2023

A Class of Regression Models for Pairwise Comparisons of Forensic Handwriting Comparison Systems

Cami M. Fuglsby

A CLASS OF REGRESSION MODELS FOR PAIRWISE COMPARISONS OF FORENSIC HANDWRITING COMPARISON SYSTEMS

BY

CAMI M. FUGLSBY

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science & Statistics

South Dakota State University

2023

DISSERTATION ACCEPTANCE PAGE
Cami Fuglsby

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Christopher Saunders
Advisor Date

Kurt Cogswell
Department Head Date

Nicole Lounsbery, PhD
Director, Graduate School Date

To my parents, Dale and Vicki, and to my brother, Brett, for all of your help, patience, and understanding while I completed this.

To Ghost for keeping me company all these years.

ACKNOWLEDGEMENTS

I would like to acknowledge the volunteers who participated in providing the handwriting samples, and Sciometrics for providing access to FLASH ID® system to use in this research.

I would also like to acknowledge the many people who have helped me throughout my degree in many ways. Specifically, Dr. JoAnn Buscaglia for her mentorship about research within the forensic science field, Danica Ommen for her guidance in navigating a Ph.D. in statistics while focusing on forensics and for being my drinking buddy and hiking partner, Prof. Michael Caligiuri for displaying how to lead grants and large blocks of research, and Janean Hanka for her help and encouragement while I finished this dissertation.

To the forensic document examiner community, thank you for the many opportunities you provided for me to learn more about the discipline and for welcoming me to present my research and help write standards for examinations.

I owe my advisor, Christopher Saunders, the biggest thank you in the world and the largest bottle of red wine for the mentorship, guidance, and patience you have provided me.

Various parts of this work are supported under a grant from the National Institute of Justice, U.S. Department of Justice (Awards No. 2017-DN-BX-0148 and No. 2019-DU-BX-0011).

Funding for this research was provided by the National Science Foundation Division of Graduate Education under award number DGE-1828492: "Collaborative Research: NRT: Cyber-Physical-Social System for Understanding and Thwarting the Illicit Economy."

The views expressed are those of the authors and do not necessarily reflect the

official policy or position of the FBI or the U.S. government. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI.

TABLE OF CONTENTS

LIST OF FIGURES ix

LIST OF TABLES xii

LIST OF ALGORITHMS xiv

ABSTRACT xiv

CHAPTERS 1

1 Introduction and Overview 1

1.1 Contributions and Chapter Summaries 2

2 Dissimilarity Scores from an Automated System and Forensic Document Examiners 4

2.1 Introduction 6

2.2 Methods and Procedures 9

2.2.1 Writers and Handwriting Samples 9

2.2.2 FLASH ID® Feature Dissimilarity Scores and Population Density Curves 10

2.2.3 Writership Survey and Forensic Document Examiners 13

2.2.4 Statistical Analyses 15

2.3 Results 17

2.4 Discussion 19

2.4.1 Acknowledgments 24

3 Wasserstein Distance Score for Handwriting Feature Measurements 28

3.1 Introduction 30
3.2 Methods 33
3.2.1 Handwriting samples 33
3.2.2 Handwriting kinematic analyses 34
3.2.3 Handwriting feature dissimilarity scores and empirical cumulative distribution functions 36
3.2.4 FDE writership opinions 39
3.2.5 Statistical analyses 42
3.3 Results 45
3.3.1 The relative importance of kinematic feature sets in predicting FDE responses 45
3.3.2 Contrasts between the single-stage and two-stage predictive models 49
3.4 Discussion 50
3.5 Conclusions 54

4 Regression Modeling for the Comparison of the Euclidean Distance Score and the Wasserstein Distance Score 58

4.1 Introduction 60
4.2 Methods 64
4.2.1 Study participants and handwriting sample collection 64
4.2.2 FLASH ID® feature dissimilarity scores 65
4.2.3 MovAlyzeR® kinematic feature dissimilarity scores 66
4.2.4 Regression models of pairwise comparisons 67
4.3 Results 69
4.4 Discussion 75
4.4.1 Data formatting and dissimilarity scores 85
4.4.2 Regression models of pairwise comparisons 86

4.4.3 Concerns about the Null-space estimation 89

5 Regression Models for Comparing Metrics 94

5.1 Setting Up the Models 95

5.2 Dependency Structures and Error Terms 98

5.3 Simulation Under the Null Hypothesis 99

5.3.1 Power Study 103

5.4 Application to Handwriting Samples 106

5.4.1 FLASH ID® Shape Codes 106

5.4.2 Handwriting Kinematics 109

5.5 Wasserstein Distance Score Development 110

5.5.1 Results 111

6 Discussion and Future Research 116

6.1 Discussion 116

6.2 Future Research 118

APPENDIX 1 121

APPENDIX 2 129

APPENDIX 3 154

REFERENCES 180

CURRICULUM VITAE 183

LIST OF FIGURES

2.1 An example of the writing sample from a single subject for a single phrase. 10

2.2 Population density distribution plots from between-writer pairs (A) and within-writer pairs (B) for the phrase “Mr. Lloyd has gone to Switzerland.” Vertical lines identify the dissimilarity scores on the X-axis, while the shaded areas represent the cumulative density scores for the sample pairs displayed above each plot. 13

