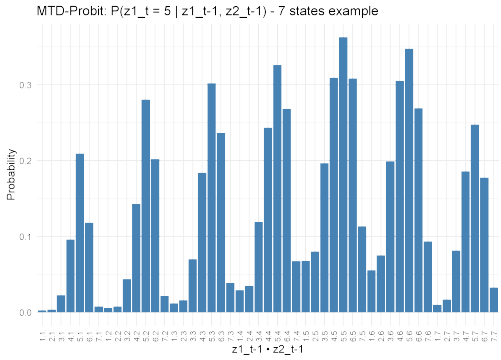

(b) $z_{1,t} = 2$



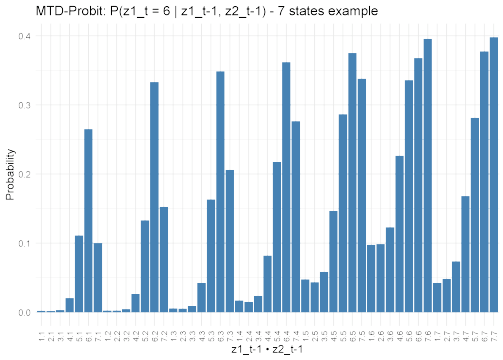

(c) $z_{1,t} = 3$



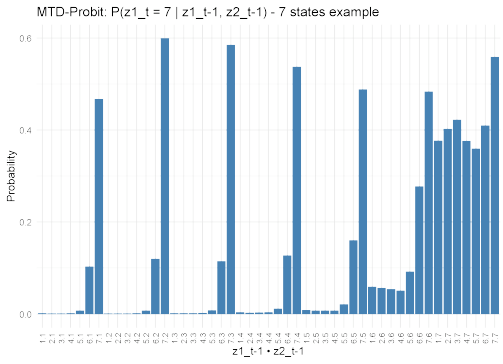

(d) $z_{1,t} = 4$




(e) $z_{1,t} = 5$




(f) $z_{1,t} = 6$




(g) $z_{1,t} = 7$

Figure A.3: MTD-Probit conditional probabilities for each state $z_{1,t}$ (7 state example).


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade NOVA de Lisboa