Université de Lausanne
Faculté De Droit, Des Sciences Criminelles Et D'administration Publique
Ecole Des Sciences Criminelles

and

Athens University of Economics and Business
School of Information Sciences and Technology
Department of Statistics

Statistical Methods for the Forensic Examination of Handwriting and Dynamic Signatures Data.

PhD THESIS of

Lampis TZAI

(Co-tutelle PhD in Forensic Science and Statistics)

Thesis Supervisors:
Professor Franco TARONI
Professor Ioannis NTZOUFRAS

Research supported by the Swiss National Science Foundation, through grant no. 100011_204554/1

Summary

Nowadays, the evaluation of forensic evidence increasingly relies on statistical methodologies, which provide a coherent framework for quantifying evidential strength and addressing uncertainty. In the domains of handwriting and dynamic signature examination, the adoption of probabilistic models is enabling more objective, transparent, and scientifically grounded evaluations. This dissertation contributes to that ongoing development through two complementary studies.

The first study introduces a novel statistical approach that aims at the identification of valid and useful patterns in handwriting examination via Bayesian modeling. Beginning with handwritten manuscript images, loop-character contours are characterized through Fourier-based features, including the first four pairs of coefficients and the surface size. Six Bayesian models are examined for such handwritten features. These models arise from two likelihood structures: (a) a Bayesian Normal model and (b) a Bayesian MANOVA model. For each likelihood, three different prior specifications are considered: a conjugate approach, a hierarchical Normal-Inverse-Wishart, and a Normal-LogNormal-LKJ prior. The Bayesian MANOVA model with hierarchical prior formulations is of primary interest because it can incorporate the within- and between-writer variability, as well as the between-character variability, which are particular distinguishing elements. Bayes factors are employed to assess the support for the competing propositions of interest (the person of interest (PoI) is the writer of the questioned document or not), aligning with the forensic international guidelines ENFSI. The study explores model comparison, the discriminative power, and a sensitivity analysis with respect to prior elicitation.

The second study addresses the evaluation of the dynamic signature. Dynamic signatures are a behavioural biometric modality that captures the temporal characteristics of the signature. Hidden Markov Models (HMMs), including both Bakis and Ergodic structures, are used to model the sequential multivariate nature of signature data and to assess the likelihood ratios associated with the competing propositions of interest (questioned signature is genuine and was made by PoI or simulated). The study explores feature selection, automated determination of the hidden states, and the use of combined discriminative dynamic features to reduce the misleading support between competing propositions.

Overall, these studies demonstrate how statistical methodologies can significantly advance forensic handwriting and dynamic signature examination. The integration of rigorous Bayesian and HMM-based techniques contributes to the development of more robust, interpretable, and scientifically grounded tools for the evaluation of evidence in both traditional and digital forensic contexts.

Résumé

De nos jours, l'évaluation des preuves judiciaires repose de plus en plus sur des méthodologies statistiques, qui offrent un cadre cohérent pour quantifier la force probante et gérer l'incertitude. Dans les domaines de l'analyse de l'écriture manuscrite et de la signature dynamique, l'adoption de modèles probabilistes permet des évaluations plus objectives, transparentes et scientifiquement fondées. Cette thèse contribue à ce développement à travers deux études complémentaires.

La première étude présente une nouvelle approche statistique visant à identifier des motifs valides et utiles dans l'examen de l'écriture manuscrite via la modélisation bayésienne. À partir d'images de manuscrits, les contours des caractères en boucle sont décrits à l'aide de caractéristiques basées sur la transformation de Fourier, incluant les quatre premières paires de coefficients et la taille de la surface. Six modèles bayésiens sont examinés pour ces caractéristiques manuscrites. Ces modèles reposent sur deux structures de vraisemblance : (a) un modèle bayésien normal et (b) un modèle bayésien MANOVA. Pour chaque vraisemblance, trois spécifications de priors différentes sont considérées : une approche conjointe, un prior hiérarchique Normal–Inverse-Wishart et un prior Normal–LogNormal–LKJ. Le modèle bayésien MANOVA avec priors hiérarchiques est d'un intérêt particulier, car il permet d'incorporer la variabilité intra- et inter-auteurs ainsi que la variabilité inter-caractères, qui constitue un élément distinctif essentiel. Les facteurs de Bayes sont utilisés pour évaluer le soutien aux propositions concurrentes d'intérêt (la personne d'intérêt (PoI) est-elle l'auteur du document contesté ou non), conformément aux recommandations internationales en matière de criminalistique de l'ENFSI. L'étude inclut la comparaison des modèles, l'évaluation du pouvoir discriminant et une analyse de sensibilité vis-à-vis de la spécification des priors.

La deuxième étude concerne l'évaluation des signatures dynamiques. Les signatures dynamiques constituent un type de biométrie comportementale qui capture les caractéristiques temporelles de la signature. Des modèles de Markov cachés (Hidden Markov Models – HMM), comprenant à la fois des structures Bakis et ergodiques, sont utilisés pour modéliser la nature séquentielle et multivariée des données de signature et pour estimer les rapports de vraisemblance associés aux propositions concurrentes (la signature contestée est-elle authentique et réalisée par la PoI ou simulée). L'étude porte sur la sélection de caractéristiques, la détermination automatique du nombre d'états cachés et l'utilisation de caractéristiques dynamiques discriminantes combinées afin de réduire les soutiens erronés entre propositions concurrentes.

Dans l'ensemble, ces études démontrent comment les méthodologies statistiques peuvent améliorer de manière significative l'examen judiciaire de l'écriture manuscrite et de la signature dynamique. L'intégration de techniques bayésiennes rigoureuses et de modèles HMM contribue au développement d'outils plus robustes, interprétables et scientifiquement fondés pour l'évaluation des preuves, tant dans les contextes judiciaires traditionnels que numériques.

Περίληψη

Στη σύγχρονη εποχή, η αξιολόγηση των εγκληματολογικών στοιχειών βασίζεται ολοένα και περισσότερο σε στατιστικές μεθόδους, οι οποίες παρέχουν ένα ολοκληρωμένο πλαίσιο για την ποσοτικοποίηση της αποδεικτικής ισχύος και τη διαχείριση της αβεβαιότητας. Στους τομείς της εξέτασης γραφικού χαρακτήρα και της δυναμικής ψηφιακής υπογραφής, η υιοθέτηση πιθανοθεωρητικών μοντέλων επιτρέπει πιο αντικειμενικές, διαφανείς και επιστημονικά τεκμηριωμένες αξιολογήσεις. Η παρούσα διατριβή συμβάλλει σε αυτήν την εξέλιξη μέσα από δύο συμπληρωματικές μελέτες.

Η πρώτη μελέτη παρουσιάζει μια νέα στατιστική προσέγγιση που αποσκοπεί στον εντοπισμό έγκυρων και χρήσιμων μοτίβων στην εξέταση γραφικού χαρακτήρα μέσω Μπεϋζιανής μοντελοποίησης. Ξεκινώντας από εικόνες χειρόγραφων κειμένων, τα περιγράμματα κυκλοειδών χαρακτήρων περιγράφονται με χαρακτηριστικά (features) βασισμένα στο μετασχηματισμό Fourier, συμπεριλαμβανομένων των τεσσάρων πρώτων ζευγών συντελεστών και του μεγέθους του χαρακτήρα. Εξετάζονται έξι Μπεϋζιανά μοντέλα για τα χαρακτηριστικά (features) αυτά, τα οποία προκύπτουν από δύο δομές πιθανοφάνειας: (α) ένα Bayesian Normal μοντέλο και (β) ένα Bayesian μοντέλο MANOVA. Για κάθε πιθανοφάνεια, εξετάζονται τρεις διαφορετικές μορφές εκ-των-προτέρων κατανομών: η συζυγής προσέγγιση, το ιεραρχικό Normal-Inverse-Wishart και το Normal-LogNormal-LKJ. Το Bayesian μοντέλο MANOVA με ιεραρχική μορφή αποτελεί το βασικό σημείο ενδιαφέροντος, καθώς λαμβάνει υπόψη τη μεταβλητότητα εντός και μεταξύ συγγραφέων, καθώς και τη μεταβλητότητα μεταξύ χαρακτήρων, στοιχεία καθοριστικής σημασίας για τη διακριτική ικανότητα. Οι παράγοντες Bayes (Bayes factors) χρησιμοποιούνται για την αξιολόγηση της υποστήριξης των υποθέσεων ενδιαφέροντος (αν το εξεταζόμενο έγγραφο γράφτηκε από το άτομο ενδιαφέροντος ή όχι), ακολουθώντας τις διεθνείς εγκληματολογικές οδηγίες της ENFSI. Η μελέτη εξετάζει επίσης τη σύγκριση μοντέλων, τη διακριτική ισχύ και μια ανάλυση ευαισθησίας ως προς την επιλογή εκ-των-προτέρων παραμέτρων.

Η δεύτερη μελέτη αφορά την αξιολόγηση των δυναμικών υπογραφών. Οι δυναμικές υπογραφές αποτελούν μια βιομετρική μέθοδο που καταγράφει τα χρονικά χαρακτηριστικά της υπογραφής. Τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models – HMMs), συμπεριλαμβανομένων τόσο των δομών Bakis όσο και των Εργοδικών, χρησιμοποιούνται για τη μοντελοποίηση της διαδοχικής πολυμεταβλητής φύσης των δεδομένων υπογραφής και για την εκτίμηση των λόγων πιθανοφάνειας που αντιστοιχούν στις υποθέσεις ενδιαφέροντος (η εξεταζόμενη υπογραφή είναι γνήσια και γράφτηκε από το ενδιαφερόμενο άτομο, ή είναι απομίμηση). Η μελέτη διερευνά την επιλογή χαρακτηριστικών, τον αυτοματοποιημένο καθορισμό του αριθμού των κρυφών καταστάσεων, καθώς και τη χρήση συνδυαστικών διακριτικών δυναμικών χαρακτηριστικών με στόχο τη μείωση της εσφαλμένης υποστήριξης μεταξύ υποθέσεων ενδιαφέροντος.

Συνολικά, οι μελέτες αυτές καταδεικνύουν πώς οι στατιστικές μεθοδολογίες μπορούν να ενισχύσουν ουσιαστικά την εγκληματολογική εξέταση γραφικού χαρακτήρα και ψηφιακής δυναμικής υπογραφής. Η ενσωμάτωση αυστηρών Bayesian τεχνικών και μοντέλων HMM συμβάλλει στην ανάπτυξη πιο αξιόπιστων, κατανοητών και επιστημονικά θεμελιωμένων εργαλείων για την αξιολόγηση εγκληματολογικών στοιχειών τόσο σε παραδοσιακά όσο και σε ψηφιακά εγκληματολογικά περιβάλλοντα.

Acknowledgments

This thesis was fulfilled across two countries, two cities, two universities, and through collaborative efforts between two departments: (a) the School of Criminal Justice at the Faculty of Law, Criminal Justice, and Public Administration of the University of Lausanne, (b) the Department of Statistics at the School of Information Sciences and Technology of the Athens University of Economics and Business. The dissertation was co-supervised by Prof. Franco Taroni and Prof. Ioannis Ntzoufras, with Prof. Silvia Bozza serving as an advisor, and was funded by the Swiss National Science Foundation (SNSF).

First of all, I would like to thank my supervisors for the opportunity that they gave me and their invaluable guidance throughout this journey. They provided guidance in my research as well as in practical aspects of life. Each of them supported my work in their own unique way. I am also deeply grateful to them and to the administrative staff for their dedication to making this collaboration possible and for managing the substantial administrative workload behind the scenes.

Moreover, I am deeply grateful to my advisor, Prof. Silvia Bozza, for her guidance at the beginning of my PhD journey, helping me to shape my research direction based on the provided literature review. Most importantly, I am especially thankful for her commitment to reviewing and correcting my papers, even during times of ill health.

It is commonly noted that research progresses by building on previous contributions. For that reason, I would like to thank Dr. Raymond Marquis for his expert advice in handwriting analysis and for providing the preprocessed handwriting data from his thesis. Furthermore, I would also like to thank Dr. Jacques Linden for sharing the dynamic signature data from his thesis. Their contributions and data were essential in building the foundation of my research.

Throughout my PhD journey, I had the opportunity to meet and exchange ideas with many scientists and researchers. I would like to express my sincere gratitude to all of them, even though it is not possible to mention each one by name. In particular, I would like to thank my academic friends - Prof. Vasilis Chasiotis, Dr. Roberto Macrì-Demartino, and his supervisor Prof. Leonardo Egidi, and Dr. Lea Anna Cozzucoli-for the insightful research discussions and the pleasant moments we shared beyond academia. I am also grateful to my lab colleagues and fellow soon-to-be PhDs: Argyro Damoulaki, Anna Nalpantidi, and Andre Ehrlich.

Finally, I wish to express my deepest gratitude to my family - my father, Grigoris Tzai, my mother, Alexandra Tzai, my sister Christina Tzai, and her little daughter, Anastasia Mousouraki - for their love and mental support throughout this journey. I am also grateful to my friend Stefanos Efthymiopoulos for being my greatest supporter since my bachelor studies, and to my partner, Eirini Kordatzi, for being by my side and inspiring me to look deepest into my feelings and to become a better person.

Research Output and Publications

This thesis is based on original research conducted at the University of Lausanne (UNIL) and the Athens University of Economics and Business (AUEB) during the joint doctoral programme in Forensic Statistics. The following publications and research outputs have been produced during this doctoral work, and material from these works has been incorporated in the thesis.

Journal Articles

[J1] Tzai, L., Ntzoufras, I., & Bozza, S. (2026). Bayesian Handwriting Evidence Evaluation using MANOVA via Fourier-Based Extracted Features. *(Under review in a statistical journal)* Available at arXiv preprint: https://arxiv.org/abs/2601.07534

[J2] Tzai, L., Bozza, S., Marquis, R., Ntzoufras, I., & Taroni, F. (2026). Semi-Automated Forensic Examination of Handwritten Character Loops. *Forensic Science International. (Published)* Available at: https://www.sciencedirect.com/science/article/abs/pii/S0379073826001532

[J3] Tzai, L., Ntzoufras, I., Taroni, F., & Bozza, S. (2026). Probabilistic HMM-Based Pipeline for Forensic Evaluation of Dynamic Signatures. *(Under review)*

[J4] Tzai, L. Ntzoufras, I., & Bozza, S. (2026) Bayesian Computation for Wishart Degrees of Freedom Estimation. *(In preparation — to be submitted)*

Conference Proceedings

[C1] Tzai, L., Ntzoufras, I., Taroni, F., & Bozza, S. (2025, June). Bayesian Modelling for Forensic Evaluation of Handwritten Loop Characters. Paper presented at the *12th Scientific Meeting of the Statistics for the Evaluation and Quality of Services Group (SVQS)*, Italian Statistical Society, Bressanone-Brixen, Italy.

Contents

Summary . . . i
Résumé . . . ii
Περίληψη . . . iii
Acknowledgments . . . iv
Research Output and Publications . . . v

1 Introduction **1**
1.1 Research Context . . . 2
1.2 Objectives . . . 3
1.3 Originality of the Research . . . 4
1.4 Thesis Outline . . . 6

2 Forensic Inference and Statistics **7**
2.1 Introduction to Forensic Interpretation . . . 8
2.2 Statistics and Evaluation of Evidence . . . 9
2.3 Reporting and Communicating Scientific Findings . . . 11
2.3.1 Pitfalls of Intuition . . . 12

3 Probabilistic Data Modeling **13**
3.1 Frequentist Data Modeling . . . 14
3.1.1 Maximum Likelihood Estimation . . . 14
3.1.2 Expectation-Maximization Algorithm . . . 16
3.2 Bayesian Data Modeling . . . 17
3.2.1 Fundamentals of Bayesian Modeling . . . 18
3.2.2 Markov Chain Monte Carlo . . . 20
3.2.3 Marginal Likelihood . . . 22
3.3 Conclusion . . . 25

4 Handwriting Examination **27**
4.1 Literature Review . . . 29
4.2 Data Acquisition, Sampling and Databases . . . 30
4.3 Image Preprocessing . . . 30
4.4 Fourier based Feature Engineering . . . 33
4.4.1 Descriptive Statistics and Visualizations . . . 35
4.5 Modeling Fourier Coefficients and Surface size . . . 37
4.5.1 Bayesian Normal Models . . . 38
4.5.2 Bayesian MANOVA . . . 40

4.6 Marginal Likelihoods Estimators . . . 43
4.6.1 Marginal Likelihoods for Model Comparison using the Full Dataset . . . 44
4.6.2 Comparisons of Marginal Likelihood Estimators . . . 44
4.6.3 Marginal Likelihoods for Handwriting Evidence Evaluation . . . 49
4.7 Experimental Results . . . 49
4.7.1 Model Comparisons per Writer . . . 50
4.7.2 Accuracy of Models . . . 52
4.8 Sensitivity Analysis of Prior Elicitation . . . 55
4.8.1 Subsampling of Background Data . . . 56
4.8.2 Specification of the Inverse-Wishart's Degrees of Freedom . . . 58
4.8.3 Specification of the Parameter of the LKJ Distribution . . . 59
4.9 Case study . . . 60
4.9.1 Preliminary Analysis . . . 61
4.9.2 Results . . . 61
4.10 Discussion and Conclusion . . . 63

5 Dynamic Signatures Examination **67**
5.1 Literature Review . . . 69
5.2 Data Acquisition, Sampling and Databases . . . 70
5.3 Dynamic Signature Features . . . 71
5.3.1 Explanatory Analysis . . . 74
5.4 Gaussian Hidden Markov Model . . . 80
5.4.1 Maximum Likelihood Approach . . . 82
5.5 Trajectory Resampling . . . 83
5.6 Goodness-of-Fit . . . 84
5.6.1 Pseudo-Residual Diagnostics . . . 84
5.6.2 Estimated Parameters . . . 86
5.6.3 Convergence . . . 88
5.7 Experiments . . . 90
5.7.1 Univariate Experiments . . . 91
5.7.2 Multi-feature Experiments . . . 93
5.7.3 Automated Selection of HMM's States . . . 95
5.7.4 Impact of the Number of Control Signatures . . . 97
5.8 Discussion and Conclusion . . . 98

6 Concluding Remarks **101**
6.1 Contributions . . . 102
6.2 Future Work . . . 103

Appendices **104**

A Handwriting Examination **105**
A.1 Assumptions Underlying the Bayes Factor . . . 105
A.2 Dummy Variables . . . 106
A.3 Analytical Representation of Fourier-based Features . . . 107
A.4 Prior Elicitation . . . 110
A.4.1 Prior Parameters Elicitation for Bayesian Normal Model . . . 110

A.4.2 Prior Parameters Estimation for Bayesian MANOVA . . . 111
A.5 Sensitivity of Prior Elicitation in Handwriting Examination . . . 111
A.5.1 Subsampling of Background Data Examples . . . 111
A.5.2 Sensitivity of the Inverse-Wishart's Degrees of Freedom . . . 113
A.5.3 Sensitivity of the Parameter of the LKJ Distribution . . . 114

B Dynamic Signature Examination **117**
B.1 Trajectory Resampling Procedure . . . 117
B.2 Q–Q Plots of Pseudo-Residuals . . . 118
B.2.1 Ergodic Structure . . . 118
B.2.2 Left-to-Right Structure . . . 121
B.3 Experimental Results per Feature and Signature . . . 123

C Wishart's Degrees of Freedom Estimation **125**
C.1 Literature Review . . . 126
C.2 Maximum Likelihood Estimation of Wishart Distribution . . . 127
C.2.1 Root-finding Algorithms for MLE of the Degrees of Freedom . . . 128
C.2.2 Simulated Annealing for Maximizing the Likelihood Function . . . 132
C.3 Bayesian Modeling of Wishart Distribution . . . 134
C.3.1 Hybrid MCMC Methods . . . 136
C.3.2 No-U-Turn Sampler (Stan) . . . 138
C.4 Experimental Results . . . 139
C.4.1 Simulated Data . . . 141
C.4.2 Simulated Data with Predefined Number of Iterations . . . 148
C.4.3 Real Datasets . . . 154
C.4.4 Real Datasets with Predefined Number of Iterations . . . 157
C.5 Discussion . . . 160

D Abbreviations **163**

Chapter 1

Introduction

This introductory chapter serves as an overview of the dissertation. Specifically, it presents and discusses the basic concepts that shape the study, providing the main background and context of the work. Additionally, this chapter acts as motivation for the research, emphasizing the importance of the chosen research paths and highlighting the value of the conducted analysis.

Initially, we define the research context by introducing the fundamental ideas that shape the discipline of forensic science (see Section 1.1). The chapter emphasizes how the evaluation of scientific objectives under propositions of interest forms the basis for a justified experts opinion approach. The inherent uncertainty in the evaluation of evidence requires producing probabilistic results grounded in scientific principles. Hence, the integration of statistical reasoning is both logical and essential in forensic practice. Statistical frameworks have significantly enhanced rational inference in forensic science by providing robust methods and minimizing subjective bias. A more detailed exploration of these conceptual foundations is provided in Chapter 2.

Secondly, we determine the main objectives of our study, highlighting the research questions that the study seeks to address (see Section 1.2). Specifically, one key focus of the study is the effective management of complex datasets, which involve large volumes of heterogeneous and interrelated information. Additionally, the study explores methodologies for integrating different categories of information in order to enable a reliable joint assessment. Final attention is given to the interpretation of the Bayes factor, aiming to enhance its use and understanding in forensic inference contexts.

Furthermore, special emphasis is placed on the originality and innovative aspects of the methodologies implemented in this research (see Section 1.3). The general contributions to existing research in the field are presented analytically. In particular, references are made to previous studies that have highlighted and critically examined current research gaps, guiding the discipline toward crucial directions.

Finally, Section 1.4 presents the overall structure of the thesis, offering a detailed outline of the subsequent chapters. This serves as a guide to the reader through the chapters' main summaries of ideas and findings.

1.1 Research Context

Forensic science is a body of scientific principles and technical methods applied within well-defined legal proceedings. Its main purpose is to demonstrate the existence or past occurrence of an event of legal interest (e.g., a crime) and to assist actors of the justice system (e.g., an investigator or a public prosecutor) in determining the role of target individuals (e.g., the perpetrator(s) of a given crime) as well as the modus operandi. Forensic science is concerned with aspects such as the investigation of crime scenes and the examination of victims and suspects, either directly or through accessory objects (e.g., clothing, tools, electronic devices). This typically involves comparative analysis of so-called 'evidential material' (also called 'findings' or 'outcomes'), followed by their evaluation within the context of the criminal event under investigation. Forensic science evaluates those findings through two stages: (i) analysis, which focuses on obtaining information from relevant items, and (ii) assessment, which draws logical deductions based on the information obtained from the analysis (Morrison, 2022).

Forensic inference and decision-making are core aspects of forensic science. They involve the systematic analysis of findings to draw coherent conclusions that will support legal proceedings (Aitken et al., 2021). Forensic inference involves reasoning about propositions of interest (e.g., the source of a recovered trace or the involvement of an individual in a criminal activity) based on incomplete information (mainly derived from scientific materials and eyewitness testimonies). Forensic decision making refers to the process of choosing among alternative courses of action (e.g., regarding test protocols) that depend on the inference process. Furthermore, forensic inference and decision-making require logical assistance because unaided human reasoning can lead to bias and error. This poses a serious risk for the justice system, as fallacious reasoning and erroneous conclusions can result in miscarriages of justice, as highlighted by UK (Law Commission, 2011) and US Institutions (PCAST, 2016). Moreover, the history of forensic science shows that cases of miscarriage of justice are not rare or isolated events, but rather a recurrent and significant problem, especially regarding the reliability and validity of some forensic methods and evidence (Vuille et al., 2017).

One of the main challenges in forensic science is dealing with uncertainty, which affects both inference and decision-making. Uncertainty can be referred as the degree of doubt or variability that exists in measurements. The available measurements are often incomplete, noisy, or subject to error. Therefore, statistics and data analysis aim to quantify and communicate the uncertainty that affects the evidence and the results (Taroni et al., 2010). Statistics, through the application of probabilities, addresses uncertainty by quantifying the likelihood of an event or outcome occurring. Consequently, probability can be utilized to model the randomness and variability of measurements, given specific assumptions or propositions of interest.

Propositions are statements or hypotheses that express a possible explanation or interpretation of a forensic event or findings[1]. In legal cases, propositions are usually formulated based on the prosecution side (H_p) and the defense side (H_d). These propositions must be evaluated by the judiciary using scientific methods, taking into account the case data, background data, and any other background information related to the scenario under investigation. Propositions can be defined at different levels of abstraction, such as the source level, the activity level, or the offense level, depending on the parties' questions of interest and the available information (Cook et al., 1998).

In order to determine which proposition to support, the Bayes factor (BF for short) is assessed. It is interpreted as a measure of the strength of the findings (e.g., measurements) in favor of the hypothesis H_p against the hypothesis H_d and mathematically is expressed based on the odds form of

[1] In this thesis the terms 'findings', 'outcomes', and 'evidence' are used as interchangeably terms.

Bayes' theorem:

$$\underbrace{\frac{P(H_p \mid e, I)}{P(H_d \mid e, I)}}_{\text{Posterior Odds}} = \underbrace{\frac{P(e \mid H_p, I)}{P(e \mid H_d, I)}}_{\text{Bayes Factor}} \times \underbrace{\frac{P(H_p \mid I)}{P(H_d \mid I)}}_{\text{Prior Odds}}, \tag{1.1}$$

where e is the realization of the findings and I is the relevant background information common to both propositions, for more details see Aitken et al. (2021, Section 2).

1.2 Objectives

In many fields of forensic science, data are obtained through forensic investigations and analysis and expressed as quantitative measurements. In this study, we characterize these measurements within a probabilistic modeling framework, due to inherent uncertainty in forensic findings; for further details, see Chapter 3. Let x represent continuous measurements based on the realization of findings, $f(\cdot)$ denotes the probability distribution, and θ the unknown parameter of $f(\cdot)$.

In many forensic science applications, the Bayes factor is operationally regarded as a likelihood ratio (LR) because the unknown parameters of the probabilistic model are treated as fixed but unknown constants estimated from the data. In contrast, within the Bayesian framework, parameters are treated as random variables with specified prior distributions. Thus, although the Bayes factor and the likelihood ratio share the same mathematical form, the key distinction lies in how the model parameters are handled. As a result, LRs provide a purely data-driven support of competing propositions, while BFs provide a fully Bayesian interpretation of how the evidence updates prior beliefs about the competing propositions. The assessment of these measurements when θ is treated as a fixed parameter is given by:

$$LR = \frac{f(x \mid H_p, I)}{f(x \mid H_d, I)} = \frac{f(x \mid \widehat{\theta}_{H_p})}{f(x \mid \widehat{\theta}_{H_d})}, \tag{1.2}$$

where $\widehat{\theta}$ represents the parameter estimates obtained under each proposition.

The assessment of these measurements using the Bayes factor (BF) and Bayesian framework is a more challenging task, as presented in Eq. 1.3.

$$BF = \frac{f(x \mid H_p, I)}{f(x \mid H_d, I)} = \frac{\int f(x \mid \theta_p)\pi(\theta_p \mid H_p)d\theta_p}{\int f(x \mid \theta_d)\pi(\theta_d \mid H_d)d\theta_d}, \tag{1.3}$$

$\pi(\cdot)$ the prior distribution of parameters θ_p and θ_d (Bozza et al., 2022).

The task becomes even more difficult for multivariate data, as the computations are very complex. The Bayes factor needs to specify a likelihood function for the data and a prior distribution for the parameters under each proposition. Then, one needs to integrate over the parameters to obtain the marginal likelihood of the data under each proposition. The Bayes factor is the ratio of these marginal likelihoods. However, the integration step is often intractable or computationally intensive, especially for high-dimensional data. Therefore, one may need to use approximation methods, such as Monte Carlo techniques, to assign the Bayes factor (Bozza et al., 2022). Moreover, the interpretation and communication of the Bayes factor can be challenging, as different scales and conventions exist for assessing the strength of the evidence. Therefore, one needs to be careful and transparent when reporting and explaining the Bayes factor to the relevant stakeholders (Biedermann et al., 2017).

The primary focus of this research is on the statistical analysis of multivariate data with a complex dependence structure. The purpose is to provide further theoretical grounds for operational methods, in particular for the analysis of new, original, and relevant features. Afterwards, the research concentrates on formulating a probabilistic approach to feature evaluation following forensic international guidelines ENFSI (Willis et al., 2015). ENFSI is asking for the computation of a Bayes factor to assess the value of any type of findings. The probabilistic approach has the potential to become part of the resources that forensic experts need to substantiate evaluations and conclusions.

This research will focus on addressing the challenges of assessing the Bayes factor (or LR) in forensic disciplines of handwriting and dynamic digital signature examinations, to facilitate:

- the effective management of complex datasets,
- the integration of diverse categories of information for the assessment of a joint probative value,
- the determination of the most suitable methodological approach,
- the comprehensive interpretation of the Bayes factor.

The secondary aim is to make the evaluation of data more automated to assist forensic scientists in their work. The project will use visuals to represent and operate the elements of probability modeling in a non-technical manner. The objective of the project is to develop an integrated modeling environment for the inference of handwriting and dynamic signature data, which will be used in practice. This is a novel, original, and multidisciplinary (forensic science, statistics, and computer science) approach that enables the discovery of valid and useful patterns in modeling and inference with impact on various areas of expertise.

The selection of forensic handwriting identification and dynamic signature verification examinations is driven by (a) their prevalent occurrence in everyday contexts, (b) the ongoing necessity for comprehensive research on data generation and formal quantitative evidence evaluation procedures (Gaborini et al., 2017; Linden et al., 2018), and (c) the data structure's resemblance to various scientific domains, such as data on toxic and doping substances, where complex dependence structures involving intra- and inter-variability are frequently observed. These structures necessitate modeling that accounts for non-constant variability and the elicitation of the chosen model's parameters.

The data from both disciplines will be utilized in a manner that facilitates their application in a more generalized form. Specifically, the processes of data preprocessing, feature extraction, statistical modeling, and the assessment of the Bayes factor can be applied to a wide range of legal documents, including legal declarations, contracts, acknowledgments of debt, wills, and testaments.

1.3 Originality of the Research

Forensic examination of handwritten documents and signatures is a critical aspect of forensic science, often used in legal investigations to verification of documents and identify potential forgeries. Consequently, there is a substantial body of research in this field. Someone can start by distinguishing between the two main branches of literature. One focuses on legal issues, that is, whether handwriting and signature examination can, as a forensic discipline, be sufficiently trusted to be admissible at trial (Galbraith et al., 1995; Faigman, 2007), which is out of the scope of this research. The second deals with the scientific status of evidence examination in both theory and legal practice (Moenssens, 1999; Risinger et al., 2002; Sita et al., 2002).

Although there is a general consensus regarding the handwriting experts, some uncertainties and criticism persist within the field (PCAST, 2016; Martire et al., 2018). The National Research Council (N.R.C., 2009) recommended that the scientific basis for handwriting comparisons needs to be strengthened, and Jones (2014) highlighted that no robust methodology has been proposed to aid examiners in addressing the significant subjectivity inherent in the field. Furthermore, Gaborini et al. (2017) contends that handwriting analysis offers limited empirical support for experts' conclusions. Thus, in 2020, the National Institute of Standards and Technology (N.I.S.T., 2020) emphasized, (a) the future of the discipline will incorporate the use of more tools of quantitative analysis to handle the examination process, (b) forensic document examiners should use appropriate models to explain the significance of results, (c) researchers should publish more studies involving the use of quantitative methods for examinations.

Izenman (2020) mentioned that there are still courts that limit a forensic expert at trial to only an explanation of similarities and dissimilarities in samples. The comparative examination of handwriting and signatures are illustrative examples of the need for the continuous evaluation of data for courts across jurisdictions (Evett, 1998; Saks et al., 2003; N.R.C., 2009; PCAST, 2016), especially in cases where a limited amount of items and samples are available.

Furthermore, the integration of technology has driven the discipline towards a greater emphasis on digital applications. For instance, many legal and industry sectors utilize digitized signatures to verify individual identities, which are extensively employed across various sectors, including finance, legal, and healthcare (Linden et al., 2018). Digital (dynamic) signatures contain detailed data that significantly enhances their security and makes them more challenging to forge. As will be discussed in Section 5.1, the accuracy of discrimination and validation in these digital signatures is remarkably high.

In the subsequent chapters, we review the literature on handwriting evidence (see Chapter 4) and digitally captured signatures (see Chapter 5). The primary focus of this research is the statistical analysis and probabilistic modeling of the data to assess the Bayes factor (or LR) for evaluating purposes. Consequently, the literature review will concentrate on this aspect. We examine studies that are pertinent to feature extraction and data modeling.

This research addresses the challenges of assessing the Bayes factor (or LR) in forensic disciplines. By integrating forensic science, statistics, and computer science, the project aims to develop a multidisciplinary framework that enhances the accuracy and reliability of forensic evaluations of handwriting evidence and dynamic digital signatures. The proposed approach incorporates probabilistic modeling and visual diagnostic tools, with the aim of supporting forensic scientists in the systematic assessment of forensic evidence. The framework is designed to handle complex, high-dimensional datasets and to integrate multiple feature types for a joint evaluation of probative value. The resulting methods are intended to provide theoretically grounded and operationally applicable tools that are consistent with established international guidelines for forensic evidence reporting.

Finally, the research gaps that this study aims to address in **handwriting evidence** are:

- Bayesian modeling of the variability of each character for the evaluation of loop characters.
- Elicitation of prior parameterizations of handwriting variability, leading to robust and reliable Bayes factors.
- Implementing of an end-to-end procedure suitable for practical forensic handwriting examinations.

Regarding **Digitally Captured Signatures**, the main contributions are as follows:

- Implementation of stochastic process approaches that leverage the dynamic nature of the data to assess the likelihood ratio.
- Optimization of the modeling for enhanced performance.
- Construction of a data-driven approach for forensic inference.

1.4 Thesis Outline

This section provides an outline of the dissertation chapters. The work analyzes handwriting data and dynamic digitally captured signatures. The dissertation also includes introductory, fundamental concepts, and concluding chapters. Furthermore, in the appendix it addresses the estimation of the degrees of freedom of the Wishart distribution, which plays a key role in the assessed Bayes factor for handwriting evaluation. The structure of the dissertation is as follows:

- **Chapter 1, Introduction,** presents the research problem and questions, explains why this work matters, and gives an overview of how the dissertation is organized.
- **Chapter 2, Forensic Inference and Statistics,** describes how statistics is used in forensic science and introduces the main frameworks used to report and communicate forensic evidence.
- **Chapter 3, Probabilistic Data Modeling,** covers the key statistical and probabilistic concepts that form the basis of the methods used throughout the dissertation.
- **Chapter 4, Handwriting Examination,** is the first applied chapter. It describes the handwriting data, extracts features using Fourier analysis, builds a Bayesian model, and evaluates the evidence using Bayes factors, including a practical case study.
- **Chapter 5, Dynamic Digitally Captured Signatures,** focuses on dynamic signature data. It extracts temporal features, applies stochastic process modeling, and assesses the evidence using a likelihood ratio.
- **Chapter 7, Conclusion,** sum up the dissertation by answering the research questions, summarizing the main findings, and pointing to directions for future work.
- **The Appendix** is divided into three parts - first two *A: Handwriting Examination*, *B: Dynamic Signature Examination*, containing additional results and technical details that support the main chapters. And third one *C: Wishart's Degrees of Freedom Estimation* contains an experimental analysis of estimation of the Wishart's degrees of freedom.

Chapter 2

Forensic Inference and Statistics

Forensic science is a multidisciplinary field that focuses on the application of scientific principles and methodologies to the examination and interpretation of evidence related to legal investigations. While numerous definitions exist, the following are widely accepted within the forensic community: (1) Forensic science is the art of working with imperfect specimens (physical traces), marks, or materials exchanged during the commission of an activity under investigation (Houck and Siegel, 2009). (2) Forensic science is the application of a scientific process and technical methods in the study of traces that are rooted in the criminal activity of litigious civil or administrative matters (Champod and Evett, 2000). (3) Forensic science is the application of scientific methods to help authorities establish disputed facts of activities (N.R.C., 2009).

The role of the forensic scientist within the justice system can be distinguished in two principal levels. At the investigative level, forensic experts provide critical information that supports law enforcement agencies in guiding investigations. At the evaluative level, they assess and quantify the uncertainty inherent in evidence examination, contributing to the interpretation of findings relevant to judicial decision-making (Aitken et al., 2021). Forensic practitioners can perform a variety of functions, including case investigators, forensic examiners, reporting officers, intelligence analysts, and expert evaluators. In any of these roles, the forensic scientist must remain unbiased and impartial. It is essential to employ accredited techniques and adhere to standard operating procedures. The scientist must ensure that results are reproducible and accurately interpretable. By fulfilling these responsibilities, forensic scientists contribute effectively to the justice system (Willis et al., 2015).

In summary, forensic science is the scientific discipline concerned with the collection, analysis, and interpretation of physical evidence pertinent to criminal or civil matters (Saferstein, 2013). Notably, the challenges of forensic science are closely related to those of statistics, as both fields emphasize the objective analysis of uncertain measurements (Taroni et al., 2006). In recent years, the integration of statistical methodologies has become increasingly important in forensic science, enhancing both the objectivity and transparency of evidence assessment. Statistical principles offer a rigorous framework for quantifying uncertainty, evaluating the strength of evidence, and informing decision-making within the legal system (Banks et al., 2020).

Hence, in this chapter, we present the concept of forensic interpretation (see Section 2.1), evaluation of evidence, and the application of Bayesian thinking in this context (see Section 2.2). Finally, we discuss the recommended practices for reporting and communicating scientific findings (see Section 2.3).

2.1 Introduction to Forensic Interpretation

Uncertainty is an inherent issue in forensic science, as every piece of evidence, testimony, or information carries some degree of uncertainty that must be addressed during forensic analysis. In forensic science, uncertainty typically pertains to past events where direct observation is no longer possible and conclusions must be drawn from available traces. Importantly, uncertainty is subjective and varies from individual to individual. Lindley (2013) stated that uncertainty describes the relationship between our personal current knowledge of the past events and what actually occurred. Probability offers a systematic means of quantifying uncertainty, providing a measure of the strength of our beliefs or confidence. Therefore, forensic inference relies on probability theory to account for uncertainty, enabling the evaluation and comparison of evidence to be more logical and robust. In this context, forensic inference can be viewed as the process of assessing the evidence supporting given propositions in the presence of uncertainty (Evett, 1998).

Building on this foundation, robustness and transparency are essential to ensure logical and unbiased forensic interpretations. Evett (1998) emphasized three principles of forensic interpretation:

1. Scientific evidence must be interpreted within a framework of circumstances (what we know, what we assume, etc.)

2. Scientific evidence can only be interpreted by considering at least two propositions (prosecutor and defence propositions)

3. The scientist assigns a probability to the evidence given propositions and relevant information.

These principles are commonly formalized through the use of the likelihood ratio, which compares the probability of the evidence under the competing propositions:

$$\frac{P(e|H_p, I)}{P(e|H_d, I)}$$

where e is the observed evidence, H_p and H_d denote the prosecution and defense propositions respectively, and I represents the relevant background information common to both propositions. The probabilities of propositions given the evidence and relevant background information are the responsibility of the court. Adhering to these principles ensures that forensic evidence is interpreted accurately and fairly. It helps avoid bias, increases transparency, provides a clear framework for analysis, and ensures that the evidence is evaluated in a scientifically rigorous manner.

A brief consideration of why Evett (1998) developed these principles shows their fundamental importance to forensic science. Historically, forensic interpretation had previously been characterized by subjective opinion and inconsistent inference, creating issues of fairness in the evaluation of evidence. For example, by mandating explicit comparisons under at least two competing propositions, Evett (1998) aimed to reduce bias and ensure that forensic scientists remained neutral evaluators.

Furthermore, it is important to recognize that propositions in forensic science can be structured hierarchically, with each level requiring different forms and amounts of background information. Typically, this hierarchy can be categorized into (i) source level, (ii) activity level, and (iii) offense level propositions (Cook et al., 1998). As one moves up the hierarchy from source to offense level, the value of the forensic scientist's opinion increases, while the interpretation at the lower levels leaves more space to the judgment of the court or juries. For a comprehensive discussion on the hierarchy of propositions in forensic interpretation, see Aitken et al. (2021, Section 5). This thesis focuses specif-

ically on source level propositions, which address questions related to the origin of handwriting data or dynamic signatures origins.

In many source-level forensic cases, the judicial objective is to achieve identification, specifically, to determine that the suspect, or person of interest, is the only possible source of the questioned (recovered) material. However, as discussed in Saks and Koehler (2008), absolute individualization is rarely attainable in practice, and such certainty is not easy achievable in real-world forensic applications. In general, identification is ultimately a decision and a matter of judicial opinion, not a scientific conclusion. Hence, forensic scientists therefore cannot aim for identification as an end goal. What forensic science can offer is a principled evaluation of the evidence. This is achieved by selecting features that effectively discriminate between sources, reducing the overlap between individuals within the relevant population and thereby strengthening the evidential value of the comparison (Stoney, 1991). Building on this principle, the primary objective of this thesis is to identify novel, informative features and integrate them within a unified probabilistic framework.

In summary, this section has explored the essential of forensic interpretation, emphasizing its role in providing context and meaning to scientific evidence. Robust forensic interpretation is essential for coherent forensic reasoning and for the reliability within the justice system. In the following section, we present the statistical foundations and the evaluation of evidence which enable practitioners to assess how strongly the evidence supports competing propositions, thereby enhancing transparency and objectivity.

2.2 Statistics and Evaluation of Evidence

Statistics shapes the foundations for the evaluation of forensic evidence. This section introduces the key statistical concepts, frameworks, and practical considerations that enable an objective assessment of evidential strength in forensic contexts.

Evaluation of evidence can be considered as the assessment of a comparison. Specifically, it involves assessing how closely the reference evidential materials correspond with materials recovered from a suspect or person of interest. Essential to this process is the formulation of competing propositions: one supporting the prosecution (H_p) and the other supporting the defense (H_d). As described in Section 2.1, at least two alternative propositions must always be articulated for a robust forensic analysis.

Let e represent the observed findings and I denote the relevant background information common to both propositions. The relationship between evidence and propositions can be captured by the odds form of Bayes' theorem as described in Eq. (1.1) and restated below:

$$\underbrace{\frac{P(H_p \mid e, I)}{P(H_d \mid e, I)}}_{\text{Posterior Odds}} = \underbrace{\frac{P(e \mid H_p, I)}{P(e \mid H_d, I)}}_{\text{Bayes Factor}} \times \underbrace{\frac{P(H_p \mid I)}{P(H_d \mid I)}}_{\text{Prior Odds}}$$

In forensic practice, the scientist's responsibility is to evaluate and report the value of the evidence, which is expressed by the Bayes factor (or likelihood ratio), namely the ratio of probabilities of observing the evidence under each competing proposition, given the relevant background information. The assessment of the posterior odds (the probability of a proposition after considering the evidence) and the assignment of prior odds (the degree of belief in a proposition before considering the evidence) both fall outside the goals of the forensic scientist. These determinations are the responsibility of the court,

which integrates the value of evidence with other case-specific information to reach a final judgment (Aitken et al., 2021, Section 2).

The Bayesian framework enables a systematic approach to updating our beliefs about the propositions as new evidence becomes available. By focusing on the BF, forensic scientists ensure that their analysis is transparent and easily updated with new information. This methodological clarity enhances both the objectivity and transparency of forensic reporting. Furthermore, the incorporation of prior information makes the Bayesian framework particularly well-suited for scenarios involving small sample sizes or complex data structures, which are common in forensic science.

In this thesis, we focus on evidence that can be represented through measurable quantities. Thus, probabilistic modeling techniques are implemented to address the evaluation of evidence (see Chapter 3). Specifically, probabilistic modeling enables the practical computation and interpretation of the Bayes factor in Bayesian modeling (Section 3.2) or the likelihood ratio in the frequentist modeling (Section 3.1). This probabilistic approach allows for the explicit quantification of uncertainty and inherent variation, enabling the systematic identification of patterns within the measurements of the evidence. This approach results in more effective discrimination between questioned and control materials, as well as the estimation of relevant error rates, such as, false positive[1] and false negative rates[2] (Bishop and Nasrabadi, 2006).

In general, all probabilistic models are based on assumptions and deal with certain limitations. As famously stated by Box and Draper (1987), "all models are wrong, but some are useful". Thus, it is essential that probabilistic models implemented in forensic analysis align with the characteristics and structure of the quantification measurements of the data under consideration. It should be awareness of potential violations of the model assumptions, as these can significantly affect the validity of the findings. Consequently, extensive validation and calibration studies must be conducted on the specific methodologies, from quantification of the data to modeling assumptions, to ensure their reliability and robustness in practical forensic applications. However, no model can achieve perfect accuracy, and understanding the limits of the model's performance is crucial for forensic interpretation.

A common approach to assessing model performance involves the systematic calculation, reporting, and interpretation of error rates, such as false positive and false negative rates. These metrics provide an empirical basis for assessing model reliability and are fundamental for estimating the discriminative power of forensic methods, namely the ability to distinguish between propositions that assume sources of the materials. Performance metrics for discriminability should be provided to contextualize a model's practical effectiveness (Bozza et al., 2022).

In summary, statistical evaluation of evidence provides a rigorous framework for quantifying the probative value of forensic findings. This framework should be supported by validated measurements, statistical models, and transparent reporting with the recognition of inherent limitations. In the following section 2.3, we discuss that transparent communication of the evaluation of evidence is critical, especially in a legal context. The reporting should include the assumptions, acknowledgment of model limitations, discussion of potential sources of error, and a clear statement of the associated uncertainties.

[1]The false positive rate is the proportion of negative cases incorrectly classified as positive.
[2]The false negative rate is the proportion of positive cases incorrectly classified as negative.

2.3 Reporting and Communicating Scientific Findings

The forensic scientist must possess the necessary knowledge to report scientific findings in court. The reporting should be both constructive and rational, by offering an expert opinion grounded in established methodology as emphasized from ENFSI (Willis et al., 2015) and N.I.S.T. (2020). This section describes the essential components involved in presenting the assessment of scientific findings by following the principles of interpretation described in Section 2.1, which provide normative guidance for forensic practice.

Firstly, it is necessary to be clear about the framework of circumstances relevant to the interpretation. Defining clearly the contextual elements relevant to the case ensures that the assessment remains meaningful and transparent. When circumstances or background information change, it may be necessary to review, update, and revise the evaluation of the evidence. The role of the forensic scientist is not to validate the circumstances of the case, but to condition the evaluation upon them, treating them as accepted background knowledge rather than conclusions to be established. Moreover, only relevant circumstances should be considered, as they directly influence the propositions of interest and the background information that is used to inform the assessments (Aitken et al., 2021, Section 2).

Secondly, it is important to clarify the purpose of the analysis. Specifically, it must be explained (i) why the examination was conducted, (ii) the technologies and analytical methods that were applied, and (iii) the reason for selecting particular items or evidence. In our context of handwriting documents and dynamic signatures, if the data are valid for evaluation, this means describing clearly how measurements were extracted, how the measurements were modeled, and what assumptions and technical considerations were involved, by referencing relevant literature. It is required that the methods should be robust and yield consistent results under similar conditions and assumptions. Furthermore, explanations should be understandable to general readers, avoiding unnecessary technical terms. Finally, the analysis should explicitly acknowledge the limitations and ensure the conclusions are properly contextualized (Hicks et al., 2022).

Thirdly, the interpretation involves selecting the relevant propositions of interest. Typically, the propositions reflecting competing hypotheses are presented by the parties of the case. The findings should be communicated by reporting the Bayes factor (or LR) and an explanation of what these values imply in plain language. Because the readers of the report most often will not have practice in dealing with numbers, it is recommended that the Bayes factor be reported to a simple verbal convention. That convention is based on the use of the word “support”. In the literature, various verbal scales have been proposed to convey the value of forensic findings (see Aitken and Taroni (1998), Evett et al. (2000), Nordgaard et al. (2012)). Regardless of the specific scale adopted, it is essential that the forensic scientist transparently reports the chosen scale in their analysis. In this work, we follow the verbal convention reported by ENFSI (Willis et al., 2015); see Table 2.1.

By adhering to these practices, forensic scientists efficiently support the court’s decision-making process. They contribute with scientifically sound and understandable expert opinions with the highest standards of transparency and integrity in forensic reporting. Finally, the value of statistical analysis in forensic science is fully understandable only when findings are clearly communicated to non-technical audiences, including legal professionals and juries. In Section 2.3.1, we discuss the common fallacies that have arisen in the reporting of scientific findings.

Bayes Factor (BF)	Log_e **BF**	**Verbal Communication**
1 to 2	0 to 0.7	No support
2 to 10	0.7 to 2.3	Weak support for the first proposition relative to the alternative
10 to 100	2.3 to 4.61	Moderate support for the first proposition relative to the alternative
100 to 1 000	4.61 to 6.91	Moderately strong support for the first proposition relative to the alternative
1 000 to 10 000	6.91 to 9.2	Strong support for the first proposition relative to the alternative
10 000 to 1 000 000	9.2 to 13.82	Very strong support for the first proposition relative to the alternative
1 000 000 >	13.82>	Extremely strong support for the first proposition relative to the alternative

Table 2.1: Qualitative scale for reporting the value of evidence in support of H_p against H_d (ENFSI Willis et al. (2015))

2.3.1 Pitfalls of Intuition

It is essential that uncertainty is represented accurately in forensic science in order to prevent both jurors and experts from committing the pitfalls of intuitions. Numerous fallacies have been observed in expert testimony and court statements. The most common fallacy is the fallacy of 'transported conditional', also called the prosecutor's fallacy or the inverse fallacy (Leung, 2002). Crucially, Bayes factors reflect the probability of the evidence given a proposition, not the probability of the proposition given the evidence.

Another commonly observed fallacy involves the probability of another match. Suppose a forensic scientist states, "The probability that a random person would match this handwriting style is 1 in a million". An incorrect interpretation would be "So the chance that someone else matches is 1 in a million". This fallacy assumes that because the match is rare, it would occur by chance only 1 in a million among all other individuals. However, for a population of size N and the random match probability is p, the probability that no one else matches is $(1-p)^N$. As N becomes large and p small (with $N \cdot p$ constant), this converges to $e^{-N \cdot p}$. For instance, if $N \cdot p = 1$, the probability of no other match is $e^{-1} \approx 0.368$, so the probability of at least one other match is about 0.632. Thus, the match probability and the probability of another match are fundamentally different (Koehler et al., 1994). This distinction is critical in forensic inference.

The defender's fallacy is another frequent source of misinterpretation (Thompson and Schumann, 2017). It arises when the non-uniqueness of a characteristic is used to argue in favour of the defendant. For example, because a characteristic is shared by other individuals in the relevant population, the probability that the defendant is the source of the evidence must be low. This reasoning is flawed because it conflates the frequency of a characteristic in the population with the posterior probability of guilt, which depends on all available evidence and prior information, not on the match probability alone. The fallacy underestimates the value of evidence by focusing solely on population size and random match frequency without incorporating the forensic context or using proper conditional probability.

For a discussion of additional interpretative fallacies and how statistical reasoning assists in reaching correct conclusions, see Aitken et al. (2021, Section 2.5). In summary, as we see in this section, the probability rules and the Bayesian framework guide us to detect fallacies and express our conclusions coherently.

Chapter 3

Probabilistic Data Modeling

Probabilistic data modeling is a statistical approach that uses probability theory to model uncertainty in data. In contrast with deterministic models, which result in exact outcomes, probabilistic models describe a range of possible outcomes through probability distributions by capturing the inherent noise, variability, and missing information.

The main objective of probabilistic models is to represent and characterize random variables[1]. Random variables can be discrete with countable outcomes (e.g., the result of rolling a die), or continuous with values that lie within a given interval of real numbers (e.g., a physical measurement). They are characterized by probability distributions that describe the likelihood of different outcomes in a random phenomenon. A probability function mathematically represents this distribution by assigning probabilities or probability densities to all possible outcomes (Ross, 2014).

Probabilistic models can characterize all random variables using joint distributions or focus on a specific variable using marginal distributions. This flexibility allows the incorporation of inherent noise and natural uncertainty in specific or all measurements within the model, which is crucial in forensic statistics applications. Hence, the most considered assumption in probabilistic modeling is that the assumed distributions sufficiently approximate the true generating mechanism of a phenomenon under study (Breiman, 2001). Such assumptions are usually based on probabilistic and logical arguments concerning the nature and function of a given random phenomenon.

For the sake of illustration, let a random variable Y (often called response) describe measurements of the random phenomenon under study. We assume that Y follows a distribution with probability function $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. Consider an independent, identically distributed (I.I.D.) sample $\boldsymbol{y} = [y_1, \ldots, y_n]$ of size n of this variable. The joint density of the sample is

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}).$$

Viewed as a function of $\boldsymbol{\theta} \in \Theta$ with $\boldsymbol{y}$ held fixed, this expression defines the likelihood function of the model, which contains all the information about $\boldsymbol{\theta}$ provided by the observed sample (Casella and Berger, 2024). The likelihood function encapsulates the information contained in the observed data about the parameter vector $\boldsymbol{\theta}$. The specification of $\boldsymbol{\theta}$ follows two main approaches: (a) the frequentist and (b) the Bayesian. In the frequentist approach (Section 3.1), $\boldsymbol{\theta}$ corresponds to a fixed but unknown parameter vector, whereas in the Bayesian approach (Section 3.2), $\boldsymbol{\theta}$ is treated as a random variable.

[1] Quantities whose possible values arise from random phenomena

3.1 Frequentist Data Modeling

Frequentist data modeling assumes that the parameters of the probability function describing the data are fixed but unknown. The probability function $f(\boldsymbol{y}|\boldsymbol{\theta})$ expresses the probability (or density) of observing particular data values given specific parameter values. When viewed as a function of the data with parameters held fixed, it represents the data-generating process. Conversely, when this same mathematical expression is considered as a function of the parameters for fixed observed data, it becomes the likelihood function, which quantifies how plausible different parameter values are given the data, and it is notated as $L(\boldsymbol{\theta}|\boldsymbol{y})$. Thus, the likelihood function is equal to the probability function but is viewed as a function of the parameter vector $\boldsymbol{\theta}$.

In this section, we present the frequentist methodological framework employed in our research. In particular, we define the concept of Maximum Likelihood Estimation (MLE) and explain how it is used to estimate model parameters by maximizing the likelihood of the observed data. To illustrate these concepts, we provide examples of likelihood-based estimation methods using the Multivariate Normal distribution in Section 3.1.1.

Furthermore, when likelihood functions are complex, numerical techniques can be used to approximate their maximization, such as the Newton-Raphson algorithm (Burden et al., 2015, Chapter 2.3), as discussed in Section C.2 for the Wishart distribution. In cases where latent variables are present, such as in Hidden Markov Models (HMM) used for dynamic signature analysis (see Section 5), iterative numerical methods like the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) provide an effective way to find maximum likelihood estimates (see Section 3.1.2).

A general key assumption of data modeling is that the chosen model sufficiently approximates the true data-generating mechanism. This assumption enables estimation and inference within the data modeling framework. However, it is critical to be evaluated through diagnostic tests, goodness-of-fit tests, and validation procedures. The latter evaluations ensure that the model represents the observed measurements as well as possible, given the complexity of the data.

3.1.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a classical frequentist approach for estimating parameters of probabilistic models. It estimates parameter values that maximize the likelihood of the observed data under the assumed model. Moreover, MLE can be interpreted in terms of loss functions, since maximizing the likelihood is equivalent to minimizing a specific loss function derived from the assumptions of the underlying probabilistic model.

Assuming a model parameter vector $\boldsymbol{\theta}$ and independence among an observed vector $\boldsymbol{y}$, the likelihood can be factorized as the product of individual probability functions, $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$. The log-likelihood function is preferred in practice due to its computational convenience. Specifically, it is converting the product of likelihoods into a sum that facilitates both analytical manipulation and numerical optimization. The log-likelihood function can be denoted as follows:

$$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}),\ \theta \in \Theta.$$

This optimization is typically carried out by solving the system of equations formed by setting the

partial derivatives of the log-likelihood with respect to each component of $\boldsymbol{\theta}$ to zero:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = 0, \quad \text{for } j = 1, \ldots, |\boldsymbol{\theta}|,$$

where $|\boldsymbol{\theta}|$ denotes the number of model parameters (i.e., the cardinality of the parameter vector $\boldsymbol{\theta}$), and $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ is assumed to be continuous. A solution to this system yields the MLE:

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\theta} \ell(\boldsymbol{\theta}).$$

This approach transforms the maximization problem into solving a set of nonlinear equations, which may require analytical manipulation or numerical optimization methods, depending on the complexity of the model and the form of the likelihood function (Wright et al., 1999).

Under regularity conditions such as the differentiability of the likelihood function, the interchangeability of differentiation, and the identifiability of the model, the maximum likelihood estimators possess important properties of consistency, asymptotic normality, and efficiency (Pfanzagl and Hamböker, 1994).

Example: Multivariate Normal Distribution

Consider multivariare measurements $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ in $\mathbb{R}^p$ modeled as independent samples from a Multivariate Normal distribution $N_p(\boldsymbol{\theta}, \boldsymbol{W})$ with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{W}$.

The likelihood function is

$$L(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{W}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2} det(\boldsymbol{W})^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{y}_i - \theta)^T \boldsymbol{W}^{-1}(\boldsymbol{y}_i - \theta)).$$

The MLEs parameter estimators have closed-form solutions (Press, 2005, Section 4):

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \quad \text{and} \quad \widehat{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{y}_i - \widehat{\boldsymbol{\theta}})(\boldsymbol{y}_i - \widehat{\boldsymbol{\theta}})^T \tag{3.1}$$

where $\boldsymbol{y} = (\underset{\sim}{y_1}, \underset{\sim}{y_2}, \ldots, \underset{\sim}{y_n})$ is, here, a $p \times n$ matrix of the response values, with each $\underset{\sim}{y_i} \in \mathbb{R}^p$ denoting the p-dimensional observation vector. In the context of handwriting examination, discussed in Chapter 4 which follows, we consider the case of multiple individuals (writers), each providing a varying number of repetitions. Accordingly, the maximum likelihood estimators must account for the hierarchical data structure. Under this model formulation, the aim is to estimate both the between-individual and within-individual covariance components (Timm, 2002, Section 4.4), which capture distinct sources of variability crucial for forensic analysis. Specifically, we assume that $n = \sum_{i=1}^{m} n_i$ observations come from m individuals with possibly different repetitions n_i per individual. The repeated measurements for individual i are denoted as $\boldsymbol{y}_{ij}$, $j = 1, \ldots, n_i$.

The maximum likelihood estimators under the assumption that the data follow a multivariate Normal distribution are as follows. The individual sample mean vectors are

$$\widehat{\boldsymbol{\theta}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{y}_{ij}. \tag{3.2}$$

where $\widehat{\boldsymbol{\theta}}_i \in \mathbb{R}^p$ is the MLE of the true mean vector $\boldsymbol{\theta}_i$ for the i-th individual. The overall sample mean remains

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}. \tag{3.3}$$

The between-individual covariance matrix estimates the variability of individual means around the overall mean:

$$\widehat{\boldsymbol{B}} = \frac{1}{m-1} \sum_{i=1}^{m} n_i (\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}})^T. \tag{3.4}$$

The within-individual covariance matrix pools the variability within individual:

$$\widehat{\boldsymbol{W}} = \frac{1}{n-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \widehat{\boldsymbol{\theta}}_i)(\boldsymbol{y}_{ij} - \widehat{\boldsymbol{\theta}}_i)^T. \tag{3.5}$$

The reader can refer to Timm (2002) and Press (2005) for further details.

3.1.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is an iterative method of two steps (E-step, M-step) for computing MLEs in models where the likelihood function is analytically intractable. Typical example is the case of the presence of latent (unobserved) variables. To formalize this, suppose we wish to maximize the observed-data likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{y})$, where $\boldsymbol{y}$ denotes the observed data and $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ is the parameter vector. The key idea of the EM algorithm is data augmentation: the observed data $\boldsymbol{y}$ are augmented with latent variables $\boldsymbol{z}$, forming the complete data $(\boldsymbol{y}, \boldsymbol{z})$. The two are linked through the marginalisation $f(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{z}$ [2]. While the observed-data likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{y}) = f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is intractable due to this marginalisation, the complete-data likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})$ is typically much easier to work with. The log-likelihood $\ell(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}) = \log f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})$ is evaluated using both the observed data $\boldsymbol{y}$ and the latent variables $\boldsymbol{z}$, as opposed to the observed-data log-likelihood $\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = \log f(\boldsymbol{y} \mid \boldsymbol{\theta})$, which involves $\boldsymbol{z}$ only implicitly through the marginalisation above. Since $\boldsymbol{z}$ is unobserved, the complete-data log-likelihood cannot be evaluated directly. Instead, the EM algorithm works with its conditional expectation given the observed data (Karlis, 2023). The algorithm alternates between the following two steps:

- **E-step (Expectation):** Compute the expected value of the complete-data log-likelihood with respect to the conditional distribution of the latent variables, given the observed data and current parameter estimates $\boldsymbol{\theta}^{(t)}$:

$$\mathbb{E}\big[\log L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}) | \boldsymbol{y}, \boldsymbol{\theta}^{(t)}\big]. \tag{3.6}$$

- **M-step (Maximization):** Maximize this expected complete-data log-likelihood with respect to the model parameters to obtain updated estimates:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\big[\log L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}) | \boldsymbol{y}, \boldsymbol{\theta}^{(t)}\big]. \tag{3.7}$$

These two steps are repeated iteratively until convergence, which is typically defined as changes in the log-likelihood or parameter estimates falling below a specified threshold.

Hidden Markov Models (HMMs) are a classic example of the application of the EM algorithm. HMMs are probabilistic models for dependent data, where each observation is conditionally dependent on a hidden (unobserved) state. A more detailed explanation is presented in the example which follows.

[2] The integral is understood in the Riemann-Stieltjes sense, which generalises ordinary integration by allowing the "measuring" of area by an arbitrary function g rather than uniform length (Rudin, 1976).

Example: Gaussian Hidden Markov Model

A common application of the EM algorithm is in estimating the parameters of a Gaussian Hidden Markov Models (HMMs) (Baum and Petrie, 1966). HMMs model time series of observable variables $\{y_t\}_{t=1}^T$ as being generated by an underlying sequence of hidden (unobserved) states $\{z_t\}_{t=1}^T$. Each hidden state corresponds to a Gaussian distribution controling the emissions[3], meaning that, given the current state, the observation is assumed to follow Normal distribution. Formally:

- The **hidden state process** z_t is a discrete-time Markov chain with initial probabilities $\pi_i = P(z_1 = i)$ and transition probabilities $a_{ij} = P(z_{t+1} = j \mid z_t = i)$. General notation for the K number of states is $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $A = (a_{ij})_{i,j=1}^K$,
- The **emission model** specifies that $y_t \mid z_t = k \sim N(\mu_k, \sigma_k^2)$, where μ_k and σ_k^2 are the mean and standard deviation for state k.

The challenge is that the sequence z_t is not observed, so we must estimate the parameters $\Theta = \left(\boldsymbol{\pi}, A, \{\mu_k, \sigma_k^2\}_{k=1}^K\right)$ based on the hidden state sequence of the data $\{y_t\}_{t=1}^T$. This is where the EM algorithm comes to provide an effective solution.

E-step Compute the distributions over hidden states given the current parameter estimates and the observations, typically using the forward-backward algorithm (Baum and Petrie, 1966).

M-step Update the parameter estimates by maximizing the expected complete-data log-likelihood, where the expected values are taken with respect to the conditional distribution of the latent variables $\boldsymbol{z}$ given the observed data from the E-step.

In this context, the EM algorithm alternates between inferring the likely state sequence (E-step) and refining the model parameters to better fit the data (M-step), iterating until convergence. For a more detailed mathematical explanation, see Section 5.4.

3.2 Bayesian Data Modeling

In this section, we introduce the Bayesian modeling framework implemented in our research. A key assumption of our approach, as in any parametric model, is that the selected model formulation accurately approximates the data-generating mechanism. In other words, we assume that the model effectively captures the underlying factors, variability and patterns. This assumption allows us to analyze and interpret the data in a meaningful way. We rigorously evaluate these assumptions through statistical testing and model fit evaluations to verify their validity; however, they remain assumptions.

We begin this chapter by presenting the fundamentals of Bayesian modeling as an introductory section of our methodological framework (see Section 3.2.1). Specifically, we define the Bayes' theorem and extend the discussion to a model-based Bayesian inference, in order to establish the theoretical context for subsequent analyses. Furthermore, different approaches of prior specification are presented and discussed, with attention given to both conjugate priors and more complex hierarchical structures. For illustrative purposes, we provide examples of conjugate and non-conjugate modeling using the Normal-Inverse-Wishart distribution.

[3] An emission distribution is a probability distribution that describes how the observable y_t is produced ("emitted") by the hidden state z_t.

We further present how hierarchical Bayesian models can overcome challenges arising from the absence of closed-form solutions for the posterior distributions of interest, with a focus on the application of Markov chain Monte Carlo (MCMC) methods (see Section 3.2.2). We discuss the most widely used MCMC methods and their corresponding software implementations, and illustrate their implementation with an example involving the hierarchical form of the Normal-Inverse-Wishart model.

Finally, we present three widely used approaches for estimating the marginal likelihood (see Section 3.2.3), each generating samples from the posterior distribution via MCMC. The focus on marginal likelihood Monte Carlo estimators arises from their critical role in Bayesian model comparison and hypothesis evaluation.

3.2.1 Fundamentals of Bayesian Modeling

Bayesian statistics differ from the frequentist framework by treating model parameters as random variables. They incorporate prior knowledge and updates it with observed data to obtain the posterior distribution (Bernardo et al., 1994). For this reason, a prior distribution $\pi(\boldsymbol{\theta})$ must be initially defined. This prior distribution expresses the information available to the analyst before any "data" are involved in the statistical analysis. Hence, the aim is to calculate the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of the parameters $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}$. According to Bayes' theorem, the posterior distribution can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})},$$

where $m(\boldsymbol{y}) = \int_\theta f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the marginal likelihood, which is of prominent importance for our analysis because it represents the overall likelihood of the data under the model and under the proposition of interest (see Chapter 2), averaged over the parameter uncertainty (Gelman et al., 1995). It is also called the model "evidence" and it plays a central role in Bayesian model comparison (Ntzoufras, 2011, Chapter 11). Calculating marginal likelihoods can be challenging because the integral is often high-dimensional and lacks a closed-form solution, requiring numerical methods such as Monte Carlo integration or approximations that we are going to present in Section 3.2.3.

A key strength of Bayesian modeling is the capability to incorporate prior information and to express uncertainty in parameter estimates. This is especially important in complex or high-dimensional problems (Press, 2005) where data may be limited and various relationships between variables may be presented. Hence, the choice of the prior plays a vital role in Bayesian inference.

Conjugate priors are "convenient" priors that lead to analytically tractable posteriors. Formally, a family of prior distributions is said to be conjugate with respect to a given likelihood function if, for any prior chosen from that family, the resulting posterior distribution also belongs to the same family. In other words, the prior and posterior distributions share the same parametric form and differ only in their hyperparameter values, which are updated upon observing the data. However, in real-world problems, the settings are more complex, so hierarchical or non-conjugate priors are often preferred as they allow for greater flexibility and adaptive regularization, although they typically do not produce closed-form expressions for the posterior distribution.

Furthermore, in Bayesian hierarchical modeling (Ntzoufras, 2011, Chapter 9) where we considered non-conjugate priors, the prior specification is critical, as it can have a substantial impact (in a positive or negative way) on the resulting posterior distribution and also on the marginal likelihood. Two commonly used approaches to specify non-informative priors are (a) Jeffreys (1998) priors and (b) objective priors (Berger et al., 2024), for more details see Consonni et al. (2018). Alternatively,

informative priors can be constructed, for example, subjective priors elicited from expert opinions (Mikkola et al., 2024). Each of these approaches has distinct characteristics and implications for modeling.

In this study, we implement a subjective prior methodology in which prior distributions are informed by background data. This approach explicitly incorporates prior knowledge regarding the characteristics of the measurements used in the study. Such a strategy is particularly advantageous when available data are limited or, as in our case, when it is necessary to assess the degree of difference between two measurements in order to determine whether the data originate from the same individual or from different individuals (Bozza et al., 2008).

In Section 3.2.1, we present an example of the multivariate Normal-Inverse-Wishart distribution using typical Bayesian notation, and we explain and use two different prior approaches: the conjugate approach and the hierarchical approach, leading to two distinct models that capture different characteristics of our problem.

Example: Normal-Inverse-Wishart

Let us consider the available multivariate measurements denoted as $\boldsymbol{y}$. These data compose a $n \times p$ matrix. Let us further denote by $\boldsymbol{\theta} \in \mathbb{R}^p$ the mean vector, and by $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ the variance-covariance matrix with elements $W_{\kappa_1 \kappa_2}$ for $\kappa_1, \kappa_2 \in \{1, 2, \ldots, p\}$. Then, given $\boldsymbol{\theta}$ and $\boldsymbol{W}$, the distribution of $\mathbf{y}$ is taken to be p-variate Normal N_p, with

$$\boldsymbol{y} \sim N_p\left(\boldsymbol{\theta}, \boldsymbol{W}\right) \tag{3.8}$$

In this modeling approach, a Normal-Inverse-Wishart (NIW) prior distribution is chosen for the parameters $(\boldsymbol{\theta}, \boldsymbol{W})$. Under this general prior set-up, and considering the sampling model distribution for $\boldsymbol{y}$, the following Bayesian model is specified:

$$\begin{aligned} \boldsymbol{y} &\sim N_p(\boldsymbol{\theta}, \boldsymbol{W}) \\ \boldsymbol{\theta} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{G}) \\ \boldsymbol{W} &\sim IW(\boldsymbol{U}, \nu), \end{aligned} \tag{3.9}$$

where IW denotes the Inverse-Wishart, $\boldsymbol{\mu}$ is the prior mean vector of $\boldsymbol{\theta}$, $\boldsymbol{W}$ is the covariance matrix, $\boldsymbol{G}$ is the prior covariance matrix of $\boldsymbol{\theta}$, while $\boldsymbol{U}$ and ν represent the scale matrix and the degrees of freedom of the Inverse-Wishart distribution that a-priori models the variability. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{W}$ are assumed to be independent a priori.

Depending on the specification of the prior variance-covariance matrix $\boldsymbol{G}$, two distinct variants of the Bayesian model can be obtained. First, by setting $\boldsymbol{G} = \boldsymbol{W} k_0^{-1}$, a conjugate prior is obtained, and the prior mean vector is denoted as $\boldsymbol{\theta}|\boldsymbol{W}$. Conversely, when $\boldsymbol{G}$ is estimated from background data as $\widehat{\boldsymbol{G}}$, or assigned its own prior distribution, the model acquires a hierarchical structure, as uncertainty about $\boldsymbol{G}$ is propagated through an additional level of the model.

Hence, the implied conjugate prior distribution for $(\boldsymbol{\theta}, \boldsymbol{W})$ is the Normal-Inverse-Wishart (NIW) distribution,

$$(\boldsymbol{\theta}, \boldsymbol{W}) \sim NIW(\boldsymbol{\mu}, k_0, \boldsymbol{U}, \nu),$$

and the probability density function (PDF) is

$$\pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{\mu}, k_0, \boldsymbol{U}, \nu) = N(\boldsymbol{\theta}|\boldsymbol{\mu}, \tfrac{1}{k_0}\boldsymbol{W}) \times IW(\boldsymbol{W}|\boldsymbol{U}, \nu)$$

where $\times$ denotes the product of densities arising from the prior independence of $\boldsymbol{\theta}$ and $\boldsymbol{W}$, the prior parameter k_0 is a parameter which controls the volume of the prior precision (and the variance) of $\boldsymbol{\theta}$. The higher k_0, the more informative is the prior distribution for $\boldsymbol{\theta}$. Moreover, $\boldsymbol{U}$ and ν represent the scale matrix and the degrees of freedom of the Inverse-Wishart distribution, respectively; see Gelman et al. (1995, Chap. 3.6) for more details. Since the NIW prior is conjugate, the posterior distribution and the marginal likelihood of $\boldsymbol{y}$ are available in closed form (Murphy, 2007) and are given by:

$$(\boldsymbol{\theta}, W)|\boldsymbol{y} \sim NIW(\boldsymbol{\mu}_n, k_n, \boldsymbol{U}_n, \nu_n), \quad \text{and} \quad m(\boldsymbol{y}) = \frac{1}{\pi^{np/2}} \frac{\Gamma_p(\nu_n/2)}{\Gamma_p(\nu/2)} \frac{|\boldsymbol{U}|^{\nu/2}}{|\boldsymbol{U}_n|^{\nu_n/2}} \left(\frac{k_0}{k_n}\right)^{p/2} \tag{3.10}$$

where n is the sample size, p is the number of variables, Γ_p is the multivariate gamma function, and

$$k_n = k_0 + n, \ \nu_n = \nu + n, \ \boldsymbol{\mu}_n = \frac{k_0}{k_0 + n}\mu + \frac{n}{k_0 + n}\bar{\boldsymbol{y}}, \ \ \boldsymbol{U}_n = \boldsymbol{U} + \boldsymbol{S} + \frac{k_0 n}{k_0 + n}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})^T$$

$\bar{\boldsymbol{y}}$ denotes the average measurements and $\boldsymbol{S} = \sum_i (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T$. Notably, the parameter $\boldsymbol{\mu}_n$ is a weighted average of the prior mean and the sample mean. Although the conjugate prior is computationally convenient, it has the disadvantage of not modeling the prior distribution of $\boldsymbol{\theta}$ fully independent of $\boldsymbol{W}$, since the two are linked through the scaling parameter k_0.