2.3 A sample scoresheet from the survey. 15

3.1 Examples of population density curves for between-writer (A) and within-writer (B) pairs. For this sample, survey item no. 33, the kinematic dissimilarity score (WDS) was 0.086 and is denoted by the vertical line. Shaded areas represent the ECDF value associated with this pair from a large pool of pairs of the same phrase written in cursive. See text for further explanation. 39

4.1 Scatterplots with individual (red) and average (black) lines of best fit for cursive (top row) and handprinting (bottom row) handwriting showing the relationship between FLASH ID® dissimilarity score (y-axis) and the dissimilarity scores for spatial-geometric (left), temporal (center), and pressure (right) features for upstrokes for the phrase "Our London business is good." The red regression lines are fit using the averaged pairwise scores-one score per pair of writers, each line representing the 33 scores with one fixed writer for a total of 33 red lines. The thick black line is the average of the red lines in each plot. 72

4.2 Simple least squares p-values for testing if $\beta_1 = 0$ 88

4.3 Writer-space p-values for testing if $\beta_1 = 0$ 90

4.4 Null-space p-Values for testing if $\beta_1 = 0$ 91

5.1 Monte Carlo simulation results of simple least-squares regression performed on sets of m-dimensional multivariate Normals. 97

5.2 ECDFs of p-values in the Writer-space as k increases, resulting from Algorithm 1. 102

5.3 ECDFs of p-values in the Null-space as k increases, resulting from Algorithm 1. 103

5.4 Close-up of the lower tail of Null-space p-values CDF. Note the colors red, orange, and green are furthest from the Uniform CDF. These correspond to a lower number of features, k that make up the response score. 103

5.5 The orientation of the shape code encoding directions. A vertex going "north" would be assigned the number "3". (Image from the FLASH ID® User Manual [16].) 107

5.6 Example of three shape codes. The images show the order of the vertices (as determined by the FLASH ID®) and the number below each image is the shape code encoding. (Image from the FLASH ID® User Manual [16].) 107
5.7 Shape code 013. 108
5.8 Shape code 113. 108
5.9 Shape code 123. 108
5.10 Shape code 0123. 108

LIST OF TABLES

2.1 Mean of examiner average responses (sd of examiner average responses) scores for 36 FDEs representing strength of support for writership determinations under the same-writer and different-writer propositions for handwriting pairs written in print and cursive style. 17

2.2 ANOVA results from the regression analysis estimating FDE strength of support for the proposition that samples within a pair were written by different writers for sample pairs written by different writers. *Empirical cumulative distribution function: the cumulative value from the probability density curve that reflects the rarity of a dissimilarity score for a given phrase and writing style. 17

3.1 Handwriting stroke features used in this study and their operational definitions.[1] 35

3.2 R^2 values for full and reduced multiple regression models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting. 48

3.3 Differences in R^2 between full and three reduced models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting. 49

3.4 Correlation coefficient (R^2) for single-stage and two-stage models and their differences for cursive and print handwriting, upstrokes and downstrokes for FDE strength of support for the prosecution or same-writer hypothesis (H1) and defense or different-writers hypothesis (H2). 50

4.1 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® spatial-geometric dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes 73

4.2 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® temporal dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes 74

4.3 Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® pen pressure dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes 75

4.4 Power study results 92

5.1 Power Study results for **5** EOV in the response score. 105

5.2 Power Study results for **25** EOV in the response score. 105

5.3 Power Study results for **50** EOV in the response score. 105

5.4 Results of modeling the FLASH ID Shape Code VOS Euclidean distance using the Writer-space and Null-space transformations. The starred p-values are all less than $\alpha = 0.05$, indicating a rejection of the null hypothesis that $\beta_1 = 0$. 109

5.5 Proportion of the slope p-values < 0.05 from the regression models predicting pairs of MovAlyzeR® feature WDS for cursive writing and handprinting sample pairs for upstrokes and downstrokes in the Writer-space and Null-space. Note that the Downstrokes have an $n = 28$ and the Upstrokes have an $n = 45$. 113

LIST OF ALGORITHMS

1 Monte Carlo Algorithm for Testing the Null Hypothesis - Increasing Features 101

2 Power Study 104

3 Wasserstein Distance Score Development 111

4 Monte Carlo algorithm for testing the null hypothesis using the WDS 114

5 Iterative Estimation of Eigenvalues 120

ABSTRACT

A CLASS OF REGRESSION MODELS FOR PAIRWISE COMPARISONS OF FORENSIC HANDWRITING COMPARISON SYSTEMS

CAMI M. FUGLSBY

2023

Handwriting analysis is a complex field largely living in forensic science and the legal realm. One task of a forensic document examiner (FDE) may be to determine the writer(s) of handwritten documents. Automated identification systems (AIS) were built to aid FDEs in their examinations. Part of the uses of these AIS (such as FISH [5] [7], WANDA [6], CEDAR-FOX [17], and FLASH ID®[2]) are to measure features about a handwriting sample and to provide the user with a numeric value of the evidence. These systems use their own algorithms and definitions of features to quantify the writing and can be considered a black-box. The outputs of two AIS are used to compare to the results of a survey of FDE writership opinions.

In this dissertation I will be focusing on the development of a response surface that characterizes the feature outputs of AIS outputs. Using a set of handwriting samples, a pairwise metric, or scoring method, is applied to each of the individual features provided by the AIS to produce sets of pairwise scores. The pairwise scores lead to a degenerate U-statistic. We use a generalized least squares method to test the null hypothesis that there is no relationship between two metrics ($\beta_1 = 0$.) Monte Carlo simulations are developed and ran to ensure the results, considering the structure of the pairwise metric, behave under the null hypothesis,

[2] `http://sciometrics.com/flashid.html` (Accessed August 6, 2020).

and to ensure the modeling will catch a relationship under the alternative hypothesis. The outcome of the significance tests helps to determine which of the metrics are related to each other.

CHAPTER 1

Introduction and Overview

The complexity of handwriting and automated identification systems (AIS) makes for interesting statistical problems. The use of AIS introduces a level of complexity as the algorithms employed behind the scenes are often unknown to the user. Before an AIS is deployed, it must go through testing. This dissertation will address the development of evaluation techniques that coincide with biometric verification techniques, leading to developing methods for metric comparison, even when the metrics are used on the outputs of different AIS.

Most AIS were developed to address a specific class of questions. For example, the FLASH ID® system is designed to be a closed-set identification system. Closed-set in the sense that it does not account for the potential of another source outside of the given set of writers; and Identification system meaning it outputs a list of sources (writers) who are ranked as most likely to have produced (written) the query object [15].

Throughout this dissertation I refer to the FLASH ID® system and the MovAlyzeR® software as opaque and transparent, respectively. I am using the following definitions,

Opaque (Black-Box): "... a system that does not reveal its internal mechanisms. In machine learning, "black box" describes models that cannot be understood by looking at their parameters..." [11]

Transparent (White-Box): "...methods and models that make the behavior and predictions of machine learning systems understandable to humans." [11]

Biometric identification systems such as FLASH ID® require verification. For biometric systems, this testing is one-to-one, or pairwise, and involves the use of a scoring method to produce a univariate score between two samples that represents either similarity or dissimilarity between the two samples the direction is chosen based off of the use of the score, however it is simple to obtain one from the other.) This style of testing lends to finding creative ways to use the AIS output for one-to-one testing [15].

The one-to-one style of testing biometric algorithms has a natural U-statistic structure of degree two. This dissertation will cover the univariate U-statistics, multivariate U-statistics, and U-processes that arise from a variety of outputs of AIS.

1.1

Contributions and Chapter Summaries

This dissertation will cover three published papers covering the interpretation of AIS system outputs. These papers focus on pairwise metric development, using the output of the biometric identification systems to behave as biometric verification systems, how to use the metrics to gain insight on a survey of FDE writership opinions, and how to use the metrics to infer relationships between different metrics or on different features that use the same metric. This block of research was completed with the same set of handwriting samples, a set of 33 writers who all wrote the same six phrases from the London Letter, each phrase repeated five times. The writers wrote on a piece of paper with a pen, the paper placed on top of a tablet with the MovAlyzeR® software[1] installed and used to measure kinematic pen movements. The physical copies of the writing samples were used in the FLASH ID® system, and the MovAlyzeR® software collected the kinematics of

[1] `https://neuroscript.net/movalyzer.php` (Initially accessed August 6th, 2020.)

the writers' handwriting.

The first paper, published in the Journal of Forensic Sciences in 2020 discusses the use of the Euclidean distance metric on pairs of Vectors of Scores (VOS). These scores are then compared to the results of a survey of FDEs to explore a potential relationship between the Euclidean distance scores and the strength of support provided by the FDEs.

The second paper, published in Forensic Science International in 2021 introduces the Wasserstein distance score (WDS) applied to the output of the MovAlyzeR® software. The set of WDS were then compared to the same survey results of FDEs to explore potential relationships between the different kinematic feature WDS and the strength of support provided by the FDEs.

The third paper, published in the Journal of Forensic Sciences in 2021 tests for relationships between the WDS and the VOS metrics. Specifically, the sets of kinematic features from the MovAlyzeR® software were compared to the FLASH ID® VOS to determine if the FLASH ID® scores are correlated with spatial-geometric, temporal, or pen pressure measurements.

Chapters 2, 3, and 4 will contain specific extended introductions. The rest of this dissertation will cover the current work on comparing the output of AIS systems and the efforts to extend the research to a more generalized version.

My contributions to this research has largely been on the development, implementation, and interpretation of the metrics and algorithms used. Specific contributions are noted throughout.

CHAPTER 2

Dissimilarity Scores from an Automated System and Forensic Document Examiners

The following paper by Fuglsby et al. [9] in the Journal of Forensic Sciences is rewritten to reflect the current notational conventions. Fuglsby et al. introduces the Euclidean distance metric used on the VOS output of the FLASH ID® system to compare to a survey of handwriting examiners. This metric is a pairwise dissimilarity score, meaning that the larger the score, the more dissimilar the pair. My main contributions to this line of research was on calculating the Euclidean distance scores from the output of the FLASH ID® system, and on the writing of the paper.

Applying the Euclidean distance to the FLASH ID® VOS output is an example of turning an adaptable one-to-many score output into a pairwise score that a biometric verification algorithm would use. The intended use of the FLASH ID® system is not to make one-to-one comparisons, and so the Euclidean distance calculated on the resulting VOS allows for pairwise comparisons between handwritten documents using this system. Note that using the Euclidean distance requires the two VOS to contain the same number of elements-of-a-vector (EOV). While this requirement is easily met for this use of the FLASH ID® system, the output of other AIS do not have to follow this restraint.