Bozza et al. (2008) for handwriting evaluation of evidence chose a independent prior setup for the model parameters $(\boldsymbol{\theta}, \boldsymbol{W})$ of the following form

$$\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{B}) \text{ and } \boldsymbol{W} \sim IW(\boldsymbol{U}, \nu). \tag{3.11}$$

The proposed model is equal to $\widehat{\boldsymbol{G}} = \boldsymbol{B} = \frac{1}{m-1} \sum_{i=1}^{m} n_i (\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\mu}})^T$ the between writers variability can be estimated from background data, and it can be fixed and independent of $\boldsymbol{W}$. Empirical evidence can show that models incorporating this variability can perform better than models that do not have this flexibility. Unfortunately, the posterior distribution and the marginal likelihood for this model can not be obtained analytically.

For most practical models, particularly those involving continuous or multidimensional data, analytic solutions for posterior distributions are typically unavailable. This necessitates the use of computational methods to approximate the posterior distribution (Robert et al., 1999), which in turn enables the approximation of the marginal likelihood. Among these computational approaches, Markov chain Monte Carlo (MCMC) methods are the most widely used, and they are presented in Section 3.2.2.

3.2.2 Markov Chain Monte Carlo

When closed-form solutions for the posterior distribution are not available, Markov chain Monte Carlo (MCMC) methods offer a practical alternative by enabling approximate inference through stochastic simulation. MCMC techniques allow us to use highly complicated models and estimate the corresponding posterior distributions with sufficient accuracy. MCMC algorithms can generate samples from the posterior distribution based on the construction of a Markov chain that eventually "converges" to the target distribution (called stationary or equilibrium), which, in our case, is the posterior distribution $\pi(\boldsymbol{\theta}|y)$. In more detail, a Markov chain is a stochastic process such that

$$\pi(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(1)}) = \pi(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}),$$

the distribution of $\boldsymbol{\theta}^{(t+1)}$ which denotes the parameter vector $\boldsymbol{\theta}$ at time sequence $t+1$ given all the

corresponding $\boldsymbol{\theta}$ values at time points $t, t-1, \ldots, 1$ depends only on the value $\boldsymbol{\theta}^{(t)}$, namely the parameter values of previous sequence t. Hence, the Markov chain under proper assumptions (irreducible, aperiodic, positive-recurrent) and as $t \to \infty$ the distribution of $\boldsymbol{\theta}^{(t)}$ converges to its equilibrium distribution, which is independent of the initial values of the chain $\boldsymbol{\theta}^{(0)}$; for details, see Gilks, Richardson and Spiegelhalter (1995) and Ntzoufras (2011, Chapter 3). Hence, the sample obtain from the MCMC output is dependent, as each value generated by the Markov chain depends on the preceding value. Some techniques address the latter issue by thinning the MCMC sample to reduce autocorrelation among retained values. Finally, we may improve the sample quality by discarding initial values that are sampled before reaching convergence to the equilibrium distribution (burn-in period).

While MCMC methods are powerful and widely applicable, they require careful assessment of the convergence to ensure that the generated samples accurately represent the target posterior. Diagnostics can be obtained by trace plots, the scale reduction factor (Gelman and Rubin (1992) statistic), and effective sample size calculations which are essential to verifying the quality and reliability of the MCMC output, see Gelman et al. (1995) for more details.

The most traditionally used MCMC techniques are the Metropolis-Hastings algorithm (Hastings, 1970) and the Gibbs sampling (Gelfand and Smith, 1990). Metropolis-Hastings algorithm proposes a new generated value based on the current state of the parameter of interest and accepts or rejects it using a criterion that guarantees convergence to the target posterior. The Gibbs Sampler is a special case of the Metropolis-Hastings algorithm where generated values are drawn sequentially from each full conditional distribution of the parameters, given the current values of the other parameters. Both approaches enable us to obtain posterior samples and therefore "estimate" the posterior distributions.

Nowadays, several software packages implementing MCMC methods are widely available, including *OpenBUGS* (Lunn et al., 2000) and *JAGS* (Plummer et al., 2003), which primarily use Gibbs sampling and, in some cases, the Metropolis-Hastings algorithm. More advanced MCMC algorithms have also been developed, such as the No-U-Turn Sampler (NUTS; Hoffman et al., 2014), which features automated tuning of Hamiltonian Monte Carlo (Duane et al., 1987) parameters and is implemented in *Stan* (Carpenter et al., 2017).

Section 3.2.2 illustrates an example of the Gibbs sampling method applied to the posterior distribution of the hierarchical Normal-Inverse-Wishart model. For readers interested in MCMC methods, we refer to Appendix C, where we present various comparisons among MCMC techniques (i.e., Metropolis-within-Gibbs, Hamiltonian Monte Carlo, etc.) used to estimate the posterior distribution of the degrees of freedom parameter of the Wishart distribution.

Gibbs Sampling Example: Normal-Inverse-Wishart

Following the hierarchical form of the Normal-Inverse-Wishart model presented in Section 3.2.1, where $\widehat{\boldsymbol{G}} = \boldsymbol{B}$, the model can be expressed as follows:

$$\begin{aligned} \boldsymbol{y} &\sim N_p(\boldsymbol{\theta}, \boldsymbol{W}), \\ \boldsymbol{\theta} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{B}), \\ \boldsymbol{W} &\sim IW(\boldsymbol{U}, \nu). \end{aligned} \tag{3.12}$$

As mentioned, the posterior distribution of this model is not available in closed form; therefore, it is necessary to implement an MCMC method. The advantage is that the full conditional distributions of the parameters can be derived in closed form (Bozza et al., 2008), allowing for the implementation

of Gibbs sampling. These conditionals can be expressed as follows:

$$\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{W} \sim N_p\left(\left(\boldsymbol{B}^{-1}+n\boldsymbol{W}^{-1}\right)^{-1}\left(\boldsymbol{B}^{-1}\boldsymbol{\mu}+n\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}\right),\quad \left(\boldsymbol{B}^{-1}+n\boldsymbol{W}^{-1}\right)^{-1}\right), \tag{3.13}$$

$$\boldsymbol{W}|\boldsymbol{y},\boldsymbol{\theta} \sim IW\left(\boldsymbol{U}+\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T,\quad \nu+n\right), \tag{3.14}$$

where n is the sample size and $\bar{\boldsymbol{y}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i$ is the sample mean of $\boldsymbol{y}$. Hence, the Gibbs sampling algorithm can be implemented as follows:

- **Initialize** the parameters $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{W}^{(0)}$ (e.g., using prior means or random values).
- **For** iteration $t=1,2,\ldots,T$ (where T is the total number of MCMC samples):
  - Calculate the posterior covariance matrix of $\boldsymbol{\theta}$: $S=\left(\boldsymbol{B}^{-1}+n\boldsymbol{W}^{-1}\right)^{-1}$
  - Calculate the posterior mean vector of $\boldsymbol{\theta}$: $M=S\left(\boldsymbol{B}^{-1}\boldsymbol{\mu}+n\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}\right)$
  - Sample from $\boldsymbol{\theta}^{(t)}\sim N_p\left(M,S\right)$ (i.e. conditional posterior $\boldsymbol{\theta}^{(t)}|\boldsymbol{y},\boldsymbol{W}^{(t-1)}$ given in (3.13))
  - Sample from $\boldsymbol{W}^{(t)}\sim IW(\boldsymbol{U}+\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta}^{(t)})(\boldsymbol{y}_i-\boldsymbol{\theta}^{(t)})^T,\quad \nu+n)$
- **End For**

After excluding the burn-in period[4], assessing convergence diagnostics, and applying thinning to eliminate autocorrelation, a sufficiently large and accurate sample from the posterior distribution is obtained. With this posterior sample available, in Section 3.2.3 we proceed to discuss methods for estimating the marginal likelihood.

3.2.3 Marginal Likelihood

The marginal likelihood (ML), or model evidence, is the likelihood of the observed data under a specific probabilistic model. The marginal likelihood can be defined as the integral of the likelihood density function over the prior density function:

$$m(\boldsymbol{y}|M_l)=\int f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{W},M_l)\pi(\boldsymbol{\theta},\boldsymbol{W}|M_l)d(\boldsymbol{\theta},\boldsymbol{W}), \tag{3.15}$$

under model M_l, the $\boldsymbol{\theta}$ and $\boldsymbol{W}$ represent model parameters, specifically, the multivariate mean and covariance matrix (i.e., location and scale parameters) of the Bayesian model described in 3.2.2.

For illustration, consider two competing models M_1 and M_2. The ratio of their marginal likelihoods defines the Bayes factor:

$$BF=\frac{m(\boldsymbol{y}|M_1)}{m(\boldsymbol{y}|M_2)}$$

The Bayes factor measures how much more the observed data support one model over the other. A value greater than 1 favors M_1, while a value less than 1 favors M_2.

When models are based on conjugate priors, the marginal likelihood may be available in closed form. However, when an analytical expression is not available, we resort to approximation methods. In such cases, three popular estimators are employed to estimate the marginal likelihood: (a) the Laplace-Metropolis estimator, (b) the Generalized Harmonic Mean, and (c) the Bridge Sampling.

[4]The burn-in period refers to the initial iterations of the MCMC sampler that are discarded before collecting posterior samples. During this phase, the chain has not yet converged and the samples are therefore not representative of the posterior.

Laplace-Metropolis

Lewis and Raftery (1997) proposed an approach for marginal likelihood estimation based on Laplace's approximation (Kass et al., 1991). The Laplace-Metropolis estimator is given by

$$\widehat{m}(\boldsymbol{y}) = (2\pi)^{\frac{d}{2}} |\Sigma_{\boldsymbol{\Psi}}|^{\frac{1}{2}} f(\boldsymbol{y}|\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{W}}) \pi(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{W}}), \tag{3.16}$$

where $\overline{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{W}}$ is the average of the model parameters from the posterior sample, d is the number of parameters.

The Laplace-Metropolis approximation is evaluated at a single high-posterior-density point, typically the posterior mode (MAP) or, in practice, a numerically stable estimate of the mode, together with a covariance estimate (e.g., the inverse negative Hessian or the posterior covariance from the MCMC output). Using posterior averages $\overline{\boldsymbol{\Theta}}$ and $\overline{\boldsymbol{W}}$ can work properly when the posterior is approximately unimodal and close to Gaussian so that the mean and the mode are nearly identical, and when the chosen parameterization makes the mean well-defined (e.g., for covariance/precision matrices, averaging should preserve positive definiteness). If the posterior is skewed, heavy-tailed, multimodal, or constrained, the posterior mean may be far from the mode and the approximation can fail. In such cases, centering the approximation at the MAP (or another mode-finding estimate) is preferred.

Generalized Harmonic Mean

The generalized harmonic mean is a Monte Carlo method, which uses samples from the posterior distribution (Gelfand and Dey, 1994). By sampling from a proper importance density, high likelihood values can be obtained more frequently, while low likelihood values are sampled less often. Thus, the marginal likelihood can be expressed as an expected value with respect to the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y})$:

$$\begin{aligned} \frac{1}{m(\boldsymbol{y})} &= \mathbb{E}_{\pi(\boldsymbol{\theta},\boldsymbol{W}|\boldsymbol{y})} \left[\frac{g(\boldsymbol{\theta}, \boldsymbol{W})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{W})\, \pi(\boldsymbol{\theta}, \boldsymbol{W})} \right] \\ &= \int \frac{g(\boldsymbol{\theta}, \boldsymbol{W})}{f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{W})\, \pi(\boldsymbol{\theta}, \boldsymbol{W})}\, \pi(\boldsymbol{\theta}, \boldsymbol{W} \mid \boldsymbol{y})\, d\boldsymbol{\theta}\, d\boldsymbol{W}. \end{aligned} \tag{3.17}$$

where $g(\boldsymbol{\theta}, \boldsymbol{W})$ is the so-called importance density that must be close to the posterior density. Hence, given a sample of the parameters $\boldsymbol{\theta}^{(t)}, \boldsymbol{W}^{(t)}$ (for $t = 1, \ldots, T$) from the posterior distribution, then the generalized harmonic mean estimator is given by

$$\widehat{m}(\boldsymbol{y}) = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{g(\boldsymbol{\theta}^{(t)}, \boldsymbol{W}^{(t)})}{f(\boldsymbol{y}|\boldsymbol{\theta}^{(t)}, \boldsymbol{W}^{(t)}) \pi(\boldsymbol{\theta}^{(t)}, \boldsymbol{W}^{(t)})} \right)^{-1}, \tag{3.18}$$

where $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{W}^{(t)}$ are sampled from $\pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y})$. Following Perrakis et al. (2014), the product of the marginal distributions is used as importance density $g(\boldsymbol{\theta}, \boldsymbol{W}) \equiv g(\boldsymbol{\theta}|\boldsymbol{y}) g(\boldsymbol{W}|\boldsymbol{y})$.

The generalized harmonic mean estimator is popular because it can estimate the marginal likelihood $m(\boldsymbol{y})$ directly from posterior draws (e.g., from MCMC), hence it is easy to implement and can be more stable than the standard harmonic-mean estimator (Newton and Raftery, 1994) when the tuning density $g(\cdot)$ is chosen well. Its main drawback is fragility: the estimate can have extremely high variance and be dominated by a few draws if the implied importance weight has heavy tails under the posterior, so results can be highly sensitive to the choice of $g(\cdot)$.

Bridge Sampling

The most known and effective Monte Carlo estimator of the marginal likelihood was introduced by Meng and Wong (1996). This method is based on the following identity:

$$m(\boldsymbol{y}) = \frac{\int h(\boldsymbol{\theta}, \boldsymbol{W}) g(\boldsymbol{\theta}, \boldsymbol{W}) f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{W}) \pi(\boldsymbol{\theta}, \boldsymbol{W}) d(\boldsymbol{\theta}, \boldsymbol{W})}{\int h(\boldsymbol{\theta}, \boldsymbol{W}) g(\boldsymbol{\theta}, \boldsymbol{W}) \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{W}) \pi(\boldsymbol{\theta}, \boldsymbol{W})}{m(\boldsymbol{y})} d(\boldsymbol{\theta}, \boldsymbol{W})}, \tag{3.19}$$

where the functions $g(\boldsymbol{\theta}, \boldsymbol{W})$ and $h(\boldsymbol{\theta}, \boldsymbol{W})$ are the so-called proposal distribution and bridge function, respectively. The proposal distribution, which serves as an importance density in Section 3.2.3, should approximate the posterior distribution and ensure adequate overlap with it. Furthermore, the function $h(\cdot)$ acts as a bridge that connects the two distributions. Therefore, it must be compatible with both distributions. Simple manipulation of (3.19) leads to the following identity:

$$m(\boldsymbol{y}) \quad = \quad \frac{E_{g(\boldsymbol{\theta}, \boldsymbol{W})} \left[h(\boldsymbol{\theta}, \boldsymbol{W}) f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{W}) \pi(\boldsymbol{\theta}, \boldsymbol{W}) \right]}{E_{\pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y})} \left[h(\boldsymbol{\theta}, \boldsymbol{W}) g(\boldsymbol{\theta}, \boldsymbol{W}) \right]}. \tag{3.20}$$

Hence, the marginal likelihood can be estimated using a Monte Carlo estimator based on the identity in (3.20). This Monte Carlo marginal likelihood estimator is given by

$$\widehat{m}(\boldsymbol{y}) = \frac{\frac{1}{T_2} \sum_{t=1}^{T_2} h(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)}) f(\boldsymbol{y}|\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)}) \pi(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)})}{\frac{1}{T_1} \sum_{t=1}^{T_1} h(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)}) g(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)})}, \tag{3.21}$$

where $(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)})$, for $t = 1, \ldots, T_1$, and $(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)})$, for $t = 1, \ldots, T_2$, are samples from the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y})$ (taken from Gibbs sampling output) and from the proposal distribution $g(\boldsymbol{\theta}, \boldsymbol{W})$, respectively. The proposal distribution $g(\cdot)$ must be close to the target posterior distribution. Function $h(\cdot)$ plays the role of the bridge that links the two distributions.

According to Meng and Wong (1996), the optimal function under the mean square error is given by $h(\boldsymbol{\theta}, \boldsymbol{W}) = \frac{T_2}{T_2 g(\boldsymbol{\theta}, \boldsymbol{W}) + T_1 \pi(\boldsymbol{\theta}, \boldsymbol{W}|\boldsymbol{y})}$. By replacing this optimal bridge function in (3.20), the bridge sampling indicator, after some manipulations, simplifies to

$$\widehat{m}_o(\boldsymbol{y}) = \frac{\frac{1}{T_2} \sum_{t=1}^{T_2} \frac{f(\boldsymbol{y}|\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)}) \, \pi(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)})}{T_2 \, g(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)}) + T_1 \, \pi(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)}|\boldsymbol{y})}}{\frac{1}{T_1} \sum_{t=1}^{T_1} \frac{g(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)})}{T_2 \, g(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)}) + T_1 \, \pi(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)}|\boldsymbol{y})}} \tag{3.22}$$

For more details and a thorough explanation of this method, see Gronau et al. (2017); Gronau, Singmann and Wagenmakers (2020).

In this work, the proposal distribution is specified at the product of two independent distributions, i.e. $g(\boldsymbol{\theta}, \boldsymbol{W}) = g_\theta(\boldsymbol{\theta}) g_W(\boldsymbol{W})$, where the proposal g_θ for $\boldsymbol{\theta}$ is taken to be a Normal distribution, while the proposal for $\boldsymbol{W}$ is specified as an Inverse-Wishart distribution. This choice is computationally convenient because it yields a proper proposal with correct support for $\boldsymbol{W}$ (positive definite) and it is straightforward. In practice, this bridge sampling can fail or become unreliable for skewed or heavy-tailed posteriors that are not well captured by a $g(\boldsymbol{\theta}, \boldsymbol{W}) = g_\theta(\boldsymbol{\theta}) g_W(\boldsymbol{W})$ approximation, or in multimodal settings where a single unimodal proposal misses important regions of posterior distribution, leading to unstable estimates.

Finally, the iterative expression of (3.22) is implemented to obtain the optimal bridge sampling estimator, as described by Meng and Wong (1996, pag. 837), where an initial guess of the marginal likelihood is updated until convergence is achieved based on a predefined tolerance level. The Bayes theorem is then applied to the joint posterior distribution, and by performing some algebraic manipulations the following expression for the marginal likelihood $\widehat{m}_o(\boldsymbol{y})^{(z+1)}$ at iteration $z+1$ is obtained:

$$\widehat{m}_o(\boldsymbol{y})^{(z+1)} = \frac{\frac{1}{T_2}\sum_{t=1}^{T_2} \frac{S(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)})}{T_2\,\widehat{m}_o(\boldsymbol{y})^{(z)} + T_1\,S(\boldsymbol{\theta}^{*(t)}, \boldsymbol{W}^{*(t)})}}{\frac{1}{T_1}\sum_{t=1}^{T_1} \frac{1}{T_2\,\widehat{m}_o(\boldsymbol{y})^{(z)} + T_1\,S(\widetilde{\boldsymbol{\theta}}^{(t)}, \widetilde{\boldsymbol{W}}^{(t)})}} \tag{3.23}$$

where $S(\boldsymbol{\theta}, \boldsymbol{W}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{W})\pi(\boldsymbol{\theta},\boldsymbol{W})}{g(\boldsymbol{\theta},\boldsymbol{W})}$. Furthermore, to obtain a reliable estimate of the marginal likelihood, Overstall and Forster (2010) proposed dividing the posterior samples from the MCMC procedure (e.g., the Gibbs sampling as in the current section example) into two parts: (a) the first part is used to specify the parameters of the proposal distribution, and (b) the second part is used in (3.23) to compute the numerator of the bridge sampling estimator.

This procedure was implemented via the *bridgesampling* package in *R* (Gronau, Singmann and Wagenmakers, 2020), which provides an efficient framework for computing normalizing constants for Bayesian modeling. Notably, the "Warp-III" (Gronau, Heathcote and Matzke, 2020) variant is implemented within the *bridgesampling* package and refers to a transformation-based strategy to further enhance the efficiency and accuracy of marginal likelihood estimation. By applying a nonlinear transformation to the posterior samples, "Warp-III" aims to better align the shapes of the posterior and proposal distributions.

3.3 Conclusion

In summary, frequentist data modeling provides a framework based on likelihoods by emphasizing point estimates, while Bayesian data modeling integrates prior knowledge with observed data to produce full posterior distributions, offering a flexible approach for quantifying uncertainty. Both frameworks are used in this thesis: the frequentist approach is applied to HMMs for modeling dynamic signature features (Chapter 5), and the Bayesian approach is used for modeling handwriting features (Chapter 4).

HMMs are adopted because they naturally represent the sequential structure of dynamic signatures by associating observations to hidden states (i.e, high-velocity pen movements, low-velocity pen movements). Therefore, they can capture distinct phases of the signing process and modeling the temporal variability.

The Bayesian approach is adopted because it can incorporate information about the writer population through the prior distribution, allowing key sources of variation, such as, between writers and within writers, to be modeled explicitly. Accounting for these variation is crucial when evaluating whether a questioned document belongs to the person of interest, because it supports not only wether the differences exists, but also an assessment of how large those differences are relative to typical population variations. This leads to clearer uncertainty quantification and more defensible interpretation, particularly when limited samples are available for a given writer.

Chapter 4

Handwriting Examination

Forensic document examiners frequently need to cope with situations involving handwritten documents whose writership is questioned. Consider the following scenario involving a handwritten document whose origin is contested. Written material originating from an individual who is suspected to be the actual source of the disputed item is collected and examined for comparative purposes. The available evidence should be evaluated under a set of hypotheses in a given framework of information put forward by parties (i.e., the prosecutor and the defense attorneys, respectively) at trial. The propositions of interest can therefore be specified as follows:

H_1: the person of interest (PoI) is the writer of the questioned document;

H_2: the person of interest (PoI) is not the writer of the questioned document.

First of all, $n_1 > 0$ characters are selected from the anonymous manuscript document; these are referred to as questioned material. The value of n_1 is constrained by the document itself, as only legible and complete instances of the target characters are retained. Then, $n_2 > 0$ characters are selected from the handwritten material originating from the person of interest; these are referred to as control or reference material. Similarly, n_2 depends on the amount of available reference material. Measurements (e.g., Fourier coefficients, see Section 4.3) made on questioned and control material are denoted by: $\{\boldsymbol{y}_w\} = (\boldsymbol{y}_{w\ell j}, w = 1, 2,\ \ell = 1, \ldots, L,\ j = 1, \ldots, n_{w\ell})$, with $\boldsymbol{y}_1$ denoting the questioned material and $\boldsymbol{y}_2$ the control material, L the number of character and $n_{w\ell}$ repetitions per character. Thus, it is crucial to provide information about the value of the evidence $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ in support of the propositions of interest. This is given by the ratio of the marginal likelihoods under the competing propositions:

$$BF = \frac{m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_1)}{m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_2)} = \frac{m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_1)}{m(\boldsymbol{y}_1|H_2)m(\boldsymbol{y}_2|H_2)}. \tag{4.1}$$

The marginal likelihoods in (4.1) can be obtained using the same probabilistic model that will be adopted to describe the handwriting data (and that will be illustrated in Section 4.5), but using different sets of measurements for prior elicitation. Specifically, the marginal likelihood in the numerator, $m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_1)$, is calculated assuming that all observations of the combined dataset $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ have a common parameter vector, since it is assumed that H_1 holds, so that data originate from the same source. On the other hand, if H_2 holds and the competing materials originate from different sources, the measurements $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ can be considered independent. Consequently, the marginal likelihood $m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_2)$ can be obtained as the product of two independent marginal likelihoods, $m(\boldsymbol{y}_1|H_2)$ and $m(\boldsymbol{y}_2|H_2)$, respectively. These latter are calculated by the same probabilistic model, but by

fitting this last one separately for each dataset $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ and by taking different parameter vectors. This assumption implies that possible disguised behavior is not considered (see Appendix A.1 for the considered BF assumptions). For more definition in identification problems in forensic science, see Ommen et al. (2017).

The data that we are going to use for this research refers to three main sources. First, a sample of 13 writers and characters *a*, *d*, *o* and *q* collected for a previous study (Marquis et al., 2006), that were selected among a population of French native writers from the School of Criminal Justice of the University of Lausanne (Switzerland) because of their habits to close loops (see Section 4.2). Second, we have considered writers from the IAM Handwriting Database, which comprises various forms of handwritten English texts. This database was initially released by Marti and Bunke (1999) at the International Conference on Document Analysis and Recognition (ICDAR) in 1999. Since the proposed image analysis procedure can actually be implemented to describe the shape of closed loops of characters, we have identified and selected 50 writers showing a substantial number of characters of interest with closed loops (see Section 4.2). These writers are considered as background information for the third available source, which is the case study of Section 4.9.

The contour shape of the characters of the above data sources was processed according to the image analysis procedure proposed by Schmittbuhl et al. (1998) for the polymorphism analysis of the piriform aperture of given skeleton bones of some primates (e.g., the mandibule) and adapted by Marquis et al. (2005) for handwriting examination purposes; for more details see Section 4.4.

For such features, six Bayesian models with two different likelihood structures and three different prior specifications are analyzed and compared.

- First, the multivariate Normal model is implemented. The distribution of available data is taken to be a multivariate Normal, with (a) the Normal-Inverse-Wishart conjugate prior setup (Section 4.5.1) and (b) a prior with a similar setup where the variance-covariance matrix is independent of the mean parameter vector (Section 4.5.1), and (c) a prior with covariance decomposition where the within-writer variability follows LogNormal-LKJ distribution (Section 4.5.1). Under this model, the character-level variability is not modeled, since one common variance-covariance matrix is considered for all observed characters. Therefore, all measurements are evaluated jointly without distinction by character type. In order to take into consideration the impact of this further layer of variability, the models have also been fitted for each character separately.

- Secondly, a Bayesian MANOVA model is proposed in Section 4.5.2. Under this second model, the type of characters is taken as dummy indicator variables (with corner-point representation; see Appendix A.2 for more details). Similarly to the first modeling approach, a conjugate prior specification (Section 4.5.2), a hierarchical Normal-Inverse-Wishart prior specification (Section 4.5.2), and a Normal-LogNormal-LKJ prior specification (Section 4.5.2) have been considered.

Since the marginal likelihoods that are needed to calculate the Bayes factor in (4.1) are not always available in analytical form, three different methods for its estimation will be considered, see Section 4.6.

In this chapter, we present the methodological framework employed in the latter forensic examination of the handwriting data. The considered datasets include samples collected from both real and simulated forensic scenarios, complemented by background material from publicly available sources such as the IAM database; see Section 4.2. The analysis begins with data preprocessing, in which scanned handwriting images are preprocessed to ensure their suitability for analysis; see Section 4.3. We then perform feature engineering to derive informative variables that enhance the discriminative power of the subsequent analysis; see Section 4.4. A primary exploratory analysis follows, providing

an initial overview and summary of the measurements and establishing the context for more refined modeling; see Section 4.4.1. Building on this foundation, we employ a Bayesian modeling approach that incorporates prior knowledge on the parameter of interest (e.g., the mean of measurements) and provides a principled framework for quantifying uncertainty; see Section 4.5. This ultimately enables the coherent evaluation of handwriting evidence through the Bayes factor, providing a principled basis for rigorous forensic inference. The experiments used to assess the error rates of the methodology, as well as the sensitivity analysis, are presented in Sections 4.7 and 4.8, respectively. Finally, a real case study is provided in Section 4.9.

4.1 Literature Review

Handwriting examination is a domain of forensic science, where document examiners are often asked to inform the actors of a legal process, who are confronted with handwritten documents whose origin is contested or unknown. A number of recent studies have explored novel methods for inspecting handwritten documents and assessing the evidence on writership problems. Johnson and Ommen (2022) developed a method that quantifies the similarity between handwritten documents using machine learning and statistical techniques. They applied score-based likelihood ratios (SLRs)[1] to assess the value of the data in two scenarios: common source and specific source[2]. Wydra and Matuszewski (2022) proposed a method that evaluates handwriting material using the likelihood ratio (LR) approach[3]. They measured the similarity between handwriting samples by the Jaccard index, which is a statistic that captures the overlap between two sets of features. Crawford et al. (2023) propose rotation-based features to extract measurements from handwritten documents, and a Bayesian hierarchical model[4] to estimate the posterior predictive probability of writership and test the posterior predictive performance when the author of the questioned document is part of a closed set of writers.

This research is going to investigate and extend the Bayesian probabilistic approach proposed by Bozza et al. (2008). In that work, specific handwritten features originating from an ad-hoc image analysis procedure proposed by Marquis et al. (2005) have been used. This technique is based on Fourier analysis and enables a precise reconstruction of the contour shape of characters' loops. Following this methodology, each character loop can be described by means of a Fourier coefficient, which can be used to characterize the shape complexity and other geometric attributes. Preliminary studies by Marquis et al. (2006) have shown that these features characterization has a good discriminating power. The value of the features is subsequently assessed by means of the Bayes factor. The use of the Bayes factor as a metric to assess the probative value of forensic findings is largely supported by operational standards and recommendations in different forensic disciplines (see, for example, the guidelines of the European Network of Forensic Science Institutes (Willis et al., 2015). A general review of Bayes factors for forensic decision analysis from an operational perspective is given by Bozza et al. (2022), and its logical foundations are described in Taroni et al. (2021).

In particular, Bozza et al. (2008) proposed a two-level random effects model, which has the ad-

[1]The SLRs evaluate the strength of the observed data under each hypothesis by comparing the probability distributions of similarity scores (Bozza et al., 2022).

[2]The common source scenario involves determining whether two sets of evidence originate from the same, but unknown, source. Conversely, the specific source scenario focuses on deciding whether a single set of evidence can be attributed to a known, specified source (Ommen et al., 2017).

[3]The LR compares the probability of the data under two competing hypotheses, providing a direct measure of how much more likely the data is under one hypothesis compared to the other. In contrast, the BF is used within a Bayesian framework to update the probability of a hypothesis based on the evidence, incorporating prior probabilities.

[4]Bayesian hierarchical models are statistical models that incorporate multiple levels of random variables to account for complex data structures like nested or hierarchical data.

vantage of modeling both within- and between-writer variability. However, as the authors themselves pointed out, the main limitation of this proposal is that it does not model the variability that characterises each type of character. This means that the model can be either applied without distinguishing between different character types (and thus losing an important source of information) or separately for each character type. This can be problematic, especially when it comes to combining conflicting evidence (i.e., evidence supporting different hypotheses). Another open issue that was underlined by the authors is the sensitivity of the Bayes factor to the degrees of freedom of the prior distribution modeling the handwriting variability.

Finally, the latter modeling was used for investigation purposes by Taroni et al. (2012, 2014). The approach Taroni et al. (2012) uses data from female and male writers to compare the likelihood ratio based methods for inferring the gender of the writer of a questioned document. Furthermore, the work of Taroni et al. (2014) computes the Bayes factor for different scenarios of inferring the gender and handedness of the writer of a questioned document, based on the analysis of the handwriting characters 'a' and 'd'.

4.2 Data Acquisition, Sampling and Databases

The first dataset used for this study refers to a sample of 13 writers collected for a previous study (Marquis et al., 2006), who were selected among a population of French native writers from the School of Criminal Justice of the University of Lausanne (Switzerland) because of their habit of closing loops. Specifically, one hundred individuals completed five documents. Each of these documents was completed on different days, and they contain ten times a series of letters of the alphabet written in the usual manner. Among the samples collected, only thirteen writers had closed *a*, *d*, *o*, and *q* loops. The samples from the other individuals were not used for the rest of the study. Among the samples from the thirteen writers selected, some loops had to be rejected due to their image noise. Among the thirteen writers selected, all of whom are right-handed and four are men.

To increase the reliability of our methodology, a larger sample size is required. To this end, we have selected 50 writers from the IAM Handwriting Database for which a large amount of data is available. This database was initially released by Marti and Bunke (1999) at the International Conference on Document Analysis and Recognition (ICDAR) in 1999. The IAM Handwriting Database comprises various forms of handwritten English texts, which have been utilized for tasks such as text recognition, writer identification, and verification. It includes unconstrained handwritten samples, scanned at a resolution of 300 dpi and stored as PNG images with 256 gray levels. The corpus of texts within the IAM Database is derived from sentences provided by the Lancaster-Oslo/Bergen (LOB) Corpus (Johansson et al. (1978)). Currently, the database encompasses contributions from 657 writers, amounting to 1539 pages of scanned texts. Since the proposed image analysis procedure can actually be implemented to describe the shape of closed loops of characters, we have identified and selected 50 writers showing a substantial number of characters of interest with closed loops.

4.3 Image Preprocessing

The preprocessing of handwriting data primarily involves an image analysis procedure, as illustrated in Figure 4.1. This process can be summarized as follows:

1. A character is digitized to an image;

2. The image is binarized;

3. The skeleton of the image is obtained to extract the contour;

4. The contours are expressed in polar coordinates.

Note that, to eliminate the influence of the size on shape analysis, the contour was normalized. All contour coordinates were adjusted so that the areas enclosed by them were equal to $1\,cm^2$. Consequently, the surface measurement (S) was extracted as an additional general feature of interest, measuring the approximated area of each loop character.

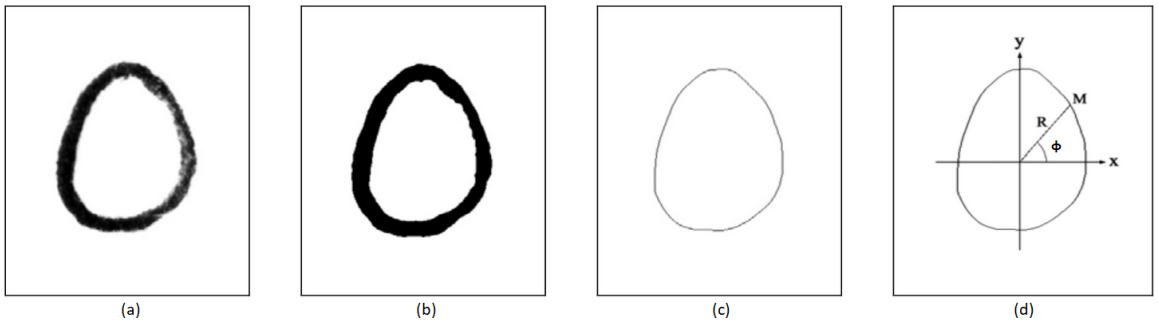


Figure 4.1: Image analysis procedure: from the original loop to polar coordinates. Adapted from Marquis et al. (2005).

This approach is recommended by Marquis et al. (2005) to enhance automation in the preprocessing pipeline. Rather than manually separating individual characters, in this study, we segment words directly from the entire text, as illustrated in Figure 4.2. Since the analysis focuses on the loop characters, letters whose written form contains a closed curved stroke (*a*, *b*, *d*, *e*, *g*, *o*, and *p*), precise segmentation is of secondary importance, provided that the resulting batches of words are reasonably isolated for further analysis.

Once a word or a discrete cluster of words is obtained, a sequence of image transformations is applied following the recommendations in Marquis et al. (2005). These transformations include binarization and skeletonization, with the intermediate results shown in Figures 4.3 and 4.4. The final stage of image preprocessing involves identifying closed contours within the skeletonized image.

Subsequently, contours corresponding to character loops are manually labeled. For this purpose, images such as the one in Figure 4.5 are generated, enabling users to associate each contour with its respective letter. This manual intervention is essential for correcting any erroneous labels generated in previous steps. Finally, each labeled contour is transformed into polar coordinates in preparation for subsequent Fourier analysis, as depicted in Figure 4.6.

Figure 4.2: Word Separation

Figure 4.3: Binarization

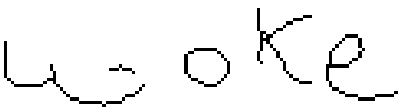


Figure 4.4: Skeletonization

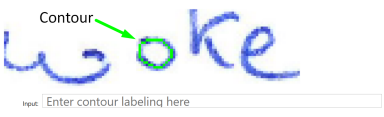


Figure 4.5: Contour Labeling

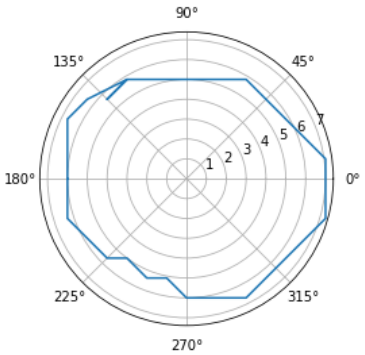


Figure 4.6: Polar representation of the contour

4.4 Fourier based Feature Engineering

From the polar coordinates, each loop character can be expressed as a Fourier series:

$$R(\phi) = a_0 + \sum_{h=1}^{H}[a_h cos(h\phi) + b_h sin(h\phi)], \qquad \phi \in [0, 2\pi). \tag{4.2}$$

In this way, each character loop can be described by a discrete function R(ϕ) representing the length of a line joining a point of the contour to the centroid, where ϕ is the angle made by this line with the horizontal axis. The contour shape can therefore be described by a series of harmonics, each one characterizing a specific contribution to the shape. Each harmonic is constructed by a pair of Fourier coefficients, i.e., a_h and b_h ($\in R$) as proposed in Thiéry (2014). Note that the original proposal by Marquis et al. (2005) envisaged the amplitude-phase form of the Fourier series:

$$R(\phi) = A_0 + \sum_{h=1}^{H}[A_h cos(h\phi - \phi_h)], \qquad \phi \in [0, 2\pi) \tag{4.3}$$

where the amplitude A_h ($\in R^+$) and the phase ϕ_h (degrees or radians) are just the polar coordinates of the coefficients a_h and b_h, and $a_h = A_h cos(\phi_h)$, $b_h = A_h sin(\phi_h)$, and $A_0 = a_0$. Fourier descriptors (A_h and ϕ_h).

In this work, a multivariate Gaussian distribution is adopted as the probabilistic model, motivated by the correlation among the extracted features and by the lack of evidence against univariate normality according to the Lilliefors test per writer. Moreover, the Fourier coefficients (a_h and b_h) exhibit more favorable values of Mardia's multivariate skewness and kurtosis, computed per writer, than the corresponding Fourier descriptors (Mardia, 1970). This remains true even after applying logarithmic or square-root transformations to the Fourier descriptors. We therefore base our subsequent analysis on the Fourier coefficients (a_h and b_h).

The harmonic contribution to the contour shape of characters is represented in Figure 4.7: the first harmonic ($h = 1$) informs about the ovate contribution to the shape, the second ($h = 2$) about the ellipticity of the shape, the third ($h = 3$) about the triangularity, the forth ($h = 4$) about the quadranqularity and the fifth ($h = 5$) about the pentagonality. A more detailed description of this image analysis procedure can be found in Schmittbuhl et al. (1998) and Marquis et al. (2005).

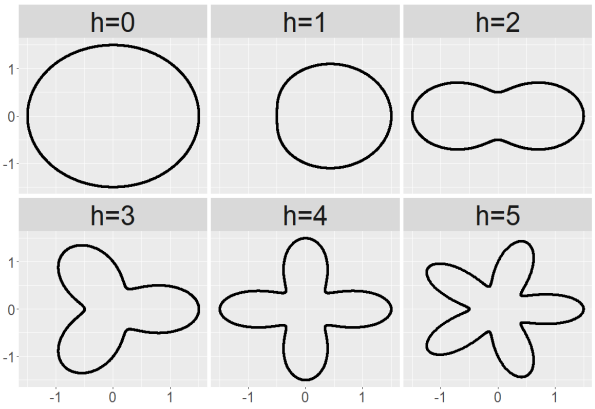


Figure 4.7: Harmonics contribution obtained by the sum of the unit circle (h=0) and the specific harmonics of interest, $\alpha_h = 0.5$ and $b_h = 0$

.

Figure 4.8 illustrates the reconstruction of a skeletonized contour of a given character by summing progressively the Fourier harmonics describing its shape.

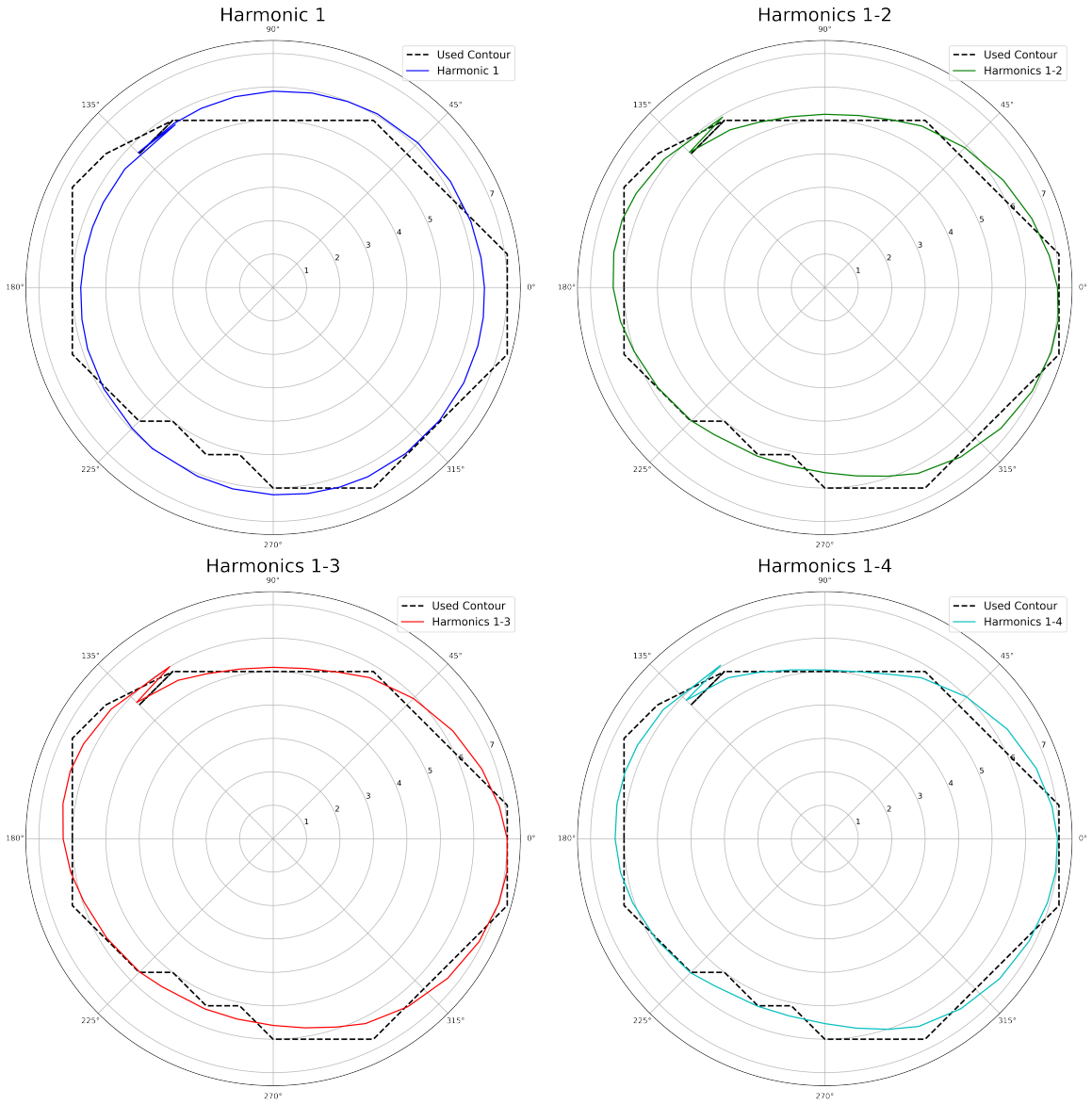


Figure 4.8: Progressive reconstruction (solid curves) of a contour (dashed curve) based on the first four Fourier harmonics.

The reconstruction of the original contours obtained by means of such a Fourier-based procedure is represented in Figure 4.9, for each writer and each character. For each analyzed loop, the average of the Fourier parameters is considered with a total number of $H = 4$ and $H = 10$ harmonics. The final image represents the average contour of loop characters. As can be observed, while some writers present loop structures characterized by marked peculiarities, others are more similar. The size is not illustrated because the original image had been normalized. This illustration does not account for within-writer variability, which will be addressed in the data modeling section (see Section 4.5).

In this project, only the first four harmonics will be retained, as suggested by Marquis et al. (2005). The global shape of each character contour can, in fact, be reconstructed without the contribution of further harmonics, as it can be observed in Figure 4.9. Furthermore, it must be underlined that the

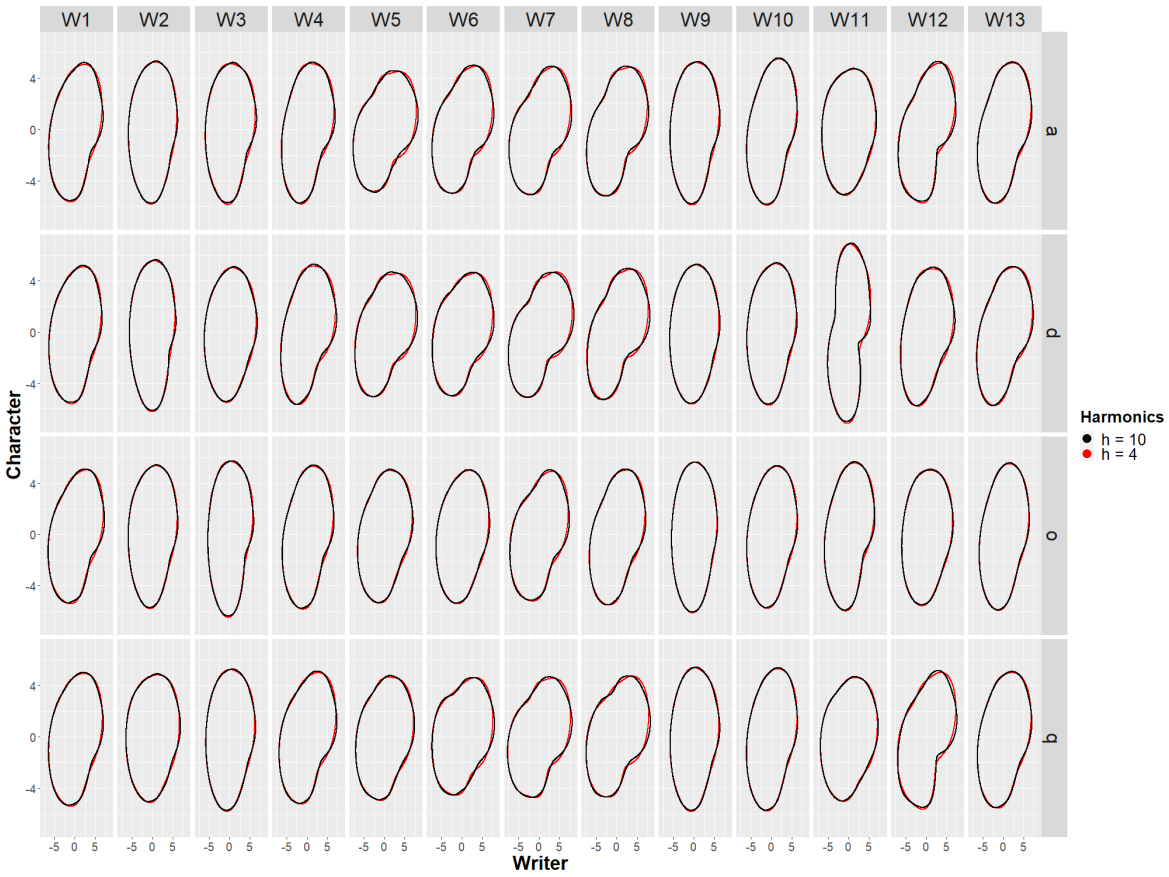


Figure 4.9: Reconstructed loop characters shown for the average Fourier coefficients per writer and character using the first four pairs of Fourier coefficients (H = 4; red curve) and the first ten pairs (H = 10; black curve); the size is not illustrated since all characters were normalized to a surface size of 1 cm^2.

constant term α_0 (or A_0 in the Amplitude-phase form) does not have an impact on the morphological characteristics of the shape of the characters under study, as demonstrated by Marquis et al. (2006). Therefore, this term is considered a nuisance parameter and is not further considered in the subsequent analyses. Thus, each character loop can be described by means of $p = 9$ variables representing the surface (S) of the character, the α_h and the b_h of the first four harmonics, $h = 1, \ldots, H$. Hence, the available background data contain the measurements of L characters collected in correspondence of m writers, and can be denoted by a four-dimensional array $\boldsymbol{\mathcal{D}}$ with elements $\mathcal{D}_{i\ell j\kappa}$, where $i = 1, \ldots, m$ (writers), $\ell = 1, \ldots, L$ (characters), $j = 1, \ldots, n_{i\ell}$ (repetition), and $\kappa = 1, \ldots, p$ (Fourier coefficients and the surface size). Note that coefficients will be standardized by dividing each value by the overall standard deviation of each Fourier coefficient.

4.4.1 Descriptive Statistics and Visualizations

In this section, we present a basic descriptive analysis of the main characteristics of the sample used for the implementation of the proposed methodology. The Fourier coefficients can be represented in Cartesian coordinates, as shown in Figure 4.10, which displays, for each harmonic, each writer and across all characters, the mean and the standard deviation of the pair of coefficients (a_h,b_h). This graphical representation can be informative of the discriminating power of each harmonic. In particular, the second harmonic measuring ellipticity seems the most informative for discriminating

purposes. For an analytical description of the available characters' features, the reader is referred to graphical illustrations in Appendix A.3 Figure A.3.1.

Finally, the Mahalanobis distance has been quantified to measure the variability between pairs of writers across all characters. The Mahalanobis distance between two writers is calculated by measuring how far the feature vectors of one writer's samples are from the mean and covariance structure of another writer's samples. From Figure 4.11, where the square roots of Mahalanobis distances are represented, it can clearly be observed that while some pairs of writers present very small distances (e.g., writers 6, 7 and 8), other pairs of writers are characterized by greater distances (e.g., writers 10 and 13), confirming more pronounced peculiarities that should make them easier to discriminate.

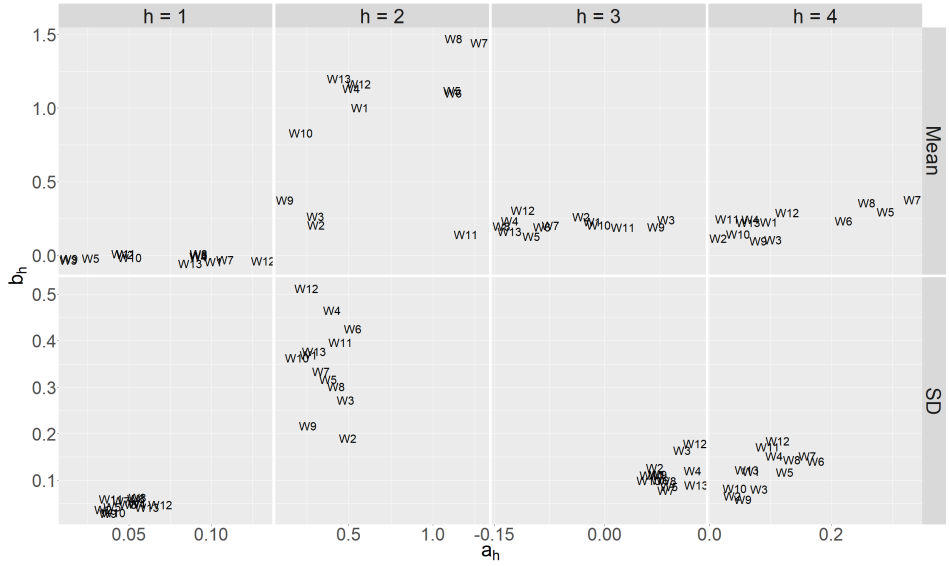


Figure 4.10: Mean and standard deviation of every pair of Fourier coefficients (a_h, b_h), for each harmonic, each writer, and across all characters.

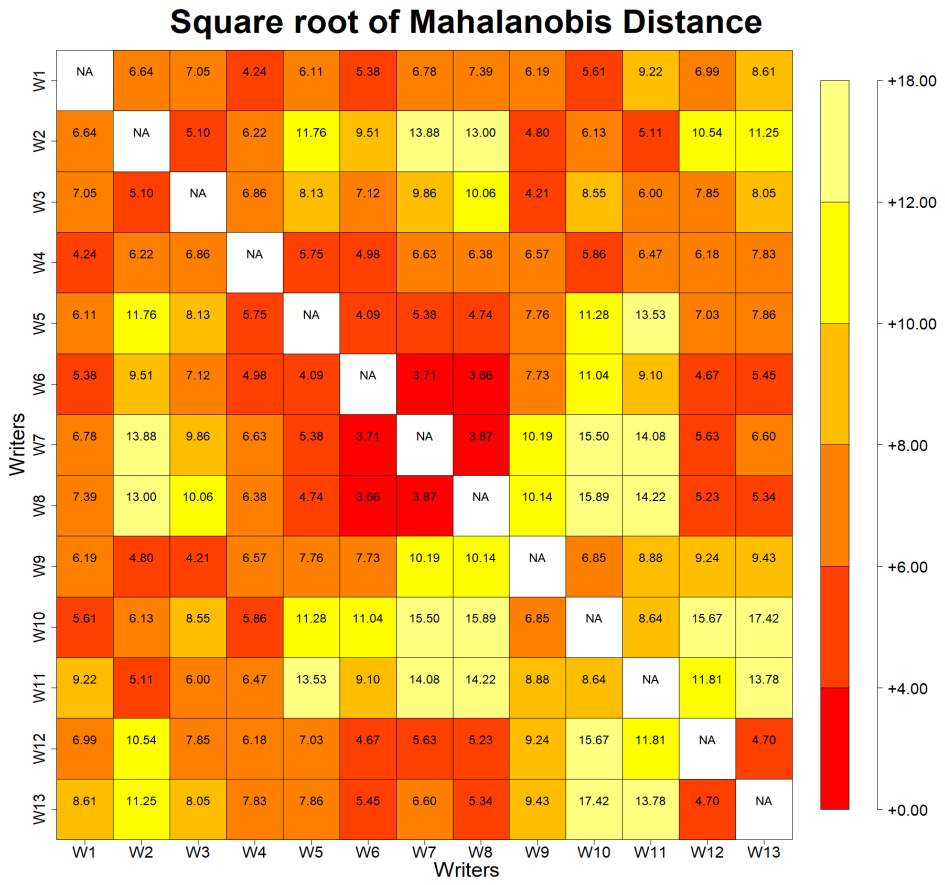


Figure 4.11: Square root of the Mahalanobis distance between writers using the Fourier coefficients and the surface size across all characters.

4.5 Modeling Fourier Coefficients and Surface size

Let us consider the available measurements on questioned and control material denoted as $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively, described in the beginning of Chapter 4, $\{\boldsymbol{y}_w\} = (\boldsymbol{y}_{w\ell j\bullet}, w = 1, 2,\ \ell = 1, \ldots, L,\ j = 1, ..., n_{w\ell})$. Following the Fourier-based features described in Section 4.4, these data form a vector of length p, corresponding to the surface size, along with four pairs of Fourier coefficients, and are denoted as $\mathbf{y}_{w\ell j\bullet}$. By $\boldsymbol{\theta}_w \in \mathbb{R}^p$ we denote the mean vector within material w, and by $\boldsymbol{W}_w \in \mathbb{R}^{p\times p}$ the variance-covariance matrix within material w with elements $W_{w\kappa_1\kappa_2}$ for $(\kappa_1, \kappa_2) \in \{1, 2, \ldots, p\}^2$. Then, given $\boldsymbol{\theta}_w$ and $\boldsymbol{W}_w$, the distribution of $\mathbf{y}_{w\ell j\bullet}$ is taken to be p-variate Normal N_p, with

$$\boldsymbol{y}_{w\ell j\bullet} \sim N_p\left(\boldsymbol{\theta}_w, \boldsymbol{W}_w\right). \tag{4.4}$$

Let us further assume that we have a dataset (or database) of manuscripts from different writers, unrelated to the case, denoted by $\boldsymbol{X}$. We refer to this as background data, which will later be used to specify model parameters (i.e., to elicit the prior distributions) where required. We further assume that the background data follow the same feature engineering process as $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Accordingly, our modeling framework involves three distinct datasets: (i) the questioned data, $\boldsymbol{y}_1$, (ii) the control data, $\boldsymbol{y}_2$, and (iii) the background data, $\boldsymbol{X}$. The questioned and control data, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, are directly used to evaluate hypotheses H_1 and H_2 via the Bayes factor (4.1), whereas the background data $\boldsymbol{X}$ are used

indirectly (via the prior) to facilitate parameter estimation in each model.

4.5.1 Bayesian Normal Models

In this first modeling approach, a Normal-Inverse-Wishart (NIW) prior distribution is chosen for the parameters $(\boldsymbol{\theta}_w, \boldsymbol{W}_w)$. Under this general prior set-up, and considering the sampling model distribution (4.4) for $\boldsymbol{y}_{w\ell j\bullet}$, the following Bayesian model is specified:

$$\begin{aligned} \boldsymbol{y}_{w\ell j\bullet} &\sim N_p(\boldsymbol{\theta}_w, \boldsymbol{W}_w) \\ \boldsymbol{\theta}_w|\boldsymbol{X} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{G}) \\ \boldsymbol{W}_w|\boldsymbol{X} &\sim IW(\boldsymbol{U}, \nu), \end{aligned} \tag{4.5}$$

where $\boldsymbol{\mu}$ is the prior mean vector of $\boldsymbol{\theta}_w$, $\boldsymbol{W}_w$ is the within-writer covariance matrix, $\boldsymbol{G}$ is the covariance matrix of $\boldsymbol{\theta}_w$, while $\boldsymbol{U}$ and ν represent the scale matrix and the degrees of freedom of the Inverse-Wishart distribution that models the within-writer variability.

Depending on the specification of the prior variance-covariance matrix $\boldsymbol{G}$, two distinct variants of the Bayesian model in (4.5) are obtained. First, by setting $\boldsymbol{G} = \boldsymbol{W}_w k_0^{-1}$, a conjugate prior is obtained (Section 4.5.1), and the prior mean vector is denoted as $\boldsymbol{\theta}_w|\boldsymbol{X}, \boldsymbol{W}_w$. Conversely, whenever the parameter $\boldsymbol{G}$ is set as the between-writers covariance matrix $\boldsymbol{B}$, a model with hierarchical prior setup is obtained (Section 4.5.1). To upgrade the latter prior approach, we perform a covariance decomposition of the within-writer covariance matrix $\boldsymbol{W}_w$, employing the LogNormal-LKJ prior approach as specified in Section 4.5.1.

Conjugate Prior: Normal-Inverse-Wishart

By assuming data follow a multivariate Normal distribution, $N_p(\boldsymbol{\theta}_w, \boldsymbol{W}_w)$, the implied conjugate prior distribution for $(\boldsymbol{\theta}_w, \boldsymbol{W}_w)$ is the Normal-Inverse-Wishart (NIW) distribution with $\boldsymbol{G} = \boldsymbol{W}_w k_0^{-1}$,

$$(\boldsymbol{\theta}_w, W_w)|\boldsymbol{X} \sim NIW(\boldsymbol{\mu}, k_0, \boldsymbol{U}, \nu),$$

for a detailed description of the model parameters, see Section 3.2.1.

Since the NIW prior is conjugate, the marginal likelihood of the available measurements $\boldsymbol{y}$ is readily available in closed form and is given by Eq. (3.10). In this work, the prior parameters $\boldsymbol{\mu}$ and $\boldsymbol{U}$ are elicited by using information from readily available background data $\boldsymbol{X}$; see Appendix A.4.1.

The degrees of freedom ν are set to be equal to $p + 2$, which is the smallest value for which the mean of the Inverse-Wishart distribution is available (Press, 2005). Finally, the prior parameter k_0 is selected based on a grid search in the (0,1) interval for all writers available in the background data. Specifically, the choice falls on the value of k_0 that maximizes the marginal likelihood in the background data.

Although the conjugate prior is the natural choice because it is computationally convenient, it has the disadvantage of not modeling the variability within-writer and between-writers separately. This can be a major limitation in this specific context, since the fundamental laws of handwriting state that (1) no one writes the same word exactly the same way twice (i.e., variability within writers), and (2) no two people write exactly the same way (i.e., variability between writers). For this reason, a second model with an independent specification of the prior is considered.

Hierarchical Extension: Normal-Inverse-Wishart

Bozza et al. (2008) chose an hierarchical prior set up for the model parameters $(\boldsymbol{\theta}, \boldsymbol{W})|\boldsymbol{X}$ of the following form

$$\boldsymbol{\theta}_w|\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{B}) \text{ and } \boldsymbol{W}_w|\boldsymbol{X} \sim IW(\boldsymbol{U}, \nu). \tag{4.6}$$

The proposed model is then given by (4.5) with $\boldsymbol{B} = \boldsymbol{G}$ to be fixed and independent of W_w.

In particular, the prior parameter $\boldsymbol{B}$ represents the variance-covariance matrix between writers. The prior parameters $\boldsymbol{\mu}$, $\boldsymbol{B}$, and $\boldsymbol{U}$ are estimated in a relatively straightforward manner from the information of the background writers (see Appendix A.4.1). Empirical evidence has shown that models incorporating the between-writers variability are performing better than models that do not have this flexibility (Bozza et al., 2008). It must be said that without the background information, the ability to effectively discriminate between writers is greatly reduced.

Unfortunately, the marginal likelihood for this model cannot be obtained analytically. Hence, bridge sampling is applied for estimation based on the output of an MCMC algorithm as it is presented and discussed in Section 4.6. In particular, MCMC algorithm No-U-Turn Sampler (NUTS) is implemented in *Stan* (see Carpenter et al., 2017) version 2.26.1, via the *RStan* interface and *R* version 4.2.1.

Covariance Decomposition: Normal-LogNormal-LKJ

A widely used and flexible approach for modeling covariance matrices in Bayesian hierarchical models is to decompose the covariance matrix into standard deviations and a correlation matrix. Specifically, the covariance matrix $\boldsymbol{W}_w$ can be expressed as:

$$\boldsymbol{W}_w = \boldsymbol{D}_w \boldsymbol{R}_w \boldsymbol{D}_w \text{ and } \boldsymbol{D}_w = Diag\left(W_{w11}^{1/2}, W_{w22}^{1/2}, \ldots, W_{wpp}^{1/2}\right); \tag{4.7}$$

where $\boldsymbol{D}_w$ is a diagonal matrix containing the standard deviations of the covariates of material w (for $w = \{1, 2\}$), while $\boldsymbol{R}_w$ is the corresponding correlation matrix. The diagonal elements of $\boldsymbol{W}_w$ specify the corresponding elements of $\boldsymbol{D}_w$ while $\boldsymbol{R}_w$ is obtained by standardizing $\boldsymbol{W}_w$, that is, by simply setting $\boldsymbol{R}_w = \boldsymbol{D}_w^{-1} \boldsymbol{W}_w \boldsymbol{D}_w^{-1}$.

In this decomposition, the standard deviations $\boldsymbol{D}_w$ are typically assigned independent priors, such as LogNormal distribution, or Half-Cauchy distribution for less informative prior specifications. Therefore, this alternative parametrization allows for a more flexible modeling of marginal variances. The correlation matrix $\boldsymbol{R}_w$ is modeled separately using the Lewandowski-Kurowicka-Joe (LKJ) distribution (Lewandowski et al., 2009), which provides a non-conjugate prior specifically designed for correlation matrices. The LKJ distribution has a shape parameter $\eta > 0$ controlling the strength of the prior towards the identity matrix (independence) and probability function given by:

$$\begin{aligned} f(\boldsymbol{R}; \eta) &= \mathcal{C} \times |\boldsymbol{R}|^{\eta - 1} \text{ with} \\ \log(\mathcal{C}) &= \sum_{\kappa=1}^{p} (2\eta - 2 + p - \kappa) \log 2 + \sum_{\kappa=1}^{p-1} (p - \kappa) \log \mathcal{B}(\eta + \tfrac{p-\kappa-1}{2},\ \eta + \tfrac{p-\kappa-1}{2}); \end{aligned}$$

where $\boldsymbol{R}$ is a $p \times p$ positive-definite matrix with unit diagonal, $|\boldsymbol{R}|$ denotes the determinant of the matrix $\boldsymbol{R}$, the normalizing constant $\mathcal{C}$ ensures that the $f(\boldsymbol{R}; \eta)$ integrates to one over the space of $p \times p$ correlation matrices, p is the dimension of the correlation matrix $\boldsymbol{R}$, κ is an index running from 1 to p (or $p - 1$ in the second sum) and $\mathcal{B}()$ is the Beta function.

Hence, the Bayesian model is specified:

$$\begin{aligned} \boldsymbol{y}_{w\ell j\bullet} &\sim N_p(\boldsymbol{\theta}_w, \boldsymbol{W}_w), \quad \boldsymbol{W}_w = \boldsymbol{D}_w \boldsymbol{R}_w \boldsymbol{D}_w \\ \boldsymbol{\theta}_w|\boldsymbol{X} &\sim N_p(\boldsymbol{\mu}, \boldsymbol{B}) \\ D_{w\kappa\kappa}|\boldsymbol{X} &\sim \text{LogNormal}(\upsilon, \sigma) \text{ for } \kappa \in \{1, 2, \ldots, p\}, \\ \boldsymbol{R}_w &\sim LKJ(\eta) \end{aligned} \tag{4.8}$$

The LogNormal(υ, σ) distribution is parameterized by the location parameter υ and the scale parameter σ, which are estimated based on prior knowledge about standard deviations derived from the background data enabling the prior elicitation (see Appendix A.4.1) and the LKJ parameter η is set equal to 1 for non-informative prior for the correlation matrix. The LogNormal prior was selected based on prior knowledge derived from background data, enabling the elicitation of prior parameters υ and σ. The same approach for prior specification is applied to the other model parameters, as described in Section 4.5.1. The marginal likelihood for the Normal-LogNormal-LKJ prior setup is not analytically available in closed form. Hence, bridge sampling is implemented for the estimation of the marginal likelihood, as described in Section 4.6. For the extraction of posterior samples of the parameters, the MCMC algorithm No-U-Turn Sampler (NUTS) is implemented in *Stan* (see Carpenter et al., 2017) version 2.26.1, via the *RStan* interface and *R* version 4.2.1.

According to Huang and Wand (2013), the LKJ prior approach can offer better performance over the classical Inverse-Wishart prior approach, particularly in scenarios where a more flexible prior for the correlation matrix is required or when limited prior information about the covariance structure is available. Additionally, the LKJ prior is the most popular approach in the context of sparse covariance matrix estimation.

All the modeling approaches considered in Section 4.5.1 do not account for character variability since characters of different types must be either treated without any distinction, or analyzed separately. Both of the above approaches, analyzing all data but treating all characters without any distinction or analyzing each character separately, has major disadvantages. Specifically, in the first approach, treating all characters without distinction ignores not only the variability of the handwriting of characters within each writer but also the differences of characters across different writers. Alternatively, analyzing each character separately raises the challenge of aggregating individual results into a single measure of evidence for H_1 or H_2. Character-specific Bayes factors cannot be directly combined, and the situation becomes more complicated when conflicting evidence arises, that is, when different character types support different hypotheses. Hence, in Section 4.5.2 we proceed in proposing models that take into account the within characters' variability.

4.5.2 Bayesian MANOVA

In this section, a second modeling approach is provided in an attempt to overcome the critical issues mentioned above. The Bayesian MANOVA model that will now be described has the great advantage of allowing the character-level variability to be modeled as well. Specifically, the indicators of the characters are transformed into dummy variables by selecting the character a as a reference group (see Appendix A.2). Hence, a regression model is built with dummy variables of characters as predictors.

This model, in its general form, can be written as

$$\begin{aligned} \boldsymbol{y}_{w\ell j\bullet} &\sim N_p(\boldsymbol{\Theta}_w^T \boldsymbol{C}_{w\bullet j}, \boldsymbol{W}_w) \\ \boldsymbol{\Theta}_w|\boldsymbol{W}_w, \boldsymbol{X} &\sim MN_{L,p}(\boldsymbol{\mathcal{M}}, \boldsymbol{K}_0^{-1}, \boldsymbol{W}_w) \\ \boldsymbol{W}_w|\boldsymbol{X} &\sim IW(\boldsymbol{U}, \nu), \end{aligned} \tag{4.9}$$

where $MN_{L,p}$ denotes the Matrix Normal with L=4 characters and $p = 9$, the 4 pairs of Fourier coefficients, and the surface size, $\boldsymbol{C}$ is a three-dimensional $m \times L \times n_L$ array of the explanatory dummy variables for $\boldsymbol{y}$. The sub-vector $\boldsymbol{C}_{w\bullet j}$ has length L and contains the dummy variables for each writer and for each repetition of the retained L characters. Each element $C_{w\ell j}$ of $\boldsymbol{C}$ is defined as $C_{wlj} = 1$, for $l = 1$ and for all w, j, while for $l = 2, \ldots, L$ it is defined as

$$\begin{aligned} C_{w\ell j} &= 1, \quad \text{if } \ell = \ell_{wj}^{obs}; \\ C_{w\ell j} &= 0, \quad \text{otherwise}, \end{aligned}$$

where ℓ_{wj}^{obs} is the observed character for questioned or control material w under repetition j (i.e. (1,0,0,0) for *a*, (1,1,0,0) for *d*, (1,0,1,0) for *o* etc.); see at Appendix A.2 for a more detailed presentation. Moreover, $\boldsymbol{\Theta}_w$ is the $L \times p$ coefficient matrix and can be re-written as follows

$$\boldsymbol{\Theta}_w = \begin{pmatrix} \boldsymbol{\vartheta}_{w1}^T \\ \boldsymbol{\vartheta}_{w2}^T \\ \vdots \\ \boldsymbol{\vartheta}_{wL}^T \end{pmatrix}, \tag{4.10}$$

where each $\boldsymbol{\vartheta}_{wl}$, $l = 1, \ldots, L$, is a vector of length p. $\boldsymbol{W}_w$ is the within-writer covariance matrix, ν denotes the degrees of freedom, $\boldsymbol{U}$ is the scale matrix of the Inverse-Wishart distribution, $\boldsymbol{\mathcal{M}}$ is the prior mean matrix (of dimension $L \times p$ prior mean matrix), and $\boldsymbol{K}_0$ is a $L \times L$ scale matrix tuning the variance of $\boldsymbol{\Theta}_w$. The matrix Normal distribution in (4.9) can be re-written as

$$\text{vec}(\boldsymbol{\Theta}_w)|\boldsymbol{X} \sim N_{L\times p}(\text{vec}(\boldsymbol{\mathcal{M}}), \boldsymbol{W}_w \otimes \boldsymbol{K}_0^{-1})$$

where vec() vectorize of a matrix and $\otimes$ Kronecker product. Therefore, the overall variance of vec($\boldsymbol{\Theta}_w$) is simply given by $\boldsymbol{W}_w \otimes \boldsymbol{K}_0^{-1}$. However, since the focus here is pointed toward the distribution of each different type of character (i.e., each row of $\boldsymbol{\Theta}_w$), $\boldsymbol{W}_w$ will be taken as the main variance component, while $\boldsymbol{K}_0^{-1}$ will act as a variance multiplicator which will be elicited using background data.

Conjugate Prior

As for the conjugate Normal-Inverse-Wishart prior setup in Section 4.5.1, a conjugate approach is initially considered for the MANOVA model in (4.9) for computational convenience. Under this approach, the natural conjugate prior using the vectorized parameter $\boldsymbol{\Theta}_w$ is of the form:

$$\begin{aligned} \text{vec}(\boldsymbol{\Theta}_w)|\boldsymbol{W}_w, \boldsymbol{X}, \boldsymbol{K}_0 &\sim N_{L\times p}(\text{vec}(\boldsymbol{\mathcal{M}}), \boldsymbol{W}_w \otimes \boldsymbol{K}_0^{-1})) \\ \boldsymbol{W}_w|\boldsymbol{X} &\sim IW(\boldsymbol{U}, \nu), \end{aligned}$$

where $\boldsymbol{K}_0$ is considered here as a fixed parameter elicited from the background data. Under the above Bayesian formulation, the posterior distribution can be expressed as a result of the same family. Hence,

the marginal likelihood can be expressed in closed form and is given by:

$$m(\boldsymbol{y}) = \frac{1}{2\pi^{np/2}} \frac{\Gamma_p(\nu_n/2)}{\Gamma_p(\nu/2)} \frac{|\boldsymbol{U}/2|^{\nu/2}}{|\boldsymbol{U}_n/2|^{\nu_n/2}} \left(\frac{|K_0|}{|K_n|}\right)^{p/2}, \tag{4.11}$$

where n is the sample size, p is the number of variables (i.e., Fourier coefficients and the surface size),

$$\begin{aligned} \nu_n &= \nu + n \\ K_n &= C^T C + K_0 \\ \boldsymbol{U}_n &= \boldsymbol{U} + \boldsymbol{y}^T \boldsymbol{y} + \boldsymbol{\mathcal{M}}^T K_0 \boldsymbol{\mathcal{M}} - \boldsymbol{\mathcal{M}}_n^T K_n \boldsymbol{\mathcal{M}}_n, \text{ and} \\ \boldsymbol{\mathcal{M}}_n &= K_n^{-1} \left(C^T \boldsymbol{y} + K_0 \boldsymbol{\mathcal{M}}\right); \end{aligned}$$

see Rowe (2002, Chapter 8.4) and Soch (2019, Chapter 2.4). A detailed illustration of prior parameter elicitation can be found in Appendix A.4.2.