Use of an Automated System to Evaluate Feature Dissimilarities in Handwriting Under a Two-Stage Evaluative Process

Abstract

The two-stage evaluative process is an established framework utilized by forensic document examiners (FDEs) for reaching a conclusion about the source(s) of handwritten evidence. In the second, or discrimination, stage, the examiner attempts to estimate the rarity of observations in a relevant background population. Unfortunately, control samples from a relevant background population are often unavailable, leaving the FDE to reach this determination based on subjective experience. Automated handwriting feature recognition systems are capable of performing both feature comparison and discrimination, yet these systems have not been subjected to empirical validation studies. In the present study, we repurposed a commercially available automated system to generate empirical distributions for ranking feature dissimilarity scores among pairs of handwritten phrases. The blinded results of this automated process were used to survey an international cohort of 36 FDEs regarding their strength of support for same- and different-writer propositions. The survey served to cross-validate FDE decision-making under the two-stage approach. Results from the survey demonstrated a clear pattern of response consistent with ground truth. Predictive regression analyses indicated that the automated feature dissimilarity scores and the log of their cumulative distribution functions accounted for 72% of the variability in FDE opinions. This study demonstrated that feature dissimilarity scores acquired using automated processes and their distributions are closely aligned with FDE decision-making processes

Rewritten for notational clarity from the Journal of Forensic Sciences, 57, C. Fuglsby, C. Saunders, D. M. Ommen, and M. P. Caligiuri, Use of an Automated System to Evaluate Feature Dissimilarities in Handwriting Under a Two-Stage Evaluative Process, pages 2080-2086, supporting the heuristic value of the two-stage evaluative framework.

2.1

Introduction

A common approach to evidence interpretation in handwriting examination involves two stages [1-5]. The first stage in this process is described as the match or comparison stage and relies upon the examiners' observational skills to determine whether characteristics or features of the two sets of suspect evidence are indistinguishable. A set of suspect samples are deemed distinguishable if they share few features in common and have a number of discriminating elements. A discriminating element is "a relatively discrete characteristic or feature of writing that varies observably or measurably across writers and may contribute reliably to distinguishing between samples from different individuals, or conversely, support the contention of sameness within a common writer" [6]. If the suspect samples are deemed distinguishable, the evidence suggests that they share two different sources. Alternatively, if the suspect samples are deemed indistinguishable, they share many features and characteristics between them, and few (or no) discriminating elements are observed. The term indistinguishable does not indicate that the two samples do share a common source, but only describes the characteristics of the two samples with respect to the proposition that they share a common but unknown source. If the suspect samples are deemed indistinguishable, the examination proceeds to the second stage, described as the discrimination or significance stage; if two sets of evidence are considered indistinguishable from one another, the examiner attempts to estimate the rarity of the observed characteristics in a relevant background population. In a slightly more formal sense, the second stage describes the likelihood of a chance match (as described in Found and Bird [7]), or random match probability. Thus, under the two-stage approach, the probative value of finding that the suspect samples are indistinguishable is strengthened by the findings that shared combinations of features and characteristics between two individuals are extremely uncommon among members of a relevant population.

Prior research offers broad support for document examiners' proficiency for reaching accurate writership decisions based on handwriting feature analyses in closed-set laboratory experiments [8-13]. However, research on the impact of population-level information to the FDE evaluative process under a two-stage framework is lacking for several reasons. With few exceptions [14,15], the population distribution of specific combinations of handwriting features and characteristics is currently insufficient to assist the examiner in meeting the goals of the discrimination stage. Because population-level frequency distributions for hand-writing characteristics relevant to a particular case are usually unavailable to assist the examiner in this discrimination stage, in cases where samples are distinguishable, writership conclusions under the two-stage approach are largely inferential. This introduces a critical challenge to admissibility of evidence involving handwriting often prompting Daubert hearings [16-18]. As in many forensic pattern-matching disciplines, handwriting is high-dimensional and complex. Handwriting is comprised of a sequence of individual movements each with multiple temporal (e.g., movement duration and speed), spatial (e.g., horizontal and vertical size), and geometric (e.g., slant or loop area) attributes that distinguish one writing segment from another to convey meaning. The complexity of handwriting is borne out by the interactions among these attributes driven by context variability, physical constraints, and individual's natural variation. Taken together, estimating the prevalence of specific feature differences within a population of writers can be a herculean undertaking. Even if such data were publicly available to examiners, statistical procedures are needed to quantify the atypicality of these feature differences in the population distribution.

Automated feature extraction programs such as FLASH ID® (Sciometrics LLC, Chantilly, VA, USA) have advantages over their human counterparts particularly with respect to estimating the likelihood that a questioned handwriting sample came from a candidate residing within a population reference set. Moreover, with careful reprogramming and armed with a large database of features and feature differences, automated systems can be repurposed to generate population distribution functions for a numeric estimation of the rarity of feature dissimilarity scores from within a large reference set. In this way, the automated system would inform the discrimination stage of the two-stage process.

Two experiments were conducted for the present study. The purpose of the first experiment was to deploy an automated feature extraction program to generate feature dissimilarity scores and population distribution functions for ranking these feature dissimilarity scores among pairs of handwritten phrases across different phrases and styles of handwriting. We designed a specialized algorithm that customizes the output of an automated feature extraction program designed for closed-set identification [19] to first calculate feature dissimilarities between 81,180 sample pairs of print and cursive handwriting from known between- and within-writer sources. The terms between-writer and within-writer also refer to different and common source, respectively. The latter terms are more general in application and can refer to source specimens other than handwriting. Population-level distribution functions were then created from the dissimilarity scores for each unique phrase. In this way, feature dissimilarity scores are ranked according to their placement within the dissimilarity score population. Sample pairs with dissimilarity scores in the tails of the distribution would be considered "rare" for a relevant population of samples.

The aim of the second experiment was to utilize these dissimilarity scores and distribution functions to design a series of difficult-case scenarios for FDEs to evaluate. Sample pairs falling along the tails of their respective population distributions were submitted to an international cohort of FDEs to demonstrate the utility of an automated feature-based process within the two-stage evaluative framework. The second experiment served as a cross-validation of FDE decision-making under the two-stage approach.

2.2

Methods and Procedures

2.2.1

Writers and Handwriting Samples

The study recruited 33 individuals from the San Diego Sheriff's Crime Laboratory who were asked to write six phrases from the London Letter and to repeat each phrase five times using both print and cursive writing styles. This provided 60 phrases per individual writer. The six phrases from the London Letter were as follows: (i) Our London business is good; (ii) but Vienna and Berlin are quiet; (iii) Mr. Lloyd has gone to Switzerland; (iv) and I hope for good news; (v) He will be there for a week; and (vi) and then goes to Turin. Subjects wrote each of the phrases five times with an inking pen on lined paper placed on a Wacom (Intuos Pro, model PTH-660) digitizing tablet. The stimulus phrase was shown on the top of each page, and repetitions were written vertically, five per page. Seven subjects returned to the laboratory two weeks later and repeated the writing experiment. Figure 2.1 shows a page layout and writing sample from a single subject for a single phrase.

While the digitized samples were subjected to analyses of kinematic features to be used in the predictive modeling component of this research (not included here), the ink copies were used in the present study. Each page of the hard copy ink samples was scanned at 600 dpi, cropped into individual repetitions, and saved as separate 16-bit TIFF files for automated analyses. With 33 writers, each writing six phrases five times each in both cursive and print, there was a total of 13,530 sample pairs available per phrase and writing style that were used to calculate phrase-dependent population-level dissimilarity scores and their respective dissimilarity functions.

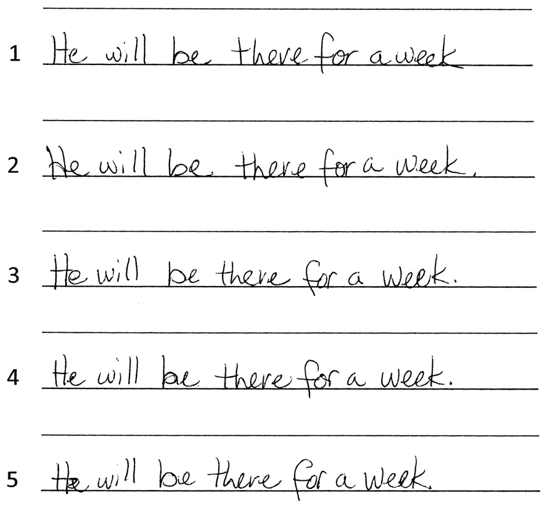

1 He will be there for a week
2 He will be there for a week.
3 He will be there for a week.
4 He will be there for a week.
5 He will be there for a week.

Figure 2.1: An example of the writing sample from a single subject for a single phrase.

2.2.2

FLASH ID$^{®}$ Feature Dissimilarity Scores and Population Density Curves

Due to the practical constraints of examining thousands of pairs of writing samples to identify cases of interest for inclusion in an FDE survey, it is desirable to use the help of an automated system to identify these pairs. All paper samples obtained during the collection process were scanned into FLASH ID$^{®}$ for the purpose of obtaining a univariate score for each pair of samples representing the level of dissimilarity between the writing contained in the samples (the higher the score is, the more dissimilar the two samples are).

FLASH ID®, an automated feature extraction program, generally serves the purpose of closed-set biometric identification [19]. FLASH ID® provides a ranking (with respect to a reference set of writing samples from 50 known writers) of candidate writers based on similarities between combinations of features from a single questioned sample and the reference set of known writing samples. For the present study, we leveraged this capability to develop a univariate omnibus dissimilarity score for comparing features from two questioned handwriting samples. Our procedure involved obtaining the Euclidean distance between the two vectors of scores (from the FLASH ID® output) to provide a univariate score reflecting the feature distance between two samples. This univariate score thus represents the dissimilarity in features between all possible pairs of samples with larger scores reflecting greater feature dissimilarity. The output of this customized algorithm consisted of dissimilarity scores for all possible pairs across the six phrases, five repeats, and two writing styles for 33 writers, leading to a total of 81,180 possible pairs for each writing style. We then derived the distributions for each phrase and writing style from the available sample pool. The cumulative density score or function served as an index of the rarity of the dissimilarity score within the relevant population of dissimilarity scores and is referred to as the empirical cumulative distribution function or ECDF.

The statistical programming language R [20] was used to compute all pairwise comparison scores for all possible between- and within-writer sample pairs and population distributions. Twenty-four separate population distributions (six phrases × two styles × two writership sources) were calculated. Forty difficult-case scenarios were then identified using the scores, ordered from largest to smallest, from the 24 distributions and included in the FDE survey.

Two types of difficult cases were included in the survey: (1) pairs that were written by the same writer, but were associated with high dissimilarity scores, and (2) pairs that were written by different writers but were associated with low dissimilarity scores. The first set of difficult cases are characterized by an unusually large dissimilarity score compared to other dissimilarity scores from within-writer pairs. The second set of difficult cases are characterized by an unusually small dissimilarity score compared to other dissimilarity scores from between-writer pairs. Thirty between-writer pairs and ten within-writer pairs were selected from this larger pool for inclusion in the survey to increase the difficulty of the survey and challenge the examiners. Figure 2.2 shows examples of population density curves for between-writer and within-writer pairs, respectively, for cursive handwriting along with their corresponding FLASH ID® dissimilarity and ECDF scores (shaded area). In the examples shown in Fig. 2.2, the dissimilarity score and ECDF value for the between-writer sample (A) were 1.83 and 0.00008, respectively. The dissimilarity score and ECDF value for the within-writer sample (B) were 5.91 and 0.96, respectively. Sample pairs for the survey were selected to represent uncommon feature dissimilarities for their respective sources. That is, low dissimilarity scores are unusual for between-writer samples, whereas higher dissimilarity scores are unusual for within-writer samples. Forty such pairs having FLASH ID® dissimilarity scores residing near the tails of their respective distributions were used in the FDE survey.

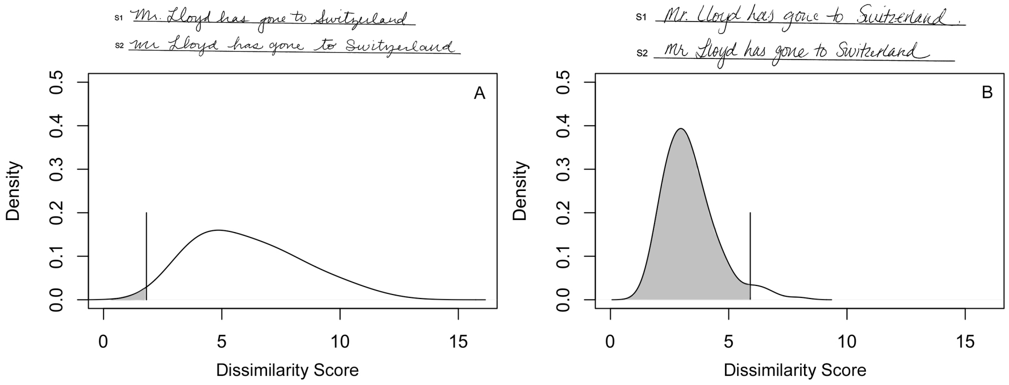


Figure 2.2: Population density distribution plots from between-writer pairs (A) and within-writer pairs (B) for the phrase "Mr. Lloyd has gone to Switzerland." Vertical lines identify the dissimilarity scores on the X-axis, while the shaded areas represent the cumulative density scores for the sample pairs displayed above each plot.

2.2.3

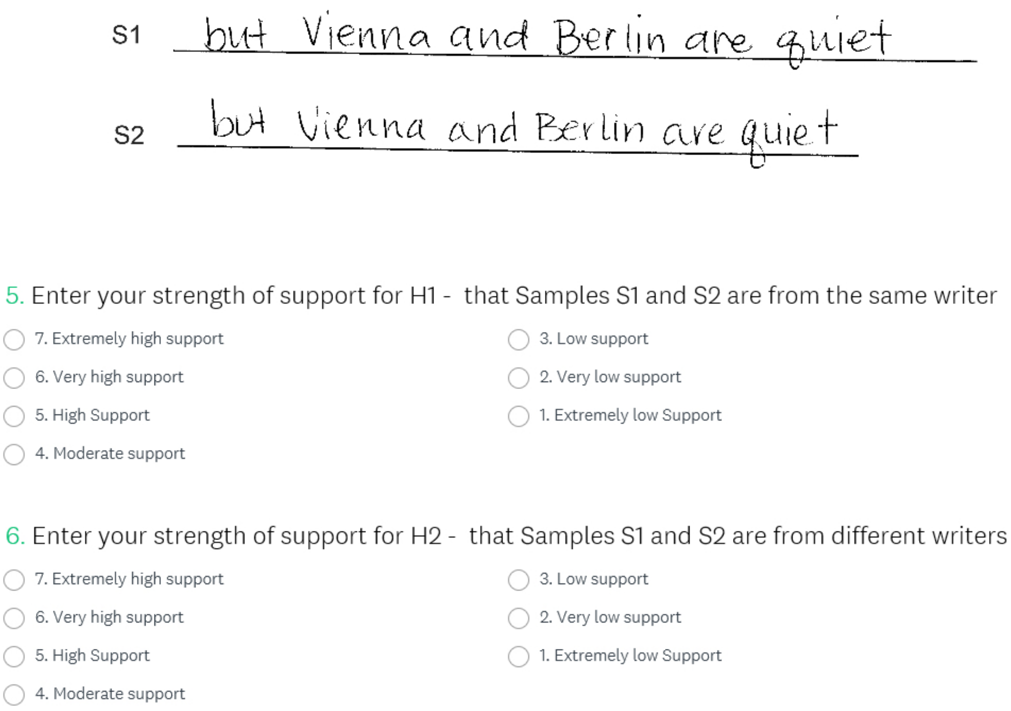
Writership Survey and Forensic Document Examiners

We designed a writership survey consisting of difficult-case scenarios to obtain FDE strength of support for same-writer and different-writer propositions. For the purpose of this study, we considered two difficult-case scenarios: (1) when the probability of observing a small dissimilarity score between two samples of handwriting from unknown sources drawn from a relevant population is low for samples from different writers and (2) probability of observing a large dissimilarity score between two samples of handwriting from unknown sources drawn from a relevant population is low for samples from the same writer.

The survey consisted of 40 sample pairs: 20 print pairs (15 between-writer and five within-writer) and 20 cursive pairs (15 between-writer and five within-writer). Pairs were presented in the survey in random order, and examiners were blinded to the writer source(s) for each pair. Five of the 40 pairs were repeated in the survey with sample order reversed for the purpose of testing examiner repeatability. Each survey item required examiners to score their strength of support for two propositions. Proposition 1 (H1) pertained to the samples being written by the same writer. Proposition 2 (H2) pertained to the samples being written by different writers. Examiners indicated their strength of support using a 7-point scale rating from extremely strong support "7" to extremely low support "1." An example of a survey pair with the scoresheet is shown in Fig. 2.3. For each respondent to the survey, there were 90 strength-of-support scores available for analysis (two from each of 40 sample pairs and two from each of the five repeated pairs).

Email requests were sent to 60 FDEs from North America, Europe, and Australia or New Zealand to participate in a writership survey. Of the 60, 41 FDEs submitted responses to the survey (68.3% response rate). Six were from North America, nine from Australia/New Zealand, and 26 from European countries. In addition to writership judgments, the FDEs provided de-identified information about their experience and work environment to the study. Of the 41 examiners participating in the survey, 37 (90.2%) worked in government laboratories; 33 (80.5%) reported that at least 75% of their casework involved handwriting; and 31 (75.6%) reported having been in practice for at least 10 years. This component of the research was reviewed and approved by the University of California San Diego Institutional Review Board. The average time to complete the 90-item survey was 66 minutes.

Five randomly selected survey items were duplicated to test FDE repeatability. Absolute difference scores between FDE strength of support for duplicated sample pairs were calculated for each of the 41 FDEs. The distribution of the average absolute differences from all 41 FDEs revealed a bimodal distribution with a cut-point located at 1.5. Based on this profile, we considered scores of 1.5 or larger to reflect inconsistent performance across repeated items of the survey. Five FDEs had scores of 1.5 or larger and were therefore excluded from further analyses. There were no differences in demographic characteristics between the five excluded examiners and the remaining 36 examiners.



S1 but Vienna and Berlin are quiet

S2 but Vienna and Berlin are quiet

5. Enter your strength of support for H1 - that Samples S1 and S2 are from the same writer

○ 7. Extremely high support
○ 6. Very high support
○ 5. High Support
○ 4. Moderate support
○ 3. Low support
○ 2. Very low support
○ 1. Extremely low Support

6. Enter your strength of support for H2 - that Samples S1 and S2 are from different writers

○ 7. Extremely high support
○ 6. Very high support
○ 5. High Support
○ 4. Moderate support
○ 3. Low support
○ 2. Very low support
○ 1. Extremely low Support

Figure 2.3: A sample scoresheet from the survey.

2.2.4

Statistical Analyses

In order to explore patterns in FDE response, the survey included both cursive and print sample pairs written by the same writer and sample pairs written by different writers. For 36 FDEs, we collected two strength-of-support scores from each survey item: one registering support for the same-writer proposition and one registering support for the different-writer proposition. The scores were averaged to create a summary statistic for each examiner. We then calculated a sample mean and sample standard deviation of the summary statistics across the examiners. These steps were followed for the 20 cursive and 20 print pairings, each consisting of five within-writer and 15 between-writer pairs. Paired t-tests were conducted to test significance of a difference in examiner average strength-of-support scores arising from known same-writer versus different-writer survey items (for both print and cursive items) for each proposition. We found the distributions to be non-normal; however, due to the boundedness of the sample space of the observations, and for the sample sizes we have for this experiment, the paired t-test is robust to departures from normality.

To examine whether FDE responses were linked in any way to the dissimilarity and probability scores derived from the automated system, we used multiple linear regression models. The regression models were tested for estimating FDE strength-of-support scores for the same-writer (H1) and different-writer (H2) proposition separately for each style of handwriting. Due to the small number of within-writer samples in the survey, models were run using only the between-writer sample pairs (15 pairs for each writing style). Two explanatory variables were tested in each of the four models: the FLASH ID® dissimilarity score and the log of the between-writer ECDF.

We hypothesized that examiners would register stronger support for the common source proposition when the sample pairs came from the same writer compared to different writers and register stronger support for the different source proposition when the sample pairs came from different writers compared to the same writer. We also hypothesized that stronger FDE support for a given proposition would be associated with more extreme dissimilarity scores and lower ECDFs within the distribution corresponding to the considered proposition. Support for these hypotheses would demonstrate that a feature-based automated system could be deployed to perform a two-stage evaluation of handwriting evidence.

	Same-writer Proposition		Different-writer Proposition	
Source	Cursive	Print	Cursive	Print
Same writer	4.25 (0.86)	3.89 (0.73)	2.91 (0.68)	3.14 (0.66)
Different writers	3.17 (0.68)	3.16 (0.64)	3.90 (0.64)	3.78 (0.64)
Difference	1.08	0.73	-1.00	-0.64
Paired T-statistic	9.34	7.73	-10.85	-8.06
p-value	<0.0001	<0.0001	<0.0001	<0.0001

Table 2.1: Mean of examiner average responses (sd of examiner average responses) scores for 36 FDEs representing strength of support for writership determinations under the same-writer and different-writer propositions for handwriting pairs written in print and cursive style.

	df	SS	MS	F	p-value
Dissimilarity score	1	0.32	0.32	8.51	0.01
Log ECDF*	1	0.85	0.85	22.66	<0.001
Residuals	12	0.45	0.04		

Table 2.2: ANOVA results from the regression analysis estimating FDE strength of support for the proposition that samples within a pair were written by different writers for sample pairs written by different writers.
*Empirical cumulative distribution function: the cumulative value from the probability density curve that reflects the rarity of a dissimilarity score for a given phrase and writing style.

2.3

Results

Table 2.1 shows the means of examiner averages (with standard deviations) for FDE strength of support for the same-writer and different-writer propositions when presented with samples written by the same writer or different writers. When asked to express strength of support for the same-writer proposition, examiners expressed significantly stronger support when the sample pair came from the same writer than from different writers. When asked to express strength of support for the different-writer proposition, examiners expressed significantly stronger support when the sample pair came from the different writers than from the same writer. These patterns held for both print and cursive handwriting. Analogous results were observed from using a nonparametric Wilcoxon signed rank-sum test on the medians.

Results from the multiple linear regression models were statistically significant only for FDE responses to the proposition that the samples came from different writers (H2) when presented with samples from different writers (i.e., FDE strength of support for ground truth) and only for printed samples ($F_{2,12} = 15.58$; $p\text{-}value < 0.001$; R^2 = 0.72). Table 2.2 shows the results of the regression analysis for known different-writer samples. This two-factor model yielded estimated coefficients (±SE) of 4.69 (0.86) and -1.16 (0.24) for the dissimilarity score and log ECDF for the between-writer distribution, respectively. The predictive model indicates that larger dissimilarity scores and lower log ECDF scores predict stronger support for the proposition that two samples with unknown writer sources likely came from different writers.

Because selection of pairings included in the survey was based on dissimilarity and ECDF scores and not writer, the possibility existed that samples from one writer (albeit a different trial for a given phrase) might be used more than once when paired with samples from another writer. To reduce the effect of this sampling bias on the statistical results, we considered incorporating the two writers of each pair as random effects. To the best of our knowledge, there is no natural way to incorporate these effects within a standard random-effects model. Therefore, we fit two different random-effects models: one that treats each writer as a block effect and a second that treats each pair of writers as a distinct random effect. Both models effectively gave the same results for tests concerning the fixed effects with the magnitude of the T-statistics ranging from 5.1 to 6.7.

The between-writer model for cursive sample pairs written by different writers was not statistically significant ($F_{2,12}$ = 0.36; $p\text{-}value > 0.10$; R^2 = 0.06). Models estimating FDE support for the same- or different-writer propositions when presented with samples from within-writer had only five pairs per writing style were not considered due to insufficient statistical power.

2.4

Discussion

The present study had two main objectives. The first objective was to repurpose an automated handwriting feature extraction program to yield output scores that parallel the two-stage evaluative process consisting of a dissimilarity stage and discrimination stage. To accomplish this, we leveraged the powerful capabilities of FLASH ID® to develop a univariate omnibus feature dissimilarity score for comparing two questioned handwriting samples. Feature dissimilarity scores were calculated for all possible within- and between-writer pairings, producing 81,180 possible pairs each for print and cursive handwriting. This large pool of dissimilarity scores was used to generate densities and cumulative distribution functions for each phrase and style of handwriting. This enabled the reliable assignment of how rare a dissimilarity score from any handwriting pair was when compared against a population of dissimilarity scores. To our knowledge, this is the first study capable of quantifying the rarity of an observed difference in handwriting features between two samples within a population of writers for the explicit purpose of validating the significance stage within the two-stage framework.

The second objective of the study was to utilize these distribution functions to design a series of difficult-case scenarios for FDEs to evaluate. Sample pairs falling along the tails of their respective population distributions were submitted to an international cohort of 36 FDEs to demonstrate the utility of an automated feature-based process within the two-stage evaluative framework. The results from the survey supported our hypotheses that average examiner strength of support for a given proposition between pairs of samples of different writership (between or within) are consistent with ground truth for difficult-case scenarios. Further results from multiple regression analyses indicated that the feature dissimilarity score and the log ECDF for the between-writer distribution combine to account for 72% of the variability in FDE strength of support for the proposition that two questioned handwriting samples came from different writers. Specifically, these results indicate that larger dissimilarity scores and lower log ECDF scores for the between-writer distribution predict stronger support for the proposition that two samples with unknown writer sources likely came from different writers.

The findings of the present study inform the ongoing controversy over evidence interpretation within the forensic science community. One of the attributes of the two-stage approach is that it allows examiners to reject the proposition that two suspect samples arose from the same source without necessarily supporting the proposition that the two suspect samples share the same source. This allows an examiner to make statements along the line of "given the observed degree of similarities and dissimilarities, I am comfortable concluding that these two suspect samples do not share a common but unknown source." Depending on a number of factors, this attribute can be considered either a strength or a weakness of this approach. To address this perceived limitation of the two-stage approach, the most commonly used alternative is a likelihood ratio-based approach. A likelihood ratio-based approach evaluates the subjective likelihood of the two suspect samples given the first proposition relative to the likelihood of the two suspect samples given the second proposition [21]. The likelihood ratio-based approaches require a well-specified background population and evidence concerning how samples arise from that population. Conversely, if the results of the examination are determined solely within the first stage, then a strength of the two-stage approach is that there is no need for a specified background population. This means fewer evidential resources are needed to provide useful information to the decision-makers in this scenario