However, as pointed out previously, although the conjugate approach may be attractive for computational convenience, it has the important limitation of not modeling heterogeneity between writers (i.e, between-writers variability). Therefore, to overcome this problem, a non-conjugate version of the MANOVA model is introduced in the following section.

Hierarchical Extension: MANOVA Normal-Inverse-Wishart

Let us consider a simplified version of Model (4.9), where $\boldsymbol{K}_0^{-1}$ is set equal to the identity matrix and $\boldsymbol{W}_w$ is replaced by the between-writers covariance matrix $\boldsymbol{B}$ in the sampling distribution of $\boldsymbol{\Theta}_w$. This allows us to write

$$\boldsymbol{\Theta}_w | \boldsymbol{W}_w, \boldsymbol{X} \sim MN_{L,p}(\boldsymbol{\mathcal{M}}, \boldsymbol{I}_L, \boldsymbol{B}),$$

which simplifies to

$$\boldsymbol{\theta}_{w\ell} | \boldsymbol{X} \sim N_p(\boldsymbol{\mu}_\ell, \boldsymbol{B}), \qquad \text{for } \ell \in \{1, \ldots, L\}$$

where $\boldsymbol{\mu}_\ell$ is the prior mean vector of $\boldsymbol{\theta}_{w\ell}$. Furthermore, the common variance-covariance matrix $\boldsymbol{B}$ is replaced by $\boldsymbol{B}_\ell$, $l = 1, \ldots, L$, to consider a heteroscedastic version of the above model where, per the character level l, the between-writer variability differs. Under these setups, the final model is given by

$$\begin{aligned} \boldsymbol{y}_{w\ell j\bullet} &\sim N_p(\boldsymbol{\Theta}_w^T \boldsymbol{C}_{w\bullet j}, \boldsymbol{W}_w) \\ \boldsymbol{\theta}_{w\ell} | \boldsymbol{X} &\sim N_p(\boldsymbol{\mu}_\ell, \boldsymbol{B}_\ell) \quad \text{for } \ell \in \{1, \ldots, L\} \\ \boldsymbol{W}_w | \boldsymbol{X} &\sim IW(\boldsymbol{U}, \nu). \end{aligned} \tag{4.12}$$

The elicitation procedure described in Section 4.5.1 has been slightly adapted to specify the prior distribution parameters for the mean vector $\boldsymbol{\theta}_{w\ell}$ and the covariance matrix $\boldsymbol{W}_w$, based on the available background data. In particular, the values of the hyperparameters characterizing the prior distribution of $\boldsymbol{\theta}_\ell$ for $\ell = 2, \ldots, L$ were elicited by taking the differences between the estimated $\widehat{\boldsymbol{\mu}}_a$ which is the sample mean of the background data of character a and $\widehat{\boldsymbol{\mu}}_l$, for $l = 2, \ldots, L$ is the sample mean difference of character l from a. This approach effectively captures the mean differences between each character l and the reference character a. Further details can be found in Appendix A.4.2. As far as the scale matrix $\boldsymbol{U}$ characterizing the prior distribution of $\boldsymbol{W}_w$, the average of the variance-covariance matrix of every writer in the background data was considered. Hence, this is an initial a priori estimate

of the within-writer variability which is a posteriori estimated via $\boldsymbol{W}_w$ using the handwriting data of material w. In the same way, $\boldsymbol{B}_\ell$ is the variance-covariance matrix of character ℓ, and $\boldsymbol{B}_\ell$ is elicited from the sample variance-covariance matrices of the measurements of character ℓ for all writers in the background data. Similarly, as noted in Section 4.5.1 (Model independent prior setup of NIW), if the background information is not incorporated in the prior, then the ability to effectively discriminate between writers is significantly reduced.

The marginal likelihood in this case is not analytically available in closed form expression as in the conjugate approach of Section 4.5.2. Hence, MCMC samples are used for the bridge sampling estimation of the marginal likelihood, as it will be described in Section 4.6. In particular, the No-U-Turn Sampler (NUTS) is implemented in *Stan* (see Carpenter et al., 2017) version 2.26.1, via the *RStan* interface and *R* version 4.2.1.

Covariance Decomposition: MANOVA Normal-LogNormal-LKJ

Similar to Section 4.5.1, we decompose the within-writer variability in model (4.12) by representing the covariance structure as a function of standard deviations and the correlation matrix. Specifically, we assign an independent LogNormal prior to the standard deviations and employ the LKJ distribution as a prior for the correlation matrix. The resulting model can be expressed as follows:

$$\begin{aligned}
\boldsymbol{y}_{w\ell j\bullet} &\sim N_p(\boldsymbol{\Theta}_w^T \boldsymbol{C}_{w\bullet j}, \boldsymbol{W}_w), \quad \boldsymbol{W}_w = \boldsymbol{D}_w \boldsymbol{R}_w \boldsymbol{D}_w \\
\boldsymbol{\theta}_{w\ell}|\boldsymbol{X} &\sim N_p(\boldsymbol{\mu}_\ell, \boldsymbol{B}_\ell) \quad \text{for} \quad \ell \in \{1, \ldots, L\} \\
D_{w\kappa\kappa}|\boldsymbol{X} &\sim \text{LogNormal}(\upsilon, \sigma) \text{ for } \kappa \in \{1, 2, \ldots, p\}, \\
\boldsymbol{R}_w &\sim LKJ(\eta)
\end{aligned} \tag{4.13}$$

For the Normal–LogNormal–LKJ prior setup, the marginal likelihood is not available in closed form, similarly to Section 4.6. Therefore, we estimate it using bridge sampling, as described in Section 4.6.

4.6 Marginal Likelihoods Estimators

This section describes the marginal likelihoods and their estimation for the two main cases considered in the data analysis.

Case 1: Model comparison. We use the full dataset $\boldsymbol{\mathcal{D}}$ (Marquis et al., 2006), namely writer-specific subsets $\boldsymbol{\mathcal{D}}_i$ for $i = 1, \ldots, 13$, to evaluate and compare models M_1–M_6 defined in Section 4.5 and summarised in Table 4.1.

Case 2: Evidence evaluation. We focus on the primary forensic comparison by evaluating H_1 against H_2 via the Bayes factor in Eq. (4.1). This requires three datasets $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{X}$, which are subsets of $\boldsymbol{\mathcal{D}}$ defined at the beginning of Section 4.5. How these subsets are constructed from $\boldsymbol{\mathcal{D}}$ depends on the specific comparison scenario and is described in detail in Section 4.7.

In both cases, only subsets of $\boldsymbol{\mathcal{D}}$ are used (never the full dataset) in accordance with the principle that the same data should not be used twice: once to inform the prior and again in the likelihood, as doing so would introduce bias into the results.

Model	Description (Model and Prior Specification)	Prior Abbreviation	Section
M_1	Normal model with Conjugate Normal-Inverse-Wishart	NIW Conjugate	4.5.1
M_2	Normal model with Hierarchical Normal-Inverse-Wishart	NIW Hierarchical	4.5.1
M_3	Normal model with Normal-LogNormal-LKJ		4.5.1
M_4	MANOVA model with Conjugate Normal-Inverse-Wishart	NIW Conjugate	4.5.2
M_5	MANOVA model with Hierarchical Normal-Inverse-Wishart	NIW Hierarchical	4.5.2
M_6	MANOVA model with Normal-LogNormal-LKJ		4.5.2

Table 4.1: Overview of the models and prior specifications considered in the study.

4.6.1 Marginal Likelihoods for Model Comparison using the Full Dataset

In order to compare the proposed models, we calculate the Bayes factor for each writer as the ratio of the marginal likelihood for any given pair of models M_l and M_ξ, for $l, \xi \in \{1, \ldots, 6\}$. The Bayes factor quantifies the relative evidence of the compared models based on the available data $\boldsymbol{\mathcal{D}}_i$ of each writer, $i = 1, \ldots, m$, the prior parameters are elicited using data from the remaining writers $(m - i)$. A higher Bayes factor indicates stronger support for one model over the other.

The marginal likelihood of each model using the data $\boldsymbol{\mathcal{D}}_i$ of each writer, $i = 1, \ldots, m$, is given by

$$m(\boldsymbol{\mathcal{D}}_i|M_l) = \int f(\boldsymbol{\mathcal{D}}_i|\boldsymbol{\Theta}, \boldsymbol{W}, M_l)\pi(\boldsymbol{\Theta}, \boldsymbol{W}|M_l)d(\boldsymbol{\Theta}, \boldsymbol{W}), \tag{4.14}$$

for any model M_l (for $l = 1, 2, 3, 4, 5, 6$) and model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{W}$. Models M_1 and M_4 are based on a conjugate prior setups, which means that the marginal likelihoods are available in an analytical form. Nevertheless, this is not the case for models (M_2,M_3) and (M_5,M_6), where the marginal likelihoods are not available in a closed-form expression. In such occasions, three popular Monte Carlo estimators will be used to estimate the marginal likelihoods under modeling approaches (M_2,M_3) and (M_5,M_6): (a) the Generalized Harmonic Mean, (b) the Laplace-Metropolis estimator, and (c) the Bridge Sampling, see Section 3.2.3. Based on the comparative analysis presented in Section 4.6.2, Bridge Sampling consistently yields smaller Monte Carlo Bias and smaller Monte Carlo standard errors (MCSEs) than the other two estimators. Consequently, Bridge Sampling is adopted as the primary estimator for the log-marginal likelihoods of models (M_2,M_3) and (M_5,M_6). For a thorough description of the three Monte Carlo estimators used in this paper and a detailed comparative analysis of their Monte Carlo standard errors, refer to Section 3.2.3.

4.6.2 Comparisons of Marginal Likelihood Estimators

This section compares the precision of the three marginal likelihood estimators for the two modeling approaches. We present the marginal likelihood estimates for each considered model per writer of the dataset (Marquis et al., 2006). Specifically, we present the Monte Carlo bias analysis for the conjugate models M_1, M_4 and the Monte Carlo standard errors (MCSE) obtained with the different marginal likelihood estimators implemented in the M_2, M_3, and M_5, M_6 models is reported for each writer.

Specifically, the estimate of the logarithmic marginal likelihood is obtained, for each writer, from a total number of 30 MCMC runs with 2000 iterations after discarding an additional 1000 iterations as burn-in. Notably, for the elicitation of prior parameters, the data from the remaining writers are treated as background information; specifically, when focusing on writer i, the data from the other $m - i$ writers are utilized to inform the prior parameter estimation.

Conjugate Normal-Inverse-Wishart prior

For the Conjugate Normal-Inverse-Wishart prior of models M_1, M_4 the marginal likelihood is available is closed form Eq. (3.10) and Eq. (4.11), respectively. However, in this section we present the Monte Carlo bias analysis for the three marginal likelihood estimators. For each writer, the log-marginal likelihood (log-ML) is estimated over MCMC replications, and the bias is computed as the deviation of the estimator mean from the closed-form ground truth. For the Generalized Harmonic Mean and Bridge Sampling estimators, we consider as proposal density (importance density) the Multivariate Normal for the mean posterior parameters and Inverse-Wishart for the posterior covariance matrix parameter, for more details see 3.2.3.

For model M_1 (Table 4.2), the Laplace-Metropolis estimator exhibits a consistent negative bias across all writers, with a mean absolute bias of 1.369 log units. Laplace-Metropolis systematic underestimation is expected, as the Laplace approximation relies on a Gaussian approximation to the posterior that becomes increasingly inaccurate when the posterior of $\boldsymbol{W}$ departs from symmetry. The Generalized Harmonic Mean estimator performs substantially better, with a mean absolute bias of 0.010 log units, but systematic overestimation is observed. Bridge Sampling achieves the most symmetric bias overall with mean absolute bias of 0.012.

Writer	Truth	M_1 **Conjugate Normal-Inverse-Wishart**					
		Laplace Metropolis		**Generalized Harmonic Mean**		**Bridge Sampling**	
		log-ML	**Bias**	**log-ML**	**Bias**	**log-ML**	**Bias**
1	-1380.2	-1381.5	-1.275	-1380.2	0.005	-1380.2	-0.009
2	-1303.4	-1304.9	-1.539	-1303.4	0.020	-1303.4	-0.006
3	-1643.1	-1644.6	-1.515	-1643.1	0.014	-1643.1	0.016
4	-1872.3	-1873.8	-1.494	-1872.3	0.012	-1872.3	0.004
5	-1020.1	-1021.3	-1.179	-1020.1	0.003	-1020.1	0.015
6	-1380.2	-1381.5	-1.337	-1380.2	0.012	-1380.2	0.003
7	-1123.3	-1124.6	-1.305	-1123.3	0.010	-1123.4	-0.033
8	-1014.3	-1015.7	-1.335	-1014.3	0.003	-1014.3	0.018
9	-969.6	-971.0	-1.359	-969.6	0.013	-969.6	0.030
10	-1313.6	-1315.0	-1.401	-1313.6	0.012	-1313.6	-0.001
11	-1848.0	-1849.5	-1.447	-1848.0	0.005	-1848.0	-0.005
12	-1507.0	-1508.3	-1.296	-1507.0	0.004	-1507.0	0.000
13	-1378.5	-1379.8	-1.319	-1378.5	0.005	-1378.5	0.022
Mean |Bias|			1.369		0.010		0.012

Table 4.2: Log-marginal likelihood ($\log -ML$) estimates (mean) and Bias (deviation from the truth) for the Normal model with conjugate prior Normal-Inverse-Wishart (M_1).

For model M_4 (Table 4.3), which extends M_1 by incorporating letter-specific effects, all three estimators exhibit larger absolute biases compared to M_1, as expected given the higher-dimensional parameter space. The Laplace-Metropolis absolute bias approximately doubles to a mean of 2.510 log units, reflecting the increased effect of Gaussian approximation to the posterior $\boldsymbol{W}$ parameters. The Generalized Harmonic Mean estimator performs a mean bias of 0.135 log units, with systematic underestimation. Bridge Sampling remains robust, with a mean absolute bias of only 0.021 log units across all writers, demonstrating that it retains its accuracy even in the more complex hierarchical setting of M_4.

In summary, these results provide empirical justification for the usage of Bridge Sampling as the primary Monte Carlo estimator for the log-marginal likelihood throughout this study.

Writer	Truth	M_4 **MANOVA Conjugate Normal-Inverse-Wishart**					
		Laplace Metropolis		**Generalized Harmonic Mean**		**Bridge Sampling**	
		log-ML	**Bias**	**log-ML**	**Bias**	**log-ML**	**Bias**
1	-1359.6	-1362.0	-2.393	-1359.7	-0.108	-1359.6	0.004
2	-1128.7	-1131.2	-2.496	-1128.8	-0.185	-1128.7	-0.004
3	-1529.6	-1532.2	-2.634	-1529.7	-0.125	-1529.7	-0.064
4	-1779.5	-1782.1	-2.543	-1779.7	-0.173	-1779.5	0.006
5	-814.1	-816.5	-2.408	-814.2	-0.092	-814.1	-0.006
6	-1263.9	-1266.4	-2.508	-1264.1	-0.184	-1264.0	-0.035
7	-966.6	-969.0	-2.415	-966.7	-0.130	-966.6	-0.065
8	-881.5	-884.0	-2.478	-881.6	-0.073	-881.5	-0.008
9	-812.2	-814.9	-2.645	-812.4	-0.153	-812.2	0.000
10	-1260.0	-1262.5	-2.541	-1260.1	-0.128	-1260.0	-0.024
11	-1632.5	-1634.9	-2.466	-1632.6	-0.110	-1632.4	0.040
12	-1310.3	-1312.9	-2.639	-1310.4	-0.126	-1310.3	-0.011
13	-1223.1	-1225.5	-2.456	-1223.2	-0.168	-1223.1	0.002
Mean |Bias|			2.510		0.135		0.021

Table 4.3: Log-marginal likelihood ($\log -ML$) estimates (mean) and Bias (deviation from the truth) for the MANOVA model with conjugate prior Normal-Inverse-Wishart (M_4).

Normal-Inverse-Wishart prior

For the Normal-Inverse-Wishart prior and for the Generalized Harmonic Mean and Bridge Sampling estimators, we consider as proposal density (importance density) the Multivariate Normal for the mean posterior parameters and Inverse-Wishart for the posterior covariance matrix parameter, for more details see 3.2.3. The mean of the log-marginal likelihood estimates and the standard deviation of them (MCSE) are reported in Tables 4.4 and 4.5, for the Bayesian Normal model with the hierarchical extension of Normal-Inverse-Wishart (NIW) prior (M_2) and MANOVA with hierarchical extension of Normal-Inverse-Wishart prior (M_5), respectively.

Writer	M_2 **Normal-Inverse-Wishart**					
	Laplace Metropolis		**Generalized Harmonic Mean**		**Bridge Sampling**	
	log-ML	**MCSE**	**log-ML**	**MCSE**	**log-ML**	**MCSE**
1	-1380.8	0.092	-1380.3	0.015	-1380.3	0.014
2	-1298.8	0.096	-1298.4	0.017	-1298.4	0.014
3	-1639.9	0.099	-1639.4	0.018	-1639.4	0.013
4	-1870.3	0.103	-1869.8	0.017	-1869.8	0.011
5	-1022.4	0.101	-1021.9	0.015	-1021.9	0.014
6	-1378.7	0.087	-1378.2	0.014	-1378.2	0.017
7	-1121.3	0.107	-1120.8	0.039	-1120.8	0.012
8	-1011.1	0.103	-1010.6	0.014	-1010.6	0.013
9	-969.9	0.084	-969.5	0.015	-969.5	0.010
10	-1320.4	0.088	-1319.9	0.021	-1319.9	0.012
11	-1851.9	0.074	-1851.4	0.015	-1851.4	0.010
12	-1501.9	0.128	-1501.4	0.018	-1501.4	0.014
13	-1366.5	0.117	-1366.1	0.014	-1366.1	0.011

Table 4.4: Log-marginal likelihood ($\log -ML$) estimates (mean) and Monte Carlo standard errors (standard deviation) for the Normal model with hierarchical prior Normal-Inverse-Wishart (M_2).

Writer	M_5 MANOVA Normal-Inverse-Wishart					
	Laplace Metropolis		Generalized Harmonic Mean		Bridge Sampling	
	log-ML	MCSE	log-ML	MCSE	log-ML	MCSE
1	-1361.8	0.135	-1361.3	0.047	-1361.3	0.028
2	-1123.9	0.140	-1125.1	0.050	-1125.1	0.031
3	-1523.0	0.147	-1523.8	0.123	-1523.8	0.031
4	-1777.5	0.119	-1776.8	0.065	-1776.8	0.028
5	-818.3	0.118	-818.3	0.052	-818.3	0.028
6	-1266.6	0.145	-1266.4	0.070	-1266.4	0.028
7	-962.5	0.164	-963.3	0.059	-963.3	0.027
8	-878.0	0.171	-878.8	0.067	-878.8	0.033
9	-816.2	0.144	-817.0	0.073	-817.0	0.038
10	-1274.3	0.142	-1274.7	0.091	-1274.7	0.035
11	-1665.4	0.111	-1666.5	0.078	-1666.5	0.037
12	-1307.7	0.163	-1308.4	0.053	-1308.4	0.028
13	-1205.6	0.144	-1207.1	0.054	-1207.1	0.037

Table 4.5: Log-marginal likelihood ($\log - ML$) estimates and Monte Carlo standard errors for the MANOVA model with independent prior Normal-Inverse-Wishart (M_5).

Two main results can be observed from Tables 4.4 and 4.5. First, for all writers and both models, the marginal likelihood estimates are rather close, particularly when the Generalized Harmonic Mean and Bridge Sampling are used. Secondarily, the Bridge Sampling estimator systematically produces smaller MCSEs as expected according to scientific literature (Sinharay and Stern, 2005; Ardia et al., 2012).

Specifically, the mean difference between the Generalized Harmonic Mean and the Bridge Sampling estimates of log-ML for the NIW is small (0.006 and 0.0003 on average across all writers $\pm$ 0.003). On the other hand, the log-ML of each of these methods is found to be 0.5 units (on average $\pm$ 0.03) away from the corresponding estimate of the Laplace-Metropolis method. When comparing MCSEs, the Bridge Sampling estimator slightly outperforms the Generalized Harmonic Mean estimator, reducing the variability by 25% (on average across writers $\pm$ 21%). Furthermore, the Bridge Sampling estimator is clearly more accurate than the Laplace-Metropolis estimator, as its MCSEs is on average lower 87% ($\pm$ 2.5%) than the latter.

For the MANOVA model, again differences between the log-ML of the Generalized Harmonic Mean and the Bridge Samplings are minimal (average difference across writers equal to 0.0015 $\pm$ 0.017), while larger differences were observed between these two methods and Laplace-Metropolis estimates (average difference across writers equal to 0.52 $\pm$ 0.67). As far as MCSEs are concerned, the Bridge Sampling estimator stands out, with 77.5% (on average across all writers $\pm$ 4.5%) less variability than the Laplace-Metropolis approach and about 51% ($\pm$ 11%) less than the Generalized Harmonic Mean estimator.

Finally, based on the presented comparative analysis of different methods, in Section 4.7.1, we employ Bridge Sampling as the primary estimator for the log-marginal likelihoods of models M_2 and M_5.

Normal-LogNormal-LKJ prior

For the Normal-LogNormal-LKJ prior and for the Generalized Harmonic Mean and Bridge Sampling estimators we consider as proposal density (importance density) the Multivariate Normal for all considered parameters, we didn't proceed with further investigation in this case to find better proposal

for posterior covariance matrix, for more details see 3.2.3. The mean of the log-marginal likelihood estimates and the standard deviation of the MCSE are reported in Tables 4.6 and 4.7.

Writer	M_3 Normal-LogNormal-LKJ					
	Laplace Metropolis		Generalized Harmonic Mean		Bridge Sampling	
	log-ML	MCSE	log-ML	MCSE	log-ML	MCSE
1	-1362.4	0.345	-1363.6	0.530	-1362.8	0.098
2	-1270.1	0.232	-1271.4	0.466	-1270.3	0.098
3	-1604.6	0.285	-1605.9	0.374	-1604.5	0.112
4	-1821.2	0.279	-1822.3	0.536	-1821.1	0.118
5	-1026.0	0.255	-1027.5	0.442	-1027.4	0.072
6	-1359.2	0.251	-1360.2	0.471	-1359.6	0.078
7	-1115.5	0.314	-1116.9	0.499	-1116.6	0.109
8	-1006.8	0.239	-1008.0	0.534	-1007.8	0.106
9	-966.0	0.243	-967.4	0.392	-966.4	0.114
10	-1295.9	0.240	-1297.2	0.440	-1296.1	0.058
11	-1813.6	0.279	-1814.9	0.368	-1813.8	0.091
12	-1485.6	0.251	-1486.6	0.376	-1485.9	0.107
13	-1349.2	0.248	-1350.5	0.748	-1349.6	0.090

Table 4.6: Log-marginal likelihood ($\log -ML$) estimates (mean) and Monte Carlo standard errors (standard deviation) for the hierarchical prior Normal-LogNormal-LKJ model (M_3).

Writer	M_6 MANOVA Normal-LogNormal-LKJ					
	Laplace Metropolis		Generalized Harmonic Mean		Bridge Sampling	
	log-ML	MCSE	log-ML	MCSE	log-ML	MCSE
1	-1346.7	0.390	-1346.2	0.863	-1346.2	0.176
2	-1106.5	0.283	-1106.3	1.608	-1105.7	0.165
3	-1492.7	0.369	-1493.1	1.509	-1491.8	0.177
4	-1732.5	0.291	-1732.0	1.105	-1731.5	0.125
5	-823.8	0.240	-823.3	1.267	-824.0	0.113
6	-1247.2	0.352	-1246.6	1.183	-1246.4	0.201
7	-964.2	0.279	-963.6	0.868	-964.3	0.157
8	-879.2	0.344	-878.5	1.077	-879.2	0.160
9	-821.6	0.398	-821.5	1.159	-821.1	0.200
10	-1254.8	0.344	-1254.1	1.372	-1254.1	0.190
11	-1633.4	0.342	-1633.0	1.527	-1632.7	0.180
12	-1289.4	0.339	-1288.7	1.254	-1288.5	0.149
13	-1196.0	0.234	-1195.4	1.388	-1195.3	0.198

Table 4.7: Log-marginal likelihood ($\log -ML$) estimates and Monte Carlo standard errors for the MANOVA model with independent prior Normal-LogNormal-LKJ (M_6).

Three key observations can be made from Tables 4.6 and 4.7. First, the marginal likelihood estimates are generally close across methods. However, under this prior setup, the Laplace-Metropolis and Generalized Harmonic Mean estimators exhibit substantially higher Monte Carlo standard errors (MCSEs). Finally, as expected from previous studies (Sinharay and Stern, 2005; Ardia et al., 2012), the Bridge Sampling estimator consistently yields lower MCSEs, demonstrating its greater precision.

Specifically, for the Normal-LogNormal-LKJ model (M_3) the mean absolute difference in log-ML between the Generalized Harmonic Mean and Bridge Sampling estimates is approximately 0.81 with a standard deviation of about 0.41 across writers. The Laplace-Metropolis estimates differ less substantially from the Bridge Sampling estimates, averaging about 0.48 units difference ($\pm$ 0.42) in log-ML.

Regarding Monte Carlo standard errors, the Bridge Sampling estimator reduces variability by about 79% on average compared to the Generalized Harmonic Mean estimator, and by roughly 64% compared to the Laplace-Metropolis estimator.

For the MANOVA model (M_6) the differences in log-ML between the Generalized Harmonic Mean and Bridge Sampling estimates are averaging around 0.43 ($\pm$ 0.37). Both these methods differ more noticeably from the Laplace-Metropolis estimates, with average differences around 0.60 ($\pm$ 0.32) log-ML units. However, the Bridge Sampling estimator achieves substantial reduction in MCSE compared to others: about 86% less variability than the Generalized Harmonic Mean method and approximately 47% less than the Laplace-Metropolis estimator, averaged across writers.

Finally, based on the presented comparative analysis of different methods, in Section 4.7.1, we employ Bridge Sampling as the primary estimator for the log-marginal likelihoods of models M_3 and M_6.

4.6.3 Marginal Likelihoods for Handwriting Evidence Evaluation

The marginal likelihoods (ML) needed to assess the Bayes factor in Eq. (4.1) represent the probability of the observed data $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) = (\boldsymbol{y}_{w\ell j\bullet},\ w = 1, 2,\ \ell = 1, \ldots, L,\ j = 1, \ldots, n_{w\ell})$ under the competing propositions H_1 and H_2, for a specific probabilistic model M. Here, by M there are considered any of the six Bayesian models $\{M_1, M_2, M_3, M_4, M_5, M_6\}$ introduced in Section 4.5; see Table 4.1 for details.

For the numerator of Eq. (4.1) the marginal likelihood under the hypothesis H_1 is expressed:

$$m(\boldsymbol{y}_1, \boldsymbol{y}_2|H_1, M_l) = \int f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{\Theta}, \boldsymbol{W}, H_1, M_l)\pi(\boldsymbol{\Theta}, \boldsymbol{W}|H_1, M_l)d(\boldsymbol{\Theta}, \boldsymbol{W}). \tag{4.15}$$

For the denominator of Eq. (4.1) under a hypothesis H_2 :

$$m(\boldsymbol{y}_w|H_2, M_l) = \int f(\boldsymbol{y}_w|\boldsymbol{\Theta}, \boldsymbol{W}, H_2, M_l)\pi(\boldsymbol{\Theta}, \boldsymbol{W}|H_2, M_l)d(\boldsymbol{\Theta}, \boldsymbol{W}),\ \text{for } w \in \{1, 2\}, \tag{4.16}$$

and any model M_l (for $l = 1, 2, 3, 4, 5, 6$) and model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{W}$. If hypothesis H_1 holds (i.e., the PoI is the writer of the questioned manuscript), model parameters are assumed equal, that is $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\Theta}$ and $\boldsymbol{W}_1 = \boldsymbol{W}_2 = \boldsymbol{W}$. On the other hand, if hypothesis H_2 holds (i.e., the person of interest is not the writer of the questioned document), then the model parameters ($\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{W}_1, \boldsymbol{W}_2$) are assumed to be different across the two different sources $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$.

In our model formulations, presented in Sections 4.5.1 and 4.5.2, the parameter vector $\boldsymbol{\Theta}$, under H_2, is denoted as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for models M_1, M_2, M_3, and by $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ for models M_4, M_5, M_6, where $\boldsymbol{\Theta}_w$ is specified as in (4.10), $w = 1, 2$. Furthermore, M_1 and M_4 yield analytical marginal likelihoods under conjugate priors, those for M_2, M_3, M_5, and M_6 were approximated using bridge sampling ("Warp-III") via the *bridgesampling* *R* package (Gronau, Singmann and Wagenmakers, 2020).

4.7 Experimental Results

Following the handwriting examination problem presented in the beginning of Chapter 4, a case scenario involving a document whose origin is contested is analyzed in this section. These models arise from two likelihood structures and three different prior specifications (hence six models in total; see Table 4.1) described in Section 4.5 were compared using the dataset described in Section 4.4. The considered model approaches are the conjugate and non-conjugate versions of the Bayesian Normal

with Normal-Inverse-Wishart or Normal-LogNormal-LKJ prior setup (M_1, M_2, and M_3, respectively), and of the Bayesian MANOVA with Normal-Inverse-Wishart or Normal-LogNormal-LKJ prior setup models (M_4, M_5, M_6, respectively).

First, in Section 4.7.1, we compare $M_1, \ldots, M_6$ for each individual writer from the full available data $\boldsymbol{\mathcal{D}}_i$ of each writer $i = 1, \ldots, m$. For each writer m, this procedure enables pairwise comparisons among the models. Furthermore, when eliciting the prior parameters for the considered models, the data from all other writers (i.e., excluding the writer under evaluation) are used. Specifically, when comparing models for writer i, the data from the remaining $m - 1$ writers are utilized for prior elicitation.

Subsequently, in Section 4.7.2, the performance of the considered models is evaluated within the Bayes factor framework, as outlined in the beginning of Chapter 4. Specifically, we conduct a series of experiments designed to simulate a real-world case study. For the same-writer experiments, each writer i from the full dataset $\boldsymbol{\mathcal{D}}$ is selected, and their data are randomly divided into questioned and control sets. The Bayes factor in Eq. 4.1 is then assessed. The data from all other writers $(m - 1)$ are used as background data $\boldsymbol{X}$ to elicit the prior parameters of the models under consideration. For the different-writer experiments, data are randomly selected from two different writers from the full dataset $\boldsymbol{\mathcal{D}}$, and the same procedure as described in the same-writer experiments is followed.

4.7.1 Model Comparisons per Writer

In this section, a model comparison is performed between the six models under consideration in Table 4.1 described in Section 4.5. For each writer i, the model M_l is fitted to the data $\boldsymbol{\mathcal{D}}_i$ characterizing this writer, and the prior parameters are fitted using data from the remaining writers (excluding the i writer from the elicitation of the prior parameters). The marginal likelihoods of all models under consideration for each writer i are then estimated. The Bayes factor for comparing two models M_l and M_ξ for writer i is given by

$$\mathbf{BF}_{l,\xi}(\boldsymbol{\mathcal{D}}_i) = \frac{m(\boldsymbol{\mathcal{D}}_i \mid M_l)}{m(\boldsymbol{\mathcal{D}}_i \mid M_\xi)} \text{ for } l, \xi \in \{1, 2, 3, 4, 5, 6\},\ l \neq \xi.$$

Bayesian MANOVA vs. Normal Model. First, the Bayesian MANOVA model is compared with the Bayesian Normal model, both in the conjugate (M_4 vs. M_1) and in the non-conjugate approaches (M_5 vs. M_2), (M_6 vs. M_3) versions. Hence, for each writer, the Bayes factors $\text{BF}_{4,1}$, $\text{BF}_{5,2}$ and $\text{BF}_{6,3}$ are calculated, respectively. Note that while the marginal likelihoods that are needed to calculate $\text{BF}_{4,1}$ can be obtained analytically, bridge sampling described in Section 4.6 is implemented for the estimation of $\text{BF}_{5,2}$ and $\text{BF}_{6,3}$. Table 4.8 presents the results (in log scale) for $\text{BF}_{4,1}$ (second column), $\text{BF}_{5,2}$ (third column) and $\text{BF}_{6,3}$ (forth column).

From these results, it can be observed that $\log \text{BF}_{4,1}$, $\log \text{BF}_{5,2}$ and $\log \text{BF}_{6,3}$ are markedly higher than the value of five. These high BF values suggest very strong evidence see Table 2.1 in favor of the MANOVA formulation. Thus, with reference to the available dataset, it can be claimed that the Bayesian MANOVA with characters as predictors fits the data per writer better for any of the prior setups. Same results from the IAM dataset with 50 writers, with 44 out of 50 supporting the Bayesian MANOVA formulation.

Conjugate vs. non-Conjugate. Next, attention is given to the comparison between the conjugate and non-conjugate versions of each model, specifically considering Normal-Inverse-Wishart hierarchical priors and the Normal-LogNormal-LKJ prior vs. conjugate versions. Bayes factors $\text{BF}_{1,2}$, $\text{BF}_{1,3}$, $\text{BF}_{2,3}$, $\text{BF}_{4,5}$, $\text{BF}_{4,6}$ and $\text{BF}_{5,6}$ are then calculated to compare the two-level random effects

Writer	Normal-Inverse-Wishart		Normal-LogNormal-LKJ
	Conjugate Approach $\log BF^{*}_{4,1}$	**Hierarchical Approach** $\log BF^{**}_{5,2}$	**Covariance Decomposition** $\log BF^{**}_{6,3}$
1	20.6	19.1	16.6
2	174.7	173.4	164.5
3	113.5	115.7	89.7
4	92.8	93.1	78.8
5	205.9	203.8	202.9
6	116.3	111.7	114.2
7	156.8	157.3	152.1
8	132.8	131.8	128.8
9	157.4	152.6	144.9
10	53.6	45.3	42.3
11	215.6	184.9	181.6
12	196.7	193.0	198.3
13	155.4	159.0	154.6

** Closed form expression; ** Bridge Sampling estimate*

Table 4.8: Logarithmic Bayes factors of MANOVA models (M_3 or M_4 or M_6) vs. the Normal models (M_1 or M_2 or M_5).

model and the MANOVA model with different prior setups (conjugate, Normal-Inverse-Wishart hierarchical prior, Normal-LogNormal-LKJ prior). As already pointed out in Section 4.5, handwriting literature emphasises the importance of modeling both within and between writers' variability. The available dataset will therefore be used to assess whether the results are consistent with this theory. This is achieved by considering the comparison between the conjugate and non-conjugate versions of the Normal and MANOVA models with a Normal-Inverse-Wishart prior. Table 4.9 presents the assessed Bayes factors per writer. According to these results, for the Normal model, the Bayes factors provide evidence supporting the non-conjugate prior (and thus, the need to also account for the between-writers variability) for 9 out of 13 writers. On the other hand, for the MANOVA model with Normal-Inverse-Wishart of conjugate and non-conjugate prior setup, the Bayes factors do not show a clear preference for the non-conjugate prior, which is only supported for about half of the writers (seven out of 13). However, similar results were observed with the IAM dataset of 50 writers: 44 out of 50 supported the non-conjugate Normal-Inverse-Wishart prior, while for the MANOVA model, the results were more balanced, with 29 out of 50 supporting the conjugate MANOVA formulation. Hence, for the MANOVA model, following the principle of parsimony, we can proceed with the conjugate MANOVA model considering the Inverse–Wishart prior.

Inverse-Wishart vs. LogNormal-LKJ. Finally, we compare models M_1, M_2, M_4, M_5 (based on the Normal-Inverse-Wishart prior) with models M_3 and M_6 (their corresponding Normal-LogNormal-LKJ prior counterparts). Table 4.10 presents the assessed Bayes factors $\text{BF}_{2,3}$ and $\text{BF}_{5,6}$ for each writer. Based on the results of this table, the Bayes factors support the Normal-LogNormal-LKJ prior over the corresponding Normal-Inverse-Wishart prior for 9 out of 13 writers under the Bayesian MANOVA. The evidence is even more systematic for the Bayesian Normal hierarchical models, where the Normal-LogNormal-LKJ prior is favored for almost all writers (12 out of 13). Similar results are presented for conjugate Normal-Inverse-Wishart and Normal-LogNormal-LKJ prior in Table 4.11 for the estimated Bayes factors $\text{BF}_{1,3}$ and $\text{BF}_{4,6}$ per writer. Finally, the same results were extracted from the IAM dataset, with 42 out of 50 supporting the Normal-LogNormal-LKJ prior.

Writer	Normal-Inverse-Wishart					
	Normal $\log BF^*_{1,2}$			MANOVA $\log BF^*_{4,5}$		
	Value	Sign	Interpretation	Value	Sign	Interpretation
1	0.14	+	Bare Mention	1.60	+	Substantial
2	-4.92	-	Strong	-3.58	-	Strong
3	-3.77	-	Strong	-6.02	-	Extreme
4	-2.44	-	Substantial	-2.72	-	Substantial
5	1.85	+	Substantial	3.99	+	Strong
6	-2.13	-	Substantial	2.44	+	Substantial
7	-2.52	-	Substantial	-3.05	-	Strong
8	-3.68	-	Strong	-2.66	-	Substantial
9	-0.17	-	Bare Mention	4.69	+	Strong
10	6.30	+	Extreme	14.62	+	Extreme
11	3.33	+	Strong	34.05	+	Extreme
12	-5.45	-	Extreme	-1.72	-	Substantial
13	-12.54	-	Extreme	-16.09	-	Extreme

** Bridge Sampling estimate*

Table 4.9: Logarithmic Bayes factors (per writer) comparing the conjugate and non-conjugate approaches of the Normal-Inverse-Wishart prior for the Bayesian Normal and MANOVA models.

Writer	Normal Inverse-Wishart vs LogNormal-LKJ $\log BF^*_{2,3}$			MANOVA Inverse-Wishart vs LogNormal-LKJ $\log BF^*_{5,6}$		
	Value	Sign	Interpretation	Value	Sign	Interpretation
1	-17.65	-	Extreme	-15.10	-	Extreme
2	-28.13	-	Extreme	-19.47	-	Extreme
3	-34.46	-	Extreme	-31.75	-	Extreme
4	-48.50	-	Extreme	-45.21	-	Extreme
5	5.39	+	Extreme	6.34	+	Extreme
6	-17.67	-	Extreme	-20.14	-	Extreme
7	-3.80	-	Strong	1.47	+	Substantial
8	-2.00	-	Substantial	1.00	+	Bare mention
9	-2.94	-	Substantial	4.69	+	Strong
10	-23.91	-	Extreme	-20.98	-	Extreme
11	-37.50	-	Extreme	-34.19	-	Extreme
12	-13.76	-	Extreme	-19.12	-	Extreme
13	-15.49	-	Extreme	-11.08	-	Extreme

** Bridge Sampling estimate*

Table 4.10: Logarithmic Bayes factors (per writer) comparing the within-writer covariance prior setups (a) Inverse-Wishart or (b) LogNormal-LKJ prior of Bayesian Normal and MANOVA models.

4.7.2 Accuracy of Models

In this section, the performance of the models under consideration is measured by means of simulation studies where the propositions of interest are whether a given individual is the writer of a questioned manuscript versus the alternative proposition that the writer is an unknown individual (see beginning of Chapter 4). To this end, the models will be compared to study their ability to deliver Bayes factor values that support the correct proposition.

To assess the performance of the proposed models whenever proposition H_1 (same writer case) holds, a variety of simulated case studies were generated considering character measures from the same writer as questioned $\boldsymbol{y}_1$ and control $\boldsymbol{y}_2$ data, respectively. Therefore, the data of each writer from the available database was divided into two parts, one used as the questioned data and the other as control data. The proportion π_{split} of data used as questioned material was randomly taken between

Writer	Normal Inverse-Wishart vs LogNormal-LKJ $\log BF^*_{1,3}$			MANOVA Inverse-Wishart vs LogNormal-LKJ $\log BF^*_{4,6}$		
	Value	**Sign**	**Interpretation**	**Value**	**Sign**	**Interpretation**
1	-17.51	-	Extreme	-13.51	-	Extreme
2	-33.05	-	Extreme	-23.05	-	Extreme
3	-38.23	-	Extreme	-37.76	-	Extreme
4	-50.95	-	Extreme	-47.93	-	Extreme
5	7.24	+	Extreme	10.33	+	Extreme
6	-19.80	-	Extreme	-17.71	-	Extreme
7	-6.31	-	Extreme	-1.58	-	Substantial
8	-5.68	-	Extreme	-1.66	-	Substantial
9	-3.11	-	Strong	9.39	+	Extreme
10	-17.61	-	Extreme	-6.36	-	Extreme
11	-34.16	-	Extreme	-0.15	-	Bare Mention
12	-19.21	-	Strong	-20.85	-	Extreme
13	-28.03	-	Extreme	-27.17	-	Extreme

**Bridge Sampling estimate*

Table 4.11: Logarithmic Bayes factors (per writer) comparing the conjugate Normal-Inverse-Wishart and Normal-LogNormal-LKJ prior setups for the Bayesian Normal and MANOVA models.

0.35 and 0.65, $\pi_{split} \in (0.35, 0.65)$, while the remaining data was taken as control material. This process was repeated 100 times (using different random splits) for each of the 13 available writers, resulting in a total number of 1300 same-writer comparisons. The remaining writers serve as background data $\boldsymbol{X}$ to elicit the prior parameters for the models, namely, we use data from all writers except the one being analyzed. This made it possible to assess the false negative rate, that is, the percentage of cases giving rise to a Bayes factor less than 1 when it should be greater than 1.

A similar procedure is followed when comparing material from different writers, i.e., H_2 is true. For each pair of writers, case studies were generated by treating the measurements of the first writer as questioned data $\boldsymbol{y}_1$ and those of the second writer as control data $\boldsymbol{y}_2$. To account for the effect of sample size, the same approach as in the simulated procedure for H_1 was followed. Hence, only a random subset of the data from each writer, in each pair, was considered. Specifically, a proportion $\pi_{split} \in (0.35, 0.65)$ of the measurements was randomly selected from the first individual of each pair (i.e., the writer of the questioned document), while the proportion of $(1 - \pi_{split})$ of the measurements of the second individual of the pair was considered as the control data. This process was repeated 100 times for each of the 78 pairs of writers, resulting in a total number of 7800 different comparisons. The remaining writers serve as background data $\boldsymbol{X}$ to elicit the prior parameters for the models, namely, we use data from all writers except the two that are being analyzed. In these comparisons, the objective was to assess the false positive rate, defined as the percentage of cases in which the Bayes factor was greater than 1 when it should have been less than 1.

The logarithm of the Bayes factors (4.1), over all simulated case-scenarios, obtained for testing competing hypotheses H_1 and H_2 (see Section 4.6.3) is provided in Figures 4.12a and 4.12b for same-writer and different-writers comparisons, respectively. In these figures, the first five panels (from the left) provide the Bayes factors for the Bayesian Normal models (M_1, M_2, M_3) for characters *a*, *d*, *o*, and *q* and for all characters together. The last panel provides the Bayes factor for the MANOVA models (M_4, M_5, M_6) using all characters together. In each panel, the BFs obtained analytically using the conjugate prior are depicted in the first boxplot (from left), while Boxplots 2 and 3 represent the BFs obtained using the hierarchical Normal-Inverse-Wishart and Normal-LogNormal-LKJ prior, where the marginal likelihoods are estimated by the bridge sampling described in Section 4.6.

Looking at these box-plots, it is evident that almost all Bayes factors, regardless of the statistical modeling approach, support the correct hypothesis. Specifically, (almost) all the Bayes factors (in logarithmic scale) are well placed above zero for the first case-scenario (same writer) and below zero in most comparisons for the second case-scenario (different writers). However, for different-writer comparisons, the boxplots in Figure 4.12b exhibit greater width, as there are cases in which the handwriting styles of two writers are markedly different from one another. Furthermore, in the case of the same writers (i.e., under H_1), we observe only one false negative case for character q under NIW conjugate and the NIW hierarchical approaches, and for character d under the Normal-LogNormal-LKJ prior, while no false negatives occurred for the remaining characters and approaches.

As far as the assessment of the performance of the different methods under hypothesis H_2, the number of false positives over the total number of 7800 different writers comparisons is presented in Table 4.12, for both modeling approaches and different prior setups. Specifically, the total number of false positives ranges from a minimum of 47 to a maximum of 589 over a total number of 7800 different writer comparisons (i.e., 0.6% – 7.6%).

For individual characters, the highest false positive rates (up to 7.6% for NIW Conjugate approach) were observed for character a, while character q consistently produces the lowest rates. When considering all characters, the false positive rate drops further from 2.2%, which is the lowest false positive rate for the character q (the best performing character), to 0.9% under the Normal-LogNormal-LKJ prior setup.

Concerning the prior setups, the NIW Conjugate and NIW Hierarchical models produced similar false positive rates for most characters, although the Hierarchical version generally performs slightly better. The Normal-LogNormal-LKJ setup resulted in lower false positive rates for characters a and o, but higher rates for characters d and q compared to the NIW approaches.

The MANOVA model exhibits the smallest proportion of false positives among all methods, with rates consistently below 0.7%. This indicates superior performance for MANOVA in cross-writer discrimination when all characters are considered together. Furthermore, only minor differences were observed among the different prior setups within the MANOVA formulation.

Model	**Characters**	**Normal-Inverse-Wishart**		**Normal-LogNormal-LKJ**
		Conjugate	**Hierarchical**	
Normal	a	589 (7.6%)	560 (7.2%)	545 (7.0%)
	d	491 (6.3%)	308 (3.9%)	589 (7.6%)
	o	410 (5.3%)	363 (4.7%)	301 (3.9%)
	q	213 (2.7%)	168 (2.2%)	265 (3.4%)
	all	113 (1.4%)	86 (1.1%)	68 (0.9%)
MANOVA	all	53 (0.7%)	51 (0.7%)	47 (0.6%)

Table 4.12: Number of false positives over 7800 different writer comparisons (percentages in brackets).

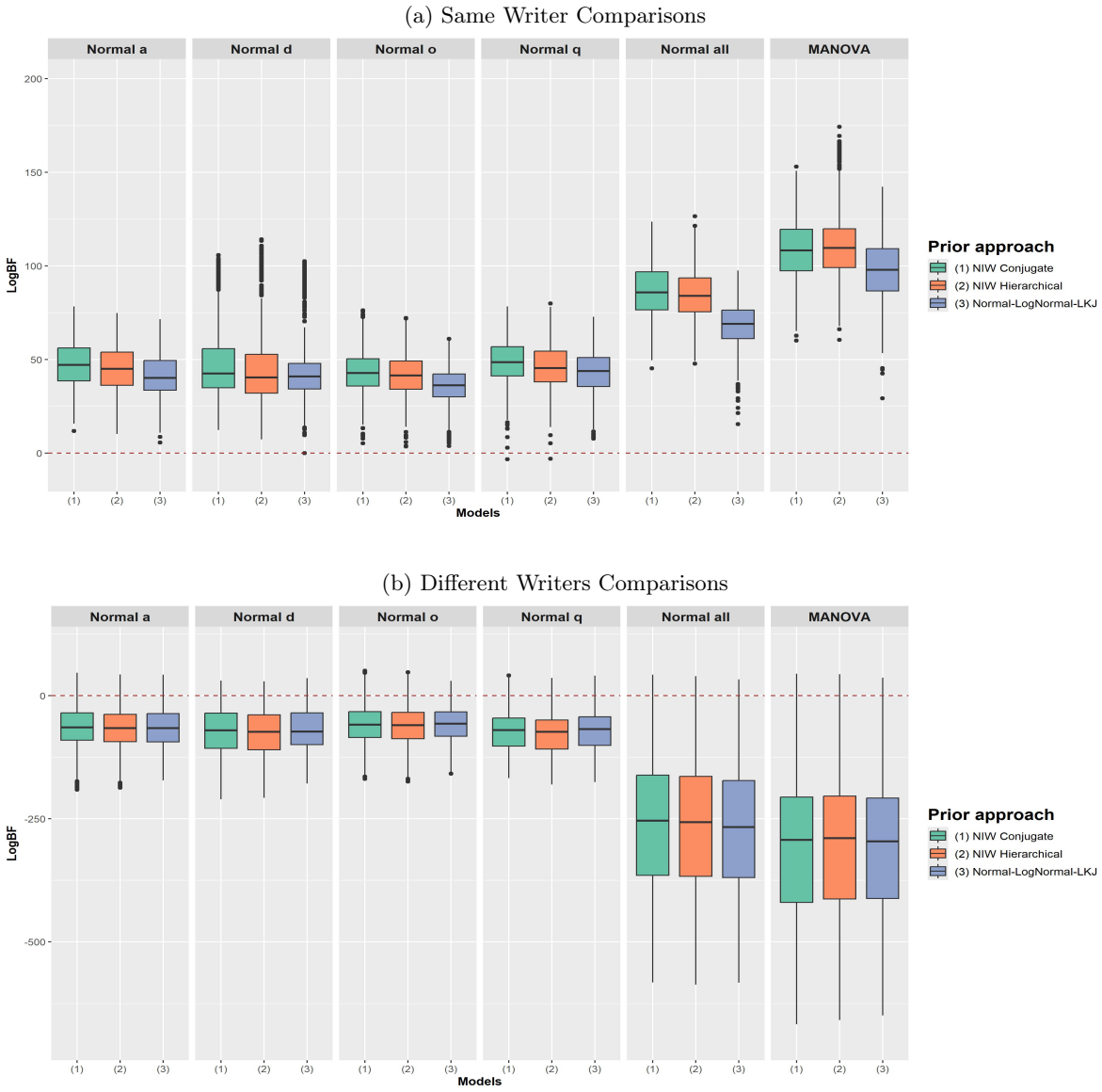


Figure 4.12: Logarithmic Bayes factors for handwriting evaluation ($\log \mathrm{BF}$) for same writer (a) and different writers comparisons (b), using the data modeling approaches described in Sections 4.5.1 and 4.5.2.

4.8 Sensitivity Analysis of Prior Elicitation

In Bayesian hierarchical modeling, prior elicitation is critical, as it can have a substantial impact on the resulting Bayes factor values. In this study, we implement a subjective prior methodology in which prior distributions are informed by empirical data derived from handwriting samples of a group of writers, different from the ones being compared. This approach explicitly incorporates prior knowledge regarding the characteristics of the handwriting features based on Fourier analysis used in the study. Such a strategy is particularly advantageous in our case, where we wish to determine whether the available handwriting evidence originates from the same writer or from a different writer.

To address the influence of this subjective prior framework, in Section 4.8.1, we implement a sensitivity analysis of the prior elicitation framework. Following the evaluation procedure described

in Section 4.7.2, namely the data of a writer or pair of writers from the full dataset $\boldsymbol{\mathcal{D}}$ are randomly selected and serve as the questioned data $\boldsymbol{y}_1$ and control data $\boldsymbol{y}_2$, while the remaining writers are considered as the background data $\boldsymbol{X}$. The sensitivity analysis is conducted by constructing the prior distributions from a random subsample of each background writer's data rather than from the entire background dataset $\boldsymbol{X}$. Specifically, for each background writer, a random subsample with replacement comprising 50% of their data is considered.

Furthermore, for the degrees of freedom of the Wishart prior in M_{1-2} and M_{4-5}, we set $\nu = p + 2$ as the default low-information value, while for the LKJ prior in M_3 and M_6, we use $\eta = 1$. Hence, in the second part of the sensitivity analysis in Sections 4.8.2 and 4.8.3, we examine how different prior values for these variance–covariance parameters affect the posterior Bayesian evidence, as measured by the Bayes factor.

4.8.1 Subsampling of Background Data

In this section, we investigate the effect of the prior specification by applying subsampling to the background data used for prior elicitation. The procedure of the randomly selected cases that are used for the evaluation of the two hypotheses (H_1 vs. H_2) is the same as the one described in Section 4.7.2. Specifically, to isolate the effect of prior elicitation, we utilize a single random data split for both the same-writer and different-writer experiments, thereby minimizing potential confounding effects arising from data partitioning. This analysis focuses on the Bayesian MANOVA model, which has been identified as the most effective and recommended approach in Section 4.7.2. To assess the robustness and stability of the Bayes factor and the resulting decision, we performed 30 iterations of random subsampling with replacement, using 50% of the background data for each writer in each evaluation case. To clarify further, three distinct datasets are considered in our analysis: (i) the questioned data $\boldsymbol{y}_1$, (ii) the control data $\boldsymbol{y}_2$, and (iii) the background data $\boldsymbol{X}$, our objective in this section is to evaluate the effect of varying the third dataset.

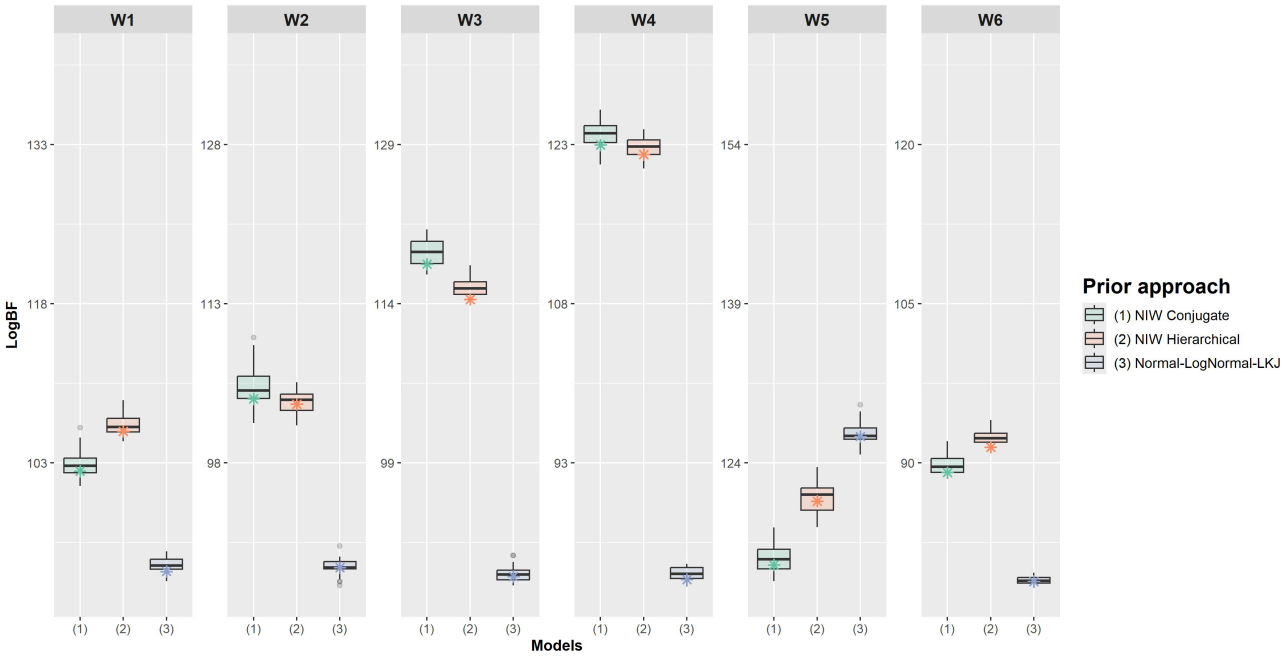
Figures 4.13 and 4.14 present the logarithmic Bayes factors (LogBF) values for handwriting evaluation across 30 different elicited priors, comparing the performance of Bayesian MANOVA models under three prior setups (conjugate, hierarchical Normal-Inverse-Wishart, Normal-LogNormal-LKJ). Figure 4.13 presents the results for same-writer comparisons, while Figure 4.14 shows comparisons between different writers. The asterisks in Figure 4.14 indicate the LogBF values computed using a prior obtained from the complete background data, serving as a reference point for the subsampled distributions. These reference points are generally located near the boxplots, indicating that our proposed approach is relatively robust (both in terms of the Bayes factor values and the resulting decision) to variations in the background data used for prior elicitation, and remain close to the Bayes factor obtained from the full dataset.

Overall, the results indicate that the LogBF values is not affected by the constructed prior elicitation and correctly support the true hypothesis in nearly all cases. Nevertheless, a certain degree of variability is observed, which is expected. For example, comparisons involving different writers (e.g., W8–W13, W7–W8) display notable variability in LogBF values of the conjugate approach, although these do not alter the overall support in favor of H_2. This pattern indicates a certain variability of the Bayes factor to the choice of background data, whereas same-writer comparisons (e.g., for W3, W4) demonstrate substantially greater stability. Further examples can be found in Appendix A.5.1.

To assess the impact of this variability, we examine four writer pairs (W5–W6, W6–W7, W6–W8, and W7–W8) that exhibit the highest similarity and were selected based on having cases with $|\log BF| < 10$. For the remaining writer pairs, the observed variability does not appear to affect the overall level of

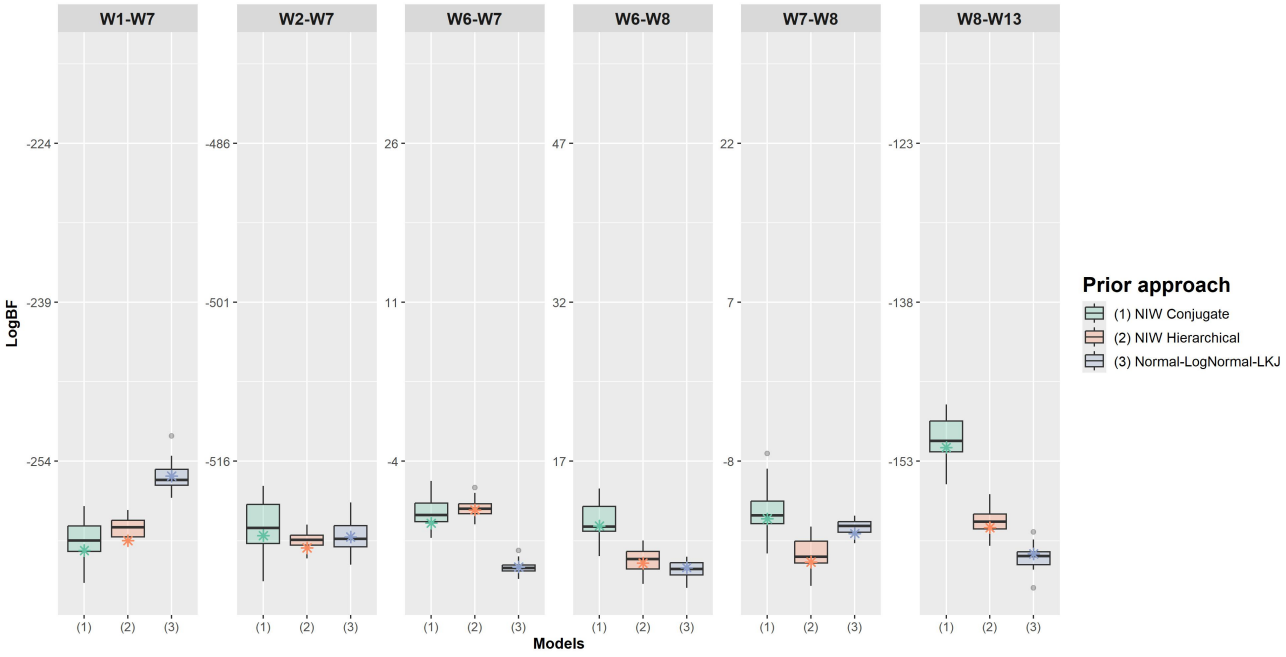
support. To elaborate further, we performed 30 random data selections from each pair of writers, following the procedure described in Section 4.7.2, namely changing the questioned and control datasets. Then, for each pair of writers, subsamples with replacement were drawn, comprising 50% of the data from the remaining writers, which served as the background dataset. In total, this procedure resulted in 3600 comparisons (case studies) for analysis.

Table 4.13 summarizes the sensitivity of support for competing hypotheses under each MANOVA prior specification. The NIW conjugate prior approach presents the highest number of cases with inconsistent support (i.e., support shifts across background data subsamples) among the 30 priors elicited from different background data subsets (9 cases, or 7.5% of 120 cases). For these nine cases, the average LogBF range is found to be 8.51 units, while the widest LogBF subsampling interval (worst case scenario) takes values in the interval $(-6.24, 4.79)$. Conversely, the Normal-LogNormal-LKJ prior setup exhibits fewer cases (2 cases, 1.67%) and smaller average differences, indicating a relatively more stable inference under prior variation. The NIW hierarchical prior shows intermediate behavior between these two setups.



indicates the log BF *using the complete background dataset for each case.*

Figure 4.13: Boxplots of Logarithmic Bayes factors (log BF) for handwriting evaluation for the same writer scenarios over different subsamples of background data for the Bayesian MANOVA approach.



**indicates the* log BF *using the complete background dataset for each case.*

Figure 4.14: Boxplots of Logarithmic Bayes factors (log BF) for handwriting evaluation for the different writers scenarios over different subsamples of background data for the Bayesian MANOVA approach.

Inconsistent Cases*			
		LogBF Summaries	
Prior setup	**Support shifts** (%)**	**Avg. range†**	**Worst case interval‡**
NIW Conjugate	9 (7.50)	8.51	(−6.24, 4.79)
NIW Hierarchical	5 (4.17)	3.83	(−2.69, 1.48)
Normal–LogNormal–LKJ	2 (1.67)	3.42	(−1.49, 3.05)

* *Results are based on 30 random subsets of four compared writer pairs and 30 random 50% subsamples of background writers' data* ($4 \times 30 \times 30 = 3600$).
** *Number of inconsistent-support cases across 30 random subsets of compared writers* ($4 \times 30 = 120$).
† *Average logBF range across inconsistent cases; range is computed over 30 background-data subsamples.*
‡ *LogBF interval* (min, max) *of the case with highest range of logBFs.*

Table 4.13: Sensitivity analysis for the most similar writer pairs (W5–W6, W6–W7, W6–W8, and W7–W8) under different prior setups.

4.8.2 Specification of the Inverse-Wishart's Degrees of Freedom

In Bayesian hierarchical modeling, the choice of priors for covariance matrices is critical, since it has a direct impact on the resulting Bayes factor values, as observed in Section 4.7.

The Inverse-Wishart prior is a traditional choice for modeling covariance matrices. It is a conjugate prior for the multivariate Normal distribution, which simplifies the computational aspects of Bayesian inference. The Inverse-Wishart prior is parameterized by a scale matrix and by the degrees of freedom, which control the prior's concentration around the scale matrix.

In this section, we present a sensitivity analysis for different values of the degrees of freedom in the approach using the Inverse-Wishart prior. In the experiments using the Inverse-Wishart prior,

the degrees of freedom ν vary from the minimum admissible value of 11[1] and from 20 to 50, with increasing in steps of 10, that is, $\nu \in \{11, 20, 30, 40, 50\}$. The Bayes factor has therefore been calculated for each value of ν, for all comparisons between characters from different writers, considering the Bayesian Normal and MANOVA models with Normal-Inverse-Wishart prior setups (conjugate and non-conjugate).

The results show that the Bayes factor is quite sensitive to the choice of degrees of freedom, which has an incremental effect on its value; see Appendix A.5.2. This occurs because higher degrees of freedom make the prior of $\boldsymbol{W}$ more concentrated within a smaller region of $\mathbb{R}^p$ (Gaborini, 2021). Therefore, for larger values of ν, the variability of the variance-covariance matrix decreases. Consequently, we expect that as ν becomes large, the resulting log-Bayes factors will approach an upper bound which will always support the hypothesis that measurements originate from the same writer (even if the measurements originate from different writers); see Appendix A.5.2 for more details. This behaviour brings similar ideas with the Jeffreys–Lindley paradox (Lindley, 1957), wherein the Bayes factor becomes dominated by the prior rather than the data. This behaviour is related to the Jeffreys-Lindley paradox (Lindley, 1957), in both cases, the Bayes factor is driven more by the prior than by the data. The difference is that here it is a prior that is too narrow, rather than too broad, that causes the problem.

4.8.3 Specification of the Parameter of the LKJ Distribution

In this section, we present a sensitivity analysis for different values of η in the LKJ prior approach. The covariance decomposition with the LKJ approach provides a more flexible alternative for modeling correlation matrices of variance-covariance matrices. The LKJ prior is parameterized by a shape parameter, η, which controls the concentration of the prior around the identity matrix. When $\eta = 1$, the LKJ prior is uniform over the space of correlation matrices, making it a non-informative prior. Higher values of η result in stronger concentration around the identity matrix, reflecting a prior belief in weaker correlations.

Let us now consider the experiments using the LKJ prior approach; the parameter η was set to values ranging from the non-informative value of one (1) to increasingly informative values of 2, 5, 10, and 20. Higher values of η result in a greater concentration of the prior around the identity matrix, reflecting the assumption that Fourier coefficients are uncorrelated (an assumption motivated by the Fourier analysis). For each value of η, the Bayes factor was computed for all pairwise comparisons between different writers, employing the two considered models: the Bayesian Normal and MANOVA models with Normal-LogNormal-LKJ prior setup. Results show that the Bayes factor is quite sensitive to the choice of η, particularly for values above five (5), where its effect becomes increasingly more marked; see Appendix A.5.3 for a related discussion. This sensitivity arises because higher values of η imply stronger prior beliefs in the absence of correlation among variables. Further details can be found in Appendix A.5.3

The empirical comparison of the two approaches for modeling the variance-covariance matrix (the Inverse-Wishart prior and the LKJ prior) suggests that the LogNormal-LKJ model generally provides more robust results, especially for the Bayesian MANOVA model. This reflects the model's flexibility arising from the use of the covariance decomposition. The LogNormal prior provides the appropriate information for the standard deviations that lie on the diagonal of the covariance matrices, while the LKJ prior controls the information about the correlation structure. Based on our analysis, we recommend using the LogNormal-LKJ prior with the use of an informative LogNormal prior derived from

[1] That is, $p + 2 = 11$ for $p = 9$.

background data measurements. This approach provides a solid foundation for the implementation of the proposed models using the LogNormal-LKJ prior setup.

4.9 Case study

In this section, a case study has been prepared to represent common casework of forensic handwriting examiners, based on a questioned acknowledgment of debt. The writer of this document, who voluntarily participated in the study, has also been requested to supply control materials for comparison purposes. This writer will be designated as a person of interest (PoI), and we will suppose that it is disputed whether that person, or another one, wrote the questioned document. Lastly, we consider the IAM database presented in Section 4.2 from various writers, which is essential for the modeling process.

Questioned and control data

The questioned handwriting materials pertain to an acknowledgment of debt, which reads as follows:

Acknowledgment of debt

I, the undersigned, Giuditta Gonzales (born on 3 June 1962), hereby declare that I owe Mr. Edoardo Vincente (born on 31 December 1964), the sum of 3,000 (three thousand) Swiss francs, this corresponds to half of the share no. 00074 in the amount of 6,000 (six thousand) Swiss francs paid on 28 November 2001 to the Fonalba cooperative housing company for the granting of apartment no. 121 at 1, rue Oscar Wilde, 3456 Lyon, France. In the event that this share (n°00074) is released (termination, move, etc.) I undertake to contact Mr. Eduardo Vincente as soon as possible and to pay the said sum. This remains an obligation. Refunds will be made by mutual agreement as soon as possible. For any dispute arising from this contract, the parties agree that French law shall apply and that the courts of Lyon shall have jurisdiction, subject to any appeal to the French Federal Court. Done at Lyon, in two copies, on November 10, 2016.

Giuditta Gonzales *Edoardo Vincente*

The PoI was asked to write the questioned document under standard conditions (i.e., sitting at a table) and using their usual handwriting, i.e., without disguising their own handwriting style or imitating that of another person. The content of the questioned document has been chosen to display a sufficient amount of the characters of interest for this study, which are *a*, *b*, *d*, *e*, *g*, *o*, and *p*.

Control handwriting samples were requested from the PoI by following procedures used in forensic casework. Each collected page reflects the natural writing style of the subject, as no specific instructions were given that could influence the writing characteristics. Variations between pages were limited to the content of the text and the writing instrument used, alternating between blue and black ballpoint pens. All documents were generated in a single session, as in casework where the PoI can be met only once. The requested control samples include a free text, followed by numerals and alphabets (repeated twice), and then dictated words, as well as the questioned text itself was dictated to the PoI. To ensure consistency, the initial free text was dictated once more at the end of the session, allowing for the detection of any alterations in the handwriting that might occur during the process. Table 4.14 gives an overview of the available data points categorized by source and letter.

Source	Character							Total
	a	***b***	***d***	***e***	***g***	***o***	***p***	
Questioned Data	56	4	19	59	6	44	4	192
Control Data	311	21	100	299	55	157	36	980
IAM Background Data	2984	412	1264	3482	582	3124	702	12550

Table 4.14: Data points by source and letter.

4.9.1 Preliminary Analysis

In cases where control material of a given person is made available to the handwriting examiner, it is common practice to conduct a preliminary analysis of all pages of the control material to check for homogeneity and detect any possible contamination with handwriting samples coming from another writer. Therefore, a comprehensive evaluation of the evidence across all pages of control data is performed. This evaluation is based on slightly modified propositions, which can be formulated as follows:

H_3: the compared pages originate from same source;

H_4: the compared pages originate from different sources.

Specifically, we calculate the Bayes factor (BF) as outlined in Eq. (4.1), using data $\left(\boldsymbol{y}_{p_i}, \boldsymbol{y}_{p_j}\right)$, where (p_i, p_j) represent pages (i, j) respectively. This calculation incorporates the model M_6 Eq. (4.13) described in Section 4.5, see Table 4.1. The results are presented in Table 4.15. Pages 2 and 6 were excluded due to the limited amount of data and the content, which consists of alphabet writing. Data did not provide support for one proposition over the other, all BF results indicate very to extremely strong support for proposition H_3 over H_4.

LogBF	Page	1	3	4	5	7	8	9	10	11	12	13
	1	-	74.2	59.9	92.7	81.8	75.7	81.9	87.3	82.7	82.9	65.3
	3		-	112.6	105.9	118.5	94.1	101.0	99.3	94.2	58.7	78.8
	4			-	90.6	118.9	103.2	98.1	93.7	113.4	75.3	87.5
	5				-	115.5	95.1	107.9	109.6	97.6	87.0	76.1
	7					-	121.1	121.9	112.1	122.5	80.0	80.4
	8						-	109.3	111.4	122.1	89.7	79.6
	9							-	127.3	114.2	104.8	87.9
	10								-	120.9	112.0	77.8
	11									-	98.3	87.9
	12										-	86.6
	13											-

Table 4.15: Evaluation of evidence of control data across all pages considering all characters of all letters per page and assessing the Bayes factor in Eq. (4.1) (in log scale) implementing model M_6 in Eq. (4.13).

4.9.2 Results

Firstly, consider the scenario where loops of characters of the questioned acknowledgements of debt are compared with loops of characters of the control material from the PoI to discriminate between propositions H_1 and H_2, as described at the beginning of Chapter 4. Table 4.16 presents the logarithmic Bayes factors for the model Normal-LogNormal-LKJ M_3, see Table 4.1 for each letter separately and all letters jointly, and for MANOVA model with Normal-LogNormal-LKJ M_6, which incorporates the between-letters variability. Note that under the Normal-LogNormal-LKJ prior, the resulting BF is

obtained using numerical approximation methods, see Section 3.2.3. Referring to the verbal equivalence of the conclusion scale reported in the ENFSI Guideline (Willis et al., 2015) or in Marquis et al. (2016), the evidence provides extremely strong support (except for letter p where strong support is obtained) to the proposition H_1 that the questioned and control data originate from the same writer, over the alternative proposition H_2.

	Normal-LogNormal-LKJ								**MANOVA**
Character	*a*	*b*	*d*	*e*	*g*	*o*	*p*	ALL	ALL
LogBF	85.3	26.9	49.0	92.9	34.1	51.5	14.3	93.3	150.8

Table 4.16: Bayes factor (in log scale) for the case where questioned data is compared with data from PoI. All letters have been used separately or jointly.

Secondly, consider the scenario where loops of characters of the questioned acknowledgement of debts are compared with loops of characters of the writer 36 of the IAM background database (Section 4.2). The choice fell on the writer showing the greatest similarity to the questioned material, based on a Mahalanobis distance of 3.8. Table 4.17 displays the logarithmic Bayes factors for the model M_3 Normal-LogNormal-LKJ, see Table 4.1 for each letter separately and all letters jointly, and for MANOVA model with Normal-LogNormal-LKJ M_6. Substantial variability is observed across letters. The evidence varies notably by letter, but all of them support the correct hypothesis H_2. When all letters are combined, the Normal–LogNormal–LKJ model yields a negative log Bayes factor (-179.4), whereas the MANOVA model provides even stronger evidence against H_1 (-322.7).

The evidence provides extremely strong support, with the sole exception of letters p and b, where strong support and moderate support have been obtained, respectively, to proposition H_2 over the alternative proposition H_1.

Figure 4.15 (top) shows the posterior distributions of the parameters associated with the b_2 Fourier coefficient from the MANOVA Normal–LogNormal–LKJ model, presented separately for questioned and control data across letters and sources. As it can be observed, most of the posterior distributions of Fourier coefficient b_2 obtained for different letters overlap when the compared material originates from the same source. Marked discrepancies can be observed when control material from the PoI is replaced by material originating from writer 36 (IAM database), that is, when comparing material originating from different sources (see Figure 4.15, bottom).

	Normal-LogNormal-LKJ								**MANOVA**
Character	*a*	*b*	*d*	*e*	*g*	*o*	*p*	ALL	ALL
LogBF	-77.6	-7.6	-35.6	-38.1	-5.1	-100.5	-18.9	-179.4	-322.7

Table 4.17: Bayes factor (in log scale) for the case where questioned data is compared with data from writer 36. All letters have been used separately or jointly.

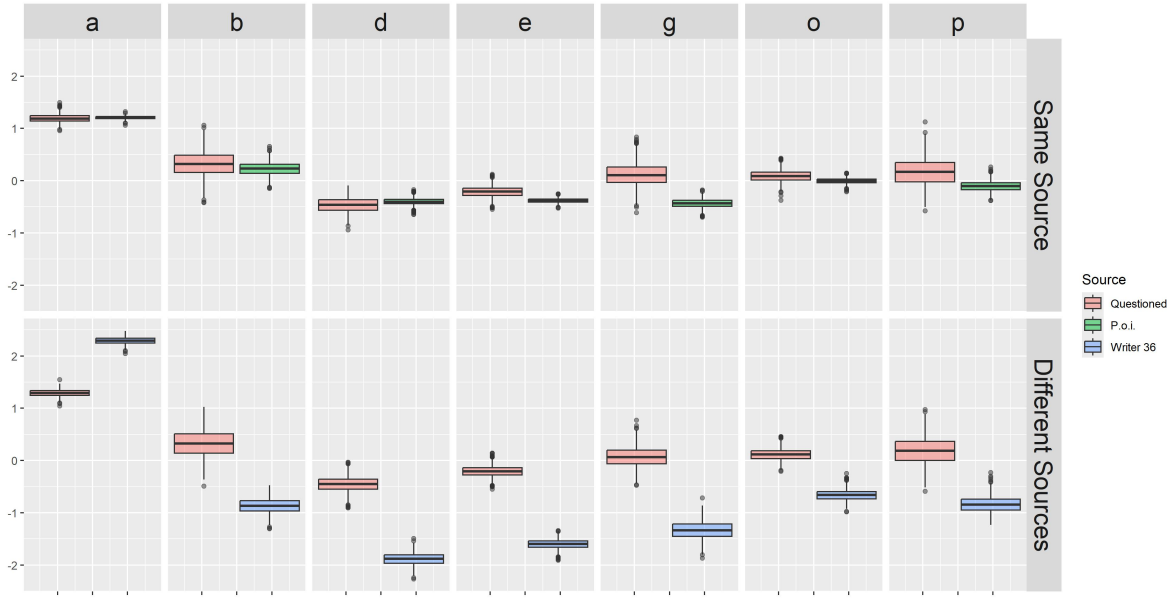


Figure 4.15: Posterior distributions of b_2 Fourier coefficient obtained under model M_6, per letter, for same source scenario (questioned data versus control data from the PoI, top), and different sources scenario (questioned data versus control data from writer 36, bottom).

4.10 Discussion and Conclusion

This study deals with the challenge of managing the uncertainty in forensic handwriting examination, which affects both inferential and decision-making processes in legal contexts. An experimental study has been conducted to facilitate (a) the management of complex multivariate data, (b) the combination of multiple sources of information (in this specific context, characters of different types) for the assessment of a joint probabilistic value, and (c) the identification of optimal modeling approaches. The proposed Bayesian statistical framework provides a comprehensive quantitative tool for uncovering valid and informative patterns in handwriting data, with the potential to substantially impact several areas of handwritten document analysis.

The proposed Bayesian probabilistic framework has been applied to a series of forensic case studies generated from real handwriting data. In particular, the performance of a two-level random effects model proposed for handwriting examination by Bozza et al. (2008) was compared with that of a Bayesian hierarchical MANOVA model, whose implementation is novel in this context. The study addressed the following five research questions:

(i) Which model fits the data better?

(ii) Which model has better discrimination performance (lower false positive and false negative rates)?

(iii) Does modeling the variability between writers have an impact on the final results?

(iv) Which character is most informative with respect to the ultimate goal of identifying the writer of a questioned handwritten manuscript?

(v) Do the results of the implemented models remain robust across different background data and prior specifications?

To address the first research question, six alternative Bayesian models have been considered and compared. These models arise from two likelihood structures: (a) a multivariate Normal model (See Section 4.5.1), and (b) a MANOVA model that accounts for character-level variability (See Section 4.5.2). For each likelihood, three different prior formulations are examined, resulting in distinct Bayesian models: (i) a conjugate Normal-Inverse-Wishart prior, (ii) a hierarchical Normal-Inverse-Wishart prior, and (iii) a Normal-LogNormal-LKJ prior specification. The results, presented in Table 4.8, showed strong support in favor of the Bayesian MANOVA model. The incorporation of the between-writers variability into the Normal-Inverse-Wishart (NIW) prior formulation appears to have a positive effect, with Bayes factors supporting this prior setup (and thus against the conjugate model) for 9 out of 13 writers, as shown in Table 4.9. For the MANOVA model with Normal-Inverse-Wishart prior, no clear advantage emerges between the conjugate and the hierarchical prior setups, as the results support both approaches (see Table 4.9). Finally, we compared the two prior setups Normal-LogNormal-LKJ prior and the Normal-Inverse-Wishart. The Bayes factors, presented in Table 4.10, support the Normal-LogNormal-LKJ prior over the Normal-Inverse-Wishart prior for most writers in both statistical models.

In order to address the research questions (ii)–(iv), a series of forensic case scenarios has been simulated by randomly selecting measurements from the available data to act as questioned and control material, respectively. Regarding the second research question, the Bayesian MANOVA model performed better than the Bayesian Normal model, with a false positive rate of around 0.6% compared to the 0.9% for the Normal model when all characters are analyzed jointly. This corresponds to a reduction in the false positive rate of approximately 33.3% compared to the Bayesian Normal model. This is not surprising, as the latter model does not take character-level variability into account. Furthermore, it was observed that employing the Normal-LogNormal-LKJ prior in the Bayesian MANOVA framework resulted in a 0.1% reduction in false positives compared to the Normal-Inverse-Wishart prior. This finding aligns with the results obtained from model fitting. However, when each character was analyzed separately using the Bayesian Normal model, the Normal-Inverse-Wishart prior demonstrated superior performance for two of the four characters considered. With respect to supporting the correct hypothesis when the handwritten material originated from the same writer (i.e., under H_1), both models consistently classified all relevant cases correctly, with the sole exception of the NIW model for character q, which exhibited an exceedingly small proportion of misclassifications ($< 0.1\%$; one over 1200 generated scenarios).

Concerning research question (iii), the results presented in Section 4.7.2 show that the hierarchical prior setup, which accounts for between-writer variability, achieves better overall performance with either Normal-LogNormal-LKJ or Normal-Inverse-Wishart prior compared to the conjugate approach (see Table 4.12). In contrast, the MANOVA model shows better results for the Normal-LogNormal-LKJ prior setup. This finding, along with the other experiments discussed in Section 4.7.2, where the accuracy of proposed models for handwriting discrimination has been tested, is in agreement with the results of the model comparison for the first research question (see Section 4.7.1). Similarly, a consistent reduction is observed between the conjugate and the non-conjugate version of the Bayesian Normal model with Normal-Inverse-Wishart prior. However, these results require further investigation and validation by ideally implementing the proposed methodology in a variety of different handwriting datasets by also involving a larger number of writers.

The analyses conducted to address research question (iv) reveal which character, among the four analyzed, is the most informative in terms of discrimination between the two hypotheses under investigation. Character q has consistently been found to produce fewer instances of misleading evidence,

with a false positive rate of 2.2%. On the other hand, the character *a* is the least informative in terms of discriminatory power, with a false positive rate reaching 7.2% with the hierarchical NIW prior. Finally, a false positive rate of 3.9 and 4.7, respectively, is observed for the characters *d* and *o*.

A sensitivity analysis was performed to evaluate the robustness of the results to prior elicitation and specification, addressing research question (v). By conducting subsampling on 50% of background writers' data of each evaluation case, we observed that the Bayes factor did not change to a degree that would alter the model support decisively. In particular, under the MANOVA model with the Normal-LogNormal-LKJ prior, the support for the tested hypotheses was reversed in only two of the 120 comparisons. These two cases involved the writers with the most similar handwriting text. This finding highlights the model's robustness in providing consistent support for or against a given hypothesis (see Section 4.8). Furthermore, the degrees of freedom parameter of the Inverse-Wishart distribution and the η parameter of the LKJ distribution play a crucial role in the behavior of the resulting Bayes factors. The Bayes factors are notably sensitive to the specification of these prior parameters. Specifically, higher values of the degrees of freedom and η lead to increasingly informative priors, which produce larger Bayes factors and thus stronger support for (the false) hypothesis H_1. Given this sensitivity, careful consideration is required when specifying the prior degrees of freedom (in the Inverse-Wishart distribution) and the prior parameter η (in the LKJ distribution). A recommended approach is to adopt weakly informative priors. This can be achieved by setting the degrees of freedom of the Inverse-Wishart distribution to its minimum value $p + 2$, and by specifying $\eta = 1$ for the LKJ distribution.

Finally, there is the open issue concerning the scale of the resulting Bayes factors. Their magnitude in this study is in accordance with the ones from previous related work (Bozza et al., 2008), but are open to critical remarks on robustness that cannot be demonstrated empirically. Similar situations have been discussed in the field of genetic (DNA) evidence (see e.g. Hopwood et al. (2012)). However, values of this magnitude can be difficult to justify, especially with such a limited database.

For future work, we suggest including neighbouring characters of loop characters in the analysis, which may provide more information and improve the accuracy of results. Such data can be modeled by a two-way MANOVA model with neighbouring characters and loop characters as dummy variables. Another possible direction for future work involves forensic casework with multiple persons of interest, where the objective is to evaluate the strength of evidence for each candidate as a potential source of the questioned material. In such scenarios, where evidential measures are assessed for multiple comparisons simultaneously, controlling the False Discovery Rate would allow practitioners to manage the expected proportion of incorrect supports among all candidates.

Chapter 5

Dynamic Signatures Examination

In dynamic signature forensic casework, legal authorities seek to determine whether a signature whose origin is questioned was made by a specific individual (the presumed source) or by someone else. Thus, the competing propositions are whether the questioned and reference materials collected from the presumed source for comparative purposes originate from the same source (H_1) or different sources (H_2). The evaluative process begins with defining these propositions, as recommended by the ENFSI guidelines (Willis et al., 2015).

For simplicity, we assume that the Person of Interest (PoI), namely the signer, has no reason to disguise his signature, and that a random match of personal information and signature is highly improbable, and are therefore excluded. However, before applying such simplifications to casework, assumptions and allegations need to be checked for compatibility. Therefore, although the simplification adopted in this study is not always justified, in most cases signature forgery remains the most plausible and relevant alternative to consider (Linden, 2022).

Thus, forensic handwriting document examiners are most often confronted with the following propositions: the signature may be genuine, or it may be a forged (simulated) signature. The first proposition suggests that the questioned signature was written by the PoI. The second proposition posits that the questioned signature was not produced by the PoI, but it is a forged (simulated) signature and produced by someone else. The propositions for signature examination can be formulated as follows:

H_1: The person of interest (PoI) produced the questioned signature;

H_2: An unknown person produced the questioned signature, trying to forge (simulate) the PoI's signature.

In the majority of dynamic signature examinations, only one dynamic signature is available; this is referred to as questioned material. Then, $n > 0$ signatures are collected from the person of interest; these are referred to as control material. Measurements of dynamic features (see Section 5.3) are extracted from questioned and control material. These features are denoted by $\boldsymbol{\mathcal{Y}}_{1:T}$ and $\{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}$ for the questioned signature and the control signatures, respectively. The index $1:T$ denotes the sequence of time points $1, 2, \ldots, T$ at which the dynamic features are recorded during the execution of a signature, where T is the total time points of feature values captured for that signature, and j the repetition of the signature (for questioned signature $j = 1$).

Following the ENFSI recommendations (Willis et al., 2015) for evaluative forensic science, the strength of evidence is quantified by a likelihood ratio (LR). The LR provides a principled probabilistic measure of how much more probable the observed signature features are under one proposition

compared to an alternative. Given the observed questioned signature features $\boldsymbol{\mathcal{Y}}_{1:T}$ and a set of n control signatures $\{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}$, the value of the evidence is expressed as

$$LR = \frac{f(\boldsymbol{\mathcal{Y}}_{1:T} \mid \{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}, H_1)}{f(\boldsymbol{\mathcal{Y}}_{1:T} \mid \{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}, H_2)}. \tag{5.1}$$

Under standard conditions in handwriting examination, the questioned and control materials are independent under the alternative hypothesis H_2. That is, if the control material originates from a different source, the probability distribution of the questioned material does not depend on the control sample. In such cases, the denominator reduces to $f(\boldsymbol{\mathcal{Y}}_{1:T} \mid H_2)$. However, in the context of alleged signature forgery, this assumption of independence cannot be maintained. A forger explicitly attempts to imitate the characteristics of a target signature, creating dependence between the questioned and control materials under H_2. Consequently, the computation of probability density under H_2 requires adequate databases containing forged signatures from which model parameters can be inferred (see Linden et al. (2021)). In particular, Linden et al. (2021) stresses the importance of using a representative general population of forged (simulated) signatures for evaluating the denominator of the LR.

In this work, $\{\boldsymbol{\mathcal{B}}_{j,1:T}\}_{j=1}^{N_{\mathcal{B}}}$ denotes the dynamic features extracted from a dataset of size $N_{\mathcal{B}}$ consisting of forged (simulated) signatures produced by individuals unrelated to the case. This background set reflects the characteristics of a general forged population, which under proposition H_2 represents the competing scenario in which someone other than the person of interest produced the questioned signature. Hence, the LR can be expressed as:

$$LR = \frac{f(\boldsymbol{\mathcal{Y}}_{1:T} | \{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}, H_1)}{f(\boldsymbol{\mathcal{Y}}_{1:T} | \{\boldsymbol{\mathcal{B}}_{j,1:T}\}_{j=1}^{N_{\mathcal{B}}}, H_2)} \tag{5.2}$$

In summary, the numerator represents the probability density of the observed features $\boldsymbol{\mathcal{Y}}_{1:T}$ under the proposition H_1 that the questioned signature is genuine and originates from the person of interest (PoI), while the denominator reflects the probability density of observing the same features under H_2, given a representative background of forged signatures.

In the following sections, the methodology used to address the evaluation of dynamic signatures is elucidated. First, a comprehensive literature review of data modeling techniques for dynamic signatures is presented in Section 5.1. The collection of dynamic signature datasets is discussed in Section 5.2. The extraction of both static and dynamic features is explained in Section 5.3. The adapted modeling which can be utilized to describe both genuine and forged signature populations within the thesis is describing in Section 5.4. The research will be focused more on stochastic process methodologies applied to the evaluation of dynamic signatures, where time is including as additional variable of interest and taking into account the multifaceted nature of dynamic signature data. To address this complexity, we adapt the stochastic process model,introducing distinct states that reflect temporal variations in the signature process. This is achieved through the use of Hidden Markov Models (HMM), which allow us to capture the evolving distribution of measurements over time inherent to dynamic signatures. We advance multi-feature stochastic modeling to the context of dynamic signatures by considering a batch of the more discriminating features in Section 5.7.2. Finally, the automated selection of the number of states in HMM using information criteria is examined in Section 5.7.3, and the impact of the number of control signatures is analyzed in Section 5.7.4.

5.1 Literature Review

Dynamic digitally captured signatures are an advanced form of electronic signature that not only capture the static image of a signature but also record various biometric parameters during the signing process. Unlike traditional static signatures, dynamic signatures are captured using specialized devices or tablets that record temporal and spatial data as the signature is produced. This data includes the speed, acceleration, pressure, and angle of the pen, providing a rich dataset that can be used for authentication and verification purposes (Linden et al., 2018). Contemporary research within this discipline has increasingly concentrated on the refinement and enhancement of AI data modeling. This focus aims to enhance predictive accuracy, computational efficiency, and generalizability across dynamic signature datasets. This review synthesizes key contributions from recent literature, focusing on methodologies such as machine/deep learning techniques, dynamic time warping (DTW), and Hidden Markov Models (HMM).

The multidimensional perspective of the dynamic signatures facilitates the distinction between genuine and forged signatures with high accuracy. This capability has been demonstrated in studies such as Chandra (2020) and Vorugunti et al. (2022), which utilize machine learning and deep learning methodologies for signature verification. Furthermore, the article of Vorugunti et al. (2023) investigates a hybrid approach that combines Convolutional Neural Networks (CNNs) and Transformer models. This approach aligns with recent trends in deep learning, where Transformers have shown significant promise in handling sequential data due to their ability to capture long-range dependencies.

Utilizing time series distance metrics on dynamic features, we can proceed to compute the Dynamic Time Warping (DTW) approach (Sakoe and Chiba, 1978). Al-Hmouz et al. (2019) proposes a probabilistic model to quantify the DTW distance for dynamic signature verification. Their approach enhances the robustness of DTW by incorporating probabilistic measures, which improves the accuracy of signature verification in the presence of variations in signing behavior. Parziale et al. (2019) introduces stability modulated DTW (SM-DTW) for signature verification. SM-DTW enhances traditional DTW by incorporating stability measures, which improve the system's resilience to intra-class variations and increase verification accuracy.

By considering stochastic process methods[1], a pioneering work by Yang et al. (1995) introduced Hidden Markov Models (HMM) in dynamic signature verification. Jahan and Farimani (2018) presents an HMM-based approach for online signature verification, focusing on velocity and hand movement directions. Their model effectively captures the dynamic aspects of signatures, providing a robust framework for distinguishing genuine signatures from forgeries. Tolosana et al. (2015) focuses on maintaining system performance over time by updating the HMM parameters based on new signature data, ensuring the system adapts to changes in signing behavior.

In a logical progression, hybrid methods have been extended to enhance the performance of the models. Miguel-Hurtado et al. (2007) combines DTW with Gaussian mixture models (GMM) for online signature verification. This hybrid approach leverages the strengths of both techniques, resulting in improved verification performance and reduced verification error rates. Fierrez and Ortega-Garcia (2008) provide a comprehensive overview of online signature verification techniques. Their work covers various methodologies, including DTW, HMM, and neural networks, highlighting the strengths and limitations of each approach in different application contexts. Tahmasebi and Pourghassem (2013) propose a signature identification system using dynamic features and HMM, combined with a K-

[1]Stochastic process is a collection of random variables indexed by time or space, representing a system that evolves over time in a probabilistic manner.

Nearest Neighbor (KNN) classifier. This approach effectively balances the complexity and identification accuracy rate of the verification process. Diaz et al. (2016) introduces a dynamic signature verification system that relies on a single real signature for training. This method addresses the challenge of limited training data by leveraging advanced feature extraction and classification techniques. The comprehensive survey by Kaur and Kumar (2023) presents an extensive overview of both online and offline signature verification methods, highlighting the evolution of feature extraction and classification techniques used in the signature verification process.

In the field of forensic science, dynamic digitally captured signatures are particularly valuable because they provide a level of detail that is not available in static signatures. The ability to analyze and to acquire data on the sequence of movements, pressure patterns, and speed offers forensic experts a method, making it more difficult to replicate or forge signatures. This makes dynamic signatures a robust solution for security-sensitive applications, where the authenticity of a signature is of utmost importance (Linden et al., 2021). Mazzolini et al. (2021) present a semi-automatic approach to signature verification in forensic settings. The proposed framework aims to provide forensic document examiners with an intuitive tool that applies Dynamic Time Warping to dynamic signatures. Okado et al. (2024) uses mathematical tools such as principal component analysis (PCA), DTW, and the Kolmogorov–Smirnov test to analyze global and local features from dynamic signatures. It aims to reduce subjectivity and increase the reproducibility of handwriting examination. Linden et al. (2022) proposed and described a Bayesian probabilistic model in an explanatory way. The paper underscores the importance of adopting a multivariate approach and highlights the role of probabilistic models in enhancing the reliability and transparency of forensic dynamic signature examination.

The objective of this research is to employ Hidden Markov Models (HMM), a stochastic probabilistic model, in the context of forensic evaluation. We focus on jointly evaluating multiple features, even under practical constraints such as the limited number of samples typically available in the routine work of questioned document experts.

5.2 Data Acquisition, Sampling and Databases

In this work we re-analyze the data collected by Linden (2022). Based on these data, four forensic publications were produced (Linden et al., 2017, 2021, 2022; Linden and Marquis, 2023). The data were gathered under natural conditions, including a desk, office chair, and horizontal surface, to reproduce a realistic scenario. The device used was a Wacom DTU-1141, with a surface area of 283 x 210 mm, a spatial resolution of 2540 lpi, a temporal resolution of 200 Hz, and pressure measured axially using 1024 discrete levels. The software was compatible with Wacom decrypted software and Windows 7 SP1. Data were recorded approximately every 5 milliseconds.

The sample data consist of two main case-related types, each originating from different populations. These types are: (1) genuine signatures (presumed source data), and (2) forged (simulated) signatures (forgery population data).

The data of the **case-related genuine signatures**, denoted by $\mathcal{G}$, were produced by three individuals, who were selected after the screening of volunteer signatures to ensure a range of styles and levels of complexity were covered, from low to high; see Linden (2022) for detailed explanation of the screening process. Participants were asked to provide signatures at irregular time intervals, with decreasing frequency, to study the variation within and between sessions over different time periods, such as days, weeks, and months. Over approximately 18 months, each participant completed at least 44 distinct acquisition sessions. Initially, participants provided 10 samples per session, but due to high

variability observed during the process, the number of samples per session was increased to 20. Each participant produced at least 750 signatures during the acquisition period. A detailed description of the number of acquired signatures per participant is provided in Table 5.1. These data mainly are used as the control group ($\boldsymbol{\mathcal{X}}$). Nevertheless, in some experiments in Section 5.7, data for a specific signature coming from $\mathcal{G}$ are also used as the questioned material ($\boldsymbol{\mathcal{Y}}$) in scenarios where both materials under investigation ($\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$) originate from the same signer.

	Signature 1	**Signature 2**	**Signature 3**
Total Signatures	770	750	800
Sessions with 10 signatures	15	13	12
Sessions with 20 signatures	31	32	34
Total sessions	46	45	46

Table 5.1: Case-related genuine dynamic signatures dataset ($\mathcal{G}$).

The data of **case-related forged (simulated) signatures**, denoted by $\mathcal{F}$, were generated by 57 additional participants acting as forgers who attempted to replicate the three original signatures of the genuine signature dataset $\mathcal{G}$. Data collection took place through a prize-based contest designed to encourage participation and competition; see Linden (2022), for a detailed description of the forgery procedure. Each forger selected which signatures to replicate and produced 10 simulated copies of each chosen signature. The number of forged (simulated) signatures per genuine signature is reported in Table 5.2. This dataset was primarily used as background data for estimating the model parameters associated with the forged population. Moreover, in the experiments reported in Section 5.7, data for a specific signature from $\mathcal{F}$ were also used as the questioned material ($\boldsymbol{\mathcal{Y}}$) in scenarios where the materials under investigation ($\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$) come from different signers. More specifically, if the questioned material ($\boldsymbol{\mathcal{Y}}$) corresponds to signature 1, then data from signatures 2 and 3 are used as background data for estimating the model parameters associated with the forged population.

	Signature 1	**Signature 2**	**Signature 3**
Forgers	26	41	16
Signatures	10	10	10
Sessions	1	1	1
Total Forgeries	260	410	160

Table 5.2: Case-related forged (simulated) dynamic signatures dataset ($\mathcal{F}$).

Additionally, a **case-unrelated** genuine signature dataset, denoted by $\mathcal{D}$, was collected from 23 different participants (20 genuine signatures each; a total of 460 observations) as background information. This dataset was used to assess the variability of signatures within a session and to normalize the case-related data $\mathcal{G}$ and $\mathcal{F}$. Participant summaries of $\mathcal{D}$ are provided in Table 5.3.

Signers	**Signatures**	**Sessions**	**Total**
23	20	1	460

Table 5.3: Case-unrelated genuine dynamic signatures ($\mathcal{D}$).

5.3 Dynamic Signature Features

Dynamic signatures represent an advanced form of biometric authentication, capturing the characteristics of an individual's signing process in real-time. These signatures encompass several features, such as the speed, pressure, and rhythm of the signing motion. This additional information enhances the

security of dynamic signatures, making them significantly more challenging to forge. The real-time nature of dynamic signatures enables immediate evaluation, providing an extra layer of security and convenience in various applications. This process is particularly efficient due to its simplicity. It only requires a small tablet and a pen, which together provide highly detailed temporal information. More recent applications require only a touch screen device, not necessarily a pen.

The digitizer of signatures captures pen data at regular or event-based intervals, recording each sampled point in a chronological sequence. These pen data are preprocessed by the driver before being transmitted to the computer. Preprocessing may involve interpolation to smooth the data and normalization of signals, such as amplifying or reducing pen pressure. The signature data consists of various input-related columns, including an index of data points, button-related columns, pen state (whether in air or on the surface), and the measured data. The measured data must include the spatial coordinates (X_t, Y_t) of each time point t, the pressure P_t, and the time stamps τ_t,

$$\left\{ (X_t, Y_t, P_t, \tau_t) \right\}_{t=1}^{T}.$$

Additional data and measurements, depending on the hardware, may include tilt and inclination. Signatures are encapsulated in signature container formats that include pen data, metadata, and a timestamp, all of which are encrypted. Once decrypted and extracted, the pen data is available as a simple table. An example of the signature data is available in Table 5.4 with timestamp difference $d\tau$ of 5 milliseconds [ms] for most time points.

Input Data			Signature Data			
Index	Button	Pen State	X	Y	P [Level]	$d\tau$ [ms]
1	1	1	15005	7600	300	0
2	1	1	15015	7605	334	5
3	1	1	15035	7615	370	5
4	1	1	15040	7613	400	5
5	1	1	15045	7612	476	5
6	0	2 (UP)	15063	7625	511	5
7	0	2 (UP)	15078	7627	480	5
8	0	2 (UP)	15120	7633	430	5
9	1	3	15150	7638	400	5
10	1	3	15167	7639	412	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Feature names: *Button: button pressed; Pen State: Stroke number; X,Y: Spatial coordinates; P: pen pressure; $d\tau$: time stamp.*

Table 5.4: Example of the raw data structure of a dynamic signature.

Based on raw data, numerous features can be derived, encompassing both static and dynamic characteristics. Features can be classified into several types, according to their level of detail (Richiardi et al., 2005). Each type of feature offers distinct advantages and challenges. *Global features* (such as total time, length, mean pressure, etc.) provide general information about the signatures. They can be associated with temporal, spatial, and aggregated information derived from dynamic features. Conversely, *Local features* (such as pressure, velocity, acceleration, etc.) offer more comprehensive and detailed insights into variations. *Global features* are simpler to use and can be aggregated based on the boolean attribute of pen state; in the literature, they are called *Segment features*. For a detailed representation of the feature that was analyzed in this thesis, refer to Table 5.5 presents the feature categories, and Table 5.6 provides the notation and mathematical formulas.

Dynamic features or *Local features* (Richiardi et al., 2005) are generally represented as a continuous (pseudo-) timeseries. These features include temporal aspects such as the velocity, acceleration,

Feature Category	**Unit**	**Examples**
Time	millisecond [ms]	total time, time pen is up/down, ...
Spatial	millimetre [mm]	total/up/down length, height, width ...
Pressure	levels [lvl]	time-series, average, ...
Differential of pressure	levels per millisecond [lvl/ms]	time-series, average, ...
Velocity	millimetre per millisecond [mm/ms]	time-series, average, ...
Acceleration	millimetre per millisecond^2 [mm/ms^2]	time-series, average, ...
Jerk	millimetre per millisecond^3 [mm/ms^3]	time-series, average, ...
Angles	degrees [^{o}]	trajectory vector angle to horizontal line, acceleration direction to horizontal line (time-series, average, ...)

Table 5.5: Feature categories of dynamic signature data.

Feature notation	**Feature description**	**Formula**
P	Pen pressure	-
$\mathrm{d}\tau_t$	Differential of timestamps (time increment)	$\tau_t - \tau_{t-1}$
$\mathrm{d}X_t$	Differential of horizontal axis	$X_t - X_{t-1}$
$\mathrm{d}Y_t$	Differential of vertical axis	$Y_t - Y_{t-1}$
$\mathrm{d}p_t$	Differential of pressure	$P_t - P_{t-1}$
Totaltime	Duration (cumulative time)	$\sum_{t=2}^{T} \mathrm{d}\tau_t = \tau_T - \tau_1$
Uptime	Cumulative duration of pen lifts	$\sum_{\text{pen up}}^{T}{}_{t=2} \mathrm{d}\tau_t$
Downtime	Cumulative duration of inking strokes	$\sum_{\text{pen down}}^{T}{}_{t=2} \mathrm{d}\tau_t$
UpTot	Ratio of pen lifts	Uptime/Totaltime
DownTot	Ratio of inking strokes	Downtime/Totaltime
DownUp	Ratio of inking strokes to pen lifts	Downtime/Uptime
TotLength	Cumulative length	$\sum_{t=2}^{T} \sqrt{(\mathrm{d}X_t)^2 + (\mathrm{d}Y_t)^2}$
DownLength	Cumulative length of inking strokes	$\sum_{t=2}^{T} \sqrt{(\mathrm{d}X_t)^2 + (\mathrm{d}Y_t)^2}$ when pen is down
UpLength	Cumulative length of pen lifts strokes	$\sum_{t=2}^{T} \sqrt{(\mathrm{d}X_t)^2 + (\mathrm{d}Y_t)^2}$ when pen is up
Width	Horizontal distance	$\max_t(X_t) - \min_t(X_t)$
Height	Vertical distance	$\max_t(Y_t) - \min_t(Y_t)$
WHRatio	Width to Height ratio	Width/Height
XY	Distance to centroid	$\sqrt{(X_t - \overline{X})^2 + (Y_t - \overline{Y})^2}$
$dp1$	First differential of pressure	$\mathrm{d}P_t/\mathrm{d}\tau_t$
$dp2$	Second differential of pressure	$\mathrm{d}(dp1)/\mathrm{d}\tau_t$
$dp3$	Third differential of pressure	$\mathrm{d}(dp2)/\mathrm{d}\tau_t$
$dx1$	Horizontal velocity	$\mathrm{d}X_t/\mathrm{d}\tau_t$
$dx2$	Horizontal acceleration	$\mathrm{d}(dx1)/\mathrm{d}\tau_t$
$dx3$	Horizontal jerk	$\mathrm{d}(dx2)/\mathrm{d}\tau_t$
$dy1$	Vertical velocity	$\mathrm{d}Y_t/\mathrm{d}\tau_t$
$dy2$	Vertical acceleration	$\mathrm{d}(dy1)/\mathrm{d}\tau_t$
$dy3$	Vertical jerk	$\mathrm{d}(dy2)/\mathrm{d}\tau_t$
$dt1$	Tangential velocity	$\sqrt{(\mathrm{d}X_t)^2 + (\mathrm{d}Y_t)^2}/\mathrm{d}\tau_t$
$dt2$	Tangential acceleration	$\mathrm{d}(dt1)/\mathrm{d}\tau_t$
$dt3$	Tangential jerk	$\mathrm{d}(dt2)/\mathrm{d}\tau_t$
TVD	Trajectory vector angle to horizontal line	$\arctan(\mathrm{d}Y_t, \mathrm{d}X_t)$
TAD	Acceleration direction to horizontal line	$\arctan(dy2, dx2)$

By _bar, _var, or _max, it indicates the mean, variance, or maximum of the feature values respectively.

Table 5.6: Feature notation, description and mathematical formulas of dynamic signature data.

and jerk, as well as pressure variations and angle-oriented information. By analyzing these dynamic elements, it is possible to gain insights into the unique behavioral patterns of the signer, which are difficult to replicate accurately.

Global features and *Segment features* (Richiardi et al., 2005), on the other hand, focus on the overall shape and structure of the signature, capturing general information that complements the dynamic features. While these features may not capture the full depth of the signing process, they offer a robust foundation for signature analysis and can be effectively combined with dynamic features for a more robust evaluation.

5.3.1 Explanatory Analysis

In this section, we present a primary analysis of the dynamic signature features. Our focus is to elucidate the principal differences among signers, considering both static and dynamic features in the analysis. In analyzing the differences between signers, we begin with comparisons of static features, as detailed in Table 5.6. The included features are time related, spatial related, and aggregated dynamic information. The dataset $\mathcal{D}$ comprises 23 signers, each providing 20 genuine signatures, as shown in Table 5.3. We conducted pairwise Bayesian t-test (see Morey et al. (2011)) for each Global and Segment feature, utilizing the 20 signatures per signer. Namely, for each pair of signers W_i and W_j, we compare their static feature values. A pairwise Bayesian t-test is used to compare the means of two groups (signers) by calculating the Bayes factor of the two hypotheses (H_0: no significant difference between the two signers for the given feature; H_1: significant difference between the two signers for the given feature). By conducting these pairwise comparisons using the Bayesian t-test, we can assess whether the static features of signatures are significantly different between the signers. The results of these tests are summarized in Table 5.7. The table presents the results of pairwise Bayesian t-tests conducted between background signers for various static features. The five more discriminate features are *Downtime, Totaltime, dt1_ bar, dt1_ var , and DownLength.* The proportions of the Bayes factors greater than 3 (indicating substantial evidence) are notably high, ranging from 0.913 to 0.968. This suggests that the majority of the writers comparisons for these features provide substantial evidence against the null hypothesis (no significant difference between the two writers for the given feature), highlighting significant differences for these features between the genuine signers. This is also illustrated in the Boxplots of *Downtime* feature and *mean velocity* feature per writer in Figures 5.1,5.2.

Feature notation	**Proportions of BF greater than 3**
Downtime	0.968
Totaltime	0.953
$dt1$_bar	0.945
$dt1$_var	0.933
DownLength	0.913

Table 5.7: Pairwise Bayesian t-test between background signers. The five more discriminate static features from the dynamic signatures are presented.

Considering the dynamic features (pseudo-timeseries), we conducted naive approach by just considering a time averaging over 20 points due to the varying available of the time points. More sophisticated approach for time series preprocessing is presented in Section 5.5. Additionally, within a single writer, there is inherent variation and time warping, as illustrated in Figure 5.3, which depicts the time series plot of *velocity* feature for two different signatures from Writer 1.

To naive compare signatories, including their dynamic characteristics, we calculate the Dynamic

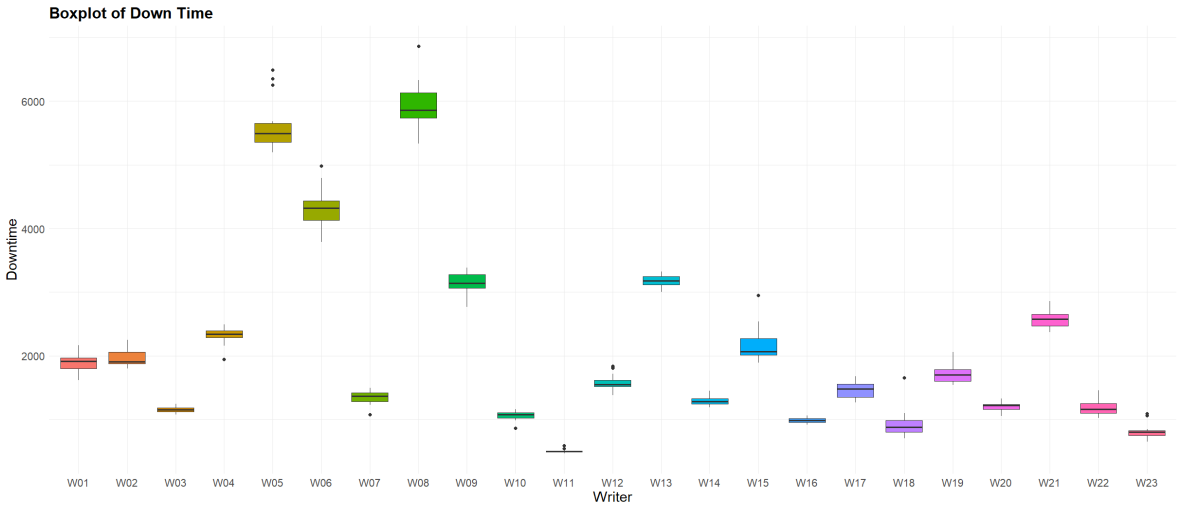


Figure 5.1: Boxplot of down time of signatures per signer of dataset $\mathcal{D}$.

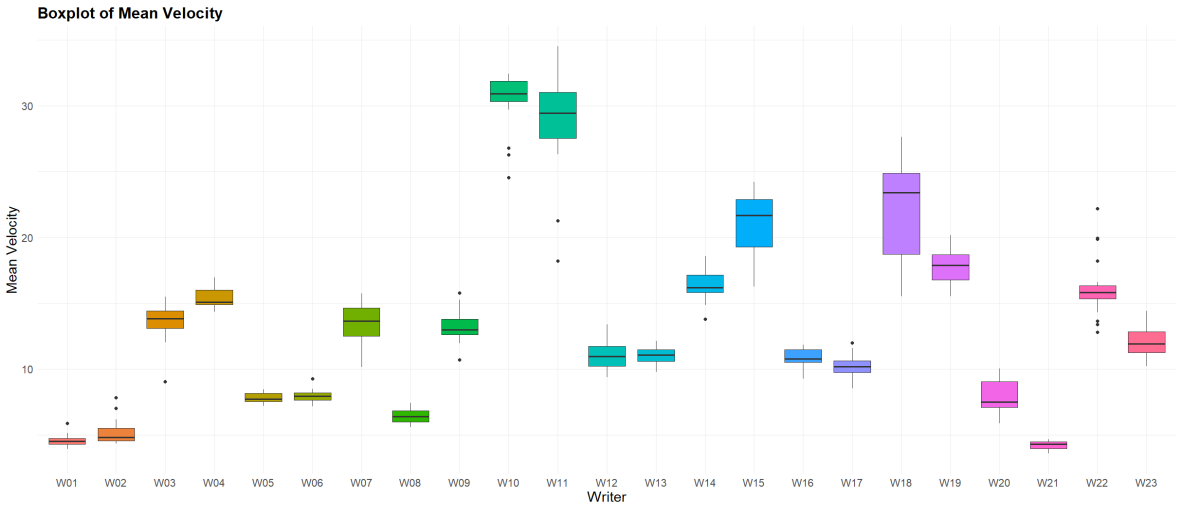


Figure 5.2: Boxplot of the average velocity of signatures per signer of dataset $\mathcal{D}$.

Time Warping (DTW) distance (Müller, 2007) within the signatures of a single signer and across different signers. The DTW distance measures the similarity between two temporal sequences that may vary in their progression over time, making it particularly useful for comparing sequences that are not perfectly aligned.

DTW was employed to compare the signatures of the same signer and different signers. For within-signer comparisons, the DTW distance was calculated between multiple pairs of different signatures from the same signer. A lower DTW distance indicates higher consistency, meaning the signatures are more similar to each other. For between-signer comparisons, the DTW distance was calculated between multiple pairs of signatures of different writers. A higher DTW distance suggests that the signatures are more dissimilar, indicating different signing styles.

Due to the significant variability observed in certain features from the same signer (within-signer variability), an alternative method was needed to distinguish the important difference between features of different signers. Thus, a preliminary approach involves calculating the differences between the DTW measures of features within a signer and across different signers. Specifically, for a given feature, the following difference is calculated:

$$Difference\ of\ DTW = abs(\overline{DTW}_{same\_signer} - \overline{DTW}_{different\_signers})$$

where $\overline{DTW}_{same\_signer}$ is the mean DTW computed between pairs of signatures from the same signer, and $\overline{DTW}_{different\_signers}$ is the mean DTW computed between pairs of signatures from different signers. By subtracting these two sets of DTW distances, we can naively identify which features distinguish the difference between signatures from different signers without being affected by the within-signer differences. However, this approach does not account for the variability of the DTW measurements.

The five features with the greatest differences in DTW distances between signatures from the same signer and different signers, are summarized in Table 5.8. These results indicating that the DTW measure is effective in distinguishing differences of signers based on dynamic features. The five features with the highest differences are the velocity and angle-oriented features, emphasizing their importance in the analysis.

It is important to note that the primary analysis does not account for the variance in the DTW distances of the features. Therefore, a probabilistic analysis that incorporates the variability of the DTW distances for each feature would be particularly beneficial, especially in cases where there is significant within-signer variability.

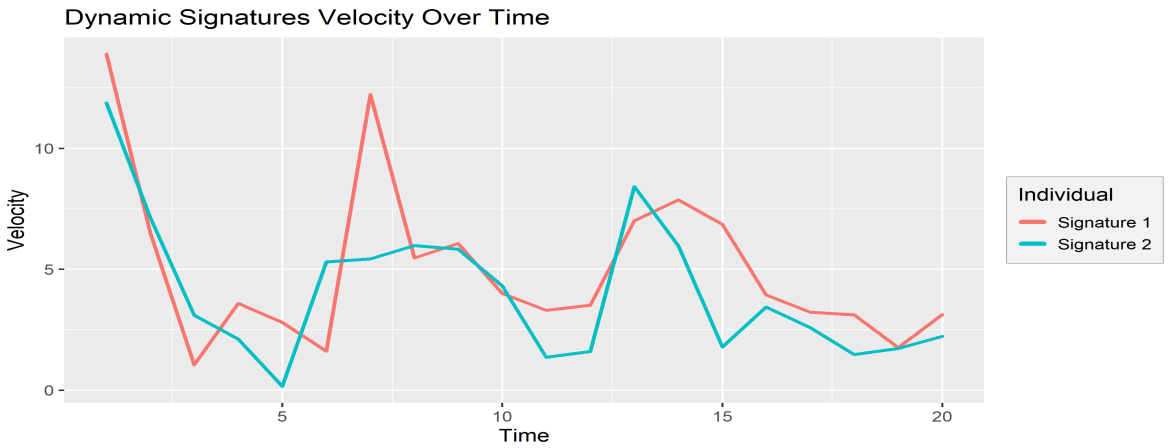


Figure 5.3: Time series of velocity dynamic feature of two signatures of writer 1

Feature notation	Same signer $\overline{DTW}$	Different signers $\overline{DTW}$	Difference of $\overline{DTW}s$
$dx1$	7.7	17.8	10.1
$dt1$	6.4	16.2	9.8
TVD	10.3	19.6	9.3
$dy1$	8.2	17.4	9.2
TAD	11.2	20.1	8.9

The five features with the largest differences of $\overline{DTW}s$ are presented.

Table 5.8: Mean of dynamic time warping (DTW) of dynamic features within and between signers.

Considering distinguishing between genuine and forged signatures by examining both static and dynamic features in our analysis presents significant challenges, primarily due to the limited availability of reference signatures, often only one or two. To initiate our analysis, we illustrate the distribution of temporal, spatial, and aggregated dynamic features for the entire dataset of genuine and forged signatures, in Figures 5.4 and 5.5. This figure reveals notable differences in several features, such as

time, pressure, velocity, and angle orientation. However, as previously mentioned, the scarcity of data in most cases limits our analysis.

Accordingly, we proceed with the comparison of dynamic features using DTW distance for each case-related signatures. Specifically, the DTW distance was calculated for each case-related signatures between multiple pairs of different genuine signatures, as well as between multiple pairs of genuine and forged signatures. However, as observed, due to the significant variability observed in certain features, either within genuine or within forged signatures, an alternative method was required. Therefore, we calculate the difference between the mean of DTW measures within genuine signatures and the mean DTW measures between genuine and forged signatures for each feature. Therefore, by considering the average DTW distance within genuine case-related signatures, we can observe some discrepancies between genuine and forged signatures.

The results for each case signature are represented separately in Tables 5.9, 5.10, and 5.11. The DTW distances for dynamic features are presented, comparing genuine signatures against genuine signatures from the same signer or genuine signatures with forged signatures. These results highlight significant feature differences between genuine and forged signatures, based on DTW measures. The five features with the greatest differences demonstrate that the *velocity* feature consistently appeared in every case-related signature, along with the *distance from the centroid (XY)* and *angle-oriented* features. Finally, as observed by Linden et al. (2021), every signature exhibits different distinguishing features.

It is crucial to reiterate that the initial analysis does not consider the variance in DTW distances of the features. Consequently, incorporating a probabilistic analysis that accounts for the variability of DTW distances for each feature would be especially advantageous.

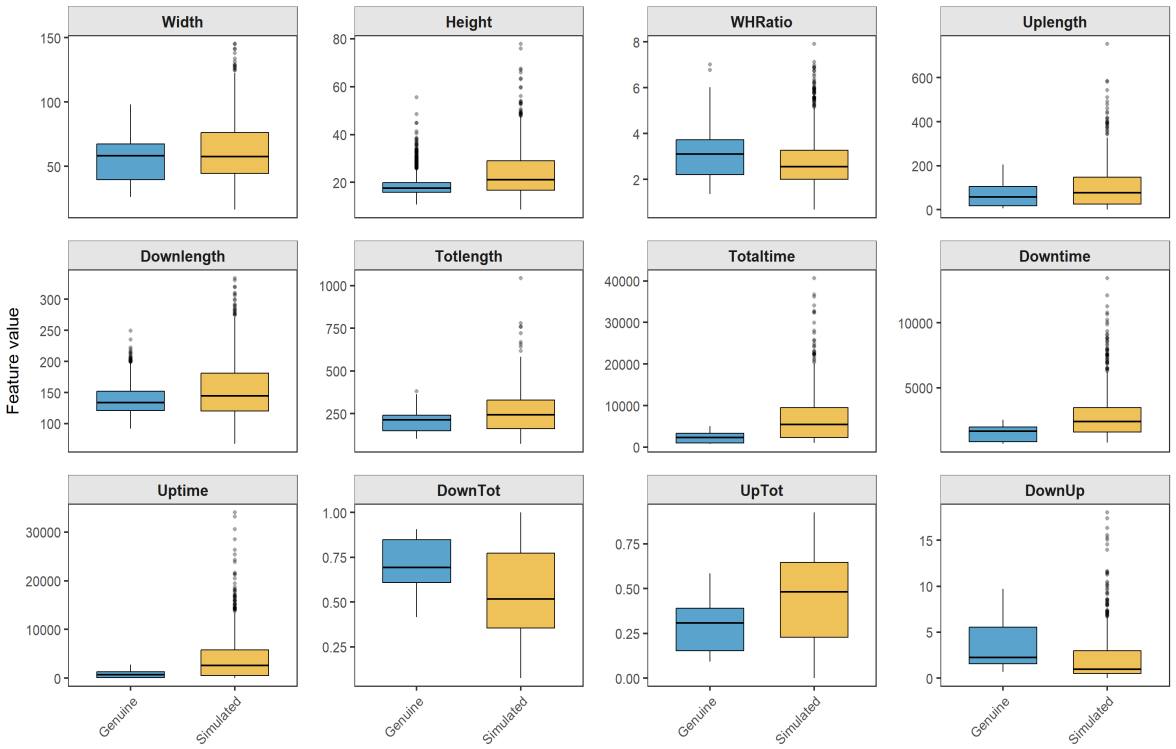


Figure 5.4: Boxplot of time and spatial category features of genuine and forged datasets.

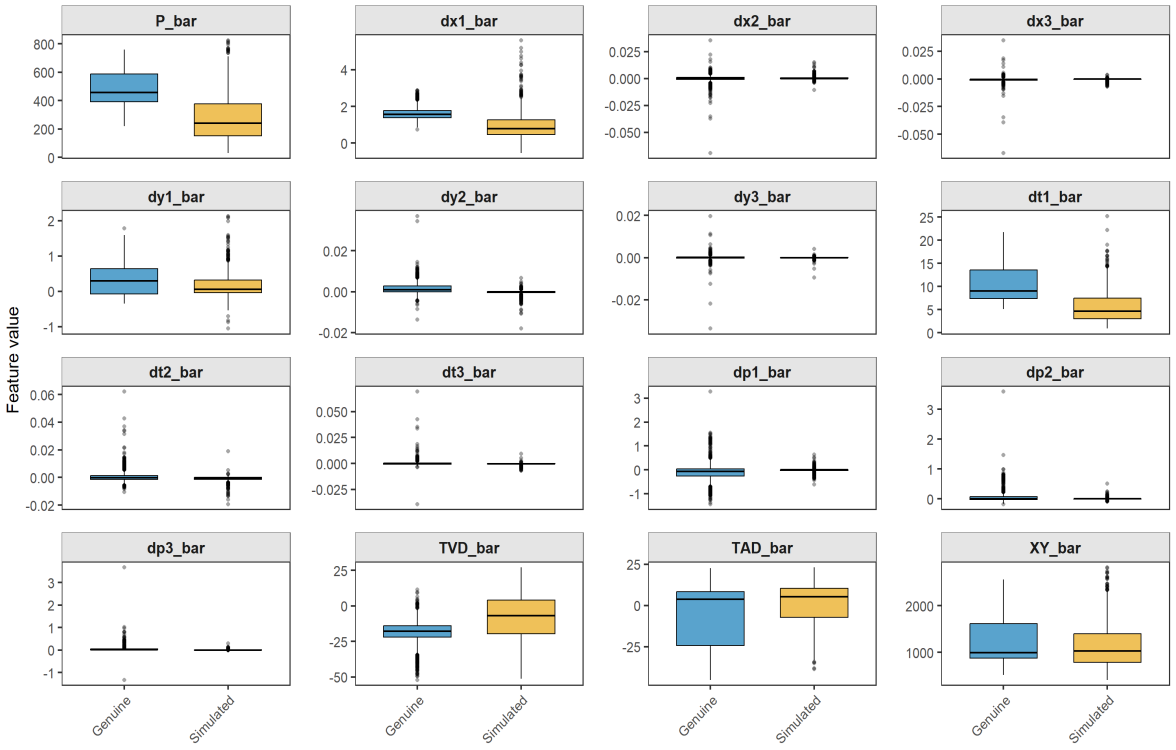


Figure 5.5: Boxplot of aggregated dynamic features of genuine and forged datasets.

Signature 1			
Feature notation	**Genuine vs Genuine** $\overline{DTW}$	**Genuine vs Forged** $\overline{DTW}$	**Difference of** $\overline{DTW}s$
XY	6.0	13.5	7.5
TVD	11.2	16.9	5.7
$dy2$	11.3	17.0	5.7
TAD	13.2	18.8	5.6
$dy1$	12.6	18.1	5.5

The five features with the largest differences of $\overline{DTW}s$ are presented.

Table 5.9: Mean dynamic time warping (DTW) values for Signature 1, comparing genuine-genuine and genuine-forged signatures.

Signature 2			
Feature notation	**Genuine vs Genuine** $\overline{DTW}$	**Genuine vs Forged** $\overline{DTW}$	**Difference of** $\overline{DTW}s$
$dy1$	10.9	19.8	6.9
$dy2$	13.9	20.7	6.8
XY	8.6	15.3	6.7
$dt1$	8.4	14.1	5.7
TAD	12.8	18.0	5.2

The five features with the largest differences of $\overline{DTW}s$ are presented.

Table 5.10: Mean dynamic time warping (DTW) values for Signature 2, comparing genuine-genuine and genuine-forged signatures.

However, a primary difficulty arises from the limited availability of questioned signatures. Consequently, it is essential to extract a detailed representation of the dynamic signature features. To illustrate a real-case scenario, we consider features from one questioned signature and two control signatures taken for comparative purposes. Figure 5.6 presents time-series plots of the features pressure, velocity, and distance from the trajectory center when the questioned signature is genuine and

Signature 3			
Feature notation	**Genuine vs Genuine** $\overline{DTW}$	**Genuine vs Forged** $\overline{DTW}$	**Difference of** $\overline{DTW}s$
$dt1$	11.3	19.6	8.3
XY	6.3	12.2	5.9
$dx1$	12.4	17.4	5.0
TVD	11.6	16.4	4.8
$dy1$	9.0	12.8	3.8

The five features with the largest differences of $\overline{DTW}s$ are presented.

Table 5.11: Mean dynamic time warping (DTW) values for Signature 3, comparing genuine-genuine and genuine-forged signatures.

originates from the same individual who provided the control signatures. The features are standardized based on the mean and standard deviation. The plots indicate that the time series are generally similar, although some time warping, temporal shifts, and differences in alignment are observed. This reflects the inherent within-signer variability. This observation underscores the value of a probabilistic analysis that explicitly incorporates the within-signer variability, which is particularly important in cases where the signer exhibits substantial intra-variability.

Figure 5.7 displays the corresponding plots when the questioned signature is a forgery. Noticeable differences emerge both in the overall shape and in the values of the time series, clearly suggesting that the signature may not originate from the same person.

In general, genuine signatures exhibit distinct peculiarities compared to forged ones across several dynamic features, such as time, pressure, velocity, and angular orientation. However, the limited availability of questioned signatures remains a major challenge and a fundamental constraint in real forensic case scenarios.

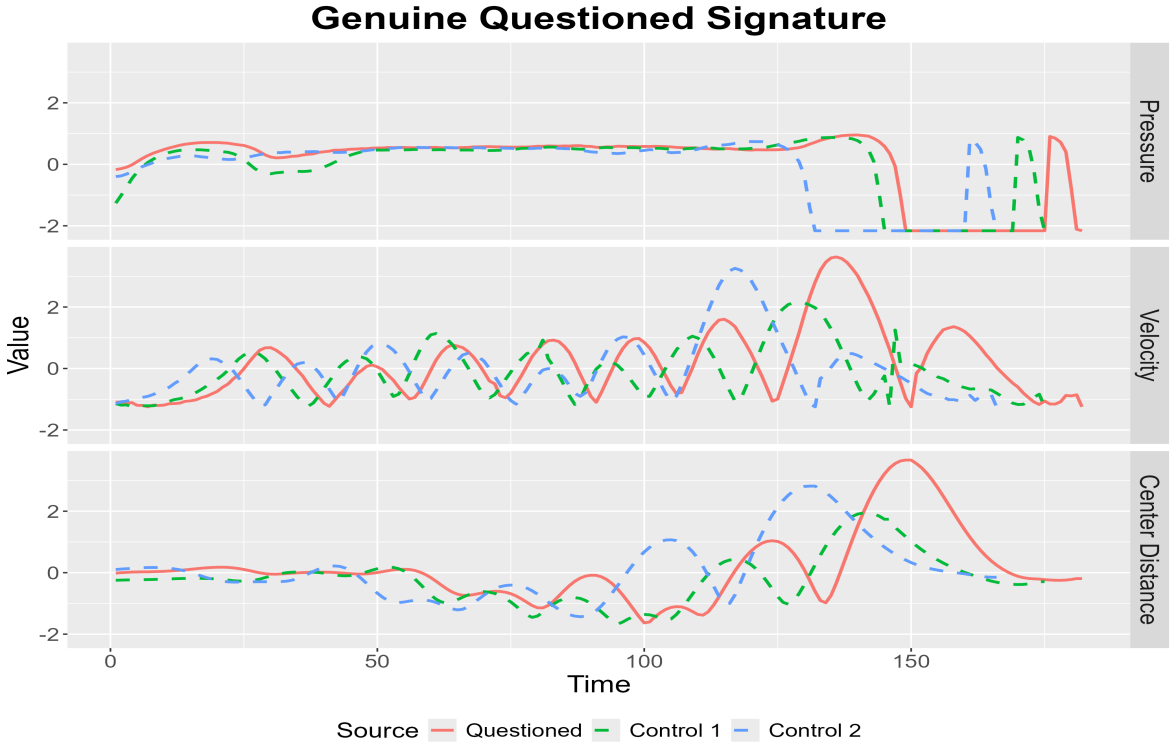


Figure 5.6: Time series of Pressure, Velocity, and Distance (from the trajectory center) for three signatures originating from the same author; First signature (red) serves as the genuine questioned signature while the two (green and blue) as control signatures.

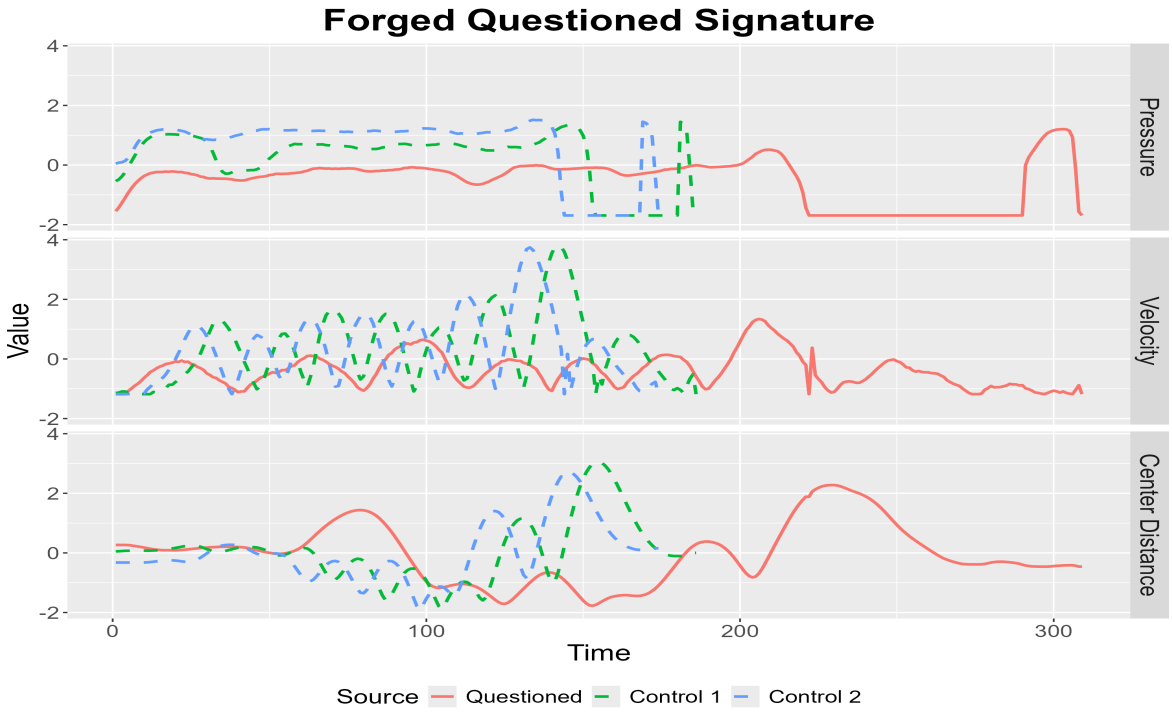


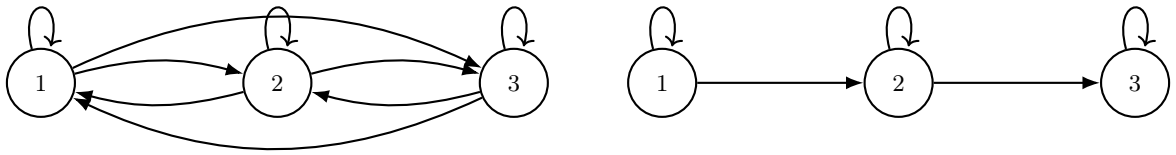
Figure 5.7: Time series of Pressure, Velocity, and Distance (from the trajectory center) for a forged signature and two control signatures; The control signatures originate from the putative author of the questioned signature.

5.4 Gaussian Hidden Markov Model

An HMM (Baum and Petrie, 1966) uses a doubly stochastic process governed by an underlying Markov chain with a finite number of states. Each state is associated with a random function. At each time point, the process is in one of the states. Given the current state, it generates an observation according to the random function corresponding to this state. The model is 'hidden' because only the sequence of response values is observed and recorded, while the underlying state that generates each dynamic feature value remains unobserved.

An HMM may have a variety of different transition structures (Bakis, 1976). A typical structure is the Ergodic HMM, in which all states are fully connected, as illustrated in the three-state example in Figure 5.8(a). However, it is possible to constrain an HMM such that only certain desired state transitions are allowed. An example of a Left-to-Right HMM with three states is represented in Figure 5.8(b), where transitions to previous states are not allowed. At a discrete time instant t, the model stays in one of the states and generates an observation. At time instant $t+1$, the model either remains in the same state or moves to a new state, according to the state transition probabilities, which are assumed to be constant with respect to t (time-homogeneous). This process continues until a final terminating state is reached at time T. The model can generate any observation of each state governed by estimated probabilities. The model is initialized by specifying the probabilities of the system occupying each state at the starting point $t = 1$.

In the case of dynamic signatures, the literature often observes that a Left-to-Right HMM structure is more suitable than an Ergodic one. This is because generating handwriting and signatures is an inherently sequential process, where the pen movement starts at one point and progresses in a temporal



(a) Fully connected / Ergodic HMM (b) Left-to-right (Bakis) HMM

Figure 5.8: Common HMM transition structures.

order until the end of the stroke. Namely, a Left-to-Right HMM (a.k.a. Bakis structure) is ideal if the feature evolves monotonically or follows a natural order. For example, some features often have clear patterns and possible states can be:

"down (writing) → hold (pause writing) → up (raise your hand) → down (writing)"

sequence, which a Left-to-Right HMM can capture perfectly. However, an ergodic HMM might "waste" parameters on modeling back transitions which will never occur.

The choice between Left-to-Right and Ergodic structures depends on the type of features used. Left-to-Right structures are particularly effective when pressure-related features are incorporated, because they preserve the temporal progression of these dynamic signals. This temporal ordering enables the HMM to distinguish between genuine and forged signatures, as forgers often struggle to reproduce the natural flow of the original signer. For angle-oriented features, Ergodic HMM is typically more flexible and realistic. An Ergodic HMM can sometimes work better when highly oscillatory features are used, since states may need to recur.

For signer i and signature j consider an observed sequence of dynamic feature (e.g., velocity, acceleration, etc.) represented by the vector $\boldsymbol{x}_{ij,1:T} = (x_{ij1}, x_{ij2}, \ldots, x_{ijT})$ and the latent states by $\boldsymbol{z}_{ij,1:T} = (z_{ij1}, z_{ij2}, \ldots, z_{ijT})$, with $z_{ijt} \in \{1, \ldots, K\}$, where K denotes the number of hidden states which is not known a priori and is treated as a model selection problem. The HMM model formulation for the time point t is given by:

$$\begin{aligned} Z_{ij1} &\sim \text{Multinomial}(1, \boldsymbol{\pi}), \\ Z_{ijt} \mid Z_{ij,t-1} &\sim \text{Multinomial}\big(1, A_{Z_{ij,t-1}}\big), \qquad t = 2, \ldots, T_{ij}, \\ X_{ijt} \mid Z_{ijt} = k &\sim N(\mu_k, \sigma_k^2), \qquad k = 1, \ldots, K \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the vector of the initial state probabilities $\pi_k = P(Z_{ij1} = k)$, $A_\ell = (a_{\ell 1}, \ldots, a_{\ell K})$ is the transition matrix with $a_{\ell k} = P(Z_{ijt} = k \mid Z_{ij,t-1} = \ell)$ *, and $\{(\mu_k, \sigma_k^2)\}_{k=1}^K$ are the Gaussian emission parameters for each state.

Hence, the joint distribution of states and observations factorizes as

$$\begin{aligned} f(\boldsymbol{x}_{ij,1:\widetilde{T}}, \boldsymbol{z}_{ij,1:\widetilde{T}} \mid \Theta) = \pi_{z_{ij1}}\, N(x_{ij1} \mid \mu_{z_{ij1}}, \sigma^2_{z_{ij1}}) \\ \times \prod_{t=2}^{\widetilde{T}} a_{z_{ij,t-1},\, z_{ijt}}\, N(x_{ijt} \mid \mu_{z_{ijt}}, \sigma^2_{z_{ijt}}), \end{aligned} \tag{5.3}$$

where

$$\Theta = (\boldsymbol{\pi}, \boldsymbol{A}, \{(\mu_k, \sigma_k^2)\}_{k=1}^K)$$

*For a Left-to-Right HMM, the transition vector of state ℓ is $A_\ell = (a_{\ell\ell}, a_{\ell,\ell+1}, \ldots, a_{\ell K})$, where $a_{\ell k} = P(Z_{ijt} = k \mid Z_{ij,t-1} = \ell)$, $a_{\ell k} = 0$ for $k < \ell$

denotes the full set of model parameters; where $A = (a_{\ell,k} \;\; \ell, k \in \{1, \ldots, K\})$.

For the estimation of the parameter vector Θ, we apply the Maximum Likelihood Approach, which is based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), also known as the Baum–Welch algorithm (Baum and Petrie, 1966) for HMMs. In the HMM model, the EM algorithm repeatedly performs two steps: first, it infers the most probable state sequence (E-step); then, it refines the model parameters to better fit the data (M-step). This process is iterated until convergence. Specifically,

- **E-step:** Compute the posterior distributions over hidden states given the current parameter estimates and the observations, typically using the forward–backward algorithm (Baum and Petrie, 1966).
- **M-step:** Update the parameter estimates by maximizing the expected complete-data log-likelihood, where the expected values are taken with respect to the posteriors from the E-step.

For a more detailed mathematical explanation, see Section 5.4.1. In our experiments, we use the *R* library *depmixS4* (Visser and Speekenbrink, 2010).

5.4.1 Maximum Likelihood Approach

A maximum likelihood estimate (MLE) of Θ maximizes

$$P(\boldsymbol{x}_{1:T} \mid \Theta) = \sum_{\mathbf{z}_{1:T}} P(\boldsymbol{x}_{1:T}, \mathbf{z}_{1:T} \mid \Theta).$$

Since direct maximization is intractable due to the sum over K^T latent trajectories, the Expectation-Maximization (EM) algorithm Section 3.1.2, also called the Baum-Welch algorithm Baum and Petrie (1966), is used. It iteratively applies:

E-step: The forward-backward algorithm to compute the posterior distribution of the hidden states given observations and current parameters. The forward probabilities are defined as:

$$\alpha_t(k) = P(\boldsymbol{x}_{1:t}, z_t = k \mid \Theta),$$

and calculated recursively by

$$\alpha_1(k) = \pi_k \, b_k(\boldsymbol{x}_1), \quad \alpha_t(k) = \sum_{j=1}^{K} \alpha_{t-1}(j) a_{jk} \, b_k(\boldsymbol{x}_t), \quad t = 2, \ldots, T,$$

where $b_k(\boldsymbol{x}_t) = P(\boldsymbol{x}_t \mid z_t = k)$ is the emission probability (Gaussian density), and the backward probabilities are defined as:

$$\beta_t(k) = P(\boldsymbol{x}_{t+1:T} \mid z_t = k, \Theta),$$

and computed recursively as

$$\beta_T(k) = 1, \quad \beta_t(k) = \sum_{j=1}^{K} a_{kj} \, b_j(\boldsymbol{x}_{t+1}) \, \beta_{t+1}(j), \quad t = T-1, \ldots, 1.$$

The posterior state occupancy probabilities are then given by

$$\gamma_t(k) = P(z_t = k \mid \boldsymbol{x}_{1:T}, \Theta) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{j=1}^{K} \alpha_t(j)\beta_t(j)}.$$

M-step: The parameters are updated by maximizing the expected complete-data log-likelihood with the posteriors $\gamma_t(k)$ serving as weights:

$$\widehat{\pi}_k = \gamma_1(k),$$

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$

where $\xi_t(i,j) = P(z_t = i, z_{t+1} = j \mid \boldsymbol{x}_{1:T}, \Theta)$ is computed from α, β and model parameters. The updated Gaussian parameters are:

$$\widehat{\mu}_k = \frac{\sum_{t=1}^{T} \gamma_t(k)\boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(k)}, \quad \widehat{\sigma}_k = \frac{\sum_{t=1}^{T} \gamma_t(k)(\boldsymbol{x}_t - \widehat{\mu}_k)^2}{\sum_{t=1}^{T} \gamma_t(k)}.$$

The recursions for forward and backward probabilities are derived using the Markov property and conditional independence assumptions. They enable efficient dynamic programming solutions to marginalize over exponentially many hidden state paths, allowing exact posterior inference and parameter learning via EM (Rabiner, 2002). In our experiments we are using the implementation of *R* library *depmixS4* Visser and Speekenbrink (2010).

5.5 Trajectory Resampling

A dynamic signature can be described by a sequence of pen coordinates (X_t, Y_t), time stamps τ_t, and pressure values P_t, and can be denoted by $\{(X_t, Y_t, P_t, \tau_t)\}_{t=1}^{T}$. As can be seen in Figures 5.6 and 5.7, these sequences often differ in length across samples. To enable comparison and modeling (here using HMMs), each trajectory must be resampled to a fixed number of points, $\widetilde{T}$. The resampling is performed using cumulative arc-length parameterization. The procedure is summarized by the following three steps and is given in more detail in Appendix B.1:

1. **Arc Length Computation:** Calculate the cumulative path length along the signature trajectory.

2. **Normalization:** Normalize the arc length to a unit interval to obtain a standardized trajectory parameter.

3. **Interpolation:** Interpolate values (e.g.,$\{(X_t, Y_t, P_t, \tau_t)\}_{t=1}^{T}$) at evenly spaced points along the normalized trajectory.

The resulting resampled trajectory is denoted by

$$\left\{\left(\widetilde{X}_t, \widetilde{Y}_t, \widetilde{P}_t, \widetilde{\tau}_t\right)\right\}_{t=1}^{\widetilde{T}},$$

with uniform arc length parameterization and fixed length $\widetilde{T}$. Put simply, we shrink the signature so that all trajectories have a uniform length. Then, we sample evenly spaced points to measure values like position, pressure, and time, making all signatures comparable; for the detailed explanation, see Appendix B.1.

This ensures that all signatures have the same dimensionality, which is crucial for subsequent modeling and likelihood evaluation. For illustration, Figure 5.9 depicts the case where the resampling

is set to $\widetilde{T} = 100$. In this case, the overall shape remains largely unchanged and exhibits only a slightly increased sharpness in certain regions.

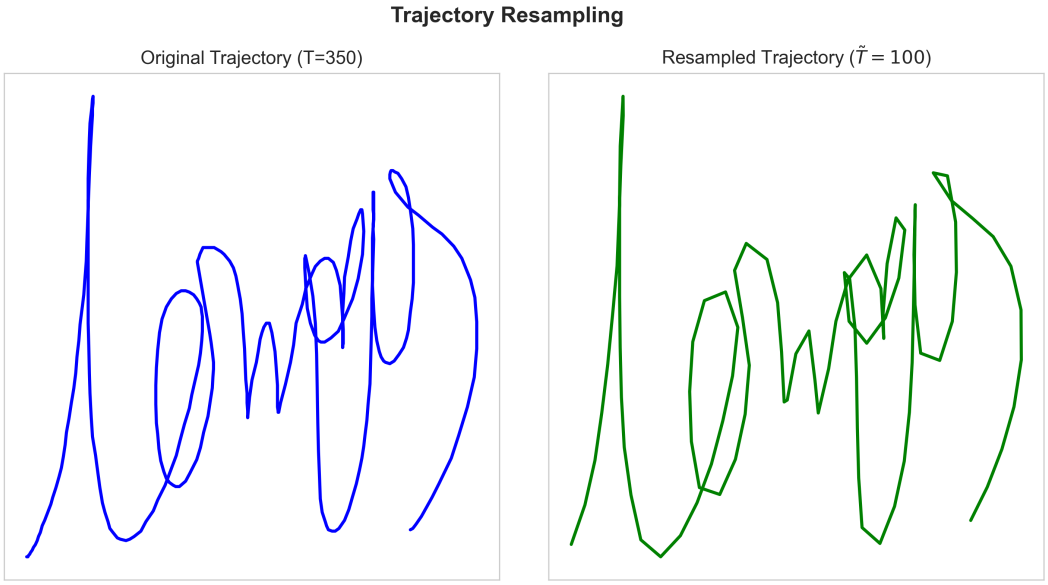


Figure 5.9: Trajectory resampling example.

Furthermore, the dynamic features were normalized according to the total time duration of each signature. Finally, feature values were standardized based on the mean and the standard deviation of the case-unrelated genuine signatures (see Section 5.2) to facilitate comparability and model fitting.

5.6 Goodness-of-Fit

This section examines the goodness-of-fit of Gaussian Hidden Markov Models applied to dynamic signature measurements. We assess the normality of within-state emission distributions through pseudo-residual diagnostics, illustrate representative fitted models with their estimated emission parameters, and discuss the convergence behaviour of the Baum-Welch algorithm under two competing topologies: Ergodic and Left-to-Right.

5.6.1 Pseudo-Residual Diagnostics

For each hidden state, the HMM assumes that observations follow a state-dependent Gaussian emission distribution:

$$X_{ijt} \mid Z_{ijt} = k \sim N(\mu_k, \sigma_k^2), \qquad k = 1, \ldots, K \tag{5.4}$$

where μ_k and σ_k^2 are estimated via the Baum-Welch algorithm.

A standard approach to assessing the goodness-of-fit of HMMs is based on *pseudo-residuals* (Zucchini et al., 2017). The central idea exploits the *Probability Integral Transform*, namely for any continuous random variable X with cumulative distribution function F, the transformed variable $U = F(X)$ follows a uniform distribution on $(0, 1)$. Applying the inverse standard normal CDF Φ^{-1} to U then yields a quantity that is standard normally distributed, $r = \Phi^{-1}(F(X)) \sim N(0, 1)$, provided that F is correctly specified.

In the context of HMMs, the emission distribution of the active state plays the role of F. Specifically, for each observation x_{ijt} the pseudo-residual is defined as

$$r_{ijt} \;=\; \Phi^{-1}\big(F_{z_{ijt}}(x_{ijt})\big)\,, \tag{5.5}$$

where $F_{z_{ijt}}$ denotes the CDF of the emission distribution of the state z_{ijt} at time t. If the model is correctly specified, the sequence $\{r_{ijt}\}_{t=1}^{\widetilde{T}}$ should behave as an independent sample from $N(0,1)$, regardless of the particular emission family assumed. Departures from normality in $\{r_{ijt}\}$ therefore directly reflect misspecification of the emission distributions. In practice, the latent states are replaced by their Viterbi [2] estimates (Viterbi, 1967) , and normality of the resulting pseudo-residuals is assessed visually through a normal Q–Q plot.

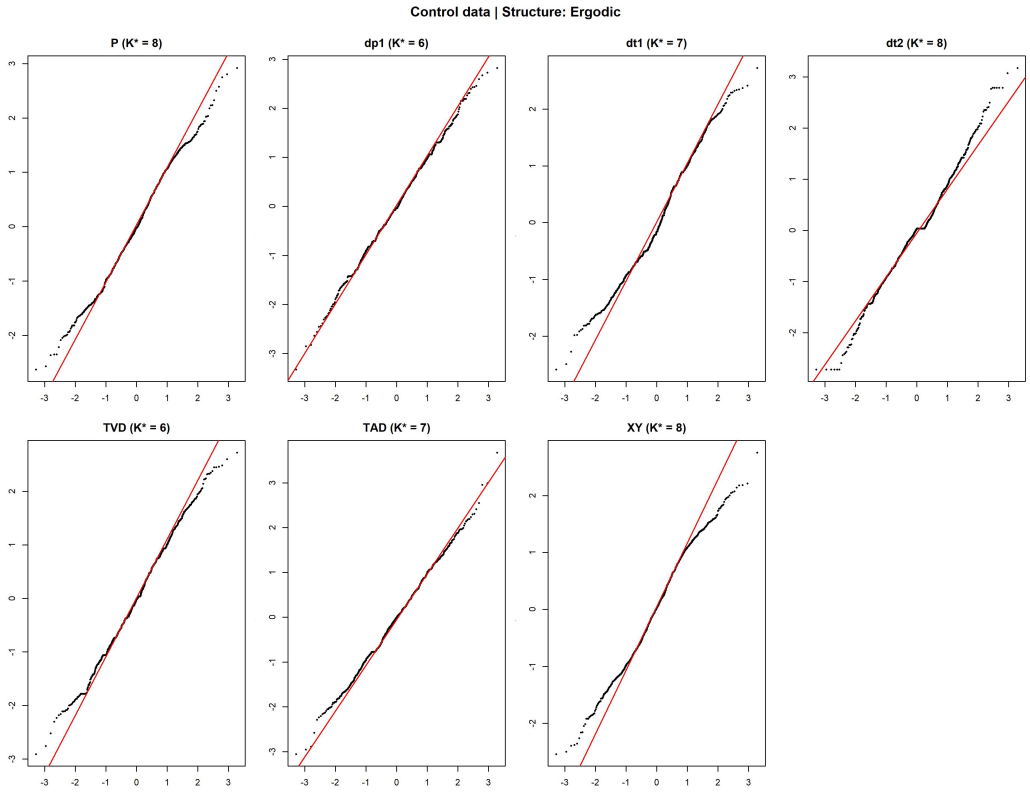
The Q–Q plots Figure 5.10 of the pseudo-residuals of some random selected signatures for the signature 1 provide a diagnostic of the Gaussian emission assumption for each feature. Overall, the fits are satisfactory, notably $dp1$ (differential of pressure), $dt1$ (velocity), TVD (trajectory angle) and TAD (acceleration direction). The empirical quantiles track the theoretical normal line closely throughout the central body of the distribution, indicating that the Gaussian emission model captures the within-state variability well. However, systematic deviations are visible in the tails across all features. The lower-left tails consistently fall below the reference line, while the upper-right tails pull above it, producing a characteristic S-shaped or positively skewed departure pattern. This is most pronounced for P (pressure) and XY (distance from the centroid), suggesting the presence of mild heavy tails or residual skewness not fully absorbed by the Gaussian states. These diagnostics suggest that the Gaussian HMM provides a reasonable first-order approximation, but that a heavier-tailed emission family, such as the Student's t distribution could potentially improve tail fit.

The Q–Q plots for the background simulated measurements of signature 2 and 3 are illustrated in Figure 5.11. The Q–Q plots reveal a noticeably different and more complex pattern compared to the control data. The most immediate observation is that the optimal K is substantially larger across all features reflecting the greater variability inherent in background forgers, whose signatures span a much wider population of writing styles and motor patterns. Despite the higher model complexity, tail deviations are more pronounced and consistent than in the control case. A characteristic sigmoid curvature is visible across nearly all features. This pattern is particularly marked for P (pressure), $dt2$ (acceleration) and XY (distance from the centroid), where the lower tail departs sharply from the reference, suggesting the presence of extreme observations that the Gaussian emission model. The features TVD (trajectory direction), TAD (acceleration direction) and $dt1$ (velocity) show a more moderate deviation.

Similar results are obtained for signatures 2 and 3 and their background simulated data, see Appendix B.2.1. Overall, these diagnostics suggest that while the Gaussian HMM provides a workable approximation for the Ergodic structure. However, the analysis motivates either a heavier-tailed emission family (e.g., Student's t) or a mixture-emission HMM as a potential improvement. Nevertheless, the overall linear trend in the central body of each Q–Q plot confirms that the Gaussian model captures the dominant structure of the data adequately for inference purposes in primary study.

On the other hand, this is not the case for the Left-to-Right structure, for which the Gaussian emission assumption appears less appropriate, the pseudo-residual Q–Q plots, provided in Appendix B.2.2, reveal more noticeable deviations from normality across the majority of features. Nevertheless, the Left-to-Right structure is considered in the analysis due to rational reasons. Human signature has

[2] Viterbi algorithm is a dynamic programming procedure that computes the globally most probable hidden state sequence given the observed values and fitted HMM.



The optimal number of states K^* is indicated in each panel title.

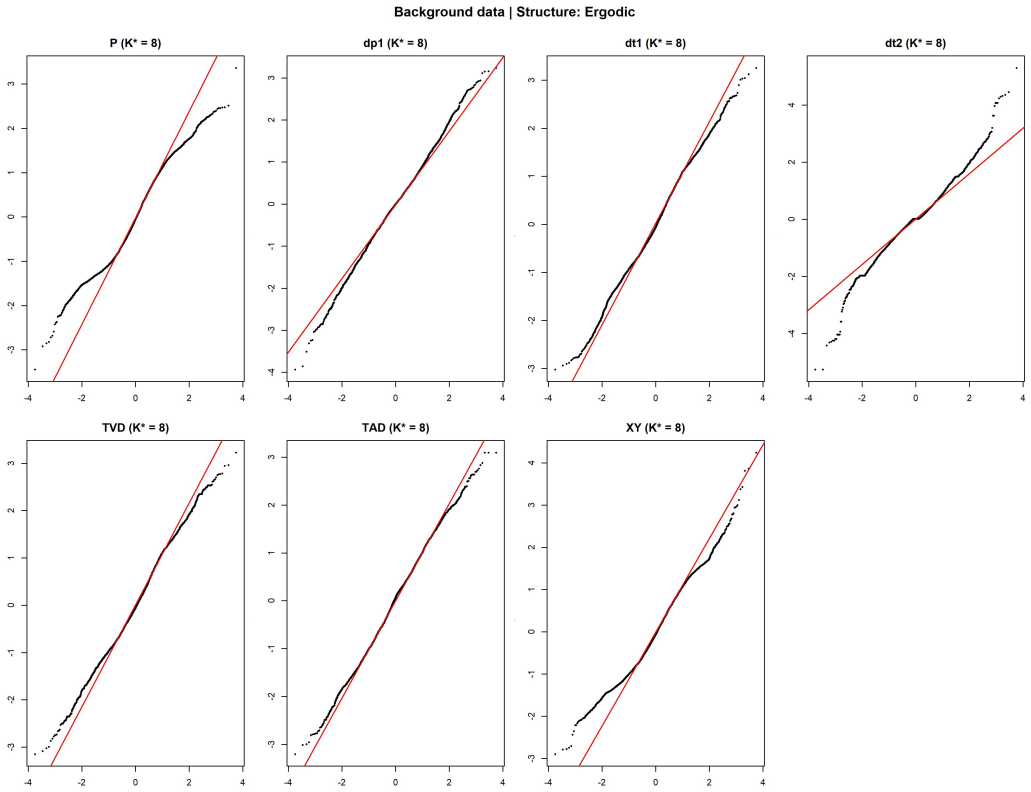
Figure 5.10: Normal Q–Q plots of the pseudo-residuals obtained from the Ergodic Gaussian HMM fitted to the signature 1 measurements, for each considered feature.

a sequential and directional process, many writers execute pen strokes in a consistent temporal order across repetitions, progressing from an initial state to a terminal one. The Left-to-Right HMM captures this directly (i.e. state 1 captures the starting dynamics of the signature, while State K models its conclusion. This contains a theoretically meaningful constraint that the Ergodic model does not explicitly represent. Accordingly, both structure are considered in this work, with the Ergodic structure serving as a flexible benchmark and the Left-to-Right structure providing an process-grounded alternative. Further research should investigate alternative model assumptions that provide a better fit to the observed signature dynamics, moving beyond the Gaussian emission constraint toward heavier-tailed or more flexible alternatives, such as the Student's t distribution, Gaussian mixture emissions, or the Generalised Normal distribution.

5.6.2 Estimated Parameters

In this section we present some examples of the estimated parameters of HMMs for both structures. Basically, Figures 5.12 and 5.13 display the fitted HMMs for signer 1 under the Ergodic and Left-to-Right structures, for the features pressure (P) and velocity ($dt1$). The Viterbi state estimates are illustrated on together with observed measurements (points), with the coloured band at each time point representing a standard deviation interval around the state-specific emission mean. The right margin of each panel reports the maximum likelihood estimates of the Gaussian emission parameters for each state k, together with the mean marginal state probability averaged over all training signatures.

Under the Ergodic structure Figure 5.12 for $K = 4$ states of P and $dt1$ indicates that the Ergodic

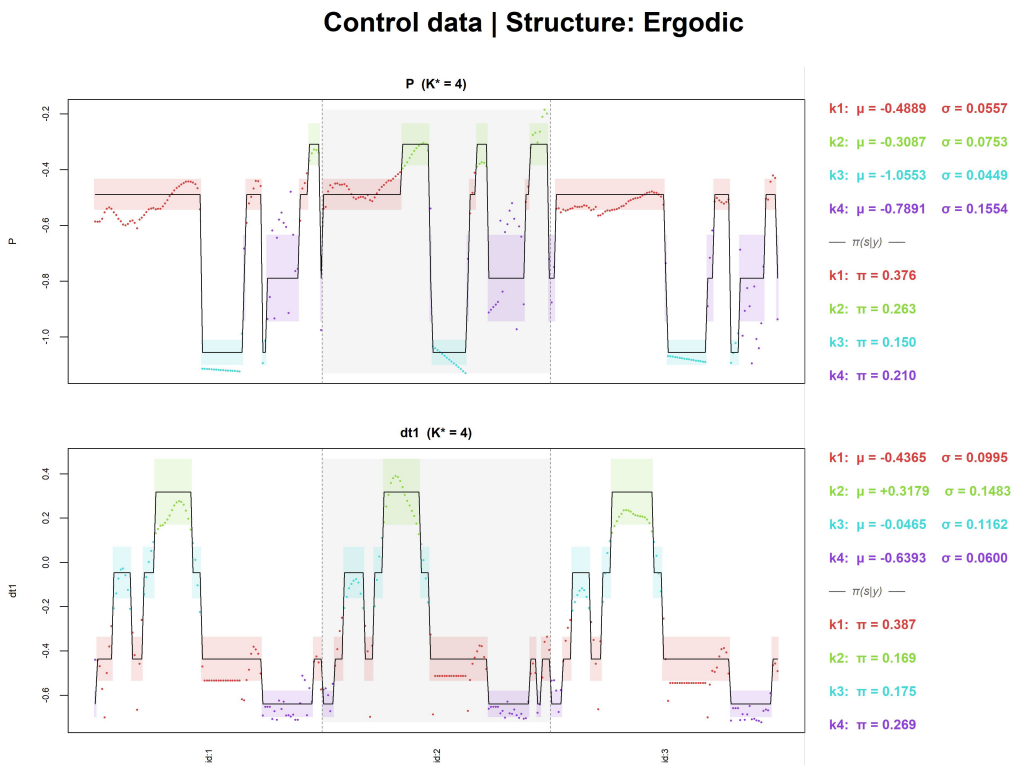


The optimal number of states K^* is indicated in each panel title.

Figure 5.11: Normal Q–Q plots of the pseudo-residuals obtained from the Ergodic Gaussian HMM fitted to the background simulated measurements of signatures 2 and 3, for each considered feature.

model splits the pressure signal into well-separated and homogeneous states. Notably for pressure (P), state $k = 3$ captures the low pressure during the signature process. The marginal probabilities reveal that the state is $k = 1$, reflecting the dominant proportion ($\pi = 0.376$) of time the writer spends in the characteristic mid-pressure. For velocity ($dt1$) under the Ergodic structure, the four states recover a symmetric structure around zero. States $k = 2$ and $k = 4$ represent the fast and slow phases respectively, while $k = 1$ and $k = 3$ cover intermediate motion. The dominance of $k = 1$ ($\pi = 0.387$) confirms that moderately velocity values, corresponding to the writer's characteristic pattern. The visual correspondence between the Viterbi state estimates and the observed measurements is strong across all three displayed signatures, becuase the black step function closely tracks the local mean of the observations, with most data points falling within the shaded bands. The high degree of cross-signature consistency in state assignments further suggest stability of the Ergodic model for this signer.

Under the Left-to-Right structure Figure 5.13 for $K = 4$ states and features P and $dt1$, the structure differs markedly. For P the ordering roughly reflecting the temporal progression of pen pressure from onset to body to lift-off. However, state $k = 4$ exhibits a substantially larger standard deviation ($\sigma_4 = 0.301$) compared to the ergodic counterpart, indicating that the Left-to-Right constraint forces one state to absorb a heterogeneous collection of pressure levels that the ergodic model would distribute across multiple fine-grained states. The Left-to-Right model thus describes the signature as spending more than half of its duration in a single broad pressure state, which is a direct consequence of the Left-to-Right constraint. For $dt1$, the picture is even more pronounced where $k = 4$ dominates $\pi = 0.827$, meaning more than 80% of the signal is attributed to a single broad state. The remaining states together cover only the remaining 17.3% of the time, and correspond to brief kinematic events, onset



Right margin: estimated emission parameters (μ_k, σ_k) and marginal probabilities π_k. Coloured bands: $\pm\sigma_k$ intervals. Black line: Viterbi fitted mean.

Figure 5.12: Ergodic HMM fitted for signer 1 for features P (pressure) and dt1 (velocity) for states $K^* = 4$. Three signatures shown.

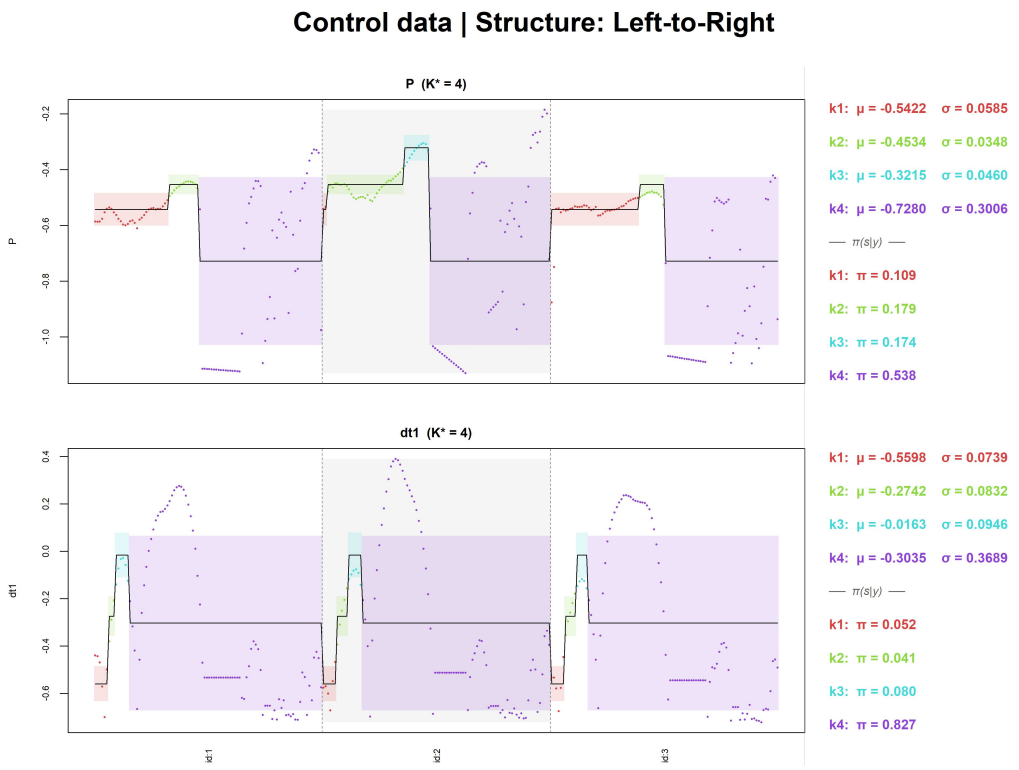
deceleration, the upstroke peak, and the short accelerating phase, that the Left-to-Right constraint isolates into the early temporal segments. The Viterbi path (Figure 5.13) is visually flatter than its ergodic counterpart, with the dominant state persisting across the majority of each signature's duration.

These results highlight a fundamental structural difference between the two structures. The Ergodic model behaves freely allocates states wherever the emission distribution best explains the data, regardless of temporal order, resulting in multiple, well-separated, low-variance states. The Left-to-Right model, by contrast, behaves as a temporal segmenter: it enforces a monotone progression through states, which compels it to concentrate most of the signal mass into a small number of broad, high-variance states that span multiple events. This structural makes the Left-to-Right model less descriptively accurate within any single signature, but more robust to the natural variability that exists across repetitions of the same signature, a property that is consequential for the forensic discrimination task, as discussed in Section 5.4

5.6.3 Convergence

Parameters of HMMs are estimated via the EM algorithm, specifically the Baum-Welch algorithm (Baum and Petrie, 1966), which iterates between computing the expected sufficient statistics of the hidden state sequence (E-step) and updating the model parameters to maximize the expected complete-data log-likelihood (M-step), as we discuss in Section 5.4.

For Ergodic HMMs, it is well established that the Baum-Welch algorithm converges to a stationary



Right margin: estimated emission parameters (μ_k, σ_k) and marginal probabilities π_k. Coloured bands: $\pm\sigma_k$ intervals. Black line: Viterbi fitted mean.

Figure 5.13: Left-to-Right HMM fitted for signer 1 for features P (pressure) and dt1 (velocity) for states $K^* = 4$. Three signatures shown.

point of the observed-data likelihood, in the sense that each iteration is guaranteed to produce a non-decreasing sequence of likelihood values (Dempster et al., 1977; Wu, 1983). However, this guarantee is one of local convergence only, the likelihood surface of an HMM is generally non-concave and may contain multiple local maxima, so the algorithm is sensitive to the choice of initial values (Rabiner, 2002). In the `depmixS4` implementation in R used in this work (Visser and Speekenbrink, 2010), convergence is declared when the relative change in the log-likelihood between successive iterations falls below a pre-specified tolerance (set here to 10^{-8}), subject to a maximum of 500 iterations.

The situation is materially different for the Left-to-Right structure, where structural constraints are imposed by fixing the transition matrix to be upper-triangular. This hard assumption of the parameter space introduces two specific convergence problems. First, the constrained likelihood surface tends to exhibit more local optima, because the upper-triangular restriction eliminates entire regions of the parameter space that might provide a path to the global maximum. Second, and more fundamentally, the theoretical guarantee of monotone likelihood improvement in EM rests on the smoothness of the complete-data log-likelihood with respect to the transition parameters (Wu, 1983), the fixed structural zeros in the Left-to-Right transition matrix introduce boundary constraints that make the objective function non-smooth at those boundaries, thereby invalidating the standard convergence proof. As a consequence, unlike the Ergodic topology for which convergence to a stationary point is assured, the Baum-Welch algorithm applied to a Left-to-Right constrained HMM carries no formal convergence guarantee, namely the algorithm may terminate at a non-stationary point or at a boundary of the parameter space.

Furthermore, as we mention the choice of initial parameter values plays a critical role in determining the quality of the solution returned by the Baum-Welch algorithm. The different starting configurations of the initial state distribution $\boldsymbol{\pi}^{(0)}$, transition matrix $\mathbf{A}^{(0)}$, and emission parameters may lead the algorithm to substantially different local optima, with correspondingly different interpretations of the latent state structure (Rabiner, 2002). Standard practical applications include random multiple restarts, k-means pre-clustering of the observations to obtain emission parameter initialisations, and perturbation of a deterministic starting point. The sensitivity to initial values is not symmetric across the two topologies considered in this work. For the Ergodic topology, the unconstrained transition matrix allows the EM algorithm considerable freedom to move through the parameter space from any starting configuration (Rabiner, 2002). For the Left-to-Right topology, the situation is considerably more sensitive. The upper-triangular constraint on the transition matrix means that the ordering of states at initialisation is consequential and then the constrained parameter space contains sharper local optima and non-smooth boundaries, as discussed above, the variance across restarts is larger for Left-to-Right model than for the Ergodic model. These considerations motivate the use of multiple random initialisations in practice (Rabiner, 2002), and they provide a principled explanation for the empirical observation discussed in Section 5.7, that the Left-to-Right structure consistently selects a smaller optimal number of states K than the Ergodic model, reflecting the reduced expressive capacity available within the constrained parameter space.

These structural differences in convergence issues and sensitivity to initial values constitute an additional argument for treating the Ergodic and Left-to-Right topologies as genuinely distinct inferential procedures rather than merely two parameterisations of the same model.

5.7 Experiments

In this section, we present an extensive experimental analysis based on the case framework described at the beginning of Chapter 5. The dataset was introduced in Section 5.2, and the Gaussian HMM methodology was outlined in Section 5.4. The experiments explore the impact of varying the HMM structure, the number of hidden states, and the different dynamic features. Specifically, we compare two HMM topologies, the Left-to-Right and the Ergodic structures (see Figure 5.8), while systematically varying the number of hidden states from 1 to 8. Each configuration is evaluated both for individual signatures and for their combination, enabling a comprehensive assessment of model performance across different materials and parameter settings.

The features considered suitable for this analysis can be discriminated into four main categories: (a) pressure-oriented features (pressure and the first derivative of pressure), (b) kinematic features (tangential velocity and acceleration), (c) spatial features (distance from the trajectory center) and (d) angle-oriented features (trajectory vector angle relative to the horizontal axis and acceleration direction relative to the horizontal axis). There are seven features in total: $p = 7$. All features were extracted after applying the trajectory resampling procedure described in Section 5.5, ensuring that all signatures contain an equal number of points, namely $\widetilde{T} = 100$.

To assess the performance of the proposed methodology, the questioned signature is evaluated under two competing hypotheses: either it is genuine, i.e., originating from the person of interest (PoI) (H_1), or it is forged, i.e., produced by someone other than the PoI (H_2). Consequently, in the experiments, the questioned signature comes either from the case-related genuine signature dataset $\mathcal{G}$ or from the case-related forged (simulated) signature dataset $\mathcal{F}$ (see Section 5.2).

5.7.1 Univariate Experiments

The experiments are conducted using the data sets described in Section 5.2 and the questioned signature is evaluated under two competing hypotheses: H_1 and H_2. Specifically, in the experiments, the questioned signature comes either from the case-related genuine signature dataset $\mathcal{G}$ or from the case-related forged (simulated) signature dataset $\mathcal{F}$ (see Section 5.2). For control data $\{\boldsymbol{x}_{j,1:\widetilde{T}}\}_{j=1}^{n}$, we use a set of $n = 10$ signatures from the case-related genuine dataset $\mathcal{G}$. For the background forged population, denoted $\{\mathcal{B}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$, we use signatures from the case-related forged dataset $\mathcal{F}$ that are unrelated to the case under study. For example, if the questioned signature corresponds to signature 1, the background forged dataset $\{\mathcal{B}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ is constructed from signatures 2, 3, etc., excluding repetitions of the signature under examination. This protocol was repeated across all experimental settings, resulting in the generation of over 283,000 distinct cases. In these experiments which the univariate framework is validated, the following LR is assessed:

$$LR = \frac{f(\boldsymbol{y}_{1:\tilde{T}}|\widehat{\Theta}_{\{\boldsymbol{x}_{j,1:\tilde{T}}\}_{j=1}^{n},H_1)}}{f(\boldsymbol{y}_{1:\tilde{T}}|\widehat{\Theta}_{\{\mathcal{B}_{j,1:\tilde{T}}\}_{j=1}^{N_{\mathcal{B}}},H_2)}} \tag{5.6}$$

where $f(\cdot)$ represents the HMM's probability density function (see Eq. 5.3), $\boldsymbol{y}_{1:\widetilde{T}}$ denotes the feature values of the questioned signature, $\widehat{\Theta}$ the estimated model parameters based on the $\{\boldsymbol{x}_{j,1:\widetilde{T}}\}_{j=1}^{n}$ feature values of n control signatures and $\{\mathcal{B}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ feature values of $N_{\mathcal{B}}$ forged signatures unrelated with the case.

To assess the performance of the proposed methodology, we therefore compute (a) the rate of false positives (i.e., the rate of cases where the signature is forged but the LR is greater than 1 and falsely supports hypothesis H_1); (b) the rate of false negatives (i.e., the rate of cases where the signature is a genuine signature from the PoI but the LR is smaller than 1 and falsely supports hypothesis H_2); and (iii) the total error rate given by the summation of the false positive rate and the false negative rate.

Table 5.12 and Figure 5.14 show the total error rates (TER) from the dynamic signature evaluation experiments conducted using Hidden Markov Models (HMM) with two different structures: Left-to-Right and Ergodic. TER indicates the total proportion of misclassifications, which is equivalent to the sum of the false positive and the false negative rates. For the results for each signature (Signature 1 to Signature 3), see Appendix B.3 Tables B.3.1, B.3.2 and B.3.3.

For all signatures, Left-to-Right structures generally show more variability in TER across states and features, with some very low TER entries (e.g., 0.043 at 5 states for TVD) but also some high spikes (e.g., 0.606 at 5 states for XY). The number of states has a great impact on model performance, with the optimal number of states varying according to the feature and structure; this highlights the need for careful model tuning.

It is common in real HMM applications, the Left-to-Right structure needs fewer states to optimal fit the data, and the Ergodic structure needs more states, especially in biometric time series, as in our case. A Left-to-Right HMM (a.k.a. Bakis structure) only allows transitions to move forward or remain in the same state. This enforces a sequential progression through states, which is ideal if the feature evolves monotonically or in a natural order (e.g., pressure increasing and then stabilizing). For example, Pressure (P) or its differential ($dp1$) often has a clear "down (writing) $\rightarrow$ hold (pause writing) $\rightarrow$ up (raise your hand) $\rightarrow$ down (writing)" sequence. However, an Ergodic structure can perform better when the number of states grows. Many features may not evolve in a strict monotonic way. A rigid Left-to-Right structure forbids backward transitions, forcing the model into an unnatural sequence and affecting log-likelihood. On the other hand, Ergodic HMMs also allow loops and revisits,

which prove useful for modeling subtle sub-states and fitting the data more efficiently, as discussed in Section 5.6.2.

Our experimental results show that pressure achieves a TER of 9.1% with the Ergodic structure (7 states) and 10.1% with the Left-to-Right HMM structure (7 states). For the differential of pressure feature, the Ergodic structure with 8 states achieves the lowest TER of 13.8%, while the Left-to-Right HMM structure reaches its minimum TER of 14.9% with 5 states. For the spatial feature (distance from the trajectory centroid), the Ergodic structure with 8 states achieves a TER of 13.9%, whereas the Left-to-Right model reaches its minimum TER of 16.1% with 2 states. As far as the kinetic features are concerned, tangential velocity achieves a TER of 7.0% with 8 states under the Ergodic structure, whereas the Left-to-Right model reaches a minimum TER of 8.0% with 5 states. Tangential acceleration yields a minimum TER of 12.7% with 8 states under Ergodic, and 12.7% with 3 states under Left-to-Right. Finally, angle-oriented features show the strongest discriminative performance, the trajectory vector direction (TVD) achieves a TER of 2.7% with 8 states under the Ergodic structure, and 4.3% with 5 states under Left-to-Right. The trajectory acceleration direction (TAD) achieves a TER of 5.8% with 8 states under Ergodic, and 4.8% with 5 states under the Left-to-Right structure.

Furthermore, certain features consistently achieve lower TER values across both structures and states. For example, the feature trajectory vector angle to the horizontal line (TVD) often yields the lowest error rates (highlighted in red). This indicates TVD might be a particularly discriminative feature for signature evaluation. While individual signature tables show specific performance nuances, the combined table aggregates results and can be used to identify robustness and generalizability trends across all signatures. The combined Table 5.12 and Figure 5.14 presenting experimental results for all analyzed signatures, suggest relatively consistent behaviour with respect to the superiority of certain features like angle-oriented and kinetic-oriented features.

Performed analyses for these three signatures indicate that the Left-to-Right HMM structure requires fewer states to capture the sequential patterns compared to the Ergodic structure. Additionally, it appears that velocity, trajectory vector angle relative to the horizontal axis, and acceleration direction relative to the horizontal axis are the most discriminative features. Next section will explore the effectiveness of the combination of these three features into a multivariate HMM.

Structure	**States**	**Total Error Rate**						
		P	**dp1**	**dt1**	**dt2**	**TVD**	**TAD**	**XY**
Left-to-Right	1	0.138	0.441	0.477	0.682	0.507	0.565	0.224
	2	0.151	0.276	0.163	0.127	0.194	0.222	0.161
	3	0.199	0.210	0.058	0.161	0.131	0.133	0.436
	4	0.148	0.173	0.099	0.189	0.095	0.058	0.412
	5	0.183	0.149	0.080	0.262	0.043	0.049	0.606
	6	0.154	0.297	0.189	0.296	0.048	0.048	0.596
	7	0.101	0.334	0.136	0.341	0.071	0.066	0.433
	8	0.125	0.408	0.110	0.519	0.079	0.143	0.434
Ergodic	1	0.138	0.441	0.477	0.682	0.507	0.565	0.224
	2	0.215	0.229	0.132	0.627	0.377	0.083	0.255
	3	0.144	0.190	0.148	0.264	0.225	0.061	0.169
	4	0.101	0.162	0.129	0.235	0.123	0.072	0.187
	5	0.101	0.185	0.119	0.248	0.058	0.069	0.178
	6	0.095	0.162	0.112	0.247	0.063	0.057	0.166
	7	0.091	0.161	0.097	0.246	0.039	0.063	0.146
	8	0.097	0.138	0.070	0.231	0.027	0.058	0.139

Table 5.12: Experimental results of dynamic signature evaluation for different HMM model structures, numbers of states, and features for all signatures.

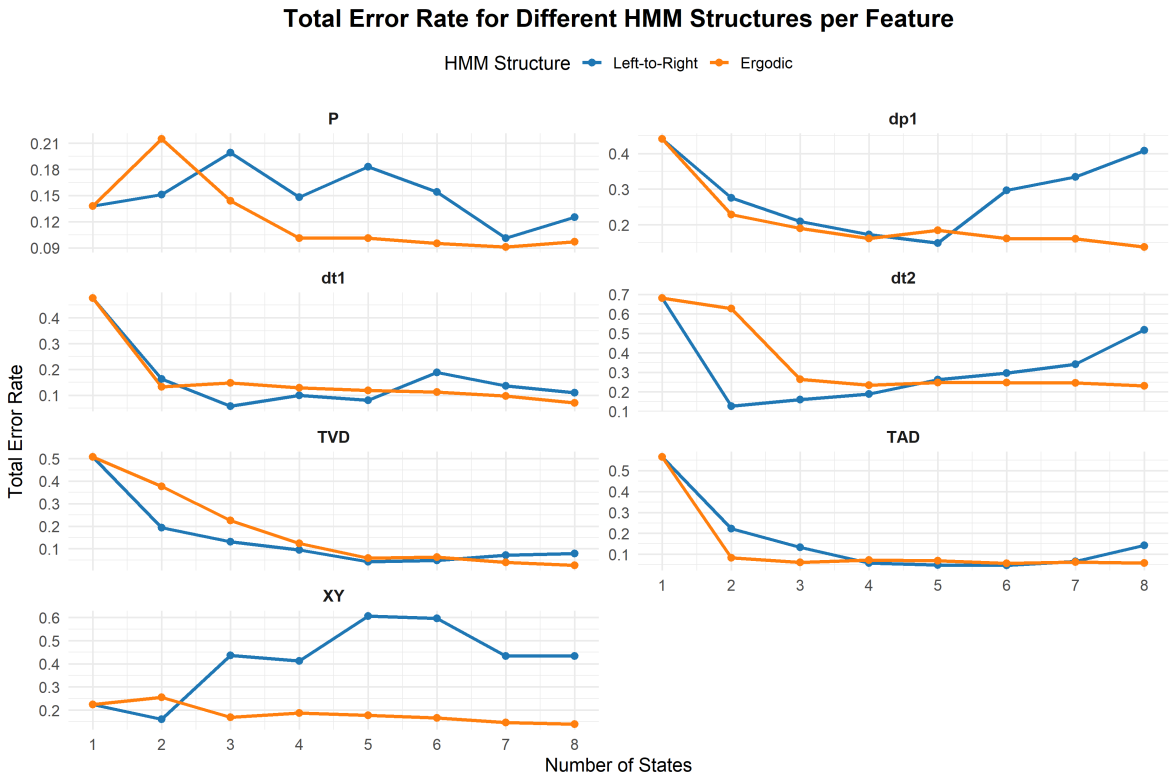


Figure 5.14: Line-plot of total error rate per number of hidden states, shown for different HMM model structures and feature across all signatures.

5.7.2 Multi-feature Experiments

In this section, we present a multi-feature approach that combines the features identified as the most discriminative in the univariate experiments of Section 5.7.1. These experiments highlight the importance of the velocity (dt1), the trajectory vector angle relative to the horizontal axis (TVD), and the acceleration direction relative to the horizontal axis (TAD) for dynamic signature evaluation.

Consequently, the observed sequence for signer i and signature j at time t can be defined as the vector

$$\boldsymbol{\mathcal{X}}_{ijt} = (\mathrm{dt1}_{ijt}, \mathrm{TVD}_{ijt}, \mathrm{TAD}_{ijt}) \in \mathbb{R}^3 \quad .$$

Assuming conditional normality, the emission distributions corresponding to the hidden state $z_{ijt} = k$ for each feature are specified by:

$$\mathrm{dt1}_{ijt} \sim N(\mu_{k,1}, \sigma^2_{k,1}), \quad \mathrm{TVD}_{ijt} \sim N(\mu_{k,2}, \sigma^2_{k,2}), \quad \mathrm{TAD}_{ijt} \sim N(\mu_{k,3}, \sigma^2_{k,3})$$

The combined emission probability for the observed vector $\boldsymbol{\mathcal{X}}_t$ given the hidden state $z_t = k$ is defined as the product of the univariate Gaussian densities:

$$f(\boldsymbol{\mathcal{X}}_{ijt} \mid z_{ijt} = k) = \prod_{d=1}^{3} N(x_{ijtd} \mid \mu_{k,d}, \sigma^2_{k,d}).$$

A key assumption in this formulation is that the observed features (dt1, TVD, and TAD) are conditionally independent given the hidden state, with each modeled by a univariate Gaussian distribution. Future work will explore the use of a multivariate Gaussian emission model to account for potential

correlations between features.

For the experiments, we proceed with the same protocol as in Section 5.7.1 with the trajectory resampling of the data as described in Section 5.5. Specifically, the questioned signature is evaluated under two competing hypotheses (H_1), or it is forged, i.e., produced by someone other than the PoI (H_2). Consequently, in the experiments, the questioned signature comes either from the case-related genuine signature dataset $\mathcal{G}$ or from the case-related forged (simulated) signature dataset $\mathcal{F}$ (see Section 5.2). For control data $\{\boldsymbol{\mathcal{X}}_{j,1:\widetilde{T}}\}_{j=1}^{n}$, we use a set of $n = 10$ signatures from the case-related genuine dataset $\mathcal{G}$. For the background forged population, denoted $\{\boldsymbol{\mathcal{B}}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$, we use signatures from the case-related forged dataset $\mathcal{F}$ that are unrelated to the case under study. For example, if the questioned signature corresponds to signature 1, the background forged dataset $\{\boldsymbol{\mathcal{B}}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ is constructed from signatures 2, 3, etc., excluding repetitions of the signature under examination.

The experiments were performed 300 times when the questioned signature originated from the genuine signature dataset $\mathcal{G}$. Analogous experiments were conducted when the questioned signature originated from the forged (simulated) dataset $\mathcal{F}$, with 260 iterations for Signature 1, 410 iterations for Signature 2, and 160 iterations for Signature 3 (see Section 5.2). In total, this resulted in more than 1700 unique cases.

For each of the 1700 case comparisons, we compare the two hypotheses using the LR, given by

$$LR = \frac{f(\boldsymbol{\mathcal{Y}}_{1:\widetilde{T}}|\widehat{\boldsymbol{\Theta}}_{\{\boldsymbol{\mathcal{X}}_{j,1:\widetilde{T}}\}_{j=1}^{n}}, H_1)}{f(\boldsymbol{\mathcal{Y}}_{1:\widetilde{T}}|\widehat{\boldsymbol{\Theta}}_{\{\boldsymbol{\mathcal{B}}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}}, H_2)}, \tag{5.7}$$

where $f(\cdot)$ represents the HMM's probability density function (see Eq. 5.3) for the model selected with the automated state selection procedure described in Section 5.7.3; the vector $\boldsymbol{\mathcal{Y}}_{1:\widetilde{T}}$ denotes the multi-feature values of the questioned signature, $\widehat{\boldsymbol{\Theta}}$ are the estimated parameters based on the $\{\boldsymbol{\mathcal{X}}_{j,1:\widetilde{T}}\}_{j=1}^{n}$ multi-feature values of $n = 10$ control signatures, and $\{\boldsymbol{\mathcal{B}}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ are the multi-feature values of $N_{\mathcal{B}}$ forged signatures unrelated to the case.

Table 5.13 presents the experimental results of dynamic signature evaluation using multi-feature HMM with differing model structures, Left-to-Right and Ergodic, by varying numbers of hidden states. The evaluation metrics reported include the False Positive Rate (FPR), False Negative Rate (FNR), and Total Error Rate (TER) for three separate signatures labeled Sig. 1, Sig. 2, and Sig. 3. The Ergodic structure for signatures 1 and 3, models with 8 and 6 states generally achieve the lowest TER values (highlighted in red), indicating superior performance for these two signatures. The Left-to-Right models lowest TER values for signatures 2, however tend to show increases in FNR, suggesting instability for specific cases. The Ergodic structure generally presents more balanced FPR and FNR values across states, with consistently low TER values, especially for Sig. 3, where the TER reaches near zero at 6 states. Furthermore, the Ergodic structure appears to require a larger number of states to reach peak performance, suggesting that increasing the maximum number of states beyond eight may yield further improvements in discrimination.

It can be observed that the differences in performance across signatures illustrate subject-specific variability, emphasizing the importance of model tuning per individual signature characteristics. In the following section, we present an automated HMM state selection based on the literature.

Structure	States	Sig. 1			Sig. 2			Sig. 3		
		FPR	FNR	TER	FPR	FNR	TER	FPR	FNR	TER
Left-to-Right	1	0.127	0.030	0.157	0.323	0.000	0.323	1.000	0.001	1.001
	2	0.262	0.000	0.262	0.143	0.000	0.143	0.000	0.009	0.009
	3	0.027	0.008	0.035	0.048	0.006	0.053	0.000	0.131	0.131
	4	0.004	0.018	0.022	0.040	0.007	0.048	0.000	0.124	0.124
	5	0.004	0.118	0.122	0.048	0.006	0.053	0.000	0.200	0.200
	6	0.000	0.096	0.096	0.031	0.010	0.041	0.000	0.091	0.091
	7	0.005	0.138	0.143	0.026	0.012	0.038	0.000	0.056	0.056
	8	0.006	0.090	0.096	0.012	0.009	0.021	0.000	0.094	0.094
Ergodic	1	0.127	0.030	0.157	0.323	0.000	0.323	1.000	0.001	1.001
	2	0.185	0.000	0.185	0.169	0.000	0.169	0.025	0.000	0.025
	3	0.154	0.000	0.154	0.090	0.003	0.093	0.006	0.005	0.012
	4	0.127	0.001	0.128	0.076	0.004	0.080	0.006	0.001	0.008
	5	0.108	0.003	0.111	0.069	0.003	0.072	0.006	0.000	0.006
	6	0.062	0.003	0.064	0.064	0.003	0.067	0.000	0.004	0.004
	7	0.058	0.004	0.062	0.052	0.006	0.058	0.000	0.007	0.007
	8	0.015	0.006	0.021	0.048	0.007	0.055	0.006	0.008	0.014

Table 5.13: Experimental results of dynamic signature evaluation with multi-feature HMM of different model structures and numbers of states.

5.7.3 Automated Selection of HMM's States

Automated state selection in HMM is commonly addressed using information criteria such as the Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Hannan-Quinn Information Criterion (HQC) (Hannan and Quinn, 1979). These approaches can determine a sufficient number of hidden states by balancing model fit and complexity. These methods are particularly advantageous when no prior information on the true number of hidden states is available. Simulation studies have shown that AIC generally outperforms other criteria, such as BIC, in scenarios with limited data or less persistent state sequences, although it may tend to slightly overestimate the optimal number of states as sample size grows (Costa and De Angelis, 2010). However, from a theoretical standpoint, BIC is the preferred criterion for HMM because it is strongly consistent as the time data points tends to infinity, it selects the true number of states with probability 1 (Csiszár and Shields, 2000), whereas other criteria does not have this consistency property (Fuh et al., 2024). In this section, focus on finding the best criteria between these for state selection in this context.

Selecting the number of states in an HMM is a critical step, as it strongly influences the ability of the model to capture the dynamics of the observed data. For each candidate number of states from 1-8 and for control measurements $\{\boldsymbol{\mathcal{X}}_{j,1:\bar{T}}\}_{j=1}^{n}$ and forged (simulated) measurements $\{\boldsymbol{\mathcal{B}}_{j,1:\bar{T}}\}_{j=1}^{N_{\mathcal{B}}}$, a HMM is fitted and its AIC, BIC, HQC value is computed according to:

$$AIC = -2\log L\left(\widehat{\boldsymbol{\Theta}}_{\{\boldsymbol{\mathcal{X}}_{j,1:\bar{T}}\}_{j=1}^{n}}\right) + 2p$$

$$BIC = -2\log L\left(\widehat{\boldsymbol{\Theta}}_{\{\boldsymbol{\mathcal{X}}_{j,1:\bar{T}}\}_{j=1}^{n}}\right) + p\log(n)$$

$$HQC = -2\log L\left(\widehat{\boldsymbol{\Theta}}_{\{\boldsymbol{\mathcal{X}}_{j,1:\bar{T}}\}_{j=1}^{n}}\right) + p\log(\log(n))$$

where $L(\widehat{\boldsymbol{\Theta}})$ is the maximized likelihood and p is the number of parameters in the model. We select the HMM states with the lowest AIC, BIC, and HQC values.

In the considered likelihood ratio framework, the numerator and denominator each consist of an HMM fitted to a distinct dataset, the control material $\{\boldsymbol{\mathcal{X}}_{j,1:T}\}_{j=1}^{n}$ and background forged population,

denoted $\{\boldsymbol{\mathcal{B}}_{j,1:\widetilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$, respectively. Since these datasets may differ in length, variability, and complexity, the optimal number of hidden states K may differ between the two models. Selecting K independently for each component does not introduce bias into the LR, provided that the same emission model and the same HMM structure are used in both. This is because the LR compares the probability of the evidence under two competing propositions.

However, the HMM transition structure must be kept consistent across numerator and denominator. This consistency is necessary because differing structures entail different transition constraints, which can bias the LR if raw likelihoods are compared directly. For example, if the numerator uses a Left-to-Right HMM with constrained transition matrix $\boldsymbol{A}$ and the denominator uses an ergodic HMM with a fully unconstrained $\boldsymbol{A}$, then the denominator can attain a higher maximum log-likelihood purely due to its greater flexibility, independently of the data. This inflates the denominator and deflates the LR artificially, introducing a systematic bias that does not reflect the strength of the evidence. Therefore, the transition structure is fixed to be identical in both components of the LR.

By allowing the state number to vary while maintaining the same model architecture, the resulting LR accurately reflects the relative support for the questioned signature originating from the genuine versus the background population, without confounding effects from model complexity or structural differences.

To assess which information criterion is most appropriate, experiments were conducted as described in Sections 5.7.1 and 5.7.2, using 10 genuine randomly selected control signatures and one questioned signature, which is either genuine or forged. Table 5.14 reports the mean and standard deviation of the log-likelihood ratio (logLR) values obtained from the dynamic signature evaluation experiments using multi-feature Hidden Markov Models (HMMs). Two HMM structures were compared: a Left-to-Right and an Ergodic. For each structure, the number of model states was selected according to three information criteria: AIC, BIC, and HQC. For the genuine questioned signatures, both model structures achieved consistently high positive logLR values, indicating strong evidence supporting genuine authorship. The Ergodic HMM produced substantially higher mean logLR values (≈ 374) than the Left-to-Right model (≈ 192). In contrast, for the forged questioned signatures, the Left-to-Right HMM produced substantially lower mean logLR values (≈ -438) than the Ergodic model (≈ -283), suggesting greater discriminative power for forged signatures. The small standard deviations across criteria confirm the robustness of the results.

Across both model structures, the differences between the criteria are minimal, indicating that AIC, BIC, and HQC select similar model complexities for this dataset. For the Left-to-Right HMM, mean logLR values vary by less than 2 units across criteria, and a similar pattern is observed for the Ergodic model. This consistency suggests that the models are well-identified and that the choice of criterion has limited practical impact on overall performance.

Structure	**Criterion**	**Genuine Questioned Sig.**		**Forged Questioned Sig.**	
		logLR		**logLR**	
		Mean	StD	Mean	StD
Left-to-Right	AIC	193.2	6.1	-438.3	9.2
	BIC	191.6	6.1	-437.5	9.2
	HQC	193.2	6.1	-438.3	9.2
Ergodic	AIC	374.3	10.5	-283.2	9.3
	BIC	374.3	10.5	-283.2	9.3
	HQC	374.3	10.5	-283.2	9.3

Table 5.14: Mean and standard deviation of the log-likelihood ratio (logLR) from dynamic signature evaluation experiments using multi-feature HMMs with two model structures. The number of states was selected according to different information criteria.

5.7.4 Impact of the Number of Control Signatures

In this section, experiments are conducted as outlined in Section 5.7.1, employing the multi-feature methodology described in Section 5.7.2 in conjunction with the automated state optimization procedures based on BIC information criteria of Section 5.7.3. The experimental protocol systematically varies the number of control signatures within the range of 5 to 30, thereby enabling a comprehensive assessment of model performance under different reference set sizes. This approach facilitates a thorough evaluation of the interplay between sample size, feature dimensionality, and state optimization in the context of multi-feature Hidden Markov Model analyses.

Figure 5.15 displays the Total Error Rate (TER) for dynamic signature evaluation as the number of control signatures increases, comparing two Hidden Markov Model (HMM) structures, Left-to-Right and Ergodic. Both structures show a sharp drop in TER from 5 to 10 control signatures, after which performance largely stabilises. This is a diminishing returns curve, the most critical gain in discrimination comes from moving from a very small reference set (5 signatures) to a moderate one (10), while additional control material beyond 10 contributes comparatively little. The Left-to-Right structure consistently outperforms the Ergodic structure in terms of average TER across all signatures; however, larger datasets are needed to empirically validate these results. Crucially, both models achieve near-optimal performance at 10 control signatures, which is a practically result which suggests that forensic examiners do not need an excessively large reference set to achieve reliable discrimination, and that 10 genuine reference signatures represent a sufficient threshold for this methodology. This has direct implications for real casework where control material is often limited

These observations reinforce the suitability of HMMs for dynamic signature evaluation, especially for applications likely to use substantial control datasets.

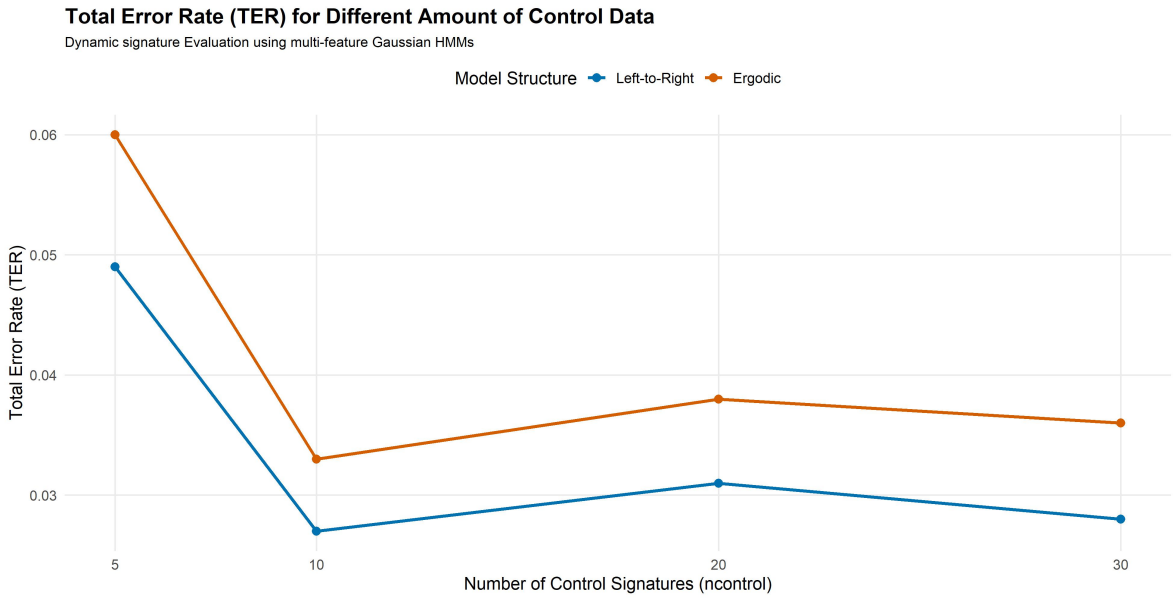


Figure 5.15: Total Error Rate (TER) of increasing the number of control signatures (5, 10, 20, 30) using Left-to-Right and Ergodic Hidden Markov Models (HMM) with automated states selection.

5.8 Discussion and Conclusion

In this chapter, we implement a Gaussian Hidden Markov Model (HMM) to evaluate dynamic signatures through probabilistic modeling. Dynamic signatures constitute a biometric modality that captures the temporal and spatial characteristics of the signing process, providing rich information about an individual's signing dynamics.

The evaluation is based on the likelihood ratio (LR) framework, in which the numerator represents the probability of observing the questioned signature's features under the hypothesis that it is genuine and produced by the person of interest (PoI), while the denominator represents the probability of observing the same features under the hypothesis that is a forged signature. This ratio quantifies how much more probable the observed signature is under the proposition that it originates from the PoI than under the proposition that it originates from the population of forged signatures.

Experiments were conducted to simulate a realistic forensic scenario involving one questioned signature and multiple genuine control signatures. The questioned signature was either genuine or simulated (forged), allowing for the assessment of false positive and false negative rates. These experiments were designed to address the following research questions:

1. Which feature possesses the greatest discriminative power?
2. Can a consistent multi-feature approach be performed?
3. What is the optimal number of HMM states?
4. Which HMM structure yields the best performance?
5. How does increasing the number of control signatures affect the support?

1) Based on the experimental results, the overall performance indicates consistent behavior among certain feature categories, particularly those related to angle orientation and kinematic properties. These features consistently yield lower Total Error Rate (TER) values, around 0.027-0.080, across different model structures and state configurations (Section 5.7.1). In particular, the trajectory vector angle relative to the horizontal axis, the acceleration direction, and the velocity magnitude frequently produce the lowest error rates. This finding suggests that such features are especially discriminative and play a crucial role in enhancing the reliability of dynamic signature evaluation.

2) Based on the previous observations, additional experiments were conducted using a multi-feature Gaussian HMM incorporating the three most discriminative features identified earlier. The results were highly promising, with the Total Error Rate (TER) ranging from 0.004 for Signature 3 to 0.021 for Signature 1 (Section 5.7.2). However, the implemented approach assumes conditional independence among the features, which may limit its ability to capture potential inter-feature correlations. Moreover, the results suggest that there is no single optimal number of states that consistently performs best across all signatures.

3) To address this limitation, Section 5.7.3 introduces an automated state-selection approach informed by literature. This method optimizes the number of states by information criteria AIC, BIC, and HQC. The differences between the criteria are minimal, indicating that AIC, BIC, and HQC select similar model complexities for this dataset. However, this procedure cannot be directly applied to structure optimization, as the Left-to-Right topology inherently involves fewer free parameters than Ergodic HMM structures.

4) To investigate which HMM structure performs better, we base our assessment on the experimental results. The Ergodic structure achieves lower TER for signatures 1 and 3, while the Left-to-Right structure performs better for signature 2. However, further investigation with a larger set of signatures is needed to draw more general empirical conclusions. Naturally, the Left-to-Right (Bakis) structure appears suitable for modeling monotonically dynamic signature features. This topology constrains transitions to proceed forward or remain in the same state, enforcing a natural sequential progression consistent with the temporal evolution of handwriting dynamics. The model trains efficiently, remains stable, and avoids spurious state reversals, effectively capturing simple feature patterns. The Ergodic HMM, in contrast, satisfies the normality assumptions of the Gaussian emission distributions. Its unconstrained transition structure provides a smoother optimisation landscape, leading to more reliable convergence of the EM algorithm compared to the Left-to-Right structure. As a result, the fitted parameters capture the distributional behaviour of the dynamic signature measurements with high fidelity, as evidenced by the goodness-of-fit analysis presented in Section 5.6.

5) In Section 5.7.4, we repeated the forensic case experiments while increasing the number of control signatures from 5 to 30. Both structures show the greatest improvement in TER when moving from 5 to 10 control signatures, with additional control signatures contributing comparatively small. Furthermore, Left-to-Right HMM outperforms Ergodic HMM, with a noticeably lower and steadily improving TER; however, this finding is based on only three signatures.

In summary, the multi-feature Gaussian HMM appears highly suitable for the forensic evaluation of dynamic signatures. Particularly, for the use of velocity and angular orientation features that capture the distinctive dynamics of the signing process. However, the Left-to-Right structure carries notable limitations: the normality assumption on emissions is restrictive, convergence to a global optimum is not guaranteed, and the sequential ordering constraint may be too rigid for some signing styles. Despite these limitations, the approach shows good performance in the experiments and it requires no assumptions about signature shape, making it broadly applicable across diverse writers and practical forensic scenarios.

Finally, future work may focus on several directions. First, the implementation of a multi-feature Gaussian HMM that explicitly accounts for feature correlations would provide a more general and realistic modeling framework. Second, further research should investigate alternative model assumptions that provide a better fit to the observed signature dynamics, moving beyond the Gaussian emission constraint. Third, further research should investigate the sensitivity of the model to background (forged) signers, ideally using a larger and more diverse dataset to draw robust conclusions. Forth, additional forensic questions should be explored, such as whether a single session with ten signatures is sufficient for reliable modeling, or whether multiple sessions and their temporal spacing should be taken into consideration.

Chapter 6

Concluding Remarks

The primary goal of this PhD thesis was to statistically analyze multivariate data with complex dependence structures derived from forensic evidence. The objective was to establish further theoretical foundations for operational methods aimed at analyzing well-extracted informative features of forensic evidence. This research focused on formulating probabilistic approaches of feature evaluation in accordance with the forensic international guidelines of ENFSI (Willis et al., 2015). Such a probabilistic framework has the potential to become an essential resource for forensic experts, supporting the substantiation of probabilistic evaluations and conclusions in cases involving comparative examinations. This framework is designed to explicitly model and quantify inherent uncertainties and variability of forensic measurements in order to provide more transparent and defensible forensic inferences.

This research addressed the challenges associated with assessing (Bayesian) likelihood ratios (Bayes factors) in forensic evaluations of handwriting and dynamic digital signature data. We established effective statistical methodologies for managing complex datasets and integrating diverse categories of information to assess a joint probative value. Furthermore, we identified the most suitable methodological approaches and provided a comprehensive framework for interpreting likelihood ratios in this forensic context. These advancements enable more justified forensic interpretations, reducing reliance on subjective judgment. By applying rigorous probabilistic models to quantify evidence, the framework enhances the transparency and consistency of forensic evidence evaluation.

Secondly, we created automated pipelines to assist forensic scientists in their work by incorporating visuals of modeling parameters within an interpretable environment. The project developed an integrated environment for the inference of handwriting and dynamic signature data, which can be applied in practice. This multidisciplinary (forensic science, statistics, and computer science) approach aimed to discover valid, meaningful and useful patterns leading to interpretable results based on the analysis of the data. These approaches can be adjusted to various forensic scientific domains, such as data on toxic and doping substances, where complex dependence structures involving intra- and inter-variability are frequently observed.

In general, this work contributes by developing probabilistic and statistical modeling frameworks for handwriting and dynamic signature data, aiming to improve the accuracy, reliability, interpretability, and practical usability of forensic evaluations according to international standards. In this way, forensic science is advanced through a comprehensive multidisciplinary approach. In Section 6.1, we present a detailed account of the contributions fulfilled by this work, and in Section 6.2, we propose potential research directions that could facilitate further advancements in the discipline.

6.1 Contributions

This thesis makes multifaceted contributions to forensic document examination by integrating statistical methodologies, implementing feature extraction techniques, and applying probabilistic modeling to evaluate handwritten and dynamic signature data. This study deals with the challenge of managing the uncertainty in forensic document examination, which affects both inferential and decision-making processes in legal contexts. An experimental study has been conducted to facilitate (a) the management of complex multivariate data, (b) the combination of multiple sources of information for the assessment of a joint probabilistic value, and (c) the identification of optimal modeling approaches. This is a promising statistical approach that enables to identify valid and useful patterns in document analysis.

We started our research journey with the traditional examination of handwritten documents. A common approach to quantify handwriting data involves digitizing the documents into images. From these images, we perform mathematical feature extraction using Fourier analysis, whereby the contours of loop characters are transformed into a Fourier series whose coefficients describe the morphological characteristics of each loop character. Following the established literature, these features are modeled using Bayesian multivariate hierarchical frameworks. We advance the existing literature by introducing a model that accounts for within-writer variability, between-writer variability, and between-character variability. Each of these three sources of variability is identified and quantified using the background data. Furthermore, we compare traditional Bayesian models employing conjugate priors with more advanced state-of-the-art hierarchical Bayesian models. Through extensive empirical experiments, we evaluate the most effective modeling approach and assess the sensitivity of the entire framework. In this way, we enhance the stability and interpretability of the examination of handwritten documents.

Building upon the latter Bayesian hierarchical modeling framework, we encounter several challenges related to the estimation of prior model parameters that play a crucial role in the sensitivity and accuracy of handwriting evidence evaluation.

As a continuation of the current era of digital transformation and the rapid expansion of online services where businesses, governments, and financial institutions increasingly rely on digital signatures, we focused on the forensic evaluation of dynamic digital signatures. Dynamic signatures capture both the spatial and temporal characteristics of the signing process, providing detailed information about the handwriting signature process. An extensive feature engineering process is performed on the raw dynamic signature data, and the resulting features are modeled using Gaussian Hidden Markov Models (HMM). This probabilistic approach employs likelihood ratios to distinguish genuine questioned signatures from forged questioned signatures. Empirical experiments identified the optimal feature sets, the most effective HMM state configurations, and the impact of the number of control signatures on evaluation stability. The proposed framework demonstrates robustness and consistency, making it suitable for practical forensic applications. It further supports automated, adaptable, and accurate signature assessment beyond visual or shape-based comparison methods. Thus, we implement a model for dynamic signature analysis that is well established in the literature and evaluate its performance in a case study of forensic evaluation.

Collectively, this research bridges forensic science, statistics, and computer science, establishing a rigorous and operationally viable framework for quantitative forensic handwriting and signature analysis. It delivers valuable theoretical insights, robust modeling approaches, and practical tools that align with international forensic guidelines, thereby advancing the accuracy, transparency, and defensibility of forensic evidence evaluations.

6.2 Future Work

Future research can extend the present study in several important directions. In general, it may focus on comparing different approaches to feature extraction or on extending the current modeling framework. Furthermore, future work should aim to improve the communication and interpretability of the probabilistic frameworks, particularly regarding the magnitude and meaning of the assessed likelihood ratios. Finally, these frameworks could be implemented in dedicated software tools to facilitate their accessibility and practical evaluation by forensic practitioners. In the following paragraphs, we present more detailed research directions based on the frameworks developed in this thesis.

One valuable direction would be to compare Fourier-based features of loop characters with deep learning representations, such as convolutional neural network (CNN) features of the whole handwritten document, to evaluate whether learned features offer improved discriminative power of handwriting document evaluations. In addition, considering Fourier-based features, one promising research path involves developing a two-way MANOVA framework that incorporates the neighboring characters of the analyzed loop character, allowing for a more nuanced analysis of contextual interactions in handwriting.

In the evaluation of handwriting data, as in many other applications, modeling the inherent within-source variability in a multivariate context remains an open issue. The Wishart distribution, as a traditional choice, tends to produce biased estimates of higher correlations between variables. A more flexible alternative involves decomposing the covariance matrix into its variance and correlation components-modeling the variances through a separate distribution and the correlations using an LKJ prior. Although this approach offers greater flexibility, it can still exhibit bias when assessing cases of near-zero correlation. Therefore, further research in this direction would be highly valuable for improving the accuracy and interpretability of multivariate probabilistic models.

In the context of dynamic signatures, implementing a multi-feature Gaussian Hidden Markov Model that explicitly accounts for correlations among features could provide a more general and realistic modeling framework. Future studies should also investigate the sensitivity of the model to background (forged) signers, ideally using a larger and more diverse dataset to ensure robust and generalizable conclusions. Another important consideration concerns the experimental design itself: it remains an open question whether a single session containing ten control signatures is sufficient for reliable modeling, or whether multiple sessions and their temporal spacing should be incorporated to improve stability and consistency. Finally, further research on more interpretable statistical and machine learning models that achieve higher accuracy would be highly beneficial, as it would improve performance without sacrificing transparency and interpretability in legal and forensic contexts.

Appendix A

Handwriting Examination

A.1 Assumptions Underlying the Bayes Factor

The Bayes factor (BF) employed for handwriting evaluation in this study is defined as follows:

$$BF = \frac{m(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid H_1)}{m(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid H_2)} = \frac{m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1) m(\boldsymbol{y}_2 \mid H_1)}{m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_2) m(\boldsymbol{y}_2 \mid H_2)} \tag{A.1}$$

This expression can be further decomposed as:

$$BF = \frac{m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1)}{m(\boldsymbol{y}_1 \mid H_2)} \times \frac{m(\boldsymbol{y}_2 \mid H_1)}{m(\boldsymbol{y}_2 \mid H_2)} \tag{A.2}$$

Under the assumption that the likelihood of $\boldsymbol{y}_1$ does not depend on $\boldsymbol{y}_2$ when H_2 holds, the Bayes factor simplifies to:

$$BF = \frac{m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1)}{m(\boldsymbol{y}_1 \mid H_2)} \tag{A.3}$$

Furthermore, if the likelihood of $\boldsymbol{y}_2$ (i.e., the control measurements) is independent of whether H_1 or H_2 is true, it follows that:

$$m(\boldsymbol{y}_2 \mid H_1) = m(\boldsymbol{y}_2 \mid H_2) \tag{A.4}$$

The marginal likelihood $m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1)$ can be expressed as:

$$m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1) = \int_\theta f(\boldsymbol{y}_1 \mid \theta) f(\theta \mid \boldsymbol{y}_2, H_1) d\theta \tag{A.5}$$

where $f(\theta \mid \boldsymbol{y}_2, H_1)$ denotes the posterior distribution of θ given $\boldsymbol{y}_2$ under H_1 (we denote model parameters equal to θ for simplicity only in this section for all considered model parameters of this study see Section 4.5).

Alternatively, using Bayes' theorem, this can be rewritten as:

$$m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1) = \frac{\int_\theta f(\boldsymbol{y}_1 \mid \theta) f(\boldsymbol{y}_2 \mid \theta) f(\theta \mid H_1) d\theta}{m(\boldsymbol{y}_2 \mid H_1)} \tag{A.6}$$

Similarly, the marginal likelihood under H_2 is given by:

$$m(\boldsymbol{y}_1 \mid H_2) = \int_\theta f(\boldsymbol{y}_1 \mid \theta) f(\theta \mid H_2) d\theta \tag{A.7}$$

Combining these results, and noting that $m(\boldsymbol{y}_2 \mid H_1) = m(\boldsymbol{y}_2 \mid H_2)$, the Bayes factor can be expressed as:

$$BF = \frac{m(\boldsymbol{y}_1 \mid \boldsymbol{y}_2, H_1)}{m(\boldsymbol{y}_1 \mid H_2)} = \frac{\int_\theta f(\boldsymbol{y}_1 \mid \theta) f(\boldsymbol{y}_2 \mid \theta) f(\theta \mid H_1) d\theta}{\int_\theta f(\boldsymbol{y}_1 \mid \theta) f(\theta \mid H_2) d\theta \int_\theta f(\boldsymbol{y}_2 \mid \theta) f(\theta \mid H_2) d\theta} \tag{A.8}$$

Note that the implemented Bayes factor assumes independence between the questioned and control material under H_2. This assumption implies that possible disguised behavior is not considered. This formulation is presented in the main text of the paper.

A.2 Dummy Variables

For each writer, the character indicator is transformed into a set of dummy variables. It omits one of the dummy variables from the equation for identifiability reasons in regression models as first introduced by Suits (1957). Moreover, the category corresponding to the omitted dummy variable serves as the reference group against which all other levels are compared. In our case the character a is selected as the reference category; see Tables A.2.1 and A.2.2.

Indicator	**Fourier Coefficients**
a	...
d	...
o	...
q	...

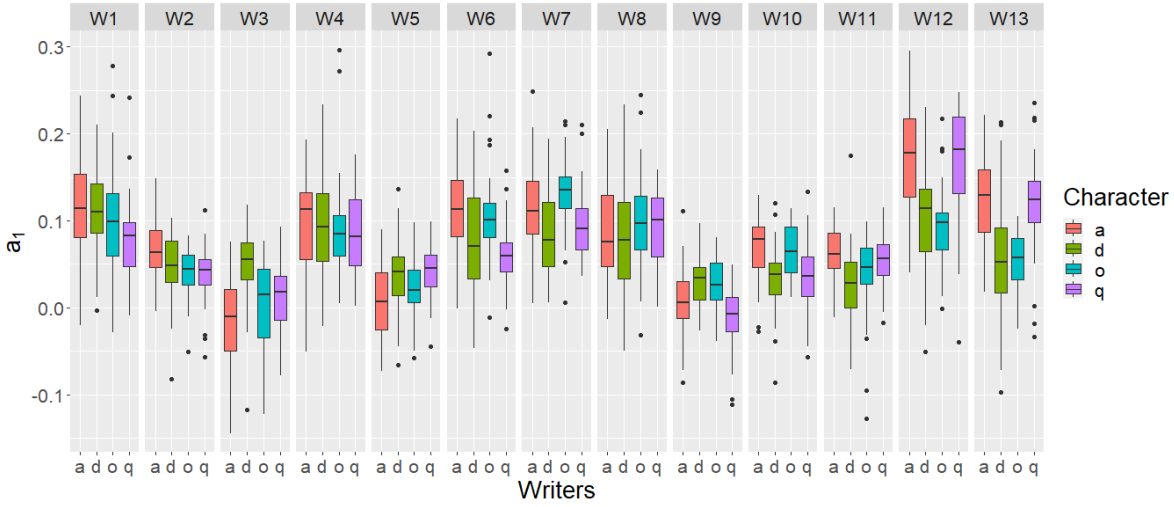
Table A.2.1: Processed Data

a	**d**	**o**	**q**	**Fourier Coefficients**
1	0	0	0	...
1	1	0	0	...
1	0	1	0	...
1	0	0	1	...

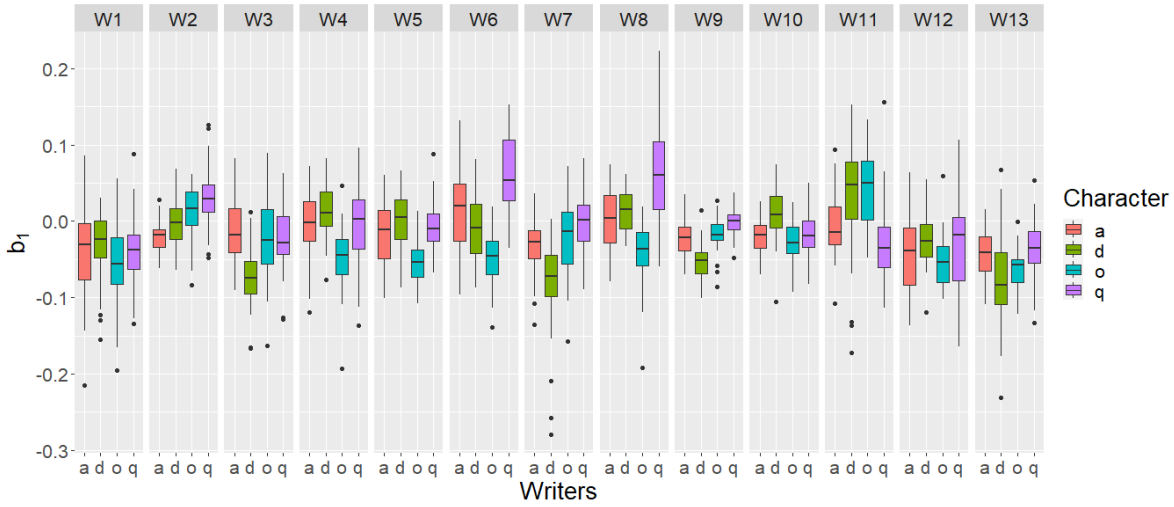
Table A.2.2: Regression Data

A.3 Analytical Representation of Fourier-based Features

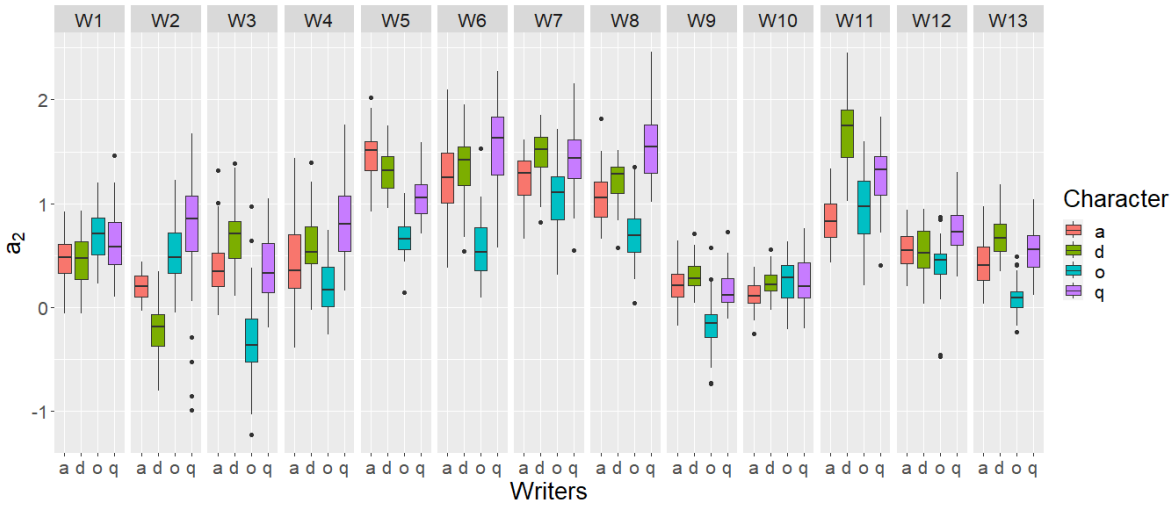
(a) Box-plots of the surface size(in cm^2) of each loop character and per writer.



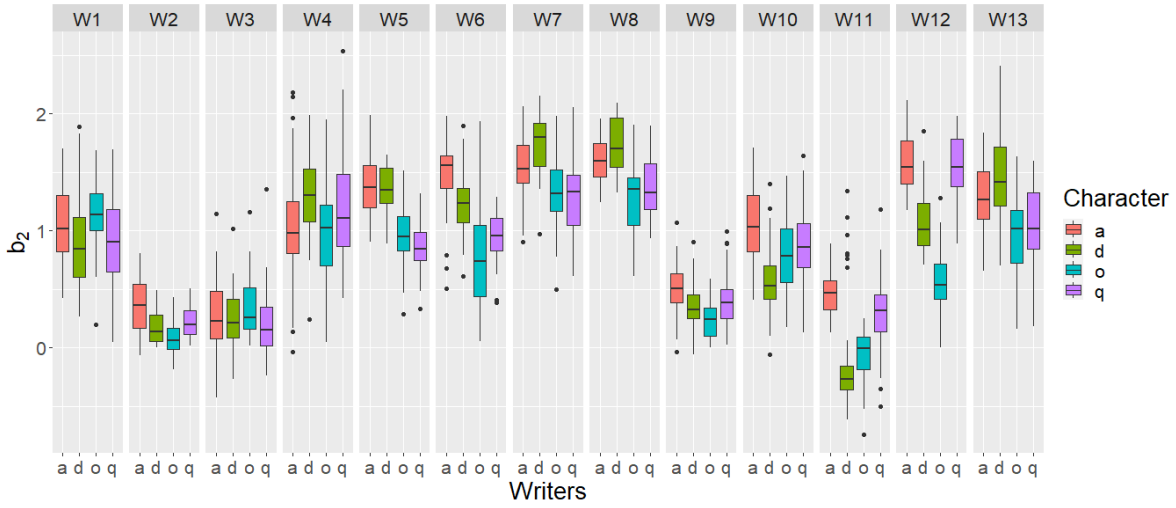
(b) Box-plots of Fourier coefficient a_1 for each loop character and writer.



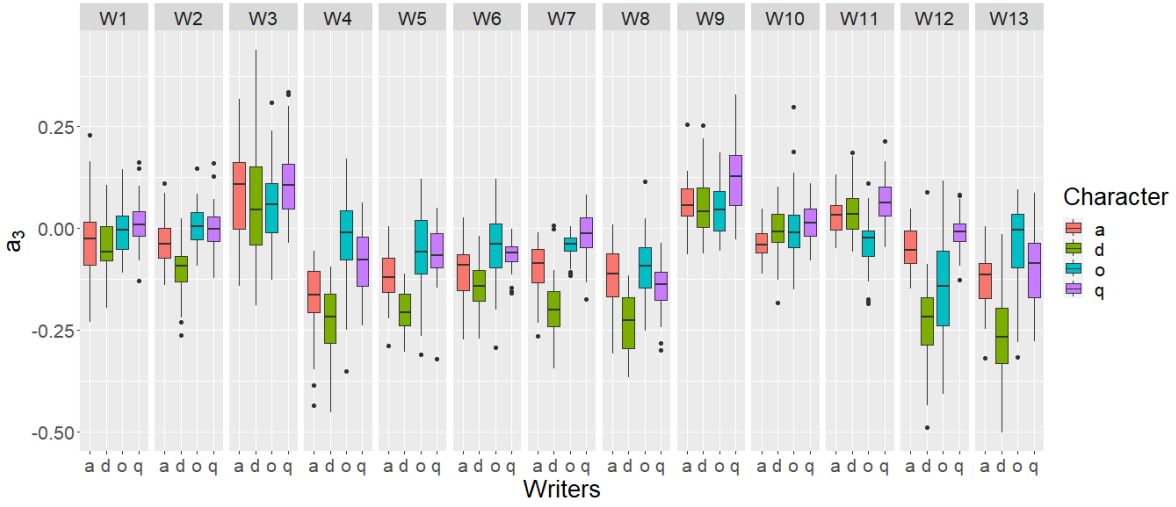
(c) Box-plots of Fourier coefficient b_1 for each loop character and writer.



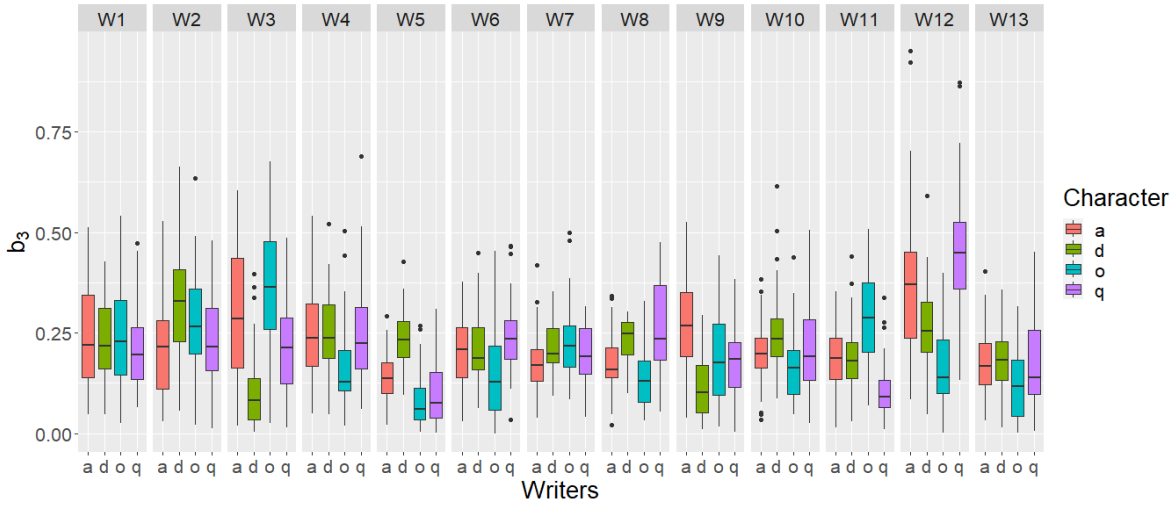
(d) Box-plots of Fourier coefficient a_2 for each loop character and writer.



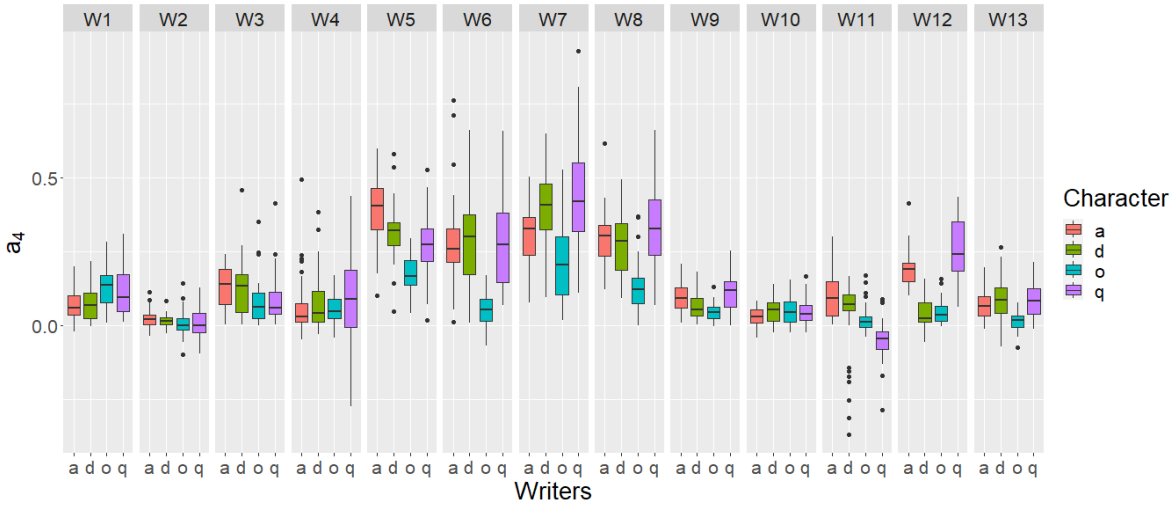
(e) Box-plots of Fourier coefficient b_2 for each loop character and writer.



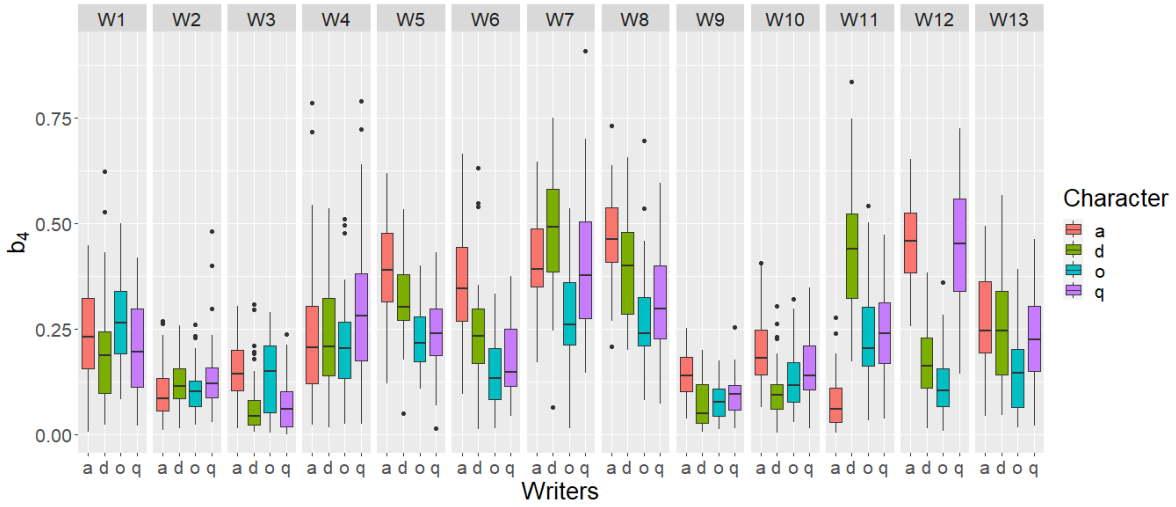
(f) Box-plots of Fourier coefficient a_3 for each loop character and writer.



(g) Grouped box-plot of the b_3 Fourier coefficient per loop character and per writer.



(h) Box-plots of Fourier coefficient a_4 for each loop character and writer.



(i) Box-plots of Fourier coefficient b_4 for each loop character and writer.

Figure A.3.1: Grouped box-plot of the surface size and the first four pairs of Fourier coefficients per loop character and per writer. The surface size is at cm^2

A.4 Prior Elicitation

A.4.1 Prior Parameters Elicitation for Bayesian Normal Model

For the Normal-Inverse-Wishart model, the prior parameters can be elicited from the background data-set. The prior mean of writer i for character ℓ, denoted by $\boldsymbol{\theta}_{i\ell}$ is given by the sample mean of the corresponding background data over all repetitions, that is

$$\widehat{\boldsymbol{\theta}}_{i\ell} = \frac{\sum_{j=1}^{n_{i\ell}} \boldsymbol{X}_{i\ell j}}{n_{i\ell}}, \tag{A.9}$$

where $i = 1, \ldots, m$, $\ell = 1, \ldots, L$, $\widehat{\boldsymbol{\theta}}_{i\ell}$ and $\boldsymbol{X}_{i\ell j}$ are vectors of length p; the elements of the latter vector are the background Fourier coefficients and the surface size of writer i for character ℓ over all repetitions $n_{i\ell}$. The overall mean of loop character ℓ is denoted by $\boldsymbol{\mu}_\ell$ is estimated as

$$\widehat{\boldsymbol{\mu}}_\ell = \sum_{i=1}^{m} \widehat{\boldsymbol{\theta}}_{i\ell} \frac{n_{i\ell}}{n_\ell} \tag{A.10}$$

where $n_\ell = \sum_{i=1}^{m} n_{i\ell}$. Equivalently, the overall mean across all characters and writers is simply given by $\widehat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{m} \sum_{\ell=1}^{L} n_{i\ell} \widehat{\boldsymbol{\theta}}_{i\ell}$; where $n = \sum_{\ell=1}^{L} n_\ell$.

The between-writer covariance matrix $\boldsymbol{B}_\ell$ for loop-character ℓ is elicited by setting it equal to the sample covariance of the corresponding background data, that is

$$\widehat{\boldsymbol{B}}_\ell = \frac{1}{m-1} \sum_{i=1}^{m} (\widehat{\boldsymbol{\theta}}_{i\ell} - \widehat{\boldsymbol{\mu}}_\ell)(\widehat{\boldsymbol{\theta}}_{i\ell} - \widehat{\boldsymbol{\mu}}_\ell)^T \tag{A.11}$$

Following the parametrization of Inverse-Wishart distribution in R and Stan, its mean is given by $E[\boldsymbol{W}_{i\ell}] = \boldsymbol{U}_\ell/(\nu - p - 1)$; where p is the number of Fourier coefficients and the surface size retained for each loop character representing our response data. Hence, parameters $\boldsymbol{U}_\ell$ (that describe the within-writer variation) is elicited by setting $\boldsymbol{U}_\ell$ equal to

$$\widehat{\boldsymbol{U}}_\ell = \widehat{\boldsymbol{W}}_{i\ell}(\nu - p - 1), \tag{A.12}$$

where $\widehat{\boldsymbol{W}}_{i\ell}$ can be estimated by

$$\widehat{\boldsymbol{W}}_{i\ell} = \frac{1}{n-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij\ell} - \widehat{\boldsymbol{\theta}}_{i\ell})(\boldsymbol{X}_{ij} - \widehat{\boldsymbol{\theta}}_{i\ell})^T \ . \tag{A.13}$$

We consider values of ν such that $\nu \geq p + 2$ so that the prior mean is well-defined; the reader can refer to Press (2005) for further details. Finally, the prior parameter k_0 is selected based on a grid search in the (0,1) interval for all writers using background data. Specifically, we choose the value of k_0 which maximizes the marginal likelihood in the background data.

For the Normal-LogNormal-LKJ prior approach, the parameters of the Normal prior are estimated consistently using the formulations presented in Equations (A.10) and (A.11). Subsequently, by extracting the diagonal elements of the estimated covariance matrix $\widehat{\boldsymbol{W}}_{i\ell}$, we derive the prior parameters of the LogNormal distribution as follows:

$$\hat{v}_\ell = \frac{1}{p} \sum_{\kappa=1}^{p} \log(diag(\widehat{\boldsymbol{W}}_{i\ell})_\kappa) \tag{A.14}$$

$$\hat{\sigma}_\ell = \frac{1}{p-1} \sum_{\kappa=1}^{p} \left(\log(diag(\widehat{\boldsymbol{W}}_{i\ell})_\kappa) - \hat{v}_\ell \right) \tag{A.15}$$

For η parameter of LKJ distribution is set equal to one.

A.4.2 Prior Parameters Estimation for Bayesian MANOVA

Similarly, for the Bayesian MANOVA model, the prior parameters are also elicited using the available background data. The elicitation of prior parameters for reference character a $\boldsymbol{\mu}_1$ and $\boldsymbol{B}_1$ of $\boldsymbol{\theta}_1$ is performed by Equations A.10–A.11 for $\ell = 1$. For the prior parameters $\boldsymbol{\mu}_\ell$ and $\boldsymbol{B}_\ell$, i.e. for $\ell = 2, 3, 4$, we use the same equation but instead of the original data $\boldsymbol{X}_{i\ell j}$ we consider the differences from character a, i.e. $d_{i\ell j} = \boldsymbol{X}_{i\ell j} - \widehat{\boldsymbol{\mu}}_1$.

The trace matrix of the Inverse-Wishart distribution $\boldsymbol{U}$ for the within-writer variation is elicited from Equations A.12 and A.13 in the same way as in the Normal model and by considering all characters together; for more details see Press (1980).

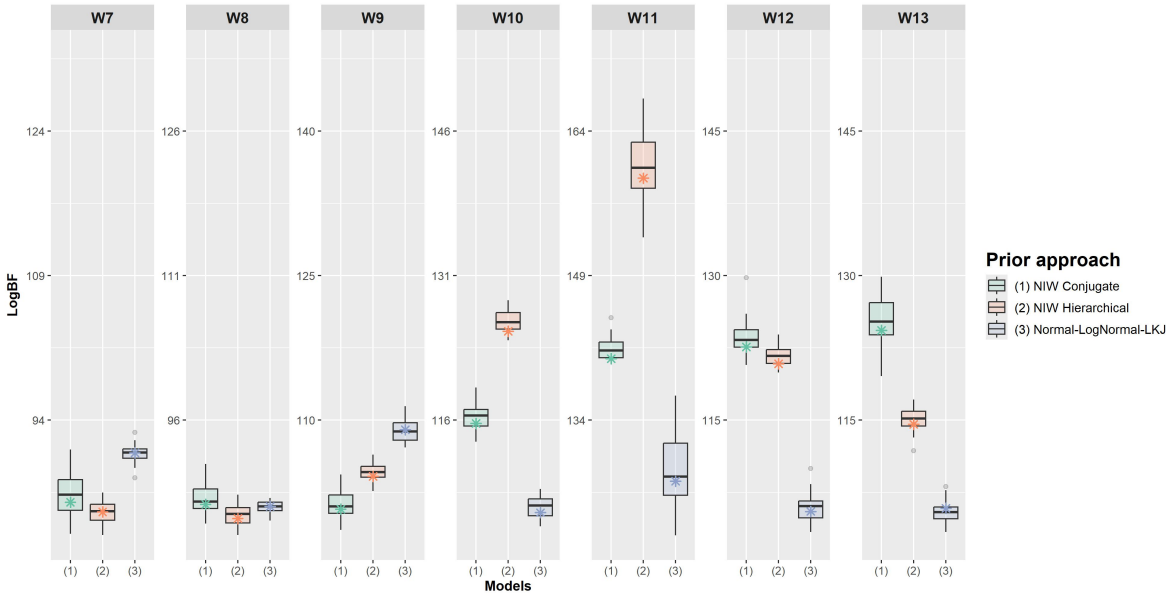
Finally, the $\boldsymbol{K}_0$ of the conjugate approach is elicited based on a grid search in $(0, 1)^L$ for all characters of background writers: the value that maximizes the marginal likelihood is selected.

For the Normal-LogNormal-LKJ prior approach, the parameters of the Normal prior are estimated consistently using the formulations presented in Equations (A.10) and (A.11) by calculating the differences as described in he beginning of this section. Furthermore, by extracting the diagonal elements of the estimated covariance matrix $\widehat{\boldsymbol{W}}$, we derive the parameters of the LogNormal distribution based on Equations (A.14), (A.15) and η parameter of LKJ distribution is set equal to one.

A.5 Sensitivity of Prior Elicitation in Handwriting Examination

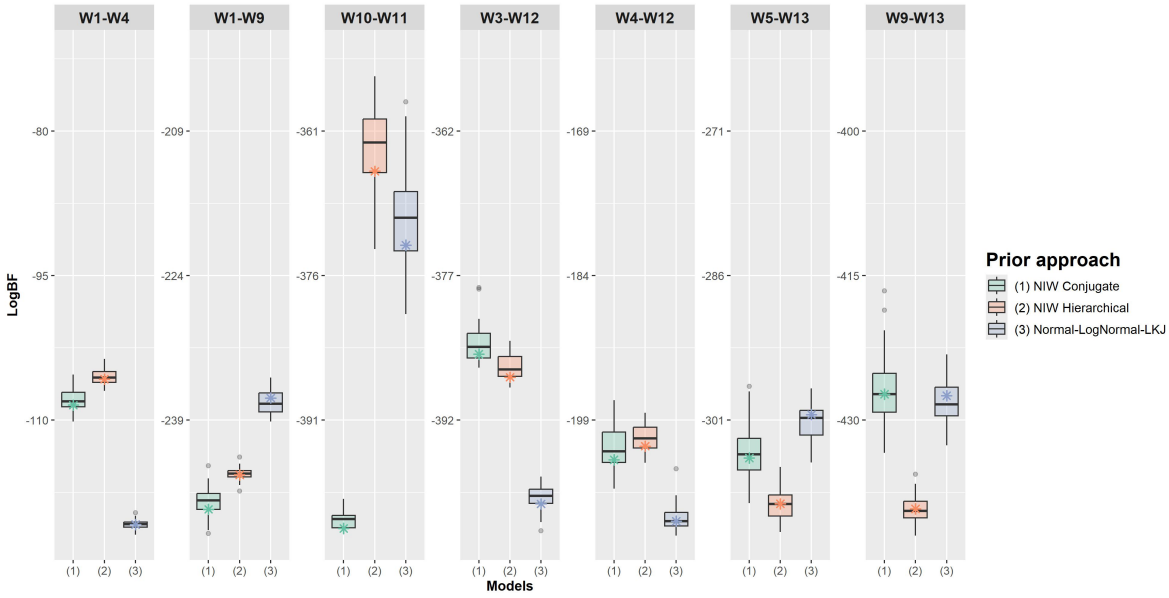
A.5.1 Subsampling of Background Data Examples

In this section, we present more examples of the implementation of subsampling to the background data associated with randomly selected cases from Section 4.7.2. To isolate the effect of prior elicitation, we utilize a single random data split for both the same-writer and different-writer experiments, thereby minimizing potential confounding effects arising from data partitioning. This analysis focuses on the Bayesian MANOVA model, which has been identified as the most effective and recommended approach for this context. For each case, we perform 30 iterations of subsampling with replacement in 50% of the background writers' data in order to ensure the robustness of the results.



indicates the log BF *using the complete background dataset for each case.*

Figure A.5.2: Boxplots of Logarithmic Bayes factors (log BF) for handwriting evaluation for the same writer scenarios over different subsamples of background data for the Bayesian MANOVA approach.



indicates the log BF *using the complete background dataset for each case.*

Figure A.5.3: Boxplots of Logarithmic Bayes factors (log BF) for handwriting evaluation for the different writers scenarios over different subsamples of background data for the Bayesian MANOVA approach.

A.5.2 Sensitivity of the Inverse-Wishart's Degrees of Freedom

In this section, a sensitivity analysis is performed to investigate the effect of the choice of the degrees of freedom of the Inverse-Wishart distribution, which models the within-writer variability. The specification of this prior parameter is of primary concern, since it has a large impact on the resulting Bayes factor values.

The sensitivity analysis was performed for values of the degrees of freedom ranging from the minimum value of 11 ($p + 2 = 11$) to 50, with a step size of 10, that is, $\nu \in \{11, 20, 30, 40, 50\}$. The Bayes factor has therefore been calculated for each value of ν, for all comparisons between characters from the same or different writers, using all models and marginal likelihood estimation methods. 10 sub-samples have been drawn for each pair of writers, following the procedure described in Section 4.7.2.

Figure A.5.1 presents the average of the Bayes factors (in log scale) for the different choices of degrees of freedom. As expected, the Bayes factor is quite sensitive to the choice of degrees of freedom, which has an incremental effect on its value. For different writers comparisons, an increase of one degree of freedom in the Inverse-Wishart distribution produces an increase of about 2.2 units in the log-Bayes factors (with $R^2 = 0.97$). Hence, the support provided by the evidence in favor of the hypothesis that the compared material originates from the same writer is (falsely) increased. This is not surprising, as the degrees of freedom play an important role in the elicitation process of the Inverse-Wishart distribution. As the degrees of freedom parameter ν increases, the prior distribution of $\boldsymbol{W}$ becomes more concentrated around a smaller region of $\mathbb{R}^p$, as discussed by Gaborini (2021). Since for larger values of the degrees of freedom, the variability of the variance-covariance matrix is reduced, it can reasonably be expected that, for large values of ν, the resulting logBF will be bounded by the case where the variance-covariance matrix is constant and fixed at the prior mean of the Inverse-Wishart distribution. In support of this, one only has to look at Figure A.5.2, where the average logarithmic BFs obtained from the comparison between writers 7 and 8 are shown. As can easily be seen from the Mahalanobis distance in Figure 4.11, these writers indeed show great similarities, and it is difficult to discriminate their style. For this pair of writers, the evidence in favour of the wrong hypothesis (i.e., H_1) still increases with degrees of freedom, but after the value of 40, it seems to be close to reaching its maximum. Moreover, it can be observed that while for small values of the degrees of freedom (lower than 30), the Bayes factor correctly supports hypothesis H_2 (i.e., the compared handwritten material originate from different writers), for larger values of the degrees of freedom, the evidence is in favour of the wrong hypothesis H_1 (i.e., the compared handwritten material originate from the same writer).

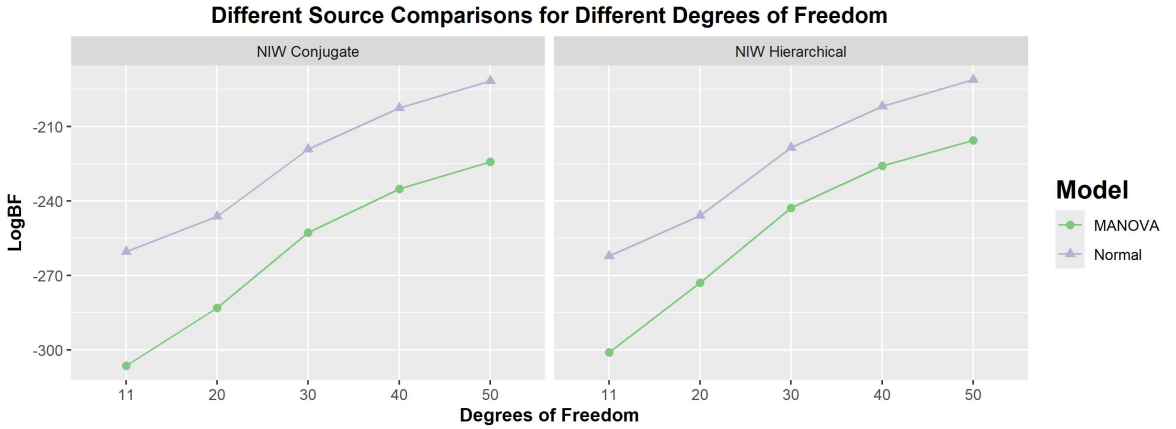


Figure A.5.1: Average logarithmic Bayes factor ($\log \mathrm{BF}$) for handwriting examination for different writers' comparisons over different degrees of freedom. The Normal model has been implemented using all character types jointly.

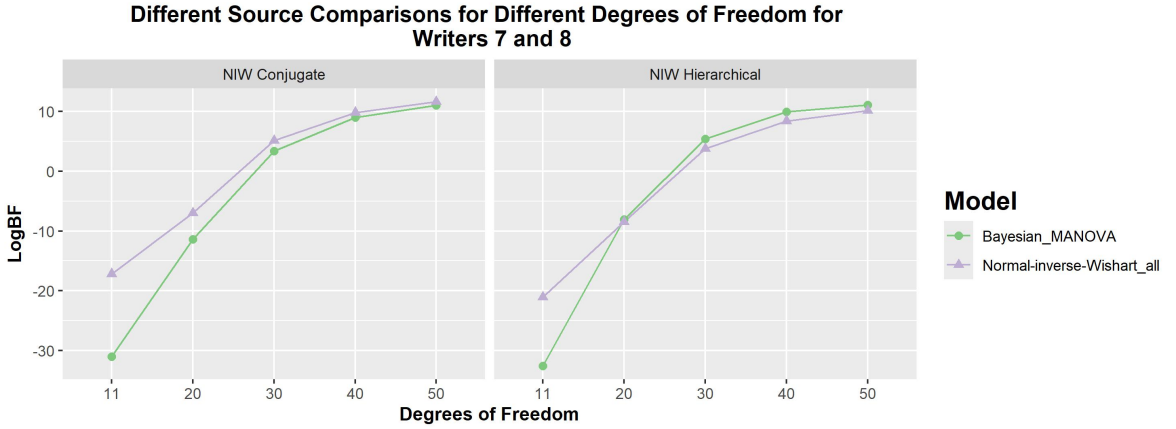


Figure A.5.2: Average logarithmic Bayes factor ($\log BF$) for handwriting examination for comparing Writers 7 and 8 over different degrees of freedom. The normal model has been implemented using all character types jointly.

A.5.3 Sensitivity of the Parameter of the LKJ Distribution

In this section, we present a series of experiments involving different writers, exploring the influence of the LKJ distribution's η parameter. The parameter η was assigned values of 1, 2, 5, 10, and 20, allowing for a clear view of its effect on the writers' comparison. Only values greater than 1 were considered, as these correspond to prior beliefs favoring the identity matrix. This approach is motivated by results from Fourier analysis, which establish that Fourier coefficients are uncorrelated. Consequently, employing higher values of η emphasizes prior structures in which variables are uncorrelated within the covariance matrix.

Figure A.5.3 illustrates how the Log Bayes factor (LogBF) varies with different values of the LKJ parameter η for the Bayesian Normal and MANOVA models. The results show that as the LKJ parameter η increases, the LogBF values for both models also increase. This indicates that higher values of η, which correspond to stronger prior beliefs favoring the identity matrix, tend to produce more positive LogBF values. This can affect the decisions, as we can observe in Figure A.5.4, which compares writers 7 and 8, where the LogBF values change from negative to positive as η increases, and from value $\eta = 20$, the evidence supports the incorrect hypothesis. This demonstrates that the

choice of prior, specifically the value of η in the LKJ distribution, can substantially affect the outcome of writers' comparison.

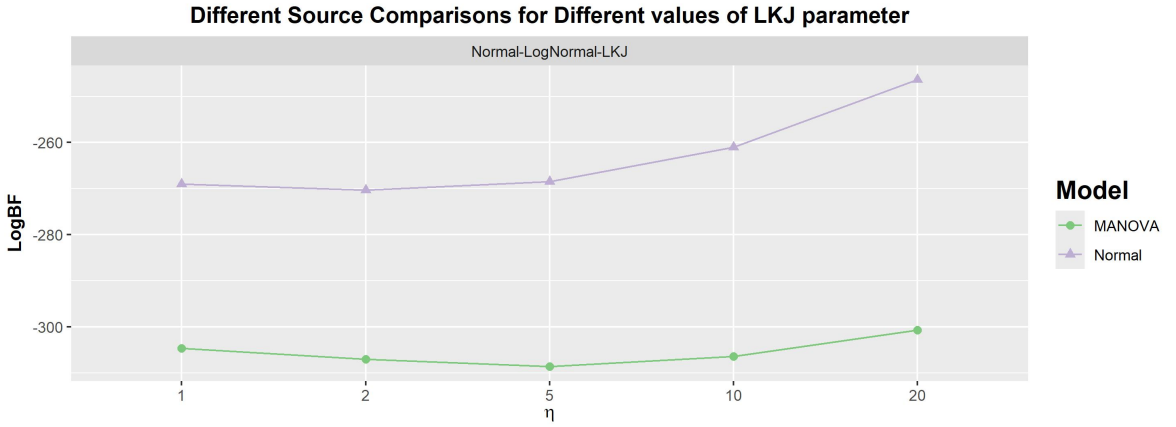


Figure A.5.3: Average logarithmic Bayes factor ($\log \mathrm{BF}$) for handwriting examination for different writer comparisons over different η values of the LKJ distribution. The Normal model has been implemented using all character types jointly.

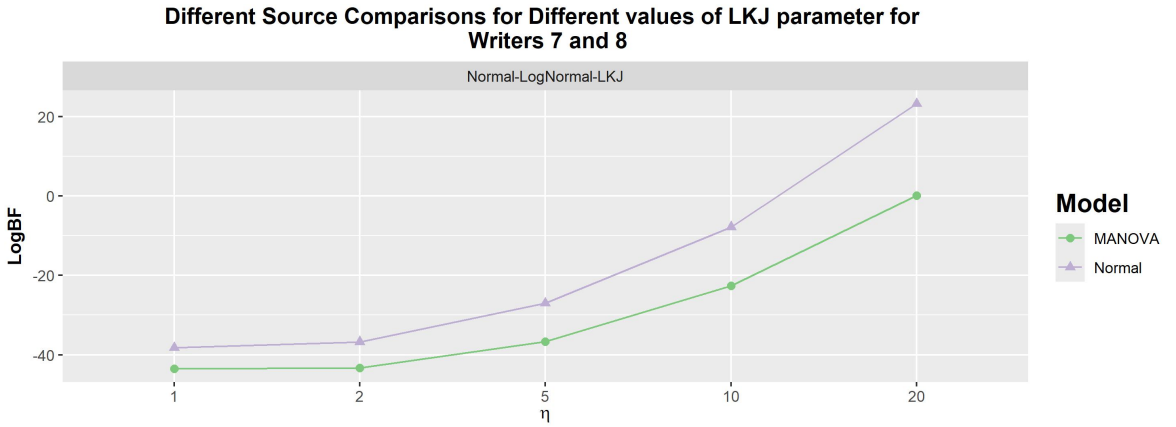


Figure A.5.4: Average logarithmic Bayes factor ($\log BF$) for handwriting examination for comparing Writers 7 and 8 over different η values of LKJ distribution. The Normal model has been implemented using all character types jointly.

Appendix B

Dynamic Signature Examination

B.1 Trajectory Resampling Procedure

Dynamic signature can be defined by a sequence of pen coordinates (X_t, Y_t), time stamps τ_t and pressure values P_t, namely $\{(X_t, Y_t, P_t, T_t)\}_{t=1}^{T}$. The resampling procedure is based on cumulative arc length parameterization:

1. **Arc length computation:** For $t = 2, \ldots, T$, compute segment distances

$$d_t = \sqrt{(X_t - X_{t-1})^2 + (Y_t - Y_{t-1})^2},$$

and the cumulative arc length

$$s_t = \sum_{j=2}^{t} d_j, \qquad s_1 = 0.$$

2. **Normalization:** Define a normalized trajectory parameter

$$u_t = \frac{s_t}{s_T}, \qquad u_i \in [0, 1],$$

for $t = 1, \ldots, T$ where s_T is the total arc length.

3. **Interpolation:** Choose $\widetilde{T}$ evenly spaced interpolation points

$$\widetilde{u}_r = \frac{r-1}{\widetilde{T}-1}, \qquad r = 1, \ldots, \widetilde{T}.$$

For each $\widetilde{u}_r$ find the index l with $u_l \leq \widetilde{u}_r \leq u_{l+1}$ and apply linear interpolation. For example, for X,

$$\widetilde{X}_r = X_l + \frac{\widetilde{u}_r - u_l}{u_{l+1} - u_l}\,(X_{l+1} - X_l),$$

and similarly for Y, P, τ:

$$\widetilde{Y}_r = Y_l + \frac{\widetilde{u}_r - u_l}{u_{l+1} - u_l}\,(Y_{l+1} - Y_l),$$

$$\widetilde{P}_r = P_l + \frac{\widetilde{u}_r - u_l}{u_{l+1} - u_l}\,(P_{l+1} - P_l), \quad \widetilde{\tau}_r = \tau_l + \frac{\widetilde{u}_r - u_l}{u_{l+1} - u_l}\,(\tau_{l+1} - \tau_l).$$

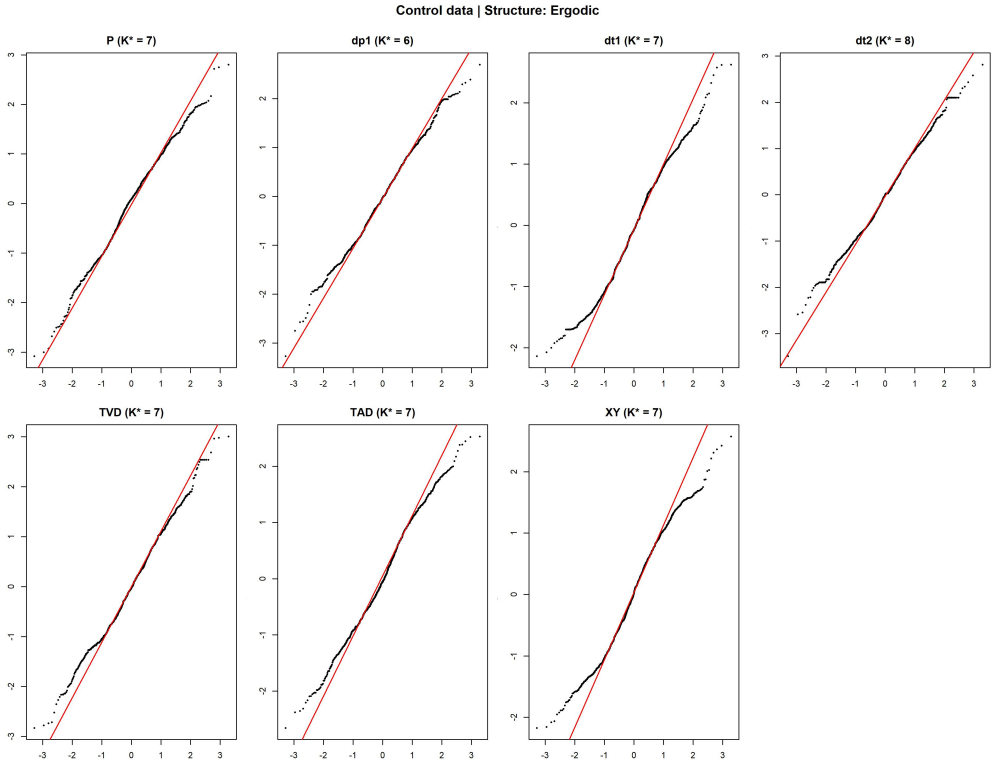
The result is a resampled trajectory

$$\{(\widetilde{X}_r, \widetilde{Y}_r, \widetilde{P}_r, \widetilde{\tau}_r)\}_{r=1}^{\widetilde{T}}$$

with uniform arc length parameterization and fixed length $\widetilde{T}$.

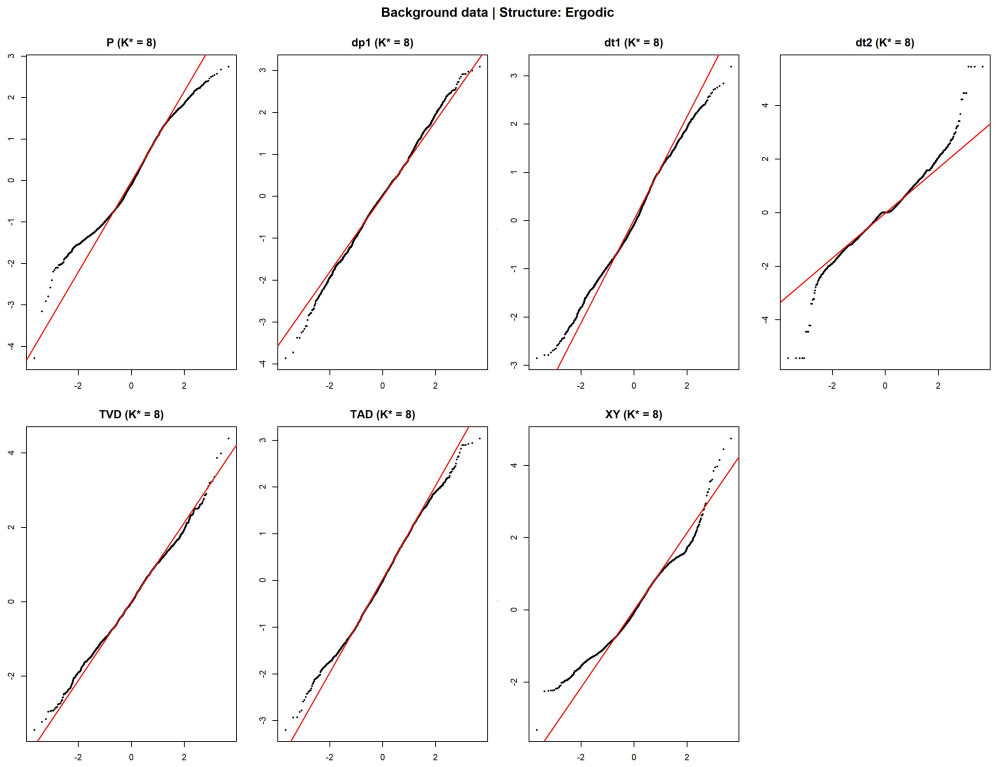
B.2 Q–Q Plots of Pseudo-Residuals

B.2.1 Ergodic Structure



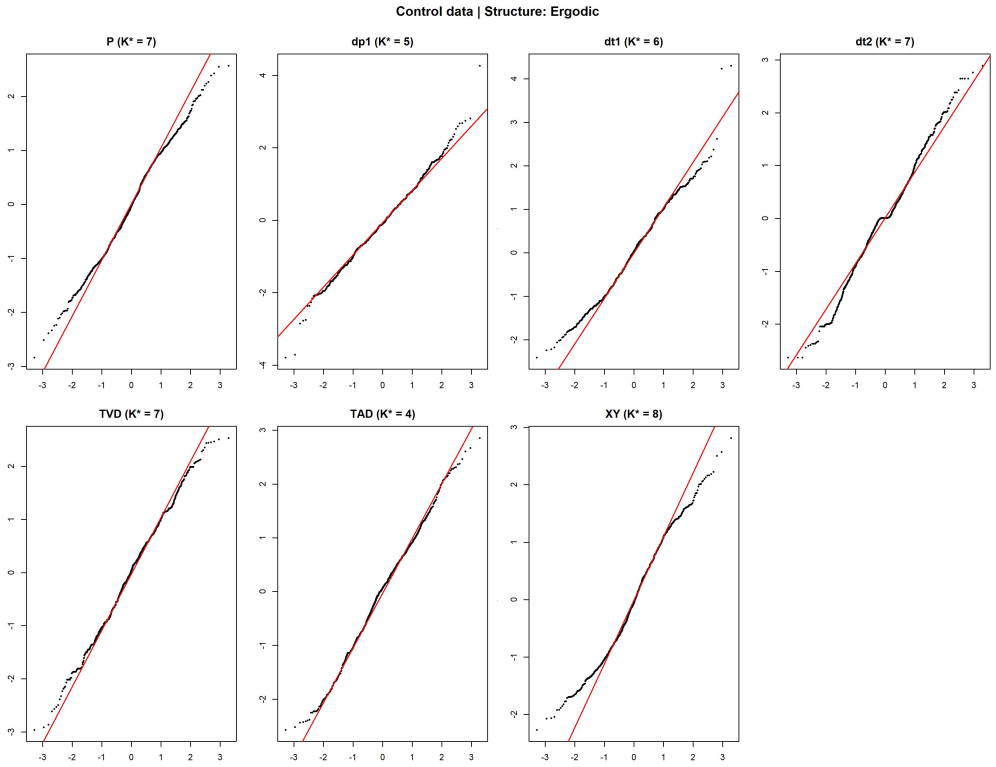
The optimal number of states K^* is indicated in each panel title.

Figure B.2.1: Normal Q–Q plots of the pseudo-residuals obtained from the Ergodic Gaussian HMM fitted to the signature 2 measurements, for each considered feature.



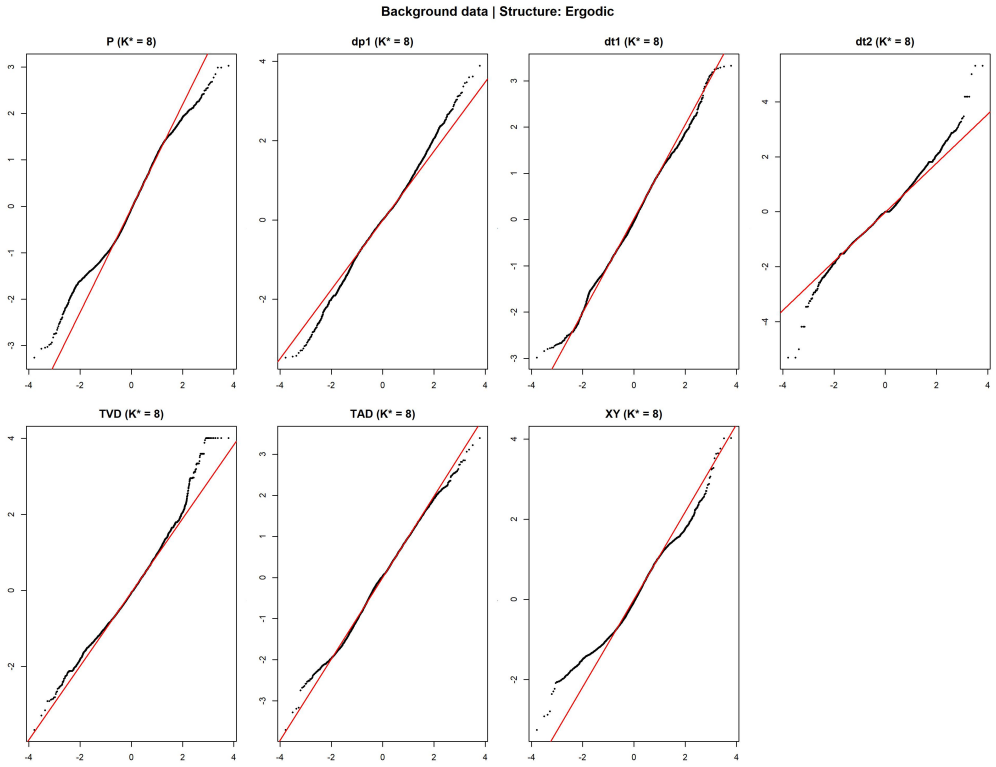
The optimal number of states K^* is indicated in each panel title.

Figure B.2.2: Normal Q–Q plots of the pseudo-residuals obtained from the Ergodic Gaussian HMM fitted to the background simulated measurements of signatures 1 and 3, for each considered feature.



The optimal number of states K^* is indicated in each panel title.

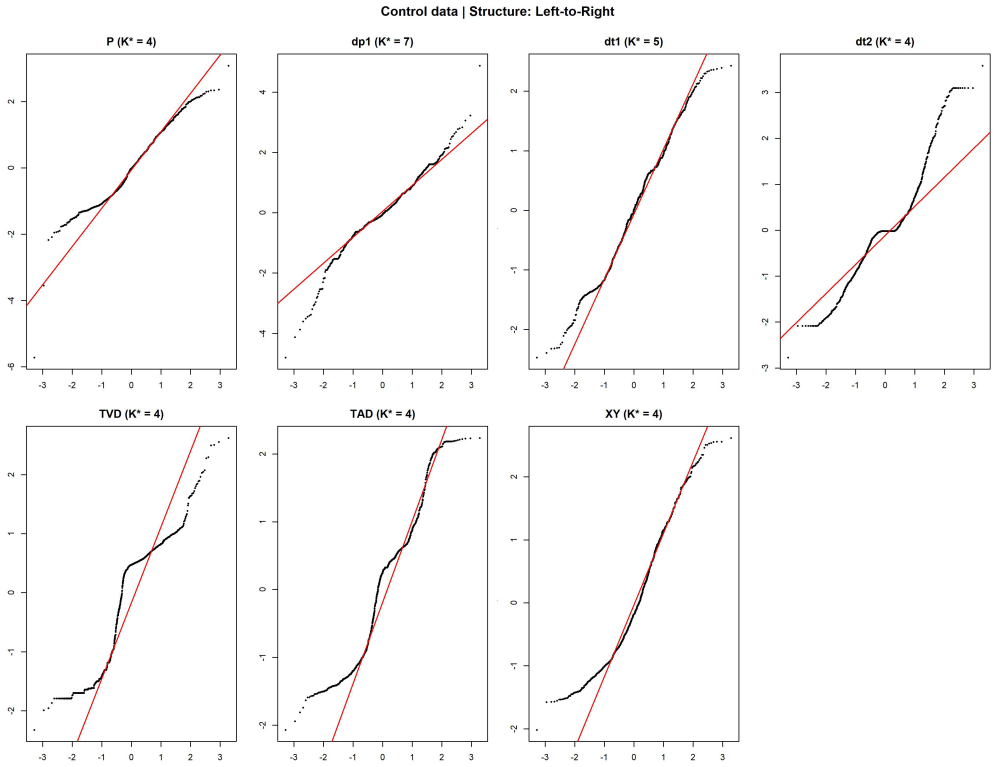
Figure B.2.3: Normal Q–Q plots of the pseudo-residuals obtained from the Ergodic Gaussian HMM fitted to the signature 3 measurements, for each considered feature.



The optimal number of states K^* is indicated in each panel title.

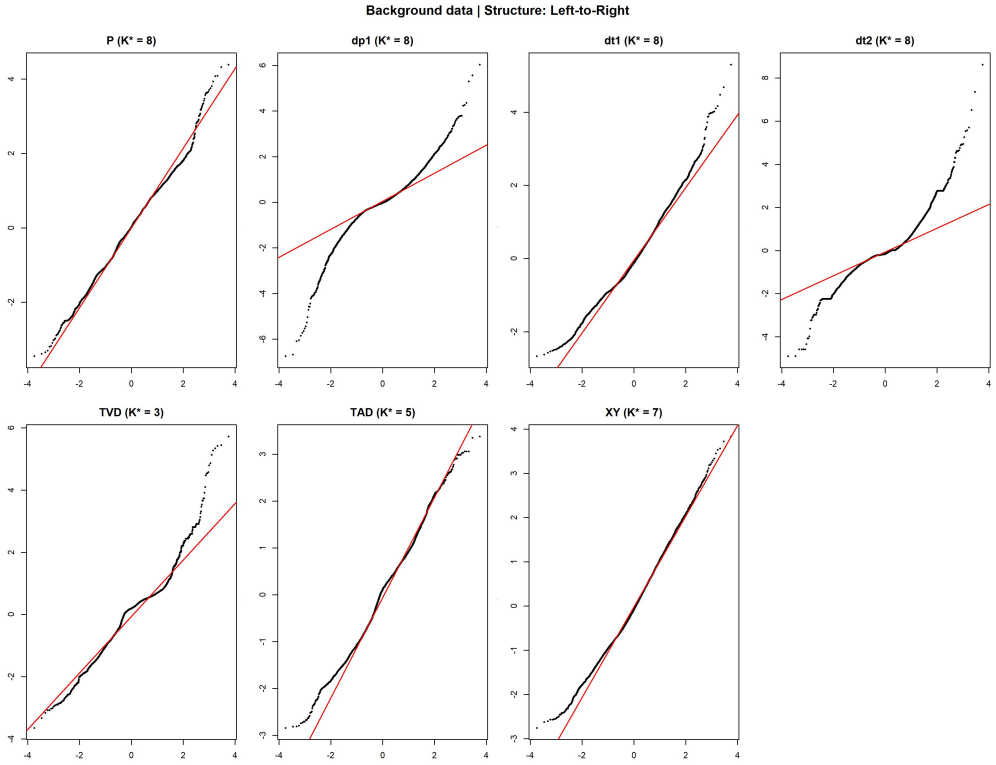
Figure B.2.4: Normal Q–Q plots of the pseudo-residuals obtained from the ergodic Gaussian HMM fitted to the background simulated measurements of signatures 1 and 2, for each considered feature.

B.2.2 Left-to-Right Structure



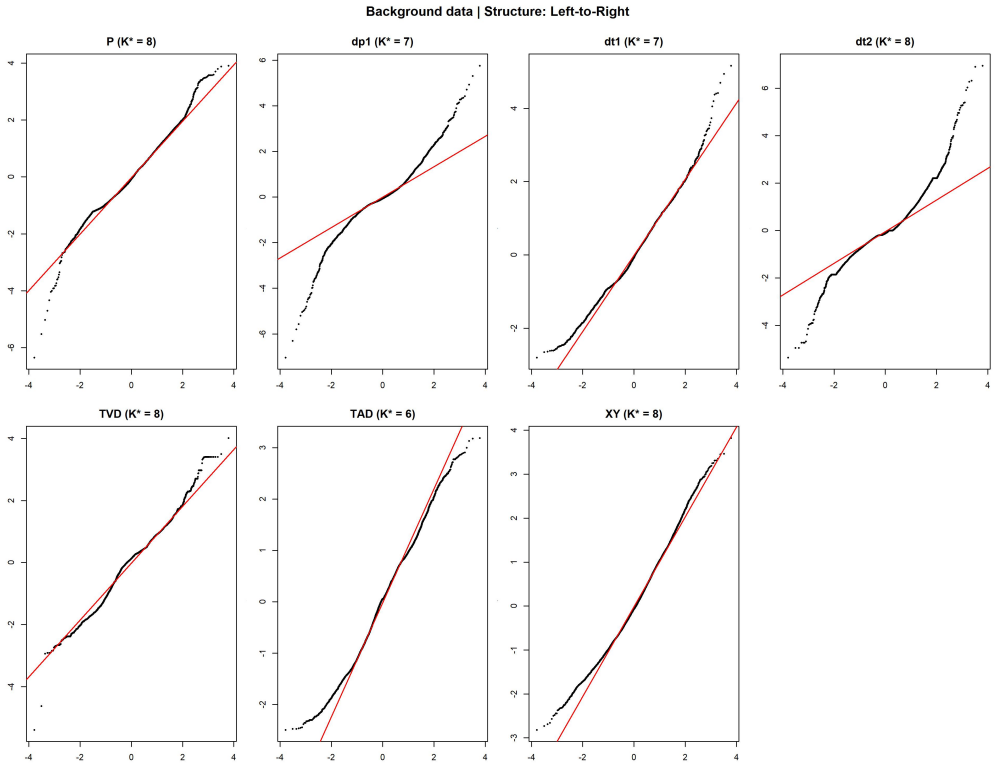
The optimal number of states K^* is indicated in each panel title.

Figure B.2.5: Normal Q–Q plots of the pseudo-residuals obtained from the Left-to-Right Gaussian HMM fitted to the signature 1 measurements, for each considered feature.



The optimal number of states K^* is indicated in each panel title.

Figure B.2.6: Normal Q–Q plots of the pseudo-residuals obtained from the Left-to-Right Gaussian HMM fitted to the background simulated measurements of signatures 2 and 3, for each considered feature.



The optimal number of states K^* is indicated in each panel title.

Figure B.2.7: Normal Q–Q plots of the pseudo-residuals obtained from the Left-to-Right Gaussian HMM fitted to the background simulated measurements of signatures 1 and 2, for each considered feature.

B.3 Experimental Results per Feature and Signature

The experiments are conducted using the data sets described in Section 5.2 and the questioned signature is evaluated under two competing hypotheses: either it is genuine, i.e., originating from the person of interest (PoI) (H_1), or it is forged, i.e., produced by someone other than the PoI (H_2). Consequently, in the experiments, the questioned signature comes either from the case-related genuine signature dataset $\mathcal{G}$ or from the case-related forged (simulated) signature dataset $\mathcal{F}$ (see Section 5.2). For control data $\{\boldsymbol{x}_{j,1:\tilde{T}}\}_{j=1}^{n}$, we use a set of $n = 10$ signatures from the case-related genuine dataset $\mathcal{G}$. For the background forged population, denoted $\{\mathcal{B}_{j,1:\tilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$, we use signatures from the case-related forged dataset $\mathcal{F}$ that are unrelated to the case under study. For example, if the questioned signature corresponds to signature 1, the background forged dataset $\{\mathcal{B}_{j,1:\tilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ is constructed from signatures 2, 3, etc., excluding repetitions of the signature under examination. This protocol was repeated across all experimental settings, resulting in the generation of over 283,000 distinct cases. In these experiments which the univariate framework is validated, the following LR is assessed:

$$LR = \frac{f(\boldsymbol{y}_{1:\tilde{T}}|\widehat{\Theta}_{\{\boldsymbol{x}_{j,1:\tilde{T}}\}_{j=1}^{n},H_1)}}{f(\boldsymbol{y}_{1:\tilde{T}}|\widehat{\Theta}_{\{\mathcal{B}_{j,1:\tilde{T}}\}_{j=1}^{N_{\mathcal{B}}},H_2)}} \tag{B.1}$$

where $\boldsymbol{y}_{1:\tilde{T}}$ the feature values of the questioned signature, $\widehat{\Theta}$ the estimated parameters based on the $\{\boldsymbol{x}_{j,1:\tilde{T}}\}_{j=1}^{n}$ feature values of n control signatures and $\{\mathcal{B}_{j,1:\tilde{T}}\}_{j=1}^{N_{\mathcal{B}}}$ feature values of $N_{\mathcal{B}}$ forged signatures unrelated with the case.

Tables B.3.1, B.3.2 and B.3.3 present total error rates (TER) from dynamic signature evaluation experiments conducted using Hidden Markov Models (HMM) with two different structures, for 1 to 8 HMM states and for each considered dynamic feature. TER indicates the total proportion of misclassifications, namely is equivalent to the summation of the false positive rate and the false negative rate.

Structure	**States**	**Total Error Rate**						
		P	**dp1**	**dt1**	**dt2**	**TVD**	**TAD**	**XY**
Left-to-Right	1	0.210	0.595	0.199	0.766	0.106	0.275	0.229
	2	0.228	0.425	0.341	0.163	0.028	0.271	0.269
	3	0.095	0.308	0.085	0.144	0.089	0.225	0.099
	4	0.145	0.228	0.073	0.078	0.061	0.115	0.081
	5	0.147	0.236	0.080	0.143	0.048	0.107	0.691
	6	0.107	0.586	0.059	0.297	0.053	0.104	0.889
	7	0.169	0.809	0.281	0.283	0.073	0.130	0.156
	8	0.187	0.788	0.040	0.956	0.107	0.341	0.170
Ergodic	1	0.210	0.595	0.199	0.766	0.106	0.275	0.229
	2	0.497	0.387	0.221	0.799	0.122	0.159	0.273
	3	0.159	0.301	0.266	0.448	0.177	0.080	0.216
	4	0.163	0.233	0.243	0.430	0.190	0.057	0.224
	5	0.201	0.256	0.185	0.484	0.083	0.068	0.181
	6	0.143	0.222	0.177	0.460	0.069	0.047	0.138
	7	0.144	0.221	0.138	0.441	0.037	0.038	0.111
	8	0.139	0.186	0.116	0.432	0.027	0.033	0.093

Table B.3.1: Experimental results of dynamic signature evaluation for different HMM model structures, numbers of states, and features for Signature 1.

Structure	States	Total Error Rate						
		P	**dp1**	**dt1**	**dt2**	**TVD**	**TAD**	**XY**
Left-to-Right	1	0.140	0.147	0.233	0.222	0.404	0.401	0.346
	2	0.101	0.284	0.140	0.119	0.228	0.169	0.200
	3	0.053	0.256	0.083	0.073	0.211	0.120	0.211
	4	0.049	0.242	0.088	0.068	0.126	0.022	0.184
	5	0.043	0.154	0.082	0.074	0.027	0.009	0.149
	6	0.042	0.237	0.075	0.058	0.045	0.019	0.163
	7	0.062	0.136	0.069	0.049	0.034	0.020	0.165
	8	0.064	0.216	0.079	0.057	0.049	0.018	0.156
Ergodic	1	0.140	0.147	0.233	0.222	0.404	0.401	0.346
	2	0.111	0.278	0.138	0.126	0.084	0.082	0.302
	3	0.087	0.257	0.157	0.096	0.097	0.103	0.266
	4	0.058	0.230	0.130	0.063	0.088	0.154	0.309
	5	0.055	0.264	0.153	0.056	0.066	0.131	0.337
	6	0.062	0.242	0.147	0.095	0.046	0.118	0.339
	7	0.061	0.218	0.142	0.100	0.046	0.131	0.315
	8	0.058	0.200	0.089	0.090	0.023	0.125	0.298

Table B.3.2: Experimental results of dynamic signature evaluation for different HMM model structures, numbers of states, and features for Signature 2.

Structure	States	Total Error Rate						
		P	**dp1**	**dt1**	**dt2**	**TVD**	**TAD**	**XY**
Left-to-Right	1	0.064	0.580	1.000	1.058	1.012	1.019	0.096
	2	0.124	0.120	0.006	0.098	0.325	0.226	0.014
	3	0.449	0.066	0.006	0.264	0.094	0.053	0.997
	4	0.251	0.049	0.137	0.421	0.097	0.035	0.970
	5	0.358	0.056	0.077	0.569	0.054	0.032	0.977
	6	0.312	0.069	0.432	0.535	0.048	0.021	0.736
	7	0.072	0.058	0.056	0.692	0.105	0.050	0.976
	8	0.125	0.221	0.211	0.543	0.082	0.071	0.976
Ergodic	1	0.064	0.580	1.000	1.058	1.012	1.019	0.096
	2	0.037	0.021	0.038	0.956	0.925	0.008	0.190
	3	0.185	0.014	0.022	0.249	0.400	0.000	0.025
	4	0.083	0.024	0.015	0.213	0.090	0.005	0.028
	5	0.047	0.036	0.018	0.203	0.024	0.009	0.017
	6	0.081	0.022	0.012	0.187	0.073	0.005	0.022
	7	0.069	0.043	0.011	0.198	0.035	0.022	0.013
	8	0.095	0.029	0.005	0.172	0.032	0.018	0.027

Table B.3.3: Experimental results of dynamic signature evaluation for different HMM model structures, numbers of states, and features for Signature 3.

Appendix C

Wishart's Degrees of Freedom Estimation

The Wishart distribution plays a fundamental role in multivariate statistics, particularly in modeling covariance matrices. However, accurately estimating its parameters can be challenging. Moreover, the Bayes factor is highly sensitive to parameter choices, especially to the specification of the degrees of freedom, as demonstrated in Section 4.8.2. Therefore, in this chapter, we investigate several estimators for the parameters of the Wishart distribution, with particular emphasis on the degrees of freedom.

The Wishart (1928) distribution is a generalization of the gamma distribution to multiple dimensions. This distribution is crucial for characterizing covariance matrices in multivariate statistics. Furthermore, in Bayesian statistics, the Wishart distribution is the conjugate prior of the inverse covariance matrix of a multivariate normal distribution (Box and Tiao, 1973).

Let $\mathbf{X}_i$ be a random sample of positive-definite matrices of size $p \times p$ and $i = 1, \ldots, m$ then:

$$\mathbf{X}_i \sim W_p(\mathbf{V}, n),$$

where m is the sample size, $\mathbf{V}$ the scale matrix ($p \times p$ fixed symmetric and positive-definite) and n the degrees of freedom ($\in \mathbb{R} \ > p \ - 1$), where $\mathbb{R}$ denotes the set of real numbers. Here, we focus on approaches for estimating the degrees of freedom n, and, secondarily, the scale matrix $\mathbf{V}$.

Under this formulation, for $n > p - 1$, the probability density function is given by

$$f_{\mathbf{x}}(\mathbf{X}) = \frac{|\mathbf{X}|^{\frac{(n-p-1)}{2}} e^{-\frac{tr(\mathbf{V}^{-1}\mathbf{X})}{2}}}{2^{\frac{np}{2}} |\mathbf{V}|^{\frac{n}{2}} \Gamma_p(\frac{n}{2})} \tag{C.1}$$

where $|\mathbf{X}|$ is the determinant of $\mathbf{X}$ and Γ_p is the multivariate gamma function (see Gupta and Nagar, 1999, Chap. 3) defined as

$$\Gamma_p\left(\frac{n}{2}\right) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right). \tag{C.2}$$

If the dimension is one ($p = 1$) and $\mathbf{V}$ (which is now a scalar) is equal to one, then this distribution is a chi-squared distribution with n degrees of freedom.

In Bayesian statistics, as mentioned above, the Wishart distribution is the conjugate prior of the precision matrix Σ^{-1} in a Gaussian model, where Σ is the covariance matrix. This gives rise to the Inverse-Wishart distribution, which results as the distribution of $\mathbf{X}^{-1}$, which is the inverse of $\mathbf{X} \sim$

$W_p(\mathbf{V}, n)$ with $\mathbf{V}$ being a positive-definite matrix of dimension $p \times p$ (see Gelman et al., 1995, Section 3.6) and n being the degrees of freedom. Under this setup, we use the notation $\mathbf{X}^{-1} \sim IW_p(\mathbf{V}^{-1}, n)$. However, this prior has several limitations. The uncertainty that characterizes the variability of data is governed by a single degree of freedom parameter, which can be restrictive as pointed out by Gelman (2006) and Gelman et al. (1995). Therefore, this chapter is dedicated to studying several approaches to modeling the degrees of freedom and improvements of their estimation.

Certainly, this study emphasizes the scenarios where the Wishart distribution is particularly suitable, such as when the assumptions of multivariate normality and the sample covariance matrix being a reliable estimator hold.

The chapter is organized into five sections. Initially, we give a brief literature review on covariance matrices in Section C.1. We start our analysis with the maximum likelihood estimation Section C.2. We discuss the Bayesian modeling analyzed with varying prior specifications for the degrees of freedom (Section C.3). Subsequently, we provide a detailed description of the MCMC algorithms, including hybrid MCMC methods (Section C.3.1) and the No-U-Turn Sampler (NUTS) implemented in Stan (Section C.3.2). The efficacy of these approaches is illustrated and compared in Section C.4 through the use of both simulated and real data. Finally, Section C.5 concludes the chapter with a discussion concerning the outcomes and the potential alternative methods applicable to the problem of finding an optimal setup for the Wishart distribution.

C.1 Literature Review

Bayesian inference for covariance matrices has been extensively studied over several decades (Box and Tiao, 1973). This review highlights the most influential contributions to the field, with a focus on prior specification strategies and their impact on statistical inference.

Leonard and Hsu (1992) laid the groundwork with their seminal paper on Bayesian inference for covariance matrices, providing a comprehensive framework for understanding the statistical properties and applications of the Wishart distribution in multivariate statistics.

Building on this foundation, Barnard et al. (2000) proposed an innovative approach to modeling covariance matrices in terms of standard deviations and correlations. Their separation strategy provided a more intuitive parameterization, allowing researchers to specify separate priors for variance components and correlation structures.

The Inverse-Wishart distribution has long been the traditional choice for covariance matrix priors due to its conjugacy with the multivariate normal likelihood. However, as Hsu et al. (2012) notes, this distribution has two major shortcomings: it applies the same confidence level to all matrix elements through a single degree of freedom parameter, and it lacks flexibility to model potential interdependencies within the covariance structure.

The field has recently witnessed growing interest in the Lewandowski-Kurowicka-Joe (LKJ) distribution (Lewandowski et al., 2009), introduced as a probability distribution over correlation matrices. The LKJ distribution offers a more flexible approach to specifying priors on correlation matrices with a single shape parameter that controls the concentration of correlations. Its implementation in popular probabilistic programming languages such as *Stan* has facilitated wider adoption in Bayesian hierarchical modeling applications.

Alvarez et al. (2014), Liu et al. (2016), and Zhang (2021) explored Bayesian inference for covariance matrices, emphasizing the use of various priors and their impact on estimation accuracy, computational efficiency, and the theoretical limitations of these priors. Their work highlighted the practical challenges

and solutions in applying Bayesian methods to real-world data within complex hierarchical models.

The evolution of Bayesian inference for covariance matrices reflects a broader trend in Bayesian statistics toward more flexible and intuitive prior specifications. The literature reveals a lack of fully Bayesian approaches for estimating the degrees of freedom in covariance matrix modeling. In this work, we investigate various computational algorithms for posterior inference of the degrees of freedom. A major challenge arises from the fact that the problem becomes significantly more complex as the data dimension increases. As we will show in the following sections, different computational methods exhibit varying levels of efficiency depending on the dimensionality of the problem.

C.2 Maximum Likelihood Estimation of Wishart Distribution

The maximum likelihood method is the most common and standard approach for estimating the parameters of a statistical model. Therefore, for $\mathbf{X}_i$ independent and identically distributed m random variables that follow a Wishart distribution with parameters $\mathbf{V}$ and n, the likelihood is expressed as:

$$f_x(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m|\mathbf{V}, n) = \prod_{i=1}^{m} f_x(\mathbf{X}_i|\mathbf{V}, n) = \prod_{i=1}^{m} \frac{|\mathbf{X}_i|^{\frac{(n-p-1)}{2}} e^{-\frac{tr(\mathbf{V}^{-1}\mathbf{X}_i)}{2}}}{2^{\frac{np}{2}}|\mathbf{V}|^{\frac{n}{2}}\Gamma_p(\frac{n}{2})} \tag{C.3}$$

The log-likelihood is then given by

$$\ell(\mathbf{V}, n) = \frac{n-p-1}{2}\sum_{i=1}^{m}\log|\mathbf{X}_i| - \frac{nmp}{2}\log 2 - m\log\Gamma_p\left(\frac{n}{2}\right) - \frac{nm}{2}\log(|\mathbf{V}|) - \frac{1}{2}\sum_{i=1}^{m} tr(\mathbf{V}^{-1}\mathbf{X}_i)$$

where $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$. The maximum likelihood estimation of $\mathbf{V}$ for a given value of the degrees of freedom can be found in a straightforward manner and it is equal to the mean of $\mathbf{X}_i$'s devided by the degrees of freedom, that is

$$\widehat{\mathbf{V}} = \frac{1}{nm}\sum_{i=1}^{m}\mathbf{X}_i = \frac{1}{n}\overline{\mathbf{X}}. \tag{C.4}$$

Equivalently, for the degrees of freedom based on Equation C.3 the maximum likelihood estimation of n given the scale matrix $\mathbf{V}$ can be found by solving the equation:

$$\frac{\partial}{\partial n}\log\Gamma_p\left(\frac{n}{2}\right) = \frac{1}{2}\overline{\log|\mathbf{X}|} - \frac{1}{2}p\log(2) - \frac{1}{2}\log|\mathbf{V}|, \tag{C.5}$$

where $\overline{\log|\mathbf{X}|}$ is is the sample mean of $\log|\mathbf{X}_i|$ being the log determinant of matrix $\mathbf{X}_i$.

The multivariate digamma function is the derivative of the logarithm of the multivariate gamma function (see Gupta and Nagar, 1999). Hence, $\frac{\partial}{\partial n}\log(\Gamma_p(\frac{n}{2})) = \frac{1}{2}\psi_p(\frac{n}{2})$ where ψ_p the multivariate digamma of function of dimension p. Based on the definition of ψ_p, this can be further simplified to

$$\frac{\partial}{\partial n}\log\left\{\Gamma_p\left(\tfrac{n}{2}\right)\right\} = \sum_{j=1}^{p}\frac{1}{2}\psi\left(\tfrac{n-j+1}{2}\right).$$

For $p = 1$, it is the same as the univariate digamma function (see Abramowitz and Stegun, 1965).

From the above, we end up in the following equation, which needs to be solved in order to obtain

the maximum likelihood estimator (MLE) of n:

$$\sum_{j=1}^{p} \psi\left(\tfrac{n-j+1}{2}\right) = \overline{\log|\mathbf{X}|} - \log|\mathbf{V}| - p\log(2)\,. \tag{C.6}$$

Moreover, the second derivative of the log-likelihood should be negative, that is $\frac{\partial^2}{\partial^2 n}\ell(n;\mathbf{X},\mathbf{V}) = -m\frac{\partial^2}{\partial^2 n}\log(\Gamma_p(\frac{n}{2})) = -m\ \psi_p'\left(\frac{n}{2}\right) < 0$, since the trigamma function $\psi_p'(n)$ is positive for any $n > 0$. Since (C.6) does not have an analytical solution, the MLE of n can be obtained by any root-finding algorithm.

The resulting maximum likelihood Eq. (C.6) provides insight into the role of the degrees of freedom, which affect the shape of the distribution. The left term of the equation contains the digamma function, an increasing function of the degrees of freedom, and the right term measures the discrepancy between the data matrices and the scale matrix parameter, adjusted by a penalty term which depends on the dimension of the Wishart distribution. Hence, larger values in the right-hand side of the equation indicate a higher variance, which implies a lower degree of freedom, and vice versa.

In Section C.2.1, which follows, we proceed with the root-finding algorithms and their implementation for the estimation of the degrees of freedom.

C.2.1 Root-finding Algorithms for MLE of the Degrees of Freedom

In this section, we proceed with numerical approaches for finding the solutions of Eq. (C.6). By considering the estimator $\widehat{\mathbf{V}} = \frac{1}{n}\overline{\mathbf{X}}$ obtained in Eq. (C.4), we replace the parameter $\mathbf{V}$ with the estimator in (C.6) and we obtain:

$$\sum_{j=1}^{p} \psi\left(\tfrac{n-j+1}{2}\right) = \overline{\log|\mathbf{X}|} - \log|\overline{\mathbf{X}}| + p\log\tfrac{n}{2} \tag{C.7}$$

Hence, the function that we need to find its roots is the following

$$f(n) = \sum_{j=1}^{p} \psi\left(\tfrac{n-j+1}{2}\right) - \overline{\log|\mathbf{X}|} + \log|\overline{\mathbf{X}}| - p\log\tfrac{n}{2} \tag{C.8}$$

We proceed by implementing two primary root-finding algorithms for Eq. (C.8): the Bisection method and the Newton-Raphson method. Following these, we will also demonstrate the simulated annealing, a probabilistic technique aimed at approximating the global optimum of a given function. For additional root-finding algorithms that may be applied to this problem, we refer the reader in Chapter 9 of Press et al. (2007).

Bisection method

The Bisection method is a root-finding method that applies to continuous functions. It is a very simple and robust method which is based on the Binary search algorithm. It begins with an interval where the function changes sign. The function is evaluated at the midpoint of the interval determining which sub-interval contains the root. The process is continued until the interval is sufficiently small.

Hence, the method needs two initial starting values with opposite function signs. To ensure this, we may define a wide interval. Nevertheless, towards, the method will result in a relatively slow convergence to the target value. In our case, we specify the lower boundary of the function of interest based on the requirement that the degrees of freedom must be greater than or equal to $p+1$. For the

Algorithm 1 Bisection Algorithm

Require: function f ▷ f the Function C.8
Require: endpoint values $L = (p + 1)$, $U = 1000$
Require: maximum iterations B
Require: tolerance tol
Ensure: $L < U$
Ensure: $(f(L) < 0 \;\&\; f(U) > 0) \;||\; (f(L) > 0 \;\&\; f(U) < 0)$
  for b *in* $1 : B$ **do**
    $C \leftarrow \frac{(L+U)}{2}$
    if $f(C) == 0 \;||\; \frac{(U-L)}{2} < tol$ **then**
      return C
    end if
    if $sign(f(C)) == sign(f(L))$ **then**
      $L \leftarrow C$
    else
      $U \leftarrow C$
    end if
  end for
  Output("Method failed.")

upper boundary, we assign a reasonably large value for degrees of freedom. The method is represented in pseudocode in Algorithm 1.

In Figure C.2.1 we can see the evolution of the Bisection algorithm with respect to the number of iterations for an illustrated dataset (denoted in the following as Simulated Dataset 1) of 30 observations for a Wishart distribution of 10×10 dimension with 45 degrees of freedom and random scale matrix. For more details in the Bisection method see in Burden et al. (2015, Chapter 2.1).

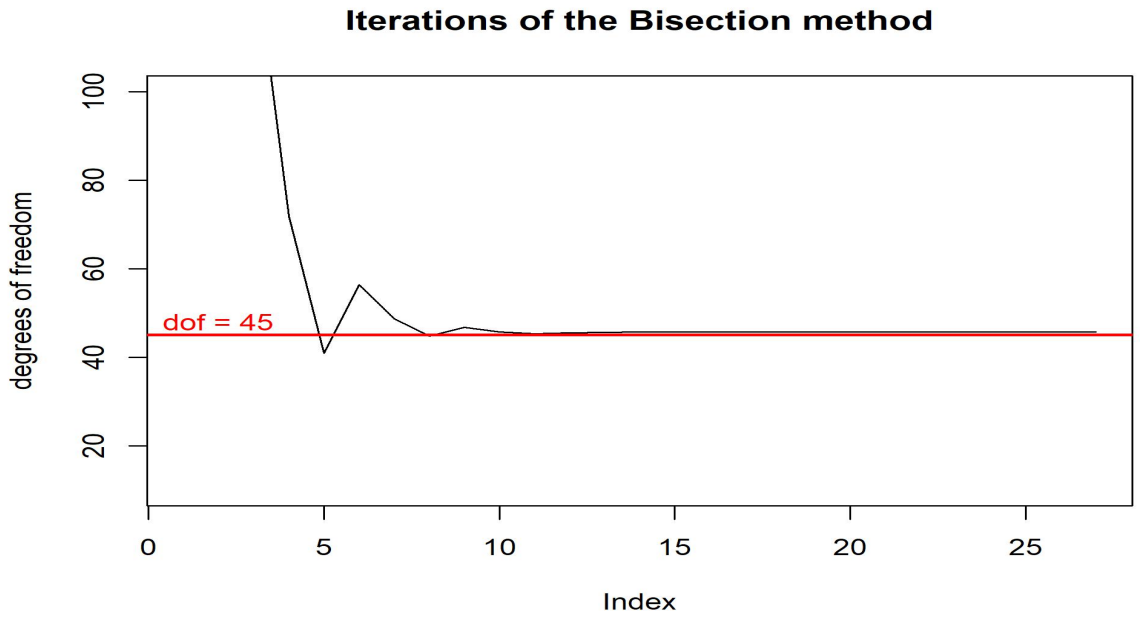


Figure C.2.1: Evolution of the Bisection algorithm with respect to the number of iterations for the Simulated Dataset 1 (30 matrices of dimension 10×10 from the Wishart distribution with 45 degrees of freedom and random scale matrix)

Newton-Raphson method

The Newton-Raphson method is another popular root-finding numerical method. The algorithm is initialized by considering an initial guess for the solution. Iteratively, this value is improved by constructing a tangent line to the curve of the function. The intersection of this tangent with the x-axis yields the next improved estimate. This iterative process continues until the desired level of accuracy is attained. The formula for finding the next improved value is the following

$$n_1 = n_0 - \frac{f(n_0)}{f'(n_0)} \tag{C.9}$$

where $f'(n)$ is the derivative of $f(n)$. In our case, the $f(n)$ is the function Eq. (C.8) and the derivative is given by $-m\ \psi_p'\left(\frac{n}{2}\right)$; where $\psi_p'(n)$ the trigamma function. The algorithm implementation for the Wishart distribution is given in the form of pseudocode in Algorithm 2.

Algorithm 2 Newton-Raphson Algorithm

Require: function f
Require: derivative function f'
Require: maximum iterations B
Require: tolerance tol
  $n \leftarrow p + 1$ ▷ p dimension of the data-matrices
  for b *in* $1 : B$ **do**
    $n_{new} = n - \frac{f(n)}{f'(n)}$
    if $abs(n - n_{new}) < tol$ **then**
      return n_{new}
    end if
  end for
  Output("Method failed.Smallest value of function is $f(n)$")

In Figure C.2.2, we can see the evolution of the Newton-Raphson algorithm with respect to the number of iterations of a 10×10 Wishart distribution with 45 degrees of freedom and a random scale matrix. For more details on the Newton-Raphson method, see Burden et al. (2015, Chapter 2.3).

Root Simulated Annealing

Simulated annealing (Kirkpatrick et al., 1983) is a stochastic technique for approximating the global optimum of a given function. Although it is typically used as an optimization method, it can be adapted for root-finding problems. This involves minimizing the function's absolute value in order to locate its roots. Essentially, we would be looking for a solution that brings the function's value as close to zero as possible. The method introduces some random components in its steps as part of the search process, and it is a stochastic variant of the hill climbing method (see Polyak, 1964). Hill climbing is an optimization algorithm that incrementally improves a candidate optimal value. If the new point is better than the current point, then the current point is replaced with the new point. This process continues for a pre-specified number of iterations.

Simulated annealing follows the same logic but with randomly generated steps. Worse values may be accepted with some probability, depending on a parameter called temperature. Hence, for any current value of the degrees of freedom n_{old}, we generate a new value from a "proposal" distribution $q(n_{new}|n_{old})$. In the following, we use the normal distribution with mean the current value n_{old} and standard deviation equal to two truncated at $p+1$ since the expected value of the Wishart distribution is not defined below this value.

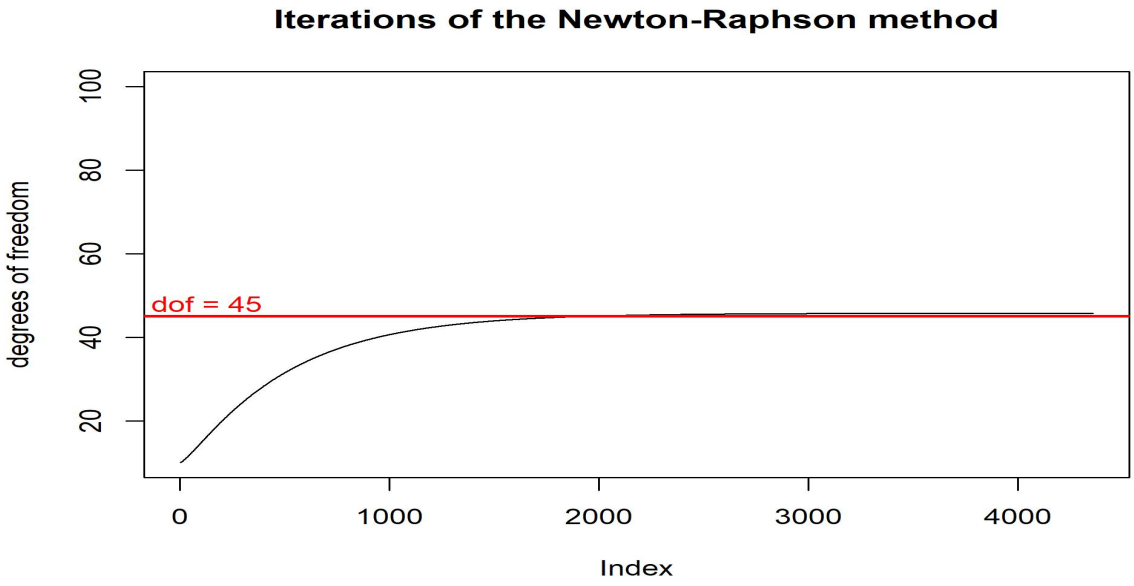


Figure C.2.2: Evolution of the Newton-Raphson algorithm with respect to the number of iterations for the Simulated Dataset 1 (30 matrices of dimension 10×10 from the Wishart distribution with 45 degrees of freedom and random scale matrix)

The initial temperature is specified as a hyperparameter. It decreases as the search progresses. A variety of temperature schedules may be employed to systematically reduce the temperature parameter from its initial value during the search process. The consequence of this temperature reduction is that worse solutions are more likely to be accepted early in the search while they have lower chance to be accepted later in the search. In this way, the process moves in the search area, with the aim to identify a promising region in the early stages. In this implementation, we use the following temperature schedule:

$$T^{(b)} = T^{(b-1)} \times (1 - \epsilon) \tag{C.10}$$

for some small value of $\epsilon > 0$ and for $b = 1, \ldots, B$; where B the total number of iterations of the algorithm; where $T^{(b)}$ is the temperature at iteration b of the algorithm.

Furthermore, a popular choice of acceptance probability function is the "Metropolis criterion" (Delahaye et al., 2019). This approach uses the temperature $T^{(b)}$ at iteration b and the difference between the previous value $n_{old} = n^{(b-1)}$ and the currently proposed value n_{new} as follows:

$$\alpha = \exp\left(-\frac{f(n_{new}) - f(n_{old})}{T}\right). \tag{C.11}$$

New solutions, denoted as n_{new}, are accepted with a probability α, which, in practice, implies acceptance when $\alpha < u$, where u is a random number drawn from a uniform distribution. As a stopping rule, we consider a small threshold value $\xi > 0$ for which the acceptance probability of the absolute function value should be smaller, that is $f_v < \xi$ (see Gall (2020)). The pseudocode for the simulated annealing, implemented in our problem, is given in Algorithm 3.

The main disadvantage of simulating annealing is the tuning of its parameters. It requires careful tuning of the initial temperature, the decreasing criterion ϵ and the stopping threshold ξ, which can be time-consuming and problem-specific.

Algorithm 3 Root Simulated Annealing

Require: function f
Require: iterations B
Require: initial value of temperature $T_0 = 1000$
Require: decreasing criterion $\epsilon = 0.01$
Require: stopping rule $\xi = 0.0001$
  $Temp \leftarrow T_0$
  $n \leftarrow p + 1$ ▷ p dimension of the data-matrices
  $fv \leftarrow abs(f(n))$
  for b *in* $1 : B$ **do**
    $n_{cand} \leftarrow N(n, 2^2)_{[(p+1),+\infty]}$ ▷ random sample from truncated normal
    $fv_{cand} \leftarrow abs(f(n_{cand}))$
    $u \leftarrow unif(0, 1)$ ▷ random sample from uniform with a and b
    $diff \leftarrow fv_{cand} - fv$
    if $fv_{cand} < fv \;||\; u < e^{-\frac{diff}{Temp}}$ **then**
      $n \leftarrow n_{cand}$
      $fv \leftarrow fv_{cand}$
    end if
    $Temp = Temp(1 - \epsilon)$
    if $(fv < \xi)$ **then** ▷ Check for convergence
      return n
    end if
  end for
  Output("Method failed. Smallest value of function is fv")

Figure C.2.3 depicts the evolution of the simulated annealing algorithm with the same dataset used in Bisection and Newton-Raphson (i.e. of sample size of 30 matrices of 10×10 dimensions from Wishart distribution with 45 degrees of freedom and random scale matrix). For more details regarding the root simulated annealing algorithm and detailed description of its convergence and behaviour in applications, see in Bertsimas and Tsitsiklis (1993), Ingber (1993), Henderson et al. (2003).

C.2.2 Simulated Annealing for Maximizing the Likelihood Function

Simulated annealing is an optimization technique tailored for nonlinear objective functions that may present challenges to conventional local search algorithms. Simulated annealing is applicable for the maximization of likelihood functions, a technique that has been effectively utilized across various domains. This includes optimization for signal parameters as delineated by Sharman (1988), in econometric contexts as demonstrated by Goffe et al. (1994), for estimating Weibull parameters as outlined by Abbasi et al. (2006), for three-parameter lognormal distributions as discussed by Vera and Díaz-García (2008), and for the generalized gamma distribution as explored by Idris and Muhammad (2022).

In this section, we will illustrate how simulated annealing can be used to estimate the maximum likelihood parameters of the Wishart distribution based on the likelihood function. To simplify the problem, we will use the relation (C.4) we obtained by optimizing the log-likelihood (C.2) with respect to $\mathbf{V}$. In this way, we will directly optimize by randomly sampling neighbors only from the degrees of freedom.

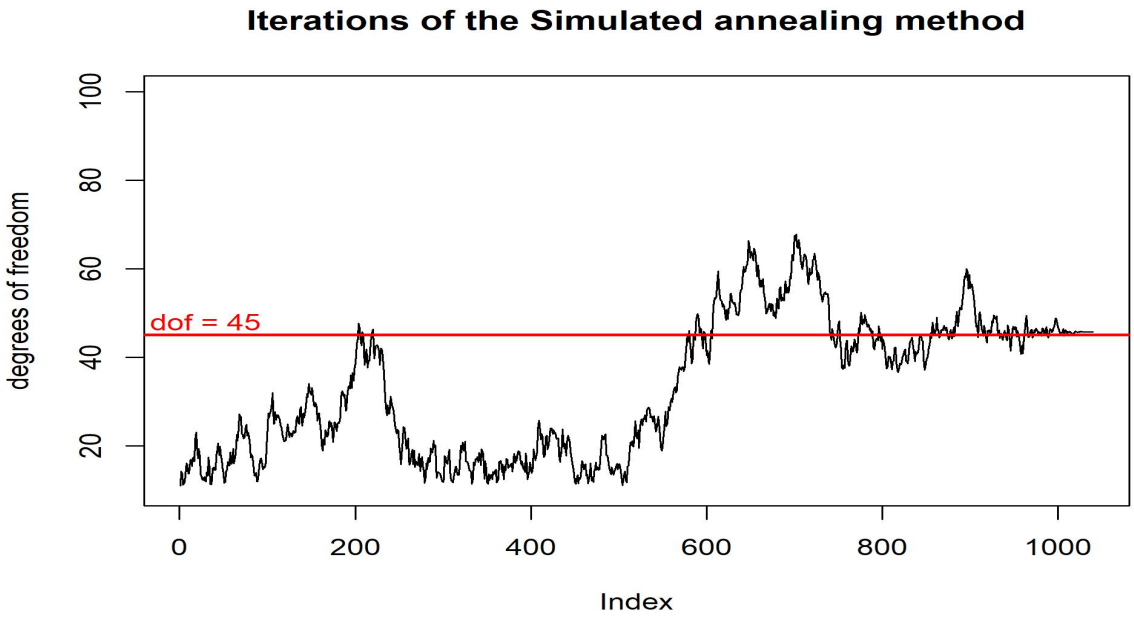


Figure C.2.3: Evolution of the root Simulated annealing algorithm with respect to the number of iterations for the Simulated Dataset 1 (30 matrices of dimension 10×10 from the Wishart distribution with 45 degrees of freedom and random scale matrix)

$$\begin{aligned}\ell'(n) = \ell\Big(\mathbf{V} = \tfrac{1}{n}\overline{\mathbf{X}},\, n\Big) =& \frac{n-p-1}{2}\sum_{i=1}^{m}\log|\mathbf{X}_i| - \frac{nmp}{2}\log 2 - m\log\Gamma_p\left(\frac{n}{2}\right) \\ &+ \frac{pnm}{2}\log n - \frac{nm}{2}\log|\overline{\mathbf{X}}| - \frac{n}{2}\sum_{i=1}^{m} tr\left(\overline{\mathbf{X}}^{-1}\mathbf{X}_i\right) \\ =& m\left[\frac{n-p-1}{2}\overline{\log|\mathbf{X}|} + \frac{np}{2}\log n - \log\Gamma_p\left(\frac{n}{2}\right) - \frac{n}{2}\Phi(\mathbf{X})\right]\end{aligned}$$

where

$$\Phi(\mathbf{X}) = p\log 2 + \log\left|\overline{\mathbf{X}}\right| + \frac{\sum_{i=1}^{m} tr\left(\overline{\mathbf{X}}^{-1}\mathbf{X}_i\right)}{m}$$

Simulated annealing is based on the (improved) log-likelihood function to determine if a new point is a better over the current one. The specific hyperparameter values for the temperature parameter and the decreasing criterion are described in the root Simulated Annealing. Due to the probabilistic nature of simulated annealing, the convergence of the algorithm to the absolute global maximum of the likelihood function is not guaranteed. However, several strategies can enhance confidence in the obtained results (Granville et al. (1994)). These methodologies, while not ensuring clear convergence to the global optimum, can significantly assist in evaluating the algorithm's convergence properties. One commonly employed stopping criterion involves terminating the algorithm when the solution remains unchanged for a predefined number of iterations (Delahaye et al. (2019)). The specific details of the implemented simulated annealing algorithm are presented in Section 4.

Figure C.2.4 presents the evolution of the simulated annealing for log-likelihood maximization with the same dataset used in Bisection and Newton-Raphson (i.e., of sample size of 30 matrices of 10×10 dimensions from Wishart distribution with 45 degrees of freedom and random scale matrix).

Algorithm 4 Maximum Likelihood Simulated Annealing

Require: Wishart probability density function f_x
Require: data $\mathbf{X}$
Require: iterations B = 10000
Require: initial value of temperature $T_0 = 1000$
Require: decreasing criterion $\epsilon = 0.01$
Require: Number of iterations for stability check $stab\_th = 1000$

$Temp \leftarrow T_0$
$stable\_count \leftarrow 0$
$m \leftarrow length(\mathbf{X})$
$n \leftarrow p + 1$ ▷ p dimension of the data-matrices

$\bar{\mathbf{X}} \leftarrow \frac{\sum_{i=1}^{m} \mathbf{X}_i}{m}$
$\mathbf{V} \leftarrow \frac{\bar{\mathbf{X}}}{n}$

$lik \leftarrow \sum_{i=1}^{m} \log\left(f_x(\mathbf{X}_i, n, \mathbf{V})\right)$
for b *in* $1 : B$ **do**
  $n_{new} \leftarrow N(n, 2^2)_{[(p+1),+\infty]}$ ▷ random sample from truncated normal
  $\mathbf{V}_{new} \leftarrow \frac{\bar{\mathbf{X}}}{n_{new}}$

  $lik_{new} \leftarrow \sum_{i=1}^{m} \log\left(f_x(\mathbf{X}_i, n_{new}, \mathbf{V}_{new})\right)$

  $u \leftarrow unif(0, 1)$ ▷ random sample from uniform with a and b
  $diff \leftarrow lik - lik_{new}$
  if $lik_{new} > lik \;||\; u < e^{-\frac{diff}{Temp}}$ **then**
    $n \leftarrow n_{new}$
    $\mathbf{V} \leftarrow \mathbf{V}_{new}$
    $lik \leftarrow lik_{new}$
    $stable\_count \leftarrow 0$
  else
    $stable\_count \leftarrow stable\_count + 1$
  end if
  $Temp = Temp(1 - \epsilon)$
  if $stable\_count > stab\_th$ **then**
    return list(n,$\mathbf{U}$)
  end if
end for
Output("Method failed.")

C.3 Bayesian Modeling of Wishart Distribution

This section explores the implementation of Bayesian inference for the Wishart distribution. In a Bayesian framework, prior distributions are specified for model parameters under estimation. Following Leonard and Hsu (1992), the conjugate prior for the scale matrix of the Wishart distribution is the Inverse-Wishart distribution. However, for the degrees of freedom parameter of the Wishart distribution, there is no known conjugate prior family.

Since the degrees of freedom refer to a positive continuous parameter, a natural prior choice is to assume a distribution that is positive definite. In this study, we analyze and compare the choices of Uniform (Unif), Exponential (Exp), Gamma, Inverse-Gamma (IGamma), and LogNormal (LNormal) distributions. The uniform distribution serves as an uninformative prior in Bayesian inference, characterized by a constant probability density across all values within its wide specified bounds $(0, 10^4)$. For the other distributions, we implement a non-informative setup by considering a very wide range

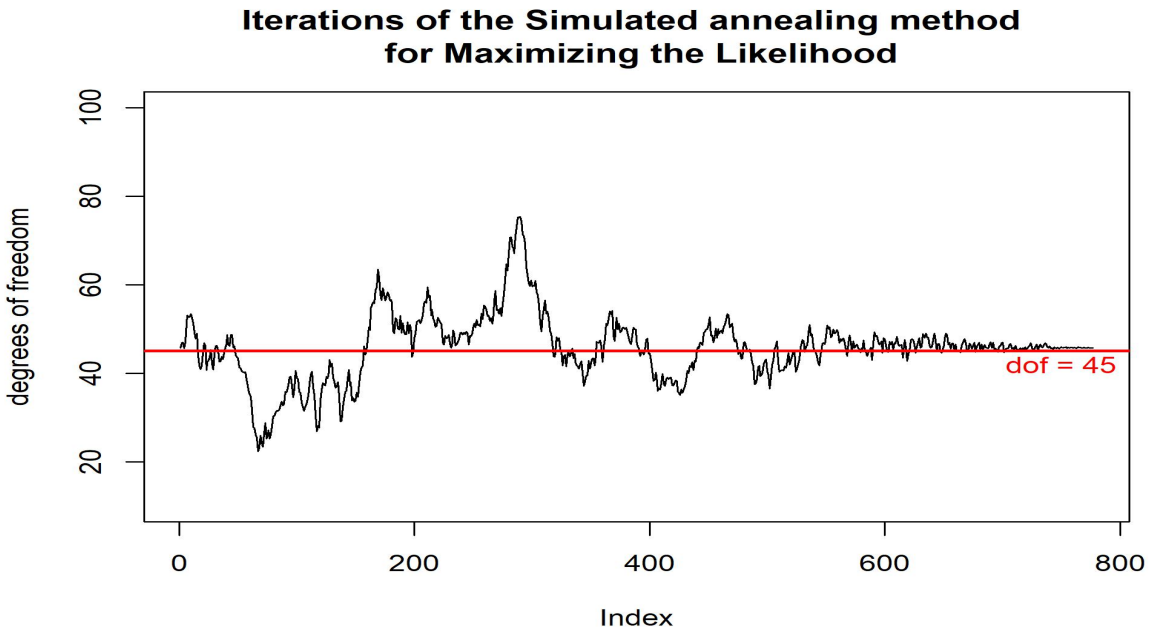


Figure C.2.4: Evolution of the Simulated annealing for maximizing the likelihood for the Simulated Dataset 1 (30 matrices of dimension 10×10 from the Wishart distribution with 45 degrees of freedom and random scale matrix)

of values like Exponential(rate $= 10^{-4}$), Gamma(shape $= 10^{-4}$, rate $= 10^{-4}$), Inverse-Gamma(shape $= 10^{-4}$, rate $= 10^{-4}$) and LogNormal(0, 5^2). Even non-informative priors carry some degree of influence, and their selection should be considered carefully in the context of Bayesian analysis. Hence, by considering the above prior setup, the complete Bayesian model can be formally described as:

$$\begin{aligned} \mathbf{X} &\sim W_p(\mathbf{V}, n) \\ \mathbf{V} &\sim IW_p(\mathbf{U}, n_v) \\ n &\sim f_n \end{aligned}$$

where $\mathbf{V}$ and n are assigned independent priors. The probability density function of the Inverse-Wishart prior, $\mathbf{V} \sim IW_p(\mathbf{U}, n_v)$, takes the following form:

$$f(\mathbf{V}; \mathbf{U}, n_v) = \frac{1}{2^{\frac{np}{2}} \Gamma_p(\frac{n_v}{2})} |\mathbf{U}|^{\frac{n_v}{2}} |\mathbf{V}|^{-\frac{1}{2}(n_v+p+1)} e^{-\frac{1}{2} tr(\mathbf{U}\mathbf{V}^{-1})} \tag{C.12}$$

The joint posterior distribution of the model based on Bayes' theorem can be written as:

$$\begin{aligned} \pi(\mathbf{V}, n|\mathbf{X}) = & \frac{f(\mathbf{X}|\mathbf{V}, n)\pi(\mathbf{V})\pi(n)}{f(\mathbf{X})} \\ \propto & \prod_{i=1}^{m} \left(\frac{|\mathbf{X}_i|^{\frac{(n-p-1)}{2}} e^{-\frac{tr(\mathbf{V}^{-1}\mathbf{X}_i)}{2}}}{2^{\frac{np}{2}} |\mathbf{V}|^{\frac{n}{2}} \Gamma_p(\frac{n}{2})} \right) \times \frac{|\mathbf{U}|^{\frac{n_v}{2}} |\mathbf{V}|^{-\frac{(n_v+p+1)}{2}} e^{-\frac{tr(\mathbf{U}\mathbf{V}^{-1})}{2}}}{2^{\frac{n_v p}{2}} \Gamma_p(\frac{n_v}{2})} \times f_n \\ \propto & \; 2^{-\frac{nmp}{2}} |\mathbf{V}|^{-\frac{nm+n_v+p+1}{2}} \left[\Gamma_p\left(\frac{n}{2}\right) \right]^{-m} \\ & \times \exp\left(\frac{(n-p-1)}{2} m \, \overline{\log|\mathbf{X}|} - \frac{1}{2} tr\left(\mathbf{V}^{-1} \left(\mathbf{U} + m\overline{\mathbf{X}} \right) \right) \right) \times f_n \end{aligned} \tag{C.13}$$

For the prior specification of the scale matrix $\boldsymbol{V}$, a non-informative prior is chosen, being $\boldsymbol{U}$ a diagonal matrix $p \times p$ of values 10^{-4}. Furthermore, we set the degrees of freedom n_v equal to the dimension p. In this manner, degrees of freedom of the Inverse-Wishart prior are set in a way that the prior variance $\mathbf{V}$ goes to infinity.

Since the posterior distribution for n cannot be expressed in a closed form regardless of the prior chosen, Markov chain Monte Carlo methods (see Section 3.2.2) are employed to address this challenge. These methods generate samples from the target distribution, allowing a good approximation of its properties.

C.3.1 Hybrid MCMC Methods

The MCMC methods can be combined to create more complex and versatile algorithms. The Metropolis-within-Gibbs algorithm (Gilks, Best and Tan, 1995) is a particularly popular example of this approach. The Gibbs sampler works iteratively by updating each random variable of interest (or group of variables) in turn, conditional on the current values of the other variables. Here the role of random variables is played by the parameters under estimation, which in the Bayesian framework are indeed random variables. On the other hand, the Metropolis-Hastings algorithm works in a different fashion, by proposing (from a proposal distribution) a new state based on the current state. This proposed new state is then accepted or rejected based on a probability that requires only the unnormalized probability density function. The combination of these two approaches leads to a hybrid method where each parameter (or block of parameters) is updated/generated from the conditional posterior distributions (which are usually univariate). When the conditional posterior distribution has a known form, a Gibbs step is applied; otherwise, a Metropolis-Hastings step is performed.

This is directly applicable to the model presented in Section C.3, where we can obtain the posterior conditional distribution for the scale matrix ($\mathbf{V}$), but not for the degrees of freedom (n), as we demonstrate in the following.

$$
\begin{aligned}
\pi(\mathbf{V}|\mathbf{X},n) \propto\ & 2^{-\frac{nmp}{2}}|\mathbf{V}|^{-\frac{nm+n_v+p+1}{2}}\left[\Gamma_p\left(\tfrac{n}{2}\right)\right]^{-m} \\
& \times \exp\left(\frac{(n-p-1)}{2}m\,\overline{\log|\mathbf{X}|}-\frac{1}{2}tr\Big(\mathbf{V}^{-1}\big(\mathbf{U}+m\overline{\mathbf{X}}\big)\Big)\right) \\
\propto\ & |\mathbf{V}|^{-\frac{nm+n_v+p+1}{2}}\exp\left(-\frac{1}{2}tr\Big(\mathbf{V}^{-1}\big(m\overline{\mathbf{X}}+\mathbf{U}\big)\Big)\right) \xrightarrow{tr(AB)=tr(BA)} \\
\propto\ & |\mathbf{V}|^{-\frac{nm+n_v+p+1}{2}}\exp\left(-\frac{1}{2}tr\Big(\big(m\overline{\mathbf{X}}+\mathbf{U}\big)\mathbf{V}^{-1}\Big)\right) \\
\sim\ & IW_p\big(m\overline{\mathbf{X}}+\mathbf{U}\,,\,nm+n_v\big).
\end{aligned}
\tag{C.14}
$$

Then the conditional distribution of $(\mathbf{V}|\mathbf{X},n)$ follows an Inverse-Wishart distribution with scale matrix $m\overline{\mathbf{X}}+\mathbf{U}$ and degrees of freedom $nm+n_v$. For the degrees of freedom the corresponding posterior conditional probability distribution is given by:

$$
\begin{aligned}
\pi(n|\mathbf{X},\mathbf{V}) \propto\ & 2^{-\frac{nmp}{2}}|\mathbf{V}|^{-\frac{nm+n_v+p+1}{2}}\left[\Gamma_p\left(\tfrac{n}{2}\right)\right]^{-m} \\
& \times \exp\left(\frac{(n-p-1)}{2}m\,\overline{\log|\mathbf{X}|}-\frac{1}{2}tr\Big(\mathbf{V}^{-1}\big(\mathbf{U}+m\overline{\mathbf{X}}\big)\Big)\right)\times f_n \\
\propto\ & \frac{\left[2^{-\frac{mp}{2}}|\mathbf{V}|^{-\frac{m}{2}}\right]^n}{\Gamma_p\left(\frac{n}{2}\right)^m}\exp\left(\frac{nm}{2}\,\overline{\log|\mathbf{X}|}\right)\times f_n.
\end{aligned}
\tag{C.15}
$$

In this study, for sampling the degrees of freedom n, we implement and compare three different

MCMC approaches: the Random Walk Metropolis (RWM), the Slice Sampling (SlS) and the Hamiltonian Monte Carlo (HMC). The pseudocode for the implemented hybrid MCMC is presented in Algorithm 5.

Algorithm 5 Hybrid MCMC

Require: data $\mathbf{X}$
Require: prior specification of degrees of freedom f_n
Require: MCMC algorithm specification within Gibbs $MCMC_{choice}$= (RWM, SlS, HMC)
Require: iterations $B = 3000$, burn-in $burn\_in = 1000$
  $p \leftarrow dim(\mathbf{X})$
  $m \leftarrow length(\mathbf{X})$
  $n_v \leftarrow p$
  $\mathbf{U} \leftarrow diag_{p\times p}(1\text{e-}4)$
  $n[1] \leftarrow p$
  for b *in* $2 : B$ **do**
    $V[b] \leftarrow W^{-1}(\sum_{i=1}^{m} \mathbf{X}_i + \mathbf{U}, n[b-1]m + n_v)$ ▷ random sample from Inverse-Wishart
    $n[b] \leftarrow MCMC_{choice}(\mathbf{X}, n[b-1], V[b], f_n)$
  end for
  return $list(n[burn\_in : B], V[burn\_in : B])$

Random Walk Metropolis within Gibbs

The Random Walk Metropolis (RWM) algorithm is a special case of the Metropolis-Hastings algorithm (see Hastings, 1970). The RWM algorithm is widely used because it's simple and can be applied to a wide range of problems. The algorithm used to obtain a sequence of random samples from a probability distribution with a simple uniform proposal, in our case $n^{(b)} \sim U(n^{(b-1)} - \delta,\, n^{(b-1)} + \delta)$, where $n^{(b)}$ is value of the degrees of freedom generated in the b$-th$ observation of the MCMC algorithm and δ is a tuning parameter controlling the range and the variance of the uniform proposal distribution. Consequently, the accept/reject mechanism behaves in the following manner: it accepts proposed values that improve the conditional posterior density function, while occasionally accepting or rejecting proposed values that reduce the corresponding posterior density.

Moreover, in the context of our study, it is imposed a constraint that the degrees of freedom must exceed $(p - 1)$, as delineated in the beginning of Chapter C. Consequently, within the confines of the algorithm, we ascertain whether the newly accepted value surpasses this threshold. For an in-depth exposition, refer to Algorithm 6.

Slice Sampling within Gibbs

Slice Sampling is an MCMC technique for drawing samples from a probability distribution by sampling uniformly from the region under the plot of the target density function (C.15). It represents an auxiliary variable technique that allows for efficient sampling without the need for tuning step size parameters. It works by defining a "slice", a horizontal level under the probability density function, and then uniformly sampling from the region above this slice within the distribution. This method ensures that the samples are distributed according to the target probability distribution without the need for complex calculations or transformations. The efficiency of Slice Sampling is highly dependent on the choice of the width of slice of the selected uniform distribution and the method used to find the interval. For detailed discussions and different specifications of the Slice Sampling algorithm, see Neal

Algorithm 6 Random Walk Metropolis (RWM)

Require: data $\mathbf{X}$
Require: n the initial value of degrees of freedom
Require: $\mathbf{V}$ scale matrix
Require: prior specification of degrees of freedom f_n
Require: analogous pdf of degrees of freedom $\pi(n|\mathbf{X}, \mathbf{V})$ ▷ function C.15
  $\delta \leftarrow 3$
  $y \leftarrow runif(n - \delta, n + \delta)$ ▷ random sample from uniform
  if $y > (p - 1)$ **then**
    $accept \leftarrow min(\frac{\pi(y|\mathbf{X},\mathbf{V})}{\pi(n|\mathbf{X},\mathbf{V})}, 1)$
    $u \leftarrow runif(0, 1)$ ▷ random sample from uniform
    if $u < accept$ **then**
      $n_{new} \leftarrow y$
    end if
  else
    $n_{new} \leftarrow n$
  end if
  return n_{new}

(2003). Moreover, we have the constraint that the degrees of freedom must exceed $(p - 1)$, for more details see the Algorithm 7.

Hamiltonian Monte Carlo within Gibbs

Hamiltonian Monte Carlo (HMC) (Neal, 2012) is an advanced MCMC algorithm that leverages the concepts of Hamiltonian mechanics to produce efficient sampling from high-dimensional probability distributions. Unlike traditional MCMC methods, HMC delimits random walk behavior and explores the target distribution more efficiently by simulating the dynamics of a particle moving through the distribution's potential energy landscape. The algorithm alternates between updating the particle's position (representing the sample) and its momentum (auxiliary variables) using Hamilton's equations, which helps in proposing new states that are far from the current state but still have a high probability of acceptance. The method uses the leapfrog integration method for numerical stability and accuracy in simulating the Hamiltonian dynamics. The Metropolis acceptance step ensures detailed balance, making the algorithm asymptotically unbiased. The potential energy is related to the negative log of the target probability density function, and the mass matrix, typically chosen as the identity matrix. The algorithm's performance is highly dependent on the choice of the tuning parameters ϵ and L. Implemented choices are detailed in Algorithm 8.

C.3.2 No-U-Turn Sampler (Stan)

The No-U-Turn Sampler (NUTS) (Hoffman et al., 2014) is an extension of Hamiltonian Monte Carlo (HMC) specifically designed to eliminate the need to set a fixed number of steps for the simulation, which is a common challenge in HMC. Implemented in the *Stan* programming language (Carpenter et al., 2017), NUTS uses a recursive algorithm to build a binary tree that efficiently explores the target distribution. The key innovation of NUTS within *Stan* is its termination criterion, which stops the simulation when it starts to double back on itself, hence the name "No-U-Turn". This criterion, along with other refinements such as a more geometrically formal termination based on momenta, direct multinomial sampling from numerical trajectories, and modified adaptation of step size and inverse metric, make *Stan*'s NUTS highly efficient for Bayesian inference.

Algorithm 7 Slice Sampling

Require: data $\mathbf{X}$
Require: n the initial value of degrees of freedom
Require: $\mathbf{V}$ scale matrix
Require: prior specification of degrees of freedom f_n
Require: analogous pdf of degrees of freedom $\pi(n|\mathbf{X}, \mathbf{V})$ ▷ function C.15
Require: width of slice $w = 1$
  Set $n_{new} \leftarrow 0$
  while $n_{new} < p - 1$ **do**
    $y_{level} \leftarrow \log(\pi(n|\mathbf{X}, \mathbf{V})) - \text{Exponential}(1)$ ▷ horizontal "slice"

    //An interval $[x_l, x_r]$ is established around the current value of n
    $x_l \leftarrow n - runif(1) \cdot w$ ▷ random sample from uniform
    $x_r \leftarrow x_l + w$

    //The interval is expanded until the value of log function C.15 at the endpoints is less than y_{level}
    while $(\log(\pi(x_l|\mathbf{X}, \mathbf{V})) > y_{level})$ $x_l \leftarrow x_l - w$
    while $(\log(\pi(x_r|\mathbf{X}, \mathbf{V})) > y_{level})$ $x_r \leftarrow x_r + w$

    $x_{pr} \leftarrow runif(x_l, x_r)$ ▷ A new sample is drawn uniformly from the interval $[x_l, x_r]$

    repeat
      $x_{\text{pr}} \leftarrow \text{runif}(x_l, x_r)$
      if $\log \pi(x_{\text{pr}}|\mathbf{X}, \mathbf{V}) > y_{\text{level}}$ **then**
        break ▷ Accept sample
      else if $x_{\text{pr}} < n$ **then**
        $x_l \leftarrow x_{\text{pr}}$
      else
        $x_r \leftarrow x_{\text{pr}}$
      end if
    until accepted
    $n_{\text{new}} \leftarrow x_{\text{pr}}$
  end while
  return n_{new}

In accordance with the Bayesian modeling framework delineated in Section C.3, we have utilized the model within the *Stan* programming environment. The No-U-Turn Sampler (NUTS), as implemented in *Stan*, benefits from meticulously calibrated optimizations, yielding rapid sampling processes. However, it is noteworthy that when dealing with high-dimensional covariance matrices, the sampling procedure exhibits a marked deceleration, potentially extending over several days, as evidenced by the empirical findings presented in Section C.4. Consequently, to circumvent such computational bottlenecks in our preliminary analysis, we have employed a multiple of MCMC methodologies.

C.4 Experimental Results

In this section, we undertake data-driven experiments to assess the quality of the point estimates for the Wishart distribution parameters derived from each MCMC algorithm and the respective prior distributions under review. Moreover, we compare the approaches in terms of iteration efficiency, effective sample size (ESS), and convergence metrics. We present results in two distinct cases. Initially, we conduct a simulation study by generating data from the Wishart distribution across a spectrum of sample sizes, dimensions, and degrees of freedom. In the second case, we compare the results and the

Algorithm 8 Hamiltonian Monte Carlo (HMC)

Require: data $\mathbf{X}$
Require: n the initial value of degrees of freedom
Require: $\mathbf{V}$ scale matrix
Require: prior specification of degrees of freedom f_n
Require: analogous pdf of degrees of freedom $\pi(n|\mathbf{X}, \mathbf{V})$ ▷ function C.15
Require: $\epsilon = 0.1$ (step size), $L = 10$(number of steps)

Sample momentum $z \sim \mathcal{N}(0, I)$ ▷ random sample from normal
Set $y \leftarrow n$ and $z^* \leftarrow z$

for ℓ in 1:L **do**
  // Leapfrog integration
  $z^* \leftarrow z^* - \frac{\epsilon}{2}\nabla log(\pi(y|\mathbf{X}, \mathbf{V}))$
  $y \leftarrow y + \epsilon \cdot z^*$
  $z^* \leftarrow z^* - \frac{\epsilon}{2}\nabla log(\pi(y|\mathbf{X}, \mathbf{V}))$
end for

if $y > p - 1$ **then**
  // Metropolis acceptance step
  $U_{\text{new}} \leftarrow -log(\pi(y|\mathbf{X}, \mathbf{V})) + \frac{1}{2}{z^*}^2$

  $U_{\text{old}} \leftarrow -log(\pi(n|\mathbf{X}, \mathbf{V})) + \frac{1}{2}z^2$
  $accept \leftarrow \min(1, \exp(U_{\text{old}} - U_{\text{new}}))$
  $u \leftarrow runif(0, 1)$ ▷ random sample from uniform
  if $u < accept$ **then**
    $n_{new} \leftarrow y$
  end if
else
  $n_{new} \leftarrow n$
end if
return n_{new}

efficiency of the methods under investigation in three real-world datasets.

We report the posterior median of the degrees of freedom n and the posterior mean of the scale matrix $\mathbf{V}$, estimated from the MCMC output, as posterior point estimates of the parameters of interest. Furthermore, only for reference purposes, we report the maximum likelihood estimates as obtained by using an ensemble method combining Bisection and Newton-Raphson algorithms as described in Section C.2.

In the simulation study, the MCMC algorithms were run for 30 seconds in scenarios for datasets with dimensions up to 15 variables and for 60 seconds in scenarios with dimensions greater than 15 variables. The estimated posterior median of the degrees of freedom was evaluated using the Percentage Error (PE), defined as the absolute difference between the true and estimated values, normalized by the true value. The PE is mathematically expressed as:

$$PE_n = \left|\frac{n - \widehat{n}}{n}\right| \times 100, \tag{C.16}$$

where $\widehat{n}$ is the estimated posterior median of the degrees of freedom.

For the estimated scale matrix with elements $\widehat{V}_{ij}$ (here the estimated posterior means of V_{ij}), we

consider the average PE of the elements of $\mathbf{V}$, that is

$$PE_{\mathbf{V}} = \frac{2}{d(d+1)} \sum_{i=1}^{p} \sum_{j=1}^{i} \left| \frac{V_{ij} - \widehat{V}_{ij}}{V_{ij}} \right| \times 100. \tag{C.17}$$

Since $\mathbf{V}$ is symmetric, in the above quantity, we consider only the elements of the diagonal and the lower triangle of the matrix in (C.17). Furthermore, using the MCMC output for the degrees of freedom, we assess the iteration efficiency of each approach, the effective sample size (ESS), the Monte Carlo standard error (MCSE), and Geweke's convergence diagnostic (Geweke, 1992), using CODA R package (Plummer et al., 2006).

For the analysis of real datasets, naturally, the true parameter values are not available. Hence, comparisons based on percentage errors are not possible. However, as in the simulation study, we can evaluate performance using the number of iterations within a fixed runtime, the effective sample size (ESS), the Monte Carlo standard error (MCSE), and the absolute value of Geweke's statistic. We assess the algorithms by comparing the results and patterns of these metrics with those from the simulation study as a reference.

C.4.1 Simulated Data

In this first experimental scenario, using simulated data from the Wishart distribution, we analyze various aspects and sources of variation. Specifically, we compare different sample sizes, values of degrees of freedom, and dimensions. For each combination, we have generated 10 simulated datasets in order to examine the variability of random covariance matrices and identify potential issues in each setting. The cases we consider are as follows:

- Sample size: $m \in \{5, 10, 30, 50, 100\}$
- Dimensions: $p \in \{2, 5, 10, 30, 50\}$
- Degrees of freedom: $n \in \{3, 6, 11, 31, 51, 101\}$.

Additionally, we need to impose constraints in order to have a properly specified probability density function. For this reason, we consider $p \leq \min\{m, n\}$.

Considering the combinations of the above values under the restriction of $p \leq \min\{m, n\}$, we end up examining 17634 different simulated data setups. We use these generated datasets to infer the strengths and weaknesses of each method. We present the aggregated results for each method and prior that yielded the best evaluation metrics, separately for setups with $p \leq 15$ and $p > 15$. The first case involves matrices of lower dimension ($p \leq 15$), with a 30-second running time. The second case focuses on higher-dimension covariance matrices with $p > 15$, with a 60-second running time. For all generated datasets, the maximum number of iterations is set to 3000.

In the context of low-dimensional ($p \leq 15$) covariance matrices with a 30-second runtime limit, the results in Table C.4.1 offer a comparison of the proposed MCMC approaches and priors. In terms of iterations, the Slice Sampling within Gibbs algorithm, using an Inverse-Gamma prior, achieved the maximum number of iterations (3000) in all implemented simulated datasets, indicating consistent performance over the runtime of 30 seconds. When considering the percentage error of the degrees of freedom (PE_n), both the Random Walk Metropolis within Gibbs, with Gamma prior, and Hamiltonian Monte Carlo (HMC) within Gibbs, with LogNormal prior, exhibited the lowest mean PE_n (13.9%), although the latter was associated with a slightly higher standard deviation. Regarding the percentage

error of the scale matrix ($PE_{\mathbf{V}}$), the effective sample size (ESS), and Monte Carlo standard error (MCSE), the implementation of NUTS in Stan with a LogNormal prior produced the best results. Finally, for the Geweke convergence diagnostic, Slice Sampling within Gibbs with an Inverse-Gamma prior exhibited the lowest mean absolute Geweke's statistic (1.021). Overall, Slice Sampling and HMC within Gibbs showed strong performance in terms of ESS and convergence, while NUTS stood out for its high ESS and low MCSE. Finally, the substantial standard deviation observed in the absolute Geweke's statistic for the NUTS approach arises from instances with larger sample sizes (sample size = 100), where the algorithm does not generate an adequate number of posterior samples.

Approach	**Prior**	**Iterations**		$\boldsymbol{PE_n}$		$\boldsymbol{PE_V}$		**ESS**		**MCSE**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	12.6	21.2	70.1	1042.1	-	-	-	-	-	-
Random Walk Metropolis within Gibbs	Gamma	2978	115	13.9	21.8	70.5	1052.1	93	82	1.275	2.718	1.152	1.175
Slice Sampling within Gibbs	IGamma	3000	0	14.7	24.2	69.2	1015.2	187	224	1.08	2.969	1.021	0.88
HMC within Gibbs	LNormal	2826	352	13.9	23.6	70.9	1053.2	221	434	1.253	2.446	1.271	1.314
NUTS (Stan)	LNormal	2181	1130	14.7	21.4	46.5	283.4	300	261	0.714	2.002	1.593	3.893

PE_n: Percentage Error of the degrees of freedom; $PE_{\mathbf{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error ; Geweke: Geweke (1992) statistic

Table C.4.1: MCMC efficiency results per method and prior for low-dimensional covariance matrices (p≤15) with a limit of 30 seconds running time.

Table C.4.2 presents the aggregated results for comparisons on high-dimensional covariance matrices ($p > 15$) within a 60-second runtime limit. The Random Walk Metropolis within Gibbs (RWM-Gibbs) with an Inverse-Gamma prior outperforms other approaches in terms of iterations, percentage error of the degrees of freedom (PE_n), and scale matrix ($PE_{\mathbf{V}}$). The Hamiltonian Monte Carlo (HMC) within Gibbs with an Exponential prior attains the highest effective sample size (ESS = 527), indicating robust performance in generating effective samples. Slice Sampling within Gibbs with Exponential prior exhibits the lowest mean values for the Monte Carlo standard error (MCSE) and mean absolute Geweke's statistic, indicating precise parameter estimates and stable convergence. A detailed illustration of MCMC performances for all different methods and priors is given in Table C.4.4.

Overall, Random Walk Metropolis within Gibbs demonstrates a slightly better performance in terms of iterations, PE_n, and $PE_{\mathbf{V}}$, while Slice Sampling within Gibbs outperforms the other methods in MCSE and Geweke diagnostics. The HMC within Gibbs stands out for its high ESS (again). The No-U-Turn Sampler (NUTS) approach, within the 60-second time limit, does not generate sufficient posterior samples, over 40 algorithm runs. For a comprehensive account of the percentage error (PE) of degrees of freedom, see Figures C.4.5; C.4.6, which includes the detailed PE boxplots for each simulated data setup.

The results discussed above, and presented in Tables C.4.1 and C.4.2, are based on priors that lead to MCMC runs with the best Monte Carlo diagnostics for each each MCMC method. Full results, for all priors, are provided in Tables C.4.3 and C.4.4. Furthermore, Appendix C.4.2 presents comparisons based on experiments conducted with a fixed number of 3000 iterations and a one-hour time limit. Even under these conditions, the NUTS approach for high-dimensional covariance matrices ($p \geq 30$) does not generate an adequate number of samples within the one-hour limit.

In summary, the evaluation of the approaches for both low and high-dimensional covariance matrices revealed distinct strengths and weaknesses. For low-dimensional matrices, NUTS (Stan) emerged as the most effective method, balancing accuracy and efficiency. However, it exhibited considerable variability

Approach	Prior	Iterations		PE_n		PE_V		ESS		MCSE		Geweke	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	1.0	0.9	61.2	57.1	-	-	-	-	-	-
Random Walk Metropolis within Gibbs	IGamma	2347	937	1.8	1.5	60.5	55.4	141	90	0.038	0.031	0.885	0.819
Slice Sampling within Gibbs	Exp	2340	1000	1.8	1.5	60.8	56.5	362	299	0.031	0.029	0.903	0.728
HMC within Gibbs	Exp	1970	1107	4.8	10.5	66.5	59.9	527	786	0.479	2.82	1.607	1.979
NUTS* (Stan)	Gamma	41	10	17.8	19.3	111.7	117.9	4	5	4.315	4.961	7.132	7.367

PE_n: Percentage Error of the degrees of freedom; PE_V: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.2: MCMC efficiency results per method and prior for high-dimensional covariance matrices (p>15) with a limit of 60 seconds running time.

in convergence diagnostics, raising concerns about potential convergence issues. In high-dimensional settings, Slice Sampling within Gibbs provided the most efficient sampling algorithm, offering a good balance between accuracy and efficiency. On the other hand, NUTS as implemented by *Stan* was unable to provide reliable posterior results in a reasonable timeframe for high-dimensional examples.

Approach	**Prior**	**Iterations**		PE_n		PE_V		**Total PE**		**ESS**		**MCSE**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	12.6	21.2	70.1	1042.1	82.7	1042.1	-	-	-	-	-	-
Random Walk Metropolis within Gibbs	Unif	2976.358	123.648	16.5	28.3	70.7	1042.2	87.1	1042.2	92.931	83.383	1.282	2.725	1.142	1.081
	Exp	2976.718	118.943	17.9	31.1	70.2	1038.3	88.1	1038.5	92.591	83.303	1.469	3.675	1.23	1.382
	Gamma	2977.837	115.430	**13.9**	21.8	70.5	1052.1	84.4	1052.1	92.697	81.762	1.275	2.718	1.152	1.175
	IGamma	2977.000	122.435	14.4	22.8	70.5	1036.8	84.9	1036.8	93.410	81.964	1.217	2.761	1.178	1.056
	LNormal	2975.746	131.154	14.7	24.3	71.7	1060.5	86.4	1060.5	92.837	83.489	1.407	3.336	1.212	1.438
Slice Sampling within Gibbs	Unif	**3000**	**0**	16.2	27.6	71.2	1065.1	87.4	1065.2	185.149	216.728	1.071	2.436	1.053	0.880
	Exp	2998.623	29.923	18.1	32.3	69.8	1024.8	87.9	1025.0	182.794	212.918	1.108	3.047	1.000	0.862
	Gamma	**3000**	**0**	14.5	22.9	70.0	1036.8	84.5	1036.8	185.378	215.682	0.98	2.139	**0.969**	0.848
	IGamma	**3000**	**0**	14.7	24.2	69.2	1015.2	83.9	1015.2	186.858	223.661	1.08	2.969	1.021	0.880
	LNormal	2998.837	30.999	14.3	22.6	70.8	1051.1	85.1	1051.1	185.060	217.658	1.032	2.464	1.003	**0.819**
HMC within Gibbs	Unif	2824.456	357.413	15.7	26.5	70.4	1041.7	86.1	1041.8	217.651	423.554	1.356	2.694	1.326	1.368
	Exp	2827.510	347.343	18.3	39.2	70.5	1040.8	88.8	1041.3	213.826	418.145	1.432	3.053	1.339	1.509
	Gamma	2825.817	350.652	14.0	**21.4**	70.3	1020.8	84.2	1020.8	220.653	432.049	1.38	2.847	1.343	1.666
	IGamma	2827.490	348.886	14.0	21.5	69.7	1015.1	83.7	1015.1	218.789	425.385	1.34	2.582	1.227	1.503
	LNormal	2826.355	352.380	**13.9**	23.6	70.9	1053.2	84.8	1053.2	221.221	434.495	1.253	2.446	1.271	1.314
NUTS (Stan)	Uniform	2173.095	1132.350	17.4	27.8	96.7	1637.1	114.1	1637.7	300.793	265.177	0.789	2.274	1.769	5.151
	Exp	2181.532	1130.783	17.4	27.8	75.7	1030.7	93.1	1031.6	**302.266**	262.592	0.79	2.403	1.790	4.697
	Gamma	2179.997	1131.470	15.6	23.8	62.4	715.8	77.9	716.7	298.769	262.089	0.771	2.243	1.629	3.822
	IGamma	2177.137	1131.429	15.5	23.5	61.2	675.2	76.7	675.8	301.648	272.272	0.742	2.068	1.764	5.628
	LNormal	2180.690	1130.373	14.7	**21.4**	**46.5**	**283.4**	**61.2**	**285.9**	300.007	**261.383**	**0.714**	**2.002**	1.593	3.893

PE_n: Percentage Error of the degrees of freedom; $PE_{\mathbf{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.3: MCMC efficiency results per method and prior for low dimensional covariance matrices (p≤15) with a limit of 30 seconds running time.

Approach	**Prior**	**Iterations**		PE_n		PE_V		**Total PE**		**ESS**		**MCSE**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	1.0	0.9	61.2	57.1	62.1	57.2	-	-	-	-	-	-
Random Walk Metropolis within Gibbs	Unif	**2417.015**	**902.557**	1.8	1.5	60.9	57.3	62.7	57.5	147.185	94.034	0.038	0.032	0.868	0.715
	Exp	2368.477	928.371	1.8	1.5	60.7	56.3	62.6	56.4	141.874	87.690	0.037	0.030	**0.777**	**0.613**
	Gamma	2349.846	933.690	**1.8**	**1.5**	61.0	57.4	62.9	57.5	138.710	90.707	0.038	0.031	0.919	0.744
	IGamma	2347.123	936.793	1.8	1.5	**60.5**	**55.4**	**62.4**	**55.5**	140.541	90.109	0.038	0.031	0.885	0.819
	LNormal	2335.554	944.078	1.8	1.5	60.8	56.6	62.6	56.7	138.496	87.495	0.039	0.032	0.994	0.743
Slice Sampling within Gibbs	Unif	2335.800	1003.190	1.8	1.5	60.9	56.8	62.7	56.9	347.006	290.238	0.032	0.031	0.895	0.726
	Exp	2340.154	1000.483	1.8	1.5	60.8	56.5	62.6	56.6	361.874	299.136	**0.031**	**0.029**	0.903	0.728
	Gamma	2352.308	985.603	1.8	1.5	60.9	56.9	62.8	57.1	366.512	303.537	0.031	0.030	0.865	0.804
	IGamma	2350.862	988.662	1.9	1.5	61.1	57.2	63.0	57.3	368.913	304.623	0.031	0.030	0.890	0.783
	LNormal	2345.062	996.688	1.8	1.5	60.9	56.6	62.7	56.7	378.870	321.733	0.031	0.031	0.875	0.656
HMC within Gibbs	Unif	1967.015	1111.115	4.1	8.8	65.3	57.6	69.4	60.3	**546.394**	795.537	0.889	4.188	14.043	85.533
	Exp	1970.308	1106.973	4.8	10.5	66.5	59.9	71.3	63.6	527.234	785.673	0.479	2.820	1.607	1.979
	Gamma	1975.738	1105.194	4.1	8.8	65.1	58.3	69.2	61.1	546.250	814.382	0.252	1.845	7.254	65.350
	IGamma	1972.908	1105.495	4.1	8.8	65.7	58.6	69.8	61.5	532.403	788.932	0.417	2.724	11.687	83.502
	LNormal	1973.446	1105.064	4.4	9.3	66.2	59.0	70.6	62.2	527.218	**766.683**	0.486	3.247	6.526	53.098
NUTS (Stan)	Uniform	41.000	9.754	17.1	19.0	106.7	107.9	123.8	114.3	3.458	4.738	4.309	4.648	9.901	27.707
	Exp	41.046	9.940	16.8	18.7	102.6	68.8	119.5	74.5	4.802	9.427	4.312	5.065	8.147	12.258
	Gamma	40.754	10.033	17.8	19.3	111.7	117.9	129.6	121.4	3.534	4.571	4.315	4.961	7.132	7.367
	IGamma	40.215	9.529	17.4	19.1	101.4	71.0	118.8	76.2	3.736	5.437	4.087	4.533	14.273	75.710
	LNormal	40.415	9.608	16.9	18.3	103.0	71.5	119.9	78.3	3.708	4.676	4.395	4.915	7.726	9.793

PE_n: Percentage Error of the degrees of freedom; $PE_{\mathbf{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.4: Aggregated results of simulated experiments with a 60-second MCMC time limit per method and prior for high dimensional covariance matrices (p>15)

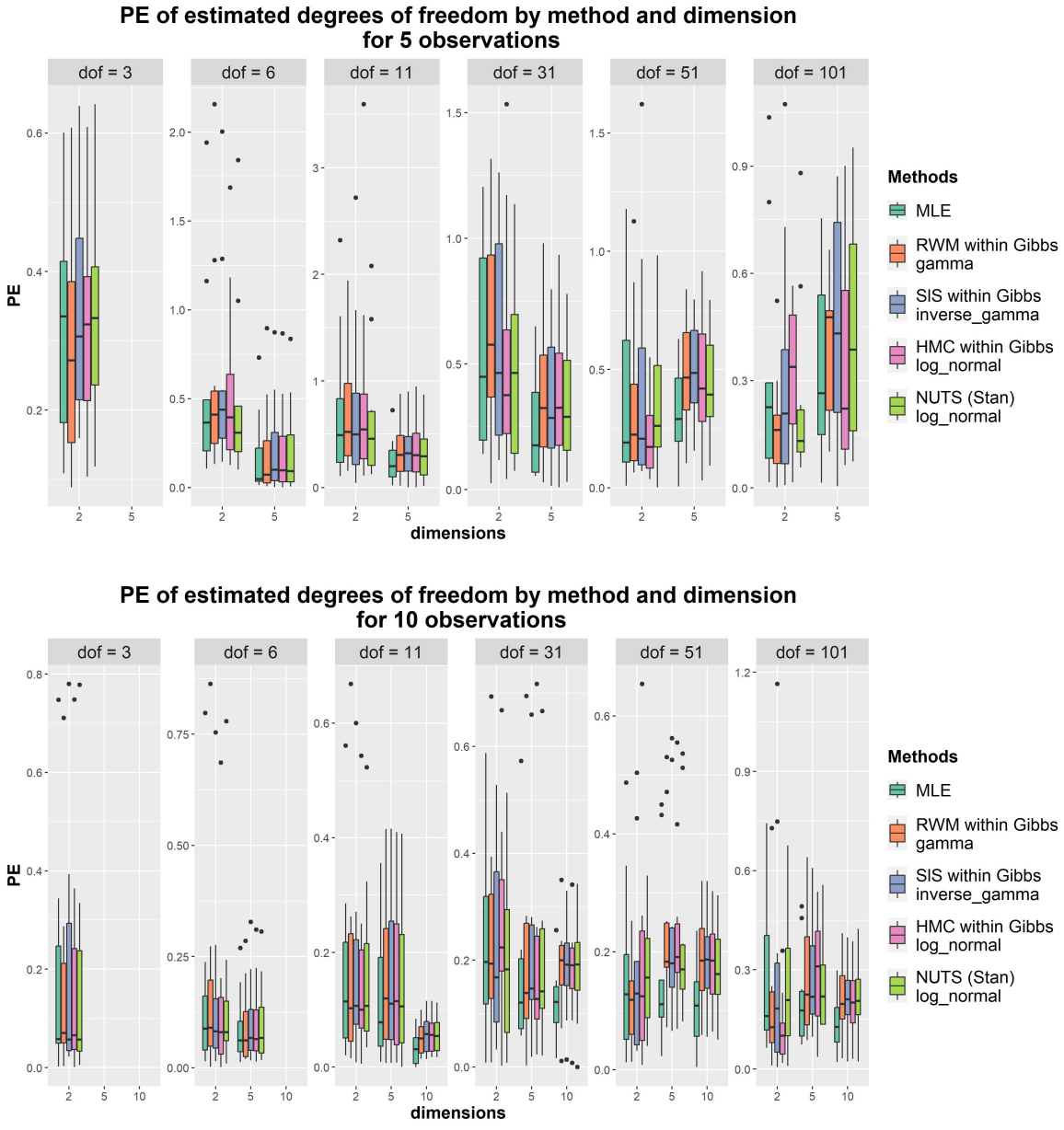
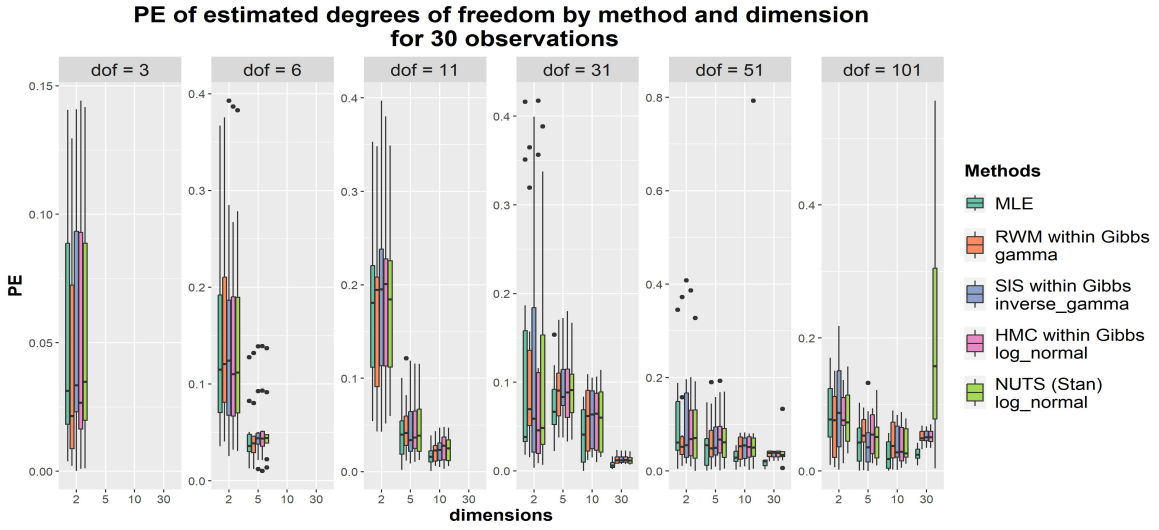
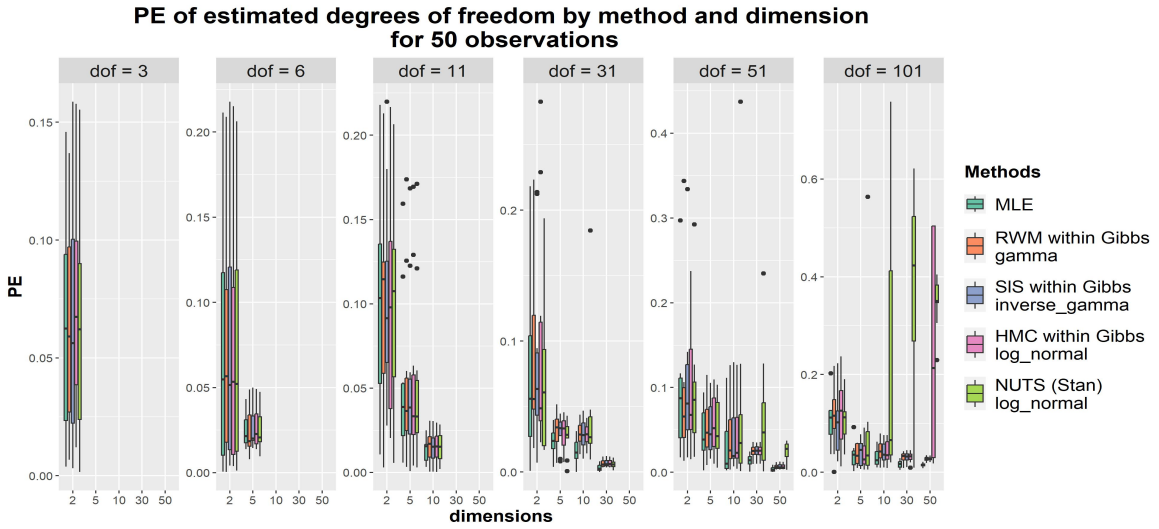
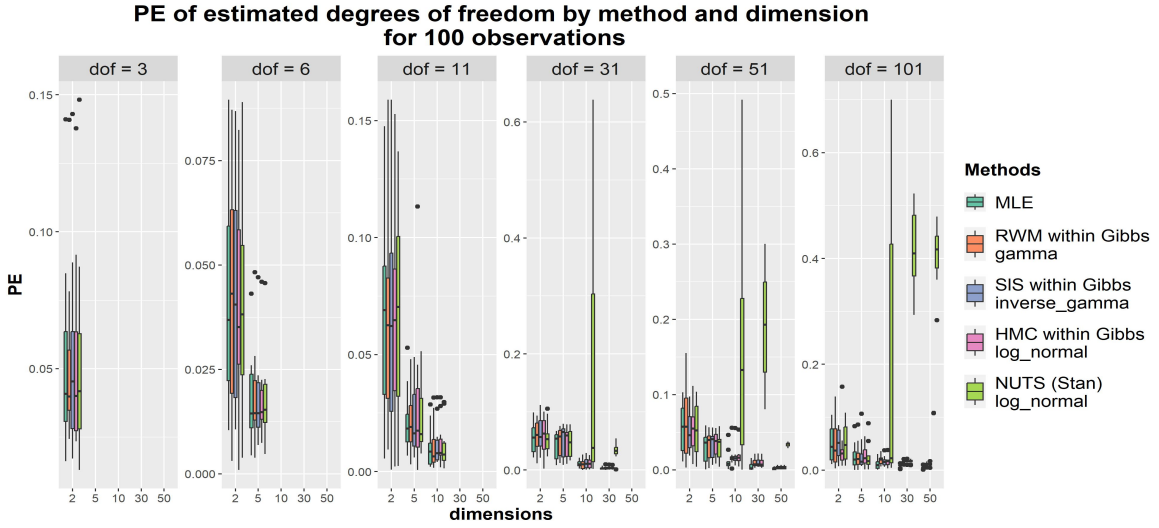


Figure C.4.5: Percentage Error of estimated degrees of freedom of each method (Random Walk Metropolis (RWM), Slice Sampling (SlS), Hamiltonian Monte Carlo (HMC) and NUTS (Stan)) of Sections C.3.1–C.3.2 per dimension, actual degrees of freedom and random covariance matrices for 5 and 10 number observations with time limit.



(a)





(c)

Figure C.4.6: Percentage Error of estimated degrees of freedom of each method (Random Walk Metropolis (RWM), Slice Sampling (SlS), Hamiltonian Monte Carlo (HMC) and NUTS (Stan)) of Sections C.3.1–C.3.2 per dimension, actual degrees of freedom and random covariance matrices for 30, 50 and 100 number observations with time limit.

C.4.2 Simulated Data with Predefined Number of Iterations

In this experimental scenario, using simulated data from the Wishart distribution, we examine the variability of each method under scenarios with a fixed number of MCMC iterations. Specifically, we conducted the same experiments as in Section C.4.1 with fixed number of MCMC iterations.

In all MCMC runs, we have used a total of 3000 iterations, with the initial 1000 iterations designated as the burn-in period. This was conducted using a single chain, primarily due to computational limitations associated with the large number of MCMC runs across different settings.

We present the aggregated results for each method using the best-performing prior under two distinct scenarios. The first scenario focuses on covariance matrices with dimensions of 15 or less, while the second focuses on high-dimensional matrices with dimensions greater than 15. Unfortunately, for the high-dimensional case, the NUTS implementation in Stan fails to complete within one hour; therefore, these results are not reported.

In the context of low-dimensional covariance matrices ($p \leq 15$), the HMC within Gibbs method demonstrated the lowest mean PE_n (13.8%), indicating superior accuracy, while the NUTS (Stan) method exhibited the highest ESS (2014.852), suggesting the most efficient sampling; see Table C.4.5. However, NUTS also showed significant variability in computation time, with a mean of 1.9 minutes and a standard deviation of 4.7 minutes. The Geweke scores indicated convergence, with NUTS achieving the lowest mean score of 0.794, while RWM within Gibbs had a mean score of 1.071 and HMC within Gibbs had a mean score of 1.290. Overall, HMC within Gibbs and NUTS (Stan) were the most effective methods, balancing accuracy and efficiency, though at the cost of computation time and convergence reliability.

Approach	Prior	Time (min)		PE_n		PE_V		ESS		Geweke	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	13.9	27.5	29.5	47.1	-	-	-	-
RWM within Gibbs	Gamma	0.131	0.150	14.7	25.2	29.3	46.0	124.599	111.908	1.071	0.997
Slice Sampling within Gibbs	Gamma	0.179	0.116	15.5	26.6	29.4	46.0	245.193	287.987	0.947	0.781
HMC within Gibbs	LNormal	0.339	0.255	13.8	22.1	29.4	46.3	304.243	563.260	1.290	1.483
NUTS * (Stan)	LNormal	1.915	4.708	14.7	25.4	29.5	46.5	2014.852	187.44	0.794	0.600

* *excluded 2 cases with running time more than one hour*
PE_n: Percentage Error of the degrees of freedom; PE_V: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.5: MCMC efficiency results for low-dimensional covariance matrices ($p \leq 15$) using 3,000 iterations with a 1,000-iteration burn-in. For each method, the prior associated with the best Monte Carlo diagnostics is presented; for all other priors, see Table C.4.7.

For high-dimensional covariance matrices ($p > 15$), the HMC within Gibbs method has the lowest mean PE_n (2.9%) and performed best in terms of sampling efficiency, with the highest ESS of 1180.9; see Table C.4.6. The Slice Sampling within Gibbs method also performed well, with a mean ESS of 645.3 and a relatively low mean PE_n (1.9%). Computation times varied, with Slice Sampling within Gibbs being the fastest (mean 0.97 minutes), followed by RWM within Gibbs (mean 1.13 minutes). The Geweke scores for Slice Sampling within Gibbs (mean 0.71) indicated convergence, surpassing those of RWM within Gibbs (mean 0.877) and HMC within Gibbs (mean 1.453), the latter potentially indicating convergence issues in some cases.

The results discussed above, and presented in Tables C.4.5 and C.4.6, are based on priors that lead to MCMC runs with the best Monte Carlo diagnostics for each each MCMC method. Full results, for

all priors, are provided in Tables C.4.7 and C.4.8.

In summary, the examination of the approaches for both low and high-dimensional covariance matrices revealed distinct strengths and weaknesses. For low-dimensional matrices, HMC within Gibbs and NUTS (Stan) emerged as the most effective methods, balancing accuracy and efficiency, though NUTS exhibited significant variability in computation time and potential convergence issues. In high-dimensional settings, HMC within Gibbs provided the most efficient sampling. Slice Sampling within Gibbs also performed well across both dimensions, offering a good balance between accuracy and efficiency. For a comprehensive distribution of the percentage error (PE) of degrees of freedom (DoF), see Figures C.4.7 and C.4.8, which includes the PE boxplots for each case.

Approach	**Prior**	**Time (min)**		$\boldsymbol{PE_n}$		$\boldsymbol{PE_V}$		**ESS**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	1.0	0.9	484.3	4210.4	-	-	-	-
RWM within Gibbs	Gamma	1.129	1.398	1.9	1.6	472.2	4073.8	248.882	111.405	0.877	0.678
Slice Sampling within Gibbs	LNormal	0.969	0.894	1.9	1.6	479.1	4163.7	645.345	450.169	0.709	0.599
HMC within Gibbs	Exp	2.038	1.852	2.9	5.0	467.0	4006.2	1180.944	1389.919	1.453	2.512

PE_n: Percentage Error of the degrees of freedom; $PE_{\boldsymbol{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.6: MCMC efficiency results for high-dimensional covariance matrices ($p > 15$) using 3,000 iterations with a 1,000-iteration burn-in. For each method, the prior associated with the best Monte Carlo diagnostics is presented; for all other priors, see Table C.4.8.

Approach	**Prior**	**Time (min)**		PE_n		PE_V		**Total PE**		**ESS**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	13.9	27.5	29.5	47.1	43.4	57.7	-	-	-	-
Random Walk Metropolis within Gibbs	Unif	**0.130**	0.150	16.9	31.5	29.8	46.2	46.7	60.0	123.893	111.244	1.068	1.135
	Exp	0.131	0.149	18.4	31.5	30.2	**45.1**	48.6	59.2	123.483	112.525	1.167	1.162
	Gamma	0.131	0.150	14.7	25.2	**29.3**	46.0	44.0	55.5	124.599	111.908	1.071	0.997
	IGamma	0.132	0.152	15.2	27.7	29.6	47.1	44.8	57.8	123.330	109.857	1.173	1.237
	LNormal	0.132	0.152	14.7	25.3	29.6	46.7	44.3	56.2	124.019	111.407	1.110	1.113
Slice Sampling within Gibbs	Unif	0.179	0.115	18.4	38.8	29.9	45.9	48.3	64.7	244.501	287.589	0.925	0.826
	Exp	0.179	0.115	20.2	40.5	30.3	46.4	50.5	66.7	241.836	288.300	0.978	0.874
	Gamma	0.179	0.116	15.5	26.6	29.4	46.0	44.9	56.6	245.193	287.987	0.947	0.781
	IGamma	0.178	**0.114**	16.6	35.1	29.5	46.3	46.1	62.2	246.712	292.194	0.996	0.901
	LNormal	0.179	0.115	15.4	27.6	29.6	47.2	45.1	58.2	247.740	291.338	0.984	0.860
HMC within Gibbs	Unif	0.338	25.5	15.4	24.3	29.6	46.5	45.0	55.7	301.844	564.988	1.335	1.716
	Exp	0.339	0.255	17.3	28.3	30.0	45.7	47.3	57.6	300.182	555.513	1.338	1.475
	Gamma	0.339	0.254	14.3	22.5	29.4	46.3	43.7	54.5	304.613	566.391	1.361	1.666
	IGamma	0.338	0.254	14.3	22.8	29.7	47.7	44.0	55.7	304.247	558.766	1.296	1.597
	LNormal	0.339	0.255	**13.8**	**22.1**	29.4	46.3	**43.2**	**54.2**	304.243	563.260	1.290	1.483
NUTS* (Stan)	Uniform	1.906	4.680	17.9	33.9	29.8	45.4	47.7	61.1	2008.639	155.109	0.845	0.658
	Exp	1.918	4.734	18.0	34.6	29.7	46.4	47.7	62.3	2008.646	148.765	0.837	0.638
	Gamma	1.898	4.613	16.0	29.4	29.4	45.5	45.3	57.8	**2017.311**	153.828	0.829	0.623
	IGamma	1.895	4.608	15.9	28.8	29.5	46.1	45.4	58.1	2012.628	**138.183**	0.847	0.637
	LNormal	1.915	4.708	14.7	25.4	29.5	46.5	44.2	56.2	2014.852	187.440	**0.794**	**0.600**

* *excluded 2 cases with running time more than one hour*

PE_n: Percentage Error of the degrees of freedom; $PE_{\mathbf{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.7: MCMC efficiency results per method and prior for low-dimensional covariance matrices ($p \leq 15$) with standard 3000 MCMC iterations with a 1000-iteration burn-in period.

Approach	**Prior**	**Time (min)**		$\boldsymbol{PE_n}$		$\boldsymbol{PE_V}$		**Total PE**		**ESS**		**Geweke**	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
MLE	-	-	-	1.0	0.9	484.3	4210.4	485.3	4210.4	-	-	-	=
Random Walk Metropolis within Gibbs	Unif	1.110	1.366	1.9	1.6	486.3	4242.4	488.2	4242.4	249.345	119.684	0.862	0.703
	Exp	1.127	1.393	1.9	1.6	489.5	4270.1	491.4	4270.1	251.714	123.566	0.882	0.628
	Gamma	1.129	1.398	1.9	1.6	472.2	4073.8	474.1	4073.8	248.882	**111.405**	0.877	0.678
	IGamma	1.130	1.401	1.9	1.6	482.8	4215.6	484.7	4215.5	242.445	112.003	0.878	0.710
	LNormal	1.127	1.381	1.9	1.6	479.5	4151.5	481.3	4151.5	246.347	115.198	0.873	0.671
Slice Sampling within Gibbs	Unif	0.970	0.896	1.9	1.6	488.2	4268.2	490.0	4268.2	635.808	437.303	0.822	0.718
	Exp	**0.964**	0.894	1.9	**1.5**	4.826	41.825	4.845	41.825	649.758	453.189	0.844	0.675
	Gamma	0.966	**0.892**	1.9	1.6	488.3	4257.6	490.2	4257.6	647.783	447.934	0.876	0.700
	IGamma	0.968	0.893	1.9	1.6	478.9	4137.2	480.7	4137.2	636.535	442.821	0.883	0.752
	LNormal	0.969	0.894	1.9	1.6	479.1	4163.7	481.0	4163.7	645.345	450.169	**0.709**	**0.599**
HMC within Gibbs	Unif	2.031	1.847	2.9	5.0	478.2	4115.8	481.2	4115.7	1159.256	1354.077	1.489	2.577
	Exp	2.038	1.852	2.9	5.0	**467.0**	**4006.2**	**470.0**	**4006.1**	1180.944	1389.919	1.453	2.512
	Gamma	2.036	1.852	2.6	2.7	490.3	4278.4	492.8	4278.4	**1202.554**	1453.842	1.310	2.089
	IGamma	2.047	1.855	2.9	5.0	483.0	4204.6	485.9	4204.5	1169.944	1364.048	1.563	2.984
	LNormal	2.043	1.844	2.9	5.0	485.1	4183.0	488.0	4182.9	1194.709	1443.282	1.446	2.093

PE_n: Percentage Error of the degrees of freedom; $PE_{\mathbf{V}}$: Percentage Error of the scale matrix; ESS: Effective Sample Size; MCSE: Monte Carlo standard error; Geweke: Geweke (1992) statistic

Table C.4.8: MCMC efficiency results per method and prior for high-dimensional covariance matrices ($p > 15$) with standard 3000 MCMC iterations with a 1000-iteration burn-in period.

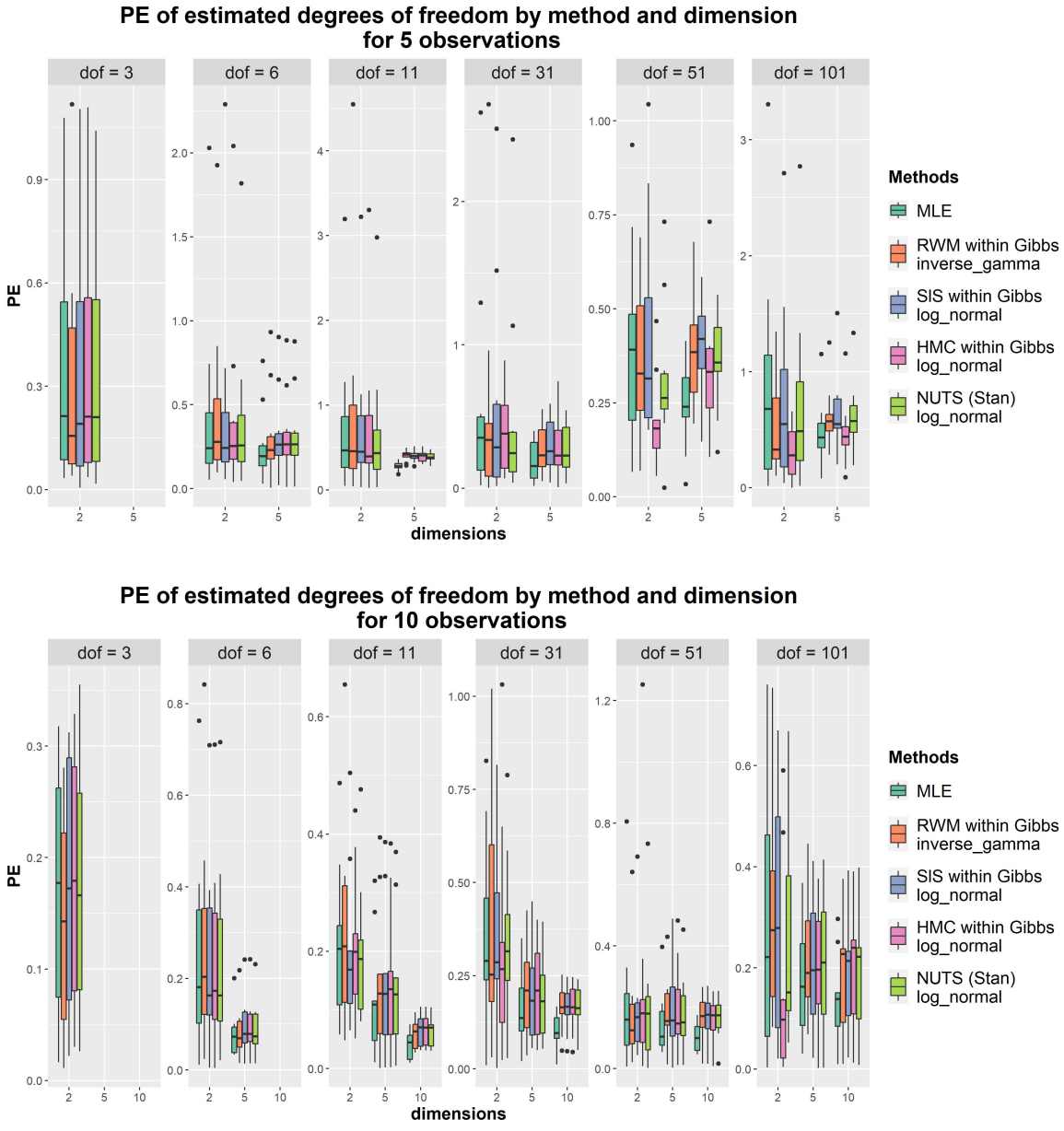
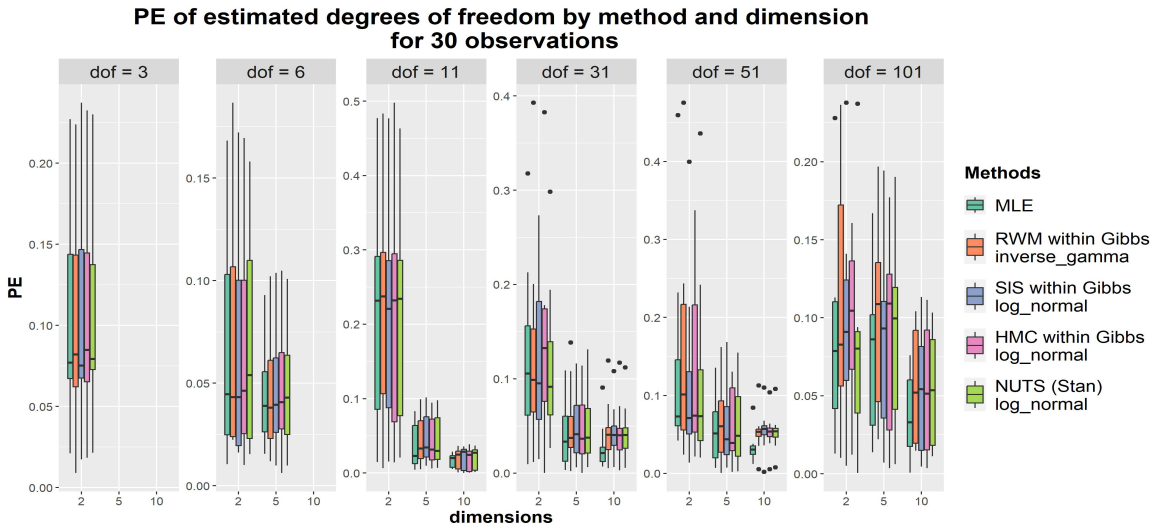
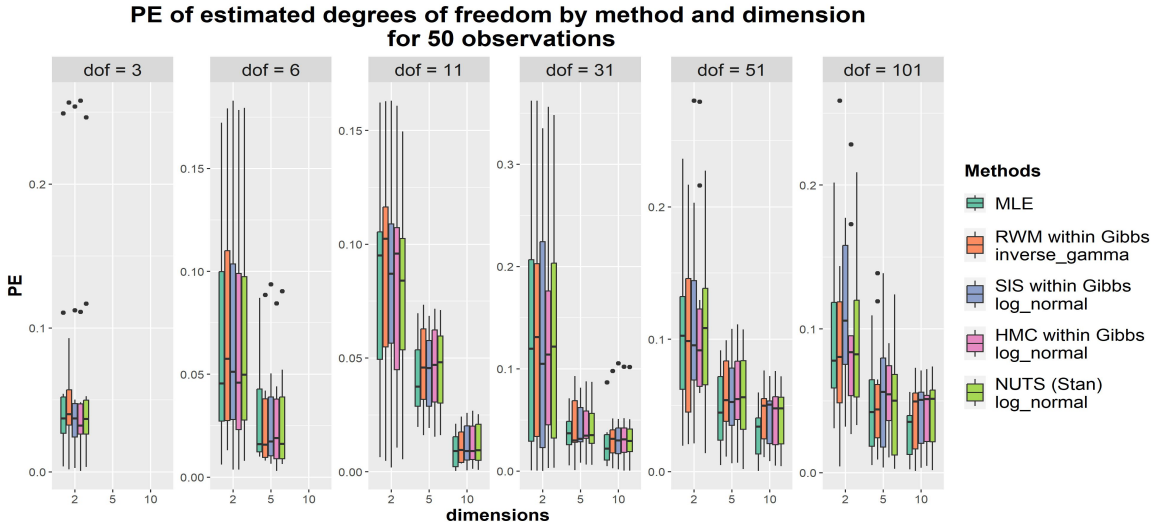
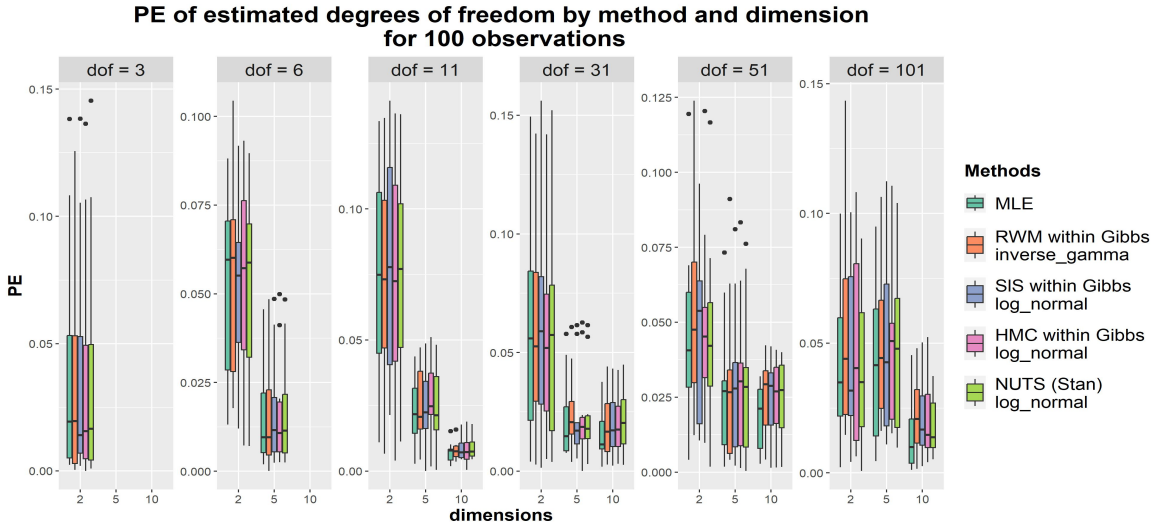


Figure C.4.7: Percentage Error of estimated degrees of freedom of each method (Random Walk Metropolis (RWM), Slice Sampling (SlS), Hamiltonian Monte Carlo (HMC) and NUTS (Stan)) of Sections C.3.1–C.3.2 per number of observations, dimension, actual degrees of freedom and random covariance matrices for 5 and 10 number observations with standard 3000 MCMC iterations.





(b)



(c)

Figure C.4.8: Percentage Error of estimated degrees of freedom of each method (Random Walk Metropolis (RWM), Slice Sampling (SlS), Hamiltonian Monte Carlo (HMC) and NUTS (Stan)) of Sections C.3.1–C.3.2 per number of observations, dimension, actual degrees of freedom and random covariance matrices for 30, 50 and 100 number observations with 3000 MCMC iterations.

C.4.3 Real Datasets

In this section, we compare the MCMC results of the investigated methods, based on their performance in three real-world datasets:

1. Air quality ($p = 2$) indexes by country and city[1].
2. The statistics of NBA players ($p = 7$) in 2022/23 for the regular season[2].
3. Quantitative characterization of morphological polymorphism of handwritten characters loops ($p = 20$); see in Marquis et al. (2006) for a detailed description of the data.

The first dataset consists of air quality indexes from different regions. Specifically, the available indexes for quantitative analysis are PM2.5, which refers to tiny particles or droplets in the air that are 2.5 micrometers or less in width, and ozone, a gas that can form in the atmosphere through a chemical reaction between sunlight and other pollutants. In this example, we consider only the Alpine countries —Switzerland, Italy, France, Germany, Austria, and Slovenia (excluding Liechtenstein due to limited data)— as homogeneous regions. Hence, for these countries, we calculate the covariance matrices based on the air quality indexes of the cities.

The second dataset, we compute the covariance matrices of 30 NBA teams based on the players' statistics. The average statistics per game considered include total rebounds, assists, steals, blocks, turnovers, personal fouls, and points.

The third dataset, which served as the primary motivation for this analysis, consists of handwritten loop characters quantified using Fourier analysis, as described by Marquis et al. (2006). These characters were subsequently modeled within a Bayesian framework, following the methodologies outlined by Bozza et al. (2008). Specifically, the dataset includes 10 harmonics of the Fourier series, which describe four characters (a, d, o, q) written by 13 different individuals. For these writers and characters, we calculated the covariance matrices based on 20 Fourier descriptors. It is worth to mentioned that in the probabilistic framework developed by Bozza et al. (2008), the estimation of degrees of freedom was identified as an open issue, which necessitated the adoption of the minimal feasible value of the parameter degrees of freedom.

Regarding the prior specification, we use the one that produced the best MCMC diagnostics, as detailed in Section C.4.1. Table C.4.9 presents the results of different MCMC approaches applied to three datasets presented in this section (Air Pollution, NBA, and Handwriting). Each algorithm was run with a 5-minute time limit. The table reports the number of iterations, degrees of freedom (DoF), effective sample size (ESS), Monte Carlo standard error (MCSE), and Geweke's statistic. The Maximum Likelihood Estimation (MLE) method provides only DoF values, with estimates closely aligning with the Bayesian point estimates. The Random Walk Metropolis (RWM) within Gibbs sampling, using a Gamma distribution, yields moderate values for ESS and Geweke's diagnostic across all generated datasets. The Slice Sampling within Gibbs, based on an Inverse-Gamma distribution, provided improved ESS and Geweke's diagnostic values, particularly for the Air Pollution dataset. The Hamiltonian Monte Carlo (HMC) within Gibbs, with a LogNormal prior, provides even higher ESS values, in accordance with the results for the simulated datasets. The No-U-Turn Sampler (NUTS) in Stan (with LogNormal prior) achieves the highest ESS for the NBA and Air Pollution datasets, indicating an efficient sampling algorithm. However, for the Handwriting dataset, where the dimension is higher, this method again did not generate a sufficient number of samples from the posterior distribution in

[1]Source: https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates
[2]Source: https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular

an efficient way (which is in accordance with our findings from the implementation in the simulated datasets of higher dimension).

Table C.4.10 presents the average Percentage Difference (PD) between the posterior means of the elements of the scale matrix for each pair of methods under consideration. The PD for comparing method l to method k is obtained using the following formula:

$$PD_{\mathbf{V}}^{(l,k)} = \frac{2}{d(d+1)} \sum_{i=1}^{p} \sum_{j=1}^{i} \left| \frac{\widehat{\mathbf{V}}_{ij}^{l} - \widehat{\mathbf{V}}_{ij}^{k}}{\widehat{\mathbf{V}}_{ij}^{l}} \right| \times 100; \tag{C.18}$$

where $\widehat{\mathbf{V}}_{ij}^{k}$ is the posterior mean of the element i, j of the scale matrix $\mathbf{V}$, using method $\{l, k\}$. In Table C.4.10, the averages of $PD_{\mathbf{V}}^{(l,k)}$ across all datasets are reported for each pair of methods under consideration. The Maximum Likelihood Estimation (MLE) method serves as the benchmark for reference.

Notably, the No-U-Turn Sampler (NUTS), as implemented in Stan, exhibits the largest discrepancies relative to other methods. This is primarily due to the handwriting dataset, where the resulting posterior sample size was small — only 41 samples were obtained within the specified time limit (see Table C.4.9).

Using Bayesian inference, we present the posterior distributions of the parameters, employing the No-U-Turn Sampler (NUTS) for the Air Pollution and NBA datasets, and the Hamiltonian Monte Carlo (HMC) within Gibbs for the Handwriting dataset. Figure C.4.9 illustrates the distribution of the posterior samples of the degrees of freedom for each dataset. The Handwriting dataset exhibits the narrowest posterior distribution of the degrees of freedom, ranging from 21 to 22, while the Air Pollution dataset shows the widest range, spanning from 5 to 30. Additionally, the posterior distributions for the degrees of freedom of the Air Pollution dataset appear to be positively skewed, whereas the NBA and Handwriting datasets display more symmetric distributions. The trace plot of the best performing method for each analyzed dataset is presented in Figure C.4.10.

Figure C.4.11 shows the 3D histograms of the posterior samples of the largest and smallest eigenvalues of the scale matrix $\mathbf{V}$ for each dataset. According to the theory, as the degrees of freedom increase, the largest and smallest eigenvalues becomes more dispersed (Bekker et al., 2017). This occurs because the largest eigenvalue corresponds to the direction of greatest variance in the data, while the smallest eigenvalue corresponds to the direction of lowest variance. Consequently, when the degrees of freedom are larger, the matrices are more informative, implying that some directions exhibit substantially greater variability than others. As a result, the range of the eigenvalues is wider.

We can extend this observation by further examining the posterior variability of the degrees of freedom. In the Air Pollution dataset, the estimated degrees of freedom are close to 14. Based on this value, we would expect considerable posterior variability in the range of eigenvalues of $\mathbf{V}$ (i.e. the difference between the largest and the smallest values). However, this is not the case in the 3D histogram (Figure C.4.11), due to the high posterior uncertainty associated with the degrees of freedom. On the other hand, in the Handwriting dataset, where the posterior distribution of the degrees of freedom is narrower — close to the lower bound of $p - 1$ (with $p = 20$), the corresponding plot is considerably wider than that of the Air Pollution dataset. Finally, in the NBA dataset, the 3D histogram in Figure C.4.11 is more spread out, with bars appearing more scattered along both axes.

Detailed further results over a fixed number of iterations (1,000 burn-in period and 3,000 iterations) are available in Appendix C.4.4.

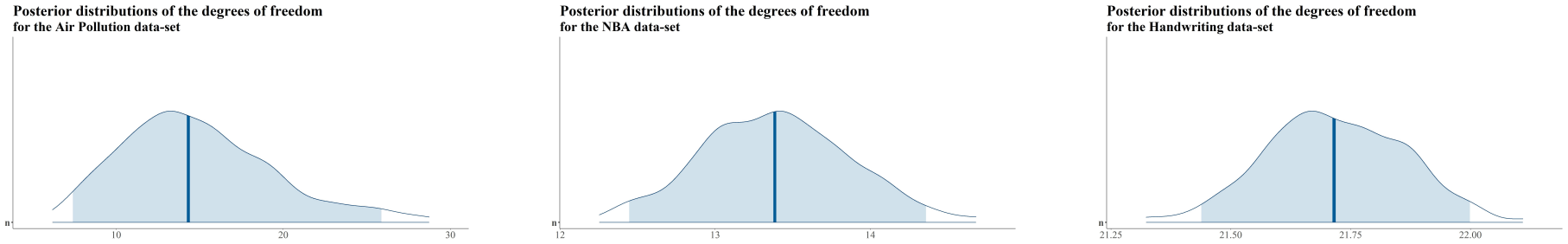


Figure C.4.9: Posterior distribution of the degrees of freedom per dataset (NUTS approach for Air Pollution and NBA dataset, HMC approach for Handwriting dataset).

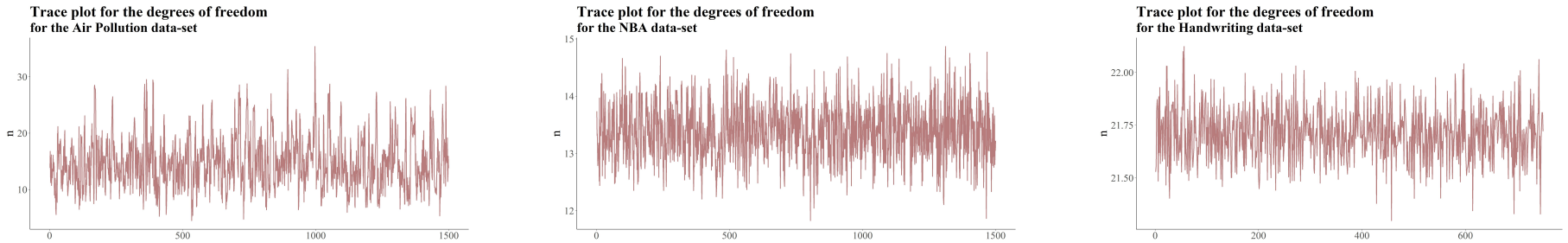


Figure C.4.10: Trace plots of the degrees of freedom per dataset (NUTS approach for Air Pollution and NBA dataset, HMC approach for Handwriting dataset).

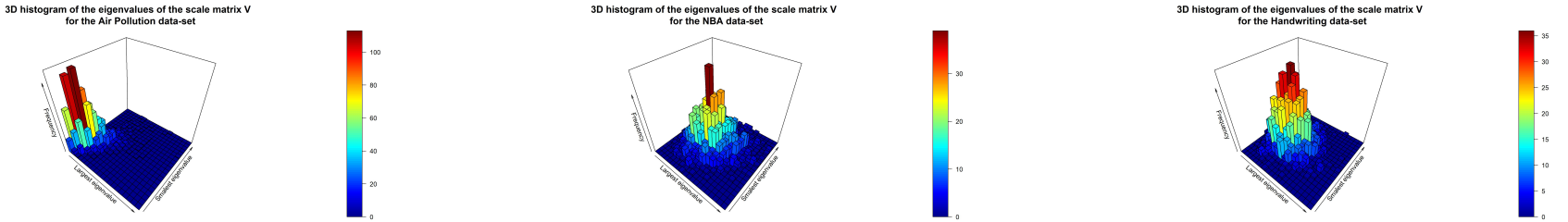


Figure C.4.11: 3D Histogram of the posterior sample of the largest and the smallest eigenvalues of the scale matrix $\mathbf{V}$ per dataset (NUTS approach for Air Pollution and NBA dataset, HMC approach for Handwriting dataset).

Approach	Air Pollution (p=2)					NBA (p=7)					Handwriting (p=20)				
	Iter	DoF	ESS	MCSE	Geweke	Iter	DoF	ESS	MCSE	Geweke	Iter	DoF	ESS	MCSE	Geweke
MLE	-	14.883	-	-	-	-	13.220	-	-	-	-	21.647	-	-	-
Random Walk Metropolis within Gibbs (Gamma)	3000	16.083	7	2.935	2.001	3000	13.390	140	0.522	0.788	3000	21.699	111	0.124	1.495
Slice Sampling within Gibbs (IGamma)	3000	15.864	88	3.833	1.456	3000	13.385	358	0.363	1.804	3000	21.728	770	0.444	0.918
HMC within Gibbs (LNormal)	3000	13.639	22	2.751	2.074	3000	13.432	716	0.459	1.233	3000	21.717	1924	0.137	1.028
NUTS-Stan (LNormal)	3000	14.339	340	2.843	2.705	3000	13.384	805	0.965	0.254	238	21.505	120	0.159	0.847

DoF: Degrees of Freedom; ESS: Effective Sample Size; MCSE: Monte Carlo Standard Error; Geweke: Geweke (1992) statistic

Table C.4.9: MCMC efficiency results per method for the real datasets of Section C.4.3 with a five-minutes runtime limit.

	MLE	RW Metropolis within Gibbs (Gamma)	Slice Sampling within Gibbs (IGamma)	HMC within Gibbs (LNormal)	NUTS-Stan (LNormal)
MLE	0	2.5	2.8	2.5	9.5
RW Metropolis within Gibbs (Gamma)	2.4	0	4.2	0.1	11.0
Slice Sampling within Gibbs (IGamma)	3.4	4.7	0	4.7	10.6
HMC within Gibbs (LNormal)	2.4	0.1	4.2	0	11.0
NUTS-Stan (LNormal)	12.2	13.3	11.3	13.3	0

RW: Random Walk

Table C.4.10: Average Percentage Difference (PD) across the real datasets of Section C.4.3 for the posterior means of the elements of the scale matrix for each pair of methods under consideration with a five-minutes runtime limit.

C.4.4 Real Datasets with Predefined Number of Iterations

In this section, we compare the MCMC results of the investigated methods using a fixed number of iterations, based on their performance in three real-world datasets. Regarding the prior specification, we use the one that produced the best MCMC diagnostics, as detailed in Section C.4.2.

Table C.4.11 presents the results of different MCMC approaches applied to three datasets: Air Pollution (p=2), NBA (p=7), and Handwriting (p=20). For each approach, the table shows the computation time, degrees of freedom (DoF), effective sample size (ESS), and Geweke diagnostic.

The Maximum Likelihood Estimation (MLE) method only provides DoF values and the estimates are very close to the Bayesian point estimates. Among the MCMC methods, the Random Walk Metropolis (RWM) within Gibbs sampling using a Gamma distribution shows moderate ESS and Geweke values across datasets. Building on this, slice Sampling within Gibbs using a LogNormal distribution improves ESS and Geweke values, especially for the Air Pollution dataset. The Hamiltonian Monte Carlo (HMC) within Gibbs using a LogNormal distribution further increases ESS, particularly for the Handwriting dataset. Finally, the No-U-Turn Sampler (NUTS) in Stan using a LogNormal distribution achieves the highest ESS across all datasets, indicating efficient sampling. However, this comes at a cost of significantly increased computation time, especially for the Handwriting dataset. Thus, the results of the real datasets are consistent with the simulation findings presented in Section C.4.2. Notably, the No-U-Turn Sampler (NUTS) implemented in Stan exhibits the largest time discrepancies when compared to other methods.

Approach	Air Pollution (p=2)				NBA (p=7)				Handwriting (p=20)			
	Time	DoF	ESS	Geweke	Time	DoF	ESS	Geweke	Time	DoF	ESS	Geweke
MLE	-	14.883	-	-	-	13.220	-	-	-	21.647	-	-
RWM within Gibbs (Gamma)	0.013	15.831	9	0.554	0.046	13.374	177	0.466	0.142	21.703	138	1.012
Slice Sampling within Gibbs (Gamma)	0.049	14.991	114	2.950	0.120	13.424	436	0.741	0.386	21.719	939	1.470
HMC within Gibbs (LNormal)	0.068	13.669	30	0.780	0.298	13.412	748	0.894	0.794	21.716	2513	0.050
NUTS-Stan (LNormal)	0.029	14.370	2000	1.118	0.414	13.380	2000	0.361	45.024	21.531	2000	0.359

Table C.4.11: MCMC efficiency results per method for the real datasets of Section C.4.3 with predefined MCMC iterations of B=3000.

	MLE	RWM within Gibbs (Gamma)	Slice Sampling within Gibbs (Gamma)	HMC within Gibbs (LNormal)	NUTS-Stan (LNormal)
MLE	0	0.014	0.046	0.014	0.058
RWM within Gibbs (Gamma)	0.014	0	0.046	0.001	0.056
Slice Sampling within Gibbs (Gamma)	0.043	0.044	0	0.044	0.032
HMC within Gibbs (LNormal)	0.014	0.001	0.046	0	0.056
NUTS-Stan (LNormal)	0.074	0.082	0.059	0.082	0

Table C.4.12: Average Percentage Difference (PD) across the real datasets of Section C.4.3 for the posterior means of the elements of the scale matrix for each pair of methods under consideration with predefined MCMC iterations of B=3000.

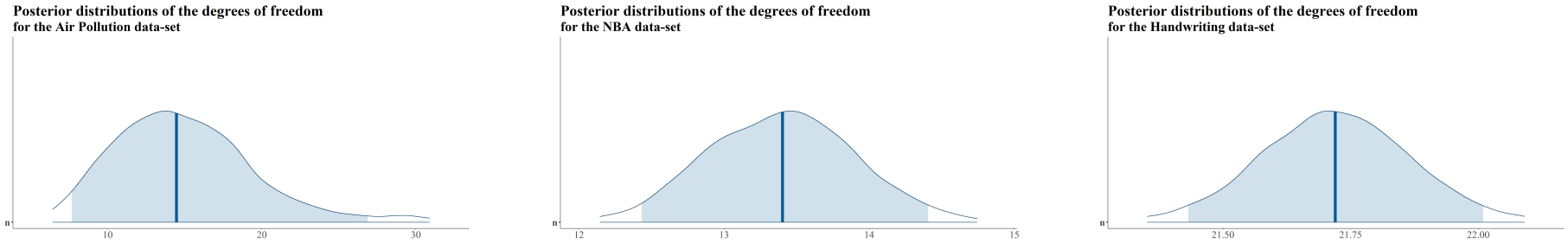


Figure C.4.12: Posterior distribution of the degrees of freedom per dataset (NUTS approach).

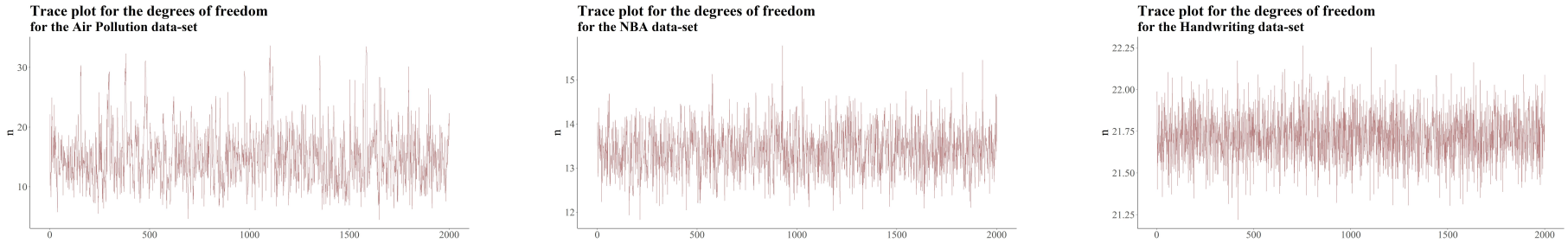


Figure C.4.13: Trace plots of the degrees of freedom per dataset (NUTS approach).

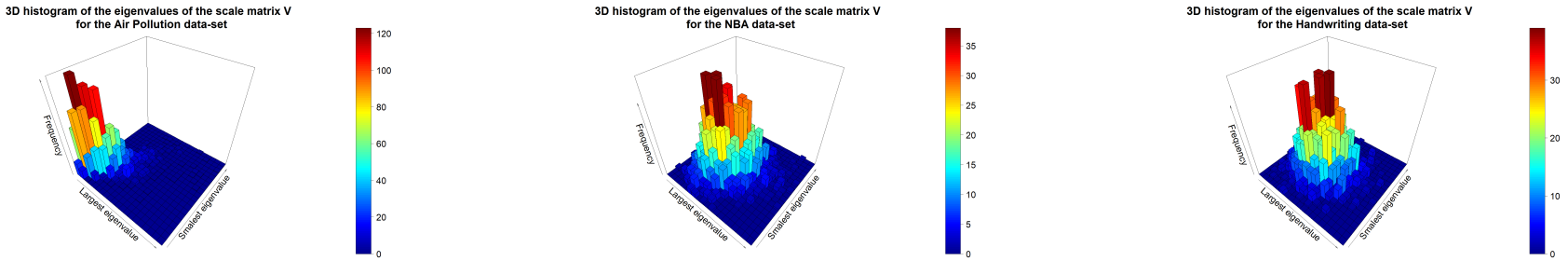


Figure C.4.14: 3D Histogram for the posterior sample of the biggest and the smallest eigenvalues of the scale matrix $\mathbf{V}$ per dataset (NUTS approach).

C.5 Discussion

In this chapter, we employed a variety of Markov chain Monte Carlo methods and explored different prior distributions for the degrees of freedom in the Bayesian modeling of the Wishart distribution. The maximum likelihood estimation with Bisection and Newton-Rapshon method are used as reference. This distribution is frequently used to model covariance matrices in multivariate settings and serves as a conjugate prior to the precision matrix of a multivariate normal distribution.

The methods considered in this study are widely recognized in the relevant literature. Specifically, we examined and compared four different MCMC implementations: (a) Within-Gibbs Random Walk Metropolis, (b) Within-Gibbs Slice Sampling, (c) Within-Gibbs Hamiltonian Monte Carlo (HMC), and (d) the No-U-Turn Sampler (NUTS) as implemented in Stan. We evaluated these methods using both simulated and real datasets, comparing their performance based on the number of iterations completed within a fixed runtime, the posterior precision of central point estimates, the effective sample size, and convergence diagnostics. Our results highlight that each method has distinct advantages and limitations in terms of robustness, accuracy, and sampling efficiency. Overall, the findings suggest that these approaches provide reliable and efficient tools for estimating the parameters of the Wishart distribution.

Identification of the best sampling method is challenging, as it depends on various factors, such as the number of observations, dimensionality, and true value of the degrees of freedom. Our simulation results indicate that different methods perform best for low-dimensional and high-dimensional covariance matrices.

For low-dimensional problems ($p \leq 15$), the No-U-Turn Sampler (NUTS) with a LogNormal prior demonstrates superior sampling efficiency, achieving a high effective sample size and very low Monte Carlo standard error (MCSE), suggesting that the Stan implementation is the optimal choice in this case. Additionally, it provides more robust and accurate estimates for the scale matrix. However, the NUTS algorithm shows convergence issues in some cases; for instance, with larger sample sizes, even of moderate size (e.g., 100 observations), it does not to generate an adequate number of posterior samples in a reasonable time frame.

For high-dimensional covariance matrices ($p > 15$), NUTS is prohibitively slow, resulting in a very low number of generated posterior samples. In this setting, HMC within Gibbs sampling with a LogNormal prior is the most efficient method in terms of generated sample size. However, in high-dimensional settings where computational time is a critical factor, slice sampling with an exponential prior seems to be the most suitable choice, after some compromises with respect the obtained effective sample size.

This (expected) result underscores the importance of choosing the appropriate method for each dataset based on the specific characteristics and the computational constraints at hand. Based on this observation, two additional aspects must be considered. First, the tuning parameters of hybrid methods can be optimized on a case-by-case basis, especially when the degrees of freedom are close to the lower bound of the Wishart distribution (that is, for $n > p - 1$). Second, we may use the Cholesky decomposition of the variance-covariance matrices. This can improve the efficiency of the sampling process, particularly in high-dimensional settings, and simplify the implementation of certain models, especially in probabilistic programming languages like Stan.

Finally, the results obtained from the real dataset closely align with the simulated results, which is particularly encouraging. This consistency highlights the robustness of our methods and their applicability to real-life problems where inference of the Wishart parameters is of crucial importance

in empirical studies.

For future work, it would be beneficial to explore the application of these methods to a broader range of real-world datasets to further validate their effectiveness and generalizability. Additionally, studying the effect of different types of prior distributions can provide valuable insights into how each choice influences posterior inference. This analysis may lead to informed recommendations on selecting the most appropriate prior for modeling the degrees of freedom. Accordingly, examining the impact of MCMC tuning parameters on the performance of the implemented computational algorithms across different settings may offer a better understanding of how these methods can be further boosted. Another promising direction is the development of more efficient algorithms and computational techniques, such as parallel computing or the Cholesky decomposition, to more effectively handle higher-dimensional datasets. Such advances could greatly enhance the practical applicability of Bayesian modeling of the Wishart distribution in complex datasets settings.

Appendix D

Abbreviations

AIC	Akaike Information Criterion
BAR	Sample Mean
BF	Bayes Factor
BIC	Bayesian Information Criterion
CNN	Convolutional Neural Network
DTW	Dynamic Time Warping
EM	Expectation-Maximization
ENFSI	European Network of Forensic Science Institutes
ESS	Effective Sample Size
FNR	False Negative Rate
FPR	False Positive Rate
GMM	Gaussian Mixture Models
HMC	Hamiltonian Monte Carlo
HMM	Hidden Markov Model
HQC	Hannan-Quinn Information Criterion
I.I.D.	Independent Identically Distributed
IW	Inverse-Wishart
LKJ	Lewandowski-Kurowicka-Joe distribution
LOGBF	Logarithmic Bayes Factor
LOGLR	Logarithmic Likelihood Ratio
LR	Likelihood Ratio
MANOVA	Multivariate Analysis of Variance
MAX	Maximum
MCMC	Markov Chain Monte Carlo
MCSE	Markov Carlo Standard Error
ML	Marginal Likelihood
MLE	Maximum Likelihood Estimation
N.I.S.T.	National Institute of Standards and Technology
NIW	Normal-Inverse-Wishart
N.R.C.	National Research Council
NUTS	No-U-Turn Sampler
PCA	Principle Component Analysis

PCAST	President's Council of Advisors on Science and Technology
PE	Percentage Error
PD	Percentage Difference
PDF	Probability Density Function
PCAST	President's Council of Advisors on Science and Technology
PoI	Person of Interest
RWM	Random Walk Metropolis
SD	Standard Deviation
SIS	Slice Sampling
SLR	Score-based Likelihood Ratio
SM-DTW	Stability Modulate Dynamic Time Warping
TER	Total Error Rate
VAR	Variance

Bibliography

Abbasi, B., Jahromi, A. H. E., Arkat, J. and Hosseinkouchack, M. (2006), 'Estimating the parameters of Weibull distribution using simulated annealing algorithm', *Applied Mathematics and Computation* **183**(1), 85–93.

Abramowitz, M. and Stegun, I. A. (1965), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Vol. 55, Courier Corporation.

Aitken, C. G. and Lucy, D. (2004), 'Evaluation of trace evidence in the form of multivariate data', *Journal of the Royal Statistical Society Series C: Applied Statistics* **53**(1), 109–122.

Aitken, C., Roberts, P. and Jackson, G. (2010), *Fundamentals of probability and statistical evidence in criminal proceedings: guidance for judges, lawyers, forensic scientists and expert witnesses*, Vol. 36, Royal Statistical Society, London.

Aitken, C. and Taroni, F. (1998), 'A verbal scale for the interpretation of evidence.', *Science & Justice* **8**, 279–281.

Aitken, C., Taroni, F. and Bozza, S. (2021), *Statistics and the evaluation of evidence for forensic scientists*, 3 edn, John Wiley and Sons, Chichester.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Al-Hmouz, R., Pedrycz, W., Daqrouq, K., Morfeq, A. and Al-Hmouz, A. (2019), 'Quantifying dynamic time warping distance using probabilistic model in verification of dynamic signatures', *Soft Computing* **23**, 407–418.

Allen, R. J. (2013), 'Taming complexity: rationality, the law of evidence and the nature of the legal system', *Law, Probability and Risk* **12**(2), 99–113.

Allen, R. J. (2017), 'The nature of juridical proof: Probability as a tool in plausible reasoning', *The International Journal of Evidence & Proof* **21**(1-2), 133–142.

Alvarez, I., Niemi, J. and Simpson, M. (2014), 'Bayesian inference for a covariance matrix', *arXiv preprint arXiv:1408.4050* .

Ardia, D., Baştürk, N., Hoogerheide, L. and Van Dijk, H. K. (2012), 'A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood', *Computational Statistics & Data Analysis* **56**(11), 3398–3414.

Bakis, R. (1976), 'Continuous speech recognition via centisecond acoustic states', *The Journal of the Acoustical Society of America* **59**(S1), S97–S97.

Banks, D. L., Kafadar, K., Kaye, D. H. and Tackett, M. (2020), *Handbook of forensic statistics*, CRC Press.

Barnard, J., McCulloch, R. and Meng, X.-L. (2000), 'Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage', *Statistica Sinica* pp. 1281–1311.

Baron, J. (2012), 'The point of normative models in judgment and decision making'.

Baum, L. E. and Petrie, T. (1966), 'Statistical inference for probabilistic functions of finite state Markov chains', *The annals of mathematical statistics* **37**(6), 1554–1563.

Bekker, A., Van Niekerk, J. and Arashi, M. (2017), 'Wishart distributions: Advances in theory with Bayesian application', *Journal of Multivariate Analysis* **155**, 272–283.

Berger, J. O., Bernardo, J.-m. and Sun, D. (2024), *Objective Bayesian Inference*, World Scientific.

Bernardo, J. M., Smith, A. F. and Berliner, M. (1994), *Bayesian theory*, Vol. 586, Wiley Online Library.

Bertsimas, D. and Tsitsiklis, J. (1993), 'Simulated annealing', *Statistical Science* **8**(1), 10–15.

Biedermann, A., Bozza, S. and Taroni, F. (2016), 'The decisionalization of individualization', *Forensic science international* **266**, 29–38.

Biedermann, A., Bozza, S. and Taroni, F. (2018), 'Analysing and exemplifying forensic conclusion criteria in terms of Bayesian decision theory', *Science & Justice* **58**(2), 159–165.

Biedermann, A., Bozza, S., Taroni, F. and Garbolino, P. (2018), 'A formal approach to qualifying and quantifying the 'goodness' of forensic identification decisions', *Law, Probability and Risk* **17**(4), 295–310.

Biedermann, A., Champod, C. and Willis, S. (2017), 'Development of european standards for evaluative reporting in forensic science: The gap between intentions and perceptions', *The International Journal of Evidence & Proof* **21**(1-2), 14–29.

Biedermann, A. and Taroni, F. (2006), 'Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data', *Forensic science international* **157**(2-3), 163–167.

Biedermann, A., Taroni, F. and Aitken, C. (2014), 'Liberties and constraints of the normative approach to evaluation and decision in forensic science: a discussion towards overcoming some common misconceptions', *Law, Probability and Risk* **13**(2), 181–191.

Biedermann, A., Taroni, F. and Champod, C. (2012), 'How to assign a likelihood ratio in a footwear mark case: an analysis and discussion in the light of R V T', *Law, Probability and Risk* **11**(4), 259–277.

Birch, I., Birch, M. and Lall, J. (2021), 'The accuracy and validity of the sheffield features of gait tool', *Science & Justice* **61**(1), 72–78.

Bishop, C. M. and Nasrabadi, N. M. (2006), *Pattern recognition and machine learning*, Vol. 4, Springer.

Box, G. E. and Draper, N. R. (1987), *Empirical model-building and response surfaces.*, John Wiley & Sons.

Box, G. E. and Tiao, G. C. (1973), *Bayesian inference in statistical analysis*, John Wiley & Sons.

Boyd, S. P. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge University Press.

Bozza, S., Broséus, J., Esseiva, P. and Taroni, F. (2014), 'Bayesian classification criterion for forensic multivariate data', *Forensic science international* **244**, 295–301.

Bozza, S., Taroni, F. and Biedermann, A. (2022), *Bayes factors for forensic decision analyses with R*, Springer. Open access: https://link.springer.com/book/10.1007/978-3-031-09839-0.

Bozza, S., Taroni, F., Marquis, R. and Schmittbuhl, M. (2008), 'Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship', *Journal of the Royal Statistical Society Series C: Applied Statistics* **57**(3), 329–341.

Breiman, L. (2001), 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)', *Statistical science* **16**(3), 199–231.

Burden, R. L., Faires, J. D. and Burden, A. M. (2015), *Numerical analysis*, Cengage Learning.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of statistical software* **76**, 1–32.

Casella, G. and Berger, R. (2024), *Statistical inference*, Chapman and Hall/CRC.

Champod, C. and Evett, I. W. (2000), 'Commentary on APA Broeders (1999)'Some observations on the use of probability scales in forensic identification', forensic linguistics 6 (2): 228–41', *The International Journal of Speech, Language and the Law* **7**(2), 238–243.

Chandra, S. (2020), 'Verification of dynamic signature using machine learning approach', *Neural Computing and Applications* **32**(15), 11875–11895.

Chandra, S. and Kumar, V. (2022), 'A novel approach to validate online signature using dynamic features based on locally weighted learning', *Multimedia Tools and Applications* **81**(28), 40959–40976.

Cheng, E. K. (2016), 'The burden of proof and the presentation of forensic results', *Harv. L. Rev. F.* **130**, 154.

Cole, S. A. (2010), 'Who speaks for science? A response to the National Academy of Sciences report on forensic science', *Law, Probability & Risk* **9**(1), 25–46.

Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I. (2018), 'Prior distributions for objective Bayesian analysis'.

Consonni, G. and Veronese, P. (2003), 'Enriched conjugate and reference priors for the Wishart family on symmetric cones', *The Annals of Statistics* **31**(5), 1491–1516.

Cook, R., Evett, I. W., Jackson, G., Jones, P. and Lambert, J. (1998), 'A hierarchy of propositions: deciding which level to address in casework', *Science & Justice* **38**(4), 231–239.

Costa, M. and De Angelis, L. (2010), 'Model selection in hidden Markov models: a simulation study'.

Crawford, A. M., Ommen, D. M. and Carriquiry, A. L. (2023), 'A rotation-based feature and Bayesian hierarchical model for the forensic evaluation of handwriting evidence in a closed set', *The Annals of Applied Statistics* **17**(2), 1127–1151.

Csiszár, I. and Shields, P. C. (2000), 'The consistency of the bic markov order estimator', *The Annals of Statistics* **28**(6), 1601–1619.

Delahaye, D., Chaimatanan, S. and Mongeau, M. (2019), Simulated annealing: From basics to applications, *in* 'Handbook of Metaheuristics', Springer, pp. 1–35.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the royal statistical society: series B (methodological)* **39**(1), 1–22.

Diaz, M., Fischer, A., Ferrer, M. A. and Plamondon, R. (2016), 'Dynamic signature verification system based on one real signature', *IEEE transactions on cybernetics* **48**(1), 228–239.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987), 'Hybrid Monte Carlo', *Physics letters B* **195**(2), 216–222.

Duchi, J. (2007), 'Properties of the trace and matrix derivatives', *Available electronically at* . **URL:** *https://web.stanford.edu/~jduchi/projects/matrix_prop.pdf*

Dyer, A. G., Found, B. and Rogers, D. (2006), 'Visual attention and expertise for forensic signature analysis', *Journal of Forensic Sciences* **51**(6), 1397–1404.

El_Rahman, S. A. (2017), 'An efficient approach for dynamic signature recognition', *International Journal of Intelligent Engineering Informatics* **5**(2), 167–190.

Evett, I. W. (1998), 'Towards a uniform framework for reporting opinions in forensic science casework', *Science & Justice* **3**(38), 198–202.

Evett, I. W., Jackson, G., Lambert, J. and McCrossan, S. (2000), 'The impact of the principles of evidence interpretation on the structure and content of statements.', *Science & justice: journal of the Forensic Science Society* **40**(4), 233–239.

Faigman, D. L. (2007), 'Anecdotal forensics, phrenology, and other abject lessons from the history of science', *Hastings LJ* **59**, 979.

Fierrez, J. and Ortega-Garcia, J. (2008), On-line signature verification, *in* 'Handbook of biometrics', Springer, Boston, pp. 189–209.

Fierrez, J., Ortega-Garcia, J., Ramos, D. and Gonzalez-Rodriguez, J. (2007), 'HMM-based on-line signature verification: Feature extraction and signature modeling', *Pattern recognition letters* **28**(16), 2325–2334.

Found, B. and Rogers, D. (2008), 'The probative character of forensic handwriting examiners' identification and elimination opinions on questioned signatures', *Forensic Science International* **178**(1), 54–60.

Fuh, C.-D., Kao, C.-L. M. and Pang, T. (2024), 'Kullback-leibler divergence and akaike information criterion in general hidden markov models', *IEEE Transactions on Information Theory* **70**(8), 5888–5909.

Gaborini, L. (2021), Bayesian Models in Questioned Handwriting and Signatures, PhD thesis, Ph. D. thesis, École des Sciences Criminelles, Université de Lausanne.

Gaborini, L., Biedermann, A. and Taroni, F. (2017), 'Towards a Bayesian evaluation of features in questioned handwritten signatures', *Science & Justice* **57**(3), 209–220.

Galbraith, O., Galbraith, C. and Galbraith, N. (1995), 'The principle of the 'Drunkard's search'as a proxy for scientific analysis: The misuse of handwriting test data in a law journal article', *International Journal of Forensic Document Examiners* **1**(1), 7–17.

Gall, J. (2020), Simulated annealing, *in* 'Computer Vision: A Reference Guide', Springer, pp. 1–5.

Gelfand, A. E. and Dey, D. K. (1994), 'Bayesian model choice: asymptotics and exact calculations', *Journal of the Royal Statistical Society: Series B (Methodological)* **56**(3), 501–514.

Gelfand, A. E. and Smith, A. F. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American statistical association* **85**(410), 398–409.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)', *Bayesian Analysis* **1**(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995), *Bayesian data analysis*, Chapman and Hall/CRC, New York.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *Annals of Applied Statistics* **2**(4), 1360–1383.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* **7**(4), 457–472.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments, *in* 'Bayesian Statistics', Vol. 4, Clarendon Press, pp. 641–649.

Gilks, W. R., Best, N. G. and Tan, K. K. (1995), 'Adaptive rejection Metropolis sampling within Gibbs sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **44**(4), 455–472.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1995), *Markov chain Monte Carlo in practice*, CRC press.

Goffe, W. L., Ferrier, G. D. and Rogers, J. (1994), 'Global optimization of statistical functions with simulated annealing', *Journal of Econometrics* **60**(1-2), 65–99.

Granville, V., Krivánek, M. and Rasson, J.-P. (1994), 'Simulated annealing: A proof of convergence', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(6), 652–656.

Gronau, Q. F., Heathcote, A. and Matzke, D. (2020), 'Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling', *Behavior research methods* **52**(2), 918–937.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J. and Steingroever, H. (2017), 'A tutorial on bridge sampling', *Journal of mathematical psychology* **81**, 80–97.

Gronau, Q. F., Singmann, H. and Wagenmakers, E.-J. (2020), 'bridgesampling: An R package for estimating normalizing constants', *Journal of Statistical Software* **92**, 1–29.

Gupta, A. K. and Nagar, D. K. (1999), *Matrix variate distributions*, Vol. 104, CRC Press.

Hannan, E. J. and Quinn, B. G. (1979), 'The determination of the order of an autoregression', *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2), 190–195.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications'.

Henderson, D., Jacobson, S. H. and Johnson, A. W. (2003), The theory and practice of simulated annealing, *in* 'Handbook of Metaheuristics', Springer, pp. 287–319.

Hicks, T., Buckleton, J., Castella, V., Evett, I. and Jackson, G. (2022), 'A logical framework for forensic DNA interpretation', *Genes* **13**(6), 957.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011), 'Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials', *Biometrics* **67**(3), 1047–1056.

Hoffman, M. D., Gelman, A. et al. (2014), 'The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.', *J. Mach. Learn. Res.* **15**(1), 1593–1623.

Hopwood, A. J., Puch-Solis, R., Tucker, V. C., Curran, J. M., Skerrett, J., Pope, S. and Tully, G. (2012), 'Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts', *Science & Justice* **52**, 185–190.

Houck, M. M. and Siegel, J. A. (2009), *Fundamentals of forensic science*, Academic Press.

Hsu, C.-W., Sinay, M. S. and Hsu, J. S. (2012), 'Bayesian estimation of a covariance matrix with flexible prior specification', *Annals of the Institute of Statistical Mathematics* **64**, 319–342.

Huang, A. and Wand, M. P. (2013), 'Simple marginally noninformative prior distributions for covariance matrices'.

Idris, A. A. and Muhammad, S. S. R. (2022), 'A simulation study on the simulated annealing algorithm in estimating the parameters of generalized gamma distribution', *Science and Technology Indonesia* **7**(1), 84–90.

Ingber, L. (1993), 'Simulated annealing: Practice versus theory', *Mathematical and Computer Modelling* **18**(11), 29–57.

Izenman, A. J. (2020), Comparing handwriting in questioned documents, *in* 'Handbook of Forensic Statistics', Chapman and Hall/CRC, New York, pp. 341–363.

Jahan, M. and Farimani, S. (2018), An HMM for online signature verification based on velocity and hand movement directions, *in* '2018 6th Iranian joint congress on fuzzy and intelligent systems (CFIS). IEEE', pp. 205–209.

Jeffreys, H. (1998), *The theory of probability*, OuP Oxford.

Johansson, S., Leech, G. N. and Goodluck, H. (1978), 'Manual of information to accompany the lancaster-oslo: Bergen corpus of british english, for use with digital computers', *Department of English, University of Oslo, Norway* .

Johnson, M. Q. and Ommen, D. M. (2022), 'Handwriting identification using random forests and score-based likelihood ratios', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **15**(3), 357–375.

Jones, J. (2014), 'The future state of handwriting examinations: a roadmap to integrate the latest measurement science and statistics', *Proceedings of the American Academy of Forensic Sciences 66th Annual Scientific Meeting* p. 521.

Karlis, D. (2023), 'Master's course in computational statistics', Lecture notes, [Athens University Economics and Business]. Department of Statistics.

Kashi, R. S., Hu, J., Nelson, W. L. and Turin, W. (1997), On-line handwritten signature verification using hidden Markov model features, *in* 'Proceedings of the fourth international conference on document analysis and recognition', Vol. 1, IEEE, pp. 253–257.

Kass, R. E., Tierney, L. and Kadane, J. B. (1991), 'Laplace's method in Bayesian analysis', *Contemporary Mathematics* **115**, 89–99.

Kaur, H. and Kumar, M. (2023), 'Signature identification and verification techniques: state-of-the-art work', *Journal of Ambient Intelligence and Humanized Computing* **14**(2), 1027–1045.

Kazmierczyk, Z. and Turner, I. J. (2022), 'Self-identification of electronically scanned signatures (ESS) and digitally constructed signatures (DCS)', *Forensic Sciences Research* **7**(2), 261–264.

Kirkpatrick, S., Gelatt Jr, C. D. and Vecchi, M. P. (1983), 'Optimization by simulated annealing', *Science* **220**(4598), 671–680.

Koehler, J. J., Chia, A. and Lindsey, S. (1994), 'The random match probability in DNA evidence: Irrelevant and prejudicial', *JURIMETRICS j.* **35**, 201.

Law Commission (2011), *Expert Evidence in Criminal Proceedings in England and Wales*, The Stationery Office, London, UK. Published 22 March 2011.

Lehmann, E. L. and Casella, G. (1998), *Theory of point estimation*, Springer.

Leonard, T. and Hsu, J. S. (1992), 'Bayesian inference for a covariance matrix', *The Annals of Statistics* **20**(4), 1669–1696.

Leung, W.-C. (2002), 'The prosecutor's fallacy—a pitfall in interpreting probabilities in forensic evidence', *Medicine, science and the law* **42**(1), 44–50.

Lewandowski, D., Kurowicka, D. and Joe, H. (2009), 'Generating random correlation matrices based on vines and extended onion method', *Journal of multivariate analysis* **100**(9), 1989–2001.

Lewis, S. M. and Raftery, A. E. (1997), 'Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator', *Journal of the American Statistical Association* **92**(438), 648–655.

Linden, J. (2022), *Forensic Examination of Dynamic Signatures*, University of Lausanne. Doctoral Dissertation.

Linden, J., Bozza, S., Marquis, R. and Taroni, F. (2022), 'Bayesian evaluation of dynamic signatures in operational conditions', *Forensic Science International* **332**, 111173.

Linden, J. and Marquis, R. (2023), 'The influence of time on dynamic signature: An exploratory data analysis', *Forensic Science International* **348**, 111577.

Linden, J., Marquis, R., Bozza, S. and Taroni, F. (2018), 'Dynamic signatures: A review of dynamic feature variation and forensic methodology', *Forensic science international* **291**, 216–229.

Linden, J., Marquis, R. and Mazzella, W. (2017), 'Forensic analysis of digital dynamic signatures: new methods for data treatment and feature evaluation', *Journal of forensic sciences* **62**(2), 382–391.

Linden, J., Taroni, F., Marquis, R. and Bozza, S. (2021), 'Bayesian multivariate models for case assessment in dynamic signature cases', *Forensic Science International* **318**, 110611.

Lindley, D. V. (1957), 'A statistical paradox', *Biometrika* **44**, 187–192.

Lindley, D. V. (2013), *Understanding uncertainty*, John Wiley & Sons.

Liu, H., Zhang, Z. and Grimm, K. J. (2016), 'Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis', *Structural Equation Modeling: A Multidisciplinary Journal* **23**(3), 354–367.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility', *Statistics and computing* **10**(4), 325–337.

Makled, R. A. and Cheng, W. (2024), 'Exploring multivariate statistics: Unveiling the power of eigenvalues in Wishart distribution analysis', *Contemporary Mathematics* pp. 4054–4063.

Mardia, K. V. (1970), 'Measures of multivariate skewness and kurtosis with applications', *Biometrika* **57**(3), 519–530.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F. and Hicks, T. (2016), 'Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings', *Science & Justice* **56**(5), 364–370.

Marquis, R., Bozza, S., Schmittbuhl, M. and Taroni, F. (2011), 'Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios', *Journal of forensic sciences* **56**, S238–S242.

Marquis, R., Schmittbuhl, M., Mazzella, W. D. and Taroni, F. (2005), 'Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of the capital character o', *Forensic science international* **150**(1), 23–32.

Marquis, R., Taroni, F., Bozza, S. and Schmittbuhl, M. (2006), 'Quantitative characterization of morphological polymorphism of handwritten characters loops', *Forensic Science International* **164**(2-3), 211–220.

Marquis, R., Taroni, F., Bozza, S. and Schmittbuhl, M. (2007), 'Size influence on shape of handwritten characters loops', *Forensic science international* **172**(1), 10–16.

Marti, U.-V. and Bunke, H. (1999), A full English sentence database for off-line handwriting recognition, *in* 'Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318)', IEEE, pp. 705–708.

Martinez-Diaz, M., Fierrez, J., Ortega-García, J. et al. (2008), Incorporating signature verification on handheld devices with user-dependent hidden Markov models, *in* 'Proc. International Conference on Frontiers in Handwriting Recognition, ICFHR', Vol. 32.

Martire, K. A., Growns, B. and Navarro, D. J. (2018), 'What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts', *Psychonomic bulletin & review* **25**, 2346–2355.

Mazzolini, D., Mignone, P., Pavan, P. and Vessio, G. (2021), 'An easy-to-explain decision support framework for forensic analysis of dynamic signatures', *Forensic Science International: Digital Investigation* **38**, 301216.

Meester, R. and Slooten, K. (2021), *Probability and forensic evidence: Theory, philosophy, and applications*, Cambridge University Press.

Meng, X.-L. and Wong, W. H. (1996), 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration', *Statistica Sinica* **6**(4), 831–860.

Miguel-Hurtado, O., Mengibar-Pozo, L., Lorenz, M. G. and Liu-Jimenez, J. (2007), On-line signature verification by dynamic time warping and Gaussian mixture models, *in* '2007 41st annual IEEE international Carnahan conference on security technology', IEEE, Ottawa, Ontario, Canada, pp. 23–29.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Abril Pla, O., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S. et al. (2024), 'Prior knowledge elicitation: The past, present, and future', *Bayesian Analysis* **19**(4), 1129–1161.

Moenssens, A. A. (1999), 'Meeting the Daubert challenge to handwriting evidence: preparing for a Daubert hearing', *Forensic Science Communications* **1**, 1–8.

Morey, R. D., Rouder, J. N., Pratte, M. S. and Speckman, P. L. (2011), 'Using MCMC chain outputs to efficiently estimate Bayes factors', *Journal of Mathematical Psychology* **55**(5), 368–378.

Morrison, G. S. (2022), 'Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science', *Forensic Science International: Synergy* **5**, 100270.

Müller, M. (2007), 'Dynamic time warping', *Information retrieval for music and motion* **1**, 69–84.

Murphy, K. P. (2007), 'Conjugate Bayesian analysis of the Gaussian distribution', *def* **1**($2\sigma 2$), 16.

Neal, R. M. (2003), 'Slice sampling', *The Annals of Statistics* **31**(3), 705–767.

Neal, R. M. (2012), 'MCMC using Hamiltonian dynamics', *arXiv preprint arXiv:1206.1901* .

Newton, M. A. and Raftery, A. E. (1994), 'Approximate Bayesian inference with the weighted likelihood bootstrap', *Journal of the Royal Statistical Society: Series B* **56**(1), 3–48.

N.I.S.T. (2020), *Expert Working Group for Human Factors in Handwriting Examination. Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach*, Department of Commerce, National Institute of Standards and Technology. https://www.nist.gov/programs-projects/forensic-handwriting-examination-and-human-factors.

Nordgaard, A., Ansell, R., Drotz, W. and Jaeger, L. (2012), 'Scale of conclusions for the value of evidence', *Law, probability & risk* **11**(1), 1–24.

N.R.C. (2009), *Strengthening forensic science in the US: a path forward*, National Academy Press, Washington D.C.

Ntzoufras, I. (2011), *Bayesian modeling using WinBUGS*, John Wiley & Sons.

Oda, H. and Komaki, F. (2023), 'Enriched standard conjugate priors and the right invariant prior for Wishart distributions', *Journal of Multivariate Analysis* **193**, 105105.

Odell, P. and Feiveson, A. (1966), 'A numerical procedure to generate a sample covariance matrix', *Journal of the American Statistical Association* **61**(313), 199–203.

Okado, J. B., da Camara e Silva, E. S. and Sily, P. D. (2024), 'Dynamic signatures: a mathematical approach to analysis', *Forensic Sciences Research* p. owae067.

Ommen, D. M., Saunders, C. P. and Neumann, C. (2017), 'The characterization of Monte Carlo errors for the quantification of the value of forensic evidence', *Journal of Statistical Computation and Simulation* **87**(8), 1608–1643.

Overstall, A. M. and Forster, J. J. (2010), 'Default Bayesian model determination methods for generalised linear mixed models', *Computational Statistics & Data Analysis* **54**(12), 3269–3288.

Parziale, A., Diaz, M., Ferrer, M. A. and Marcelli, A. (2019), 'SM-DTW: Stability modulated Dynamic Time Warping for signature verification', *Pattern Recognition Letters* **121**, 113–122.

PCAST (2016), Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods, *in* 'President's Council of Advisors on Science and Technology Report', National Academy Press, Washington, D.C.

Perrakis, K., Ntzoufras, I. and Tsionas, E. G. (2014), 'On the use of marginal posteriors in marginal likelihood estimation via importance sampling', *Computational Statistics & Data Analysis* **77**, 54–69.

Pfanzagl, J. and Hamböker, R. (1994), *Parametric statistical theory*, Walter de Gruyter.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', *R News* **6**(1), 7–11.

Plummer, M., Stukalov, A. and Denwood, M. (2016), 'rjags: Bayesian graphical models using MCMC', *R package version* **4**(6).

Plummer, M. et al. (2003), Jags: A program for analysis of bayesian graphical models using Gibbs sampling, *in* 'Proceedings of the 3rd international workshop on distributed statistical computing', Vol. 124, Vienna, Austria, pp. 1–10.

Polyak, B. T. (1964), 'Some methods of speeding up the convergence of iteration methods', *USSR Computational Mathematics and Mathematical Physics* **4**(5), 1–17.

Press, J. S. (2005), *Applied multivariate analysis: using Bayesian and frequentist methods of inference*, Courier Corporation, North Chelmsford, Massachusetts.

Press, S. J. (1980), '4 Bayesian inference in MANOVA', *Handbook of statistics* **1**, 117–132.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007), *Numerical recipes in C++: The art of scientific computing*, Vol. 2, Cambridge University Press.

Rabiner, L. R. (2002), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Raiffa, H. and Schlaifer, R. (1961), *Applied statistical decision theory*, Wiley New York.

Richiardi, J., Ketabdar, H. and Drygajlo, A. (2005), Local and global feature selection for on-line signature verification, *in* 'Eighth International Conference on Document Analysis and Recognition (ICDAR'05)', IEEE, pp. 625–629.

Risinger, M. D., Saks, M. J., Thompson, W. C. and Rosenthal, R. (2002), 'The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion', *California Law Review* **90**, 1.

Robert, C. P., Casella, G. and Casella, G. (1999), *Monte Carlo statistical methods*, Vol. 2, Springer.

Ross, S. M. (2014), *Introduction to probability models*, Academic press.

Rowe, D. B. (2002), *Multivariate Bayesian statistics: models for source separation and signal unmixing*, Chapman and Hall/CRC, New York.

Rudin, W. (1976), *Principles of Mathematical Analysis*, 3rd edn, McGraw-Hill.

Saferstein, R. (2013), *Criminalistics*, Pearson Education.

Sakoe, H. and Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43–49.

Saks, M. J. and Koehler, J. J. (2008), 'The individualization fallacy in forensic science evidence', *Vand. L. Rev.* **61**, 199.

Saks, M. J., Risinger, D. M., Rosenthal, R. and Thompson, W. C. (2003), 'Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States', *Science & Justice* **43**(2), 77–90.

Schmittbuhl, M., Le Minor, J.-M., Allenbach, B. and Schaaf, A. (1998), 'Shape of the piriform aperture in Gorilla gorilla, Pan troglodytes, and modern Homo sapiens: characterization and polymorphism analysis', *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* **106**(3), 297–310.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics* pp. 461–464.

Sharman, K. C. (1988), Maximum likelihood parameter estimation by simulated annealing, *in* 'ICASSP-88, International Conference on Acoustics, Speech, and Signal Processing', IEEE Computer Society, pp. 2741–2742.

Sinharay, S. and Stern, H. S. (2005), 'An empirical comparison of methods for computing Bayes factors in generalized linear mixed models', *Journal of Computational and Graphical Statistics* **14**(2), 415–435.

Sita, J., Found, B. and Rogers, D. K. (2002), 'Forensic handwriting examiners' expertise for signature comparison', *Journal of Forensic Sciences* **47**(5), 1117–1124.

Soch, J. (2019), 'The book of statistical proofs', https://statproofbook.github.io/.

Stoney, D. A. (1991), 'What made us ever think we could individualize using statistics?', *Journal-Forensic Science Society* **31**(2), 197–199.

Suits, D. B. (1957), 'Use of dummy variables in regression equations', *Journal of the American Statistical Association* **52**(280), 548–551.

Tahmasebi, A. and Pourghassem, H. (2013), Signature identification using dynamic and HMM features and KNN classifier, *in* '2013 International Conference on Communication Systems and Network Technologies', IEEE, Gwalior, India, pp. 201–205.

Taroni, F., Aitken, C. G., Garbolino, P. and Biedermann, A. (2006), *Bayesian networks and probabilistic inference in forensic science*, Wiley Online Library.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P. and Aitken, C. (2014), *Bayesian networks for probabilistic inference and decision analysis in forensic science*, John Wiley & Sons.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P. and Aitken, C. (2010), *Data analysis in forensic science: A Bayesian decision perspective*, John Wiley & Sons, Chichester.

Taroni, F., Garbolino, P., Bozza, S. and Aitken, C. (2021), 'The Bayes' factor: the coherent measure for hypothesis confirmation', *Law, probability and risk* **20**(1), 15–36.

Taroni, F., Marquis, R., Schmittbuhl, M., Biedermann, A., Thiéry, A. and Bozza, S. (2012), 'The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination', *Forensic Science International* **214**(1-3), 189–194.

Thiéry, A. (2014), *Développement d'un processus de quantification et d'évaluation de caractères manuscrits : théorie et applications*, University of Lausanne. Doctoral Dissertation.

Thompson, G., Team, R. C. and Thompson, M. G. (2022), 'Package 'CholWishart".

Thompson, W. C. and Schumann, E. L. (2017), Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy, *in* 'Expert evidence and scientific proof in criminal trials', Routledge, pp. 371–391.

Timm, N. H. (2002), *Applied multivariate analysis*, Springer.

Tolosana, R., Vera-Rodriguez, R., Ortega-Garcia, J. and Fierrez, J. (2015), Update strategies for HMM-based dynamic signature biometric systems, *in* '2015 IEEE International Workshop on Information Forensics and Security (WIFS)', IEEE, Rome, Italy, pp. 1–6.

Vatsa, M., Singh, R., Mitra, P. and Noore, A. (2004), Signature verification using static and dynamic features, *in* 'Neural Information Processing: 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25, 2004. Proceedings 11', Springer, pp. 350–355.

Vera, J. F. and Díaz-García, J. A. (2008), 'A global simulated annealing heuristic for the three-parameter lognormal maximum likelihood estimation', *Computational Statistics & Data Analysis* **52**(12), 5055–5065.

Visser, I. and Speekenbrink, M. (2010), 'depmixs4: an R package for hidden Markov models', *Journal of statistical Software* **36**, 1–21.

Viterbi, A. (1967), 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE transactions on Information Theory* **13**(2), 260–269.

Vorugunti, C. S., Gautam, A. and Pulabaigari, V. (2023), Osvcontramer: A hybrid CNN and Transformer based online signature verification, *in* '2023 IEEE International Joint Conference on Biometrics (IJCB)', IEEE, India, pp. 1–10.

Vorugunti, C. S., Subramanian, B., Gautam, A. and Pulabaigari, V. (2022), Impact of type of convolution operation on performance of convolutional neural networks for online signature verification, *in* 'International Conference on Frontiers in Handwriting Recognition', Springer, Slovenia, pp. 83–97.

Vuille, J., Lupària, L. and Taroni, F. (2017), 'Scientific evidence and the right to a fair trial under article 6 echr'(2017)', *Law, Probability and Risk* **16**, 55.

Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C., Sjerps, M., Lucena-Molina, J., Zadora, G., Aitken, C., Lovelock, T., Lunt, L., Champod, C., Biedermann, A., Hicks, T. and Taroni, F. (2015), 'ENFSI guideline for evaluative reporting in forensic science, Strengthening the evaluation of forensic results across Europe (STEOFRAE)', Dublin. https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

Wishart, J. (1928), 'The generalised product moment distribution in samples from a normal multivariate population', *Biometrika* pp. 32–52.

Wright, S., Nocedal, J. et al. (1999), 'Numerical optimization', *Springer Science* **35**(67-68), 7.

Wu, C. J. (1983), 'On the convergence properties of the EM algorithm', *The Annals of statistics* pp. 95–103.

Wydra, J. and Matuszewski, S. (2022), 'Likelihood ratio to evaluate handwriting evidence using similarity index', *Law, Probability and Risk* **21**(1), 21–42.

Yang, L., Widjaja, B. and Prasad, R. (1995), 'Application of hidden Markov models for signature verification', *Pattern recognition* **28**(2), 161–170.

Zhang, Z. (2021), 'A note on Wishart and inverse Wishart priors for covariance matrix', *Journal of Behavioral Data Science* **1**(2), 119–126.

Zucchini, W., MacDonald, I. L. and Langrock, R. (2017), *Hidden Markov Models for Time Series: An Introduction Using R*, 2nd edn, CRC Press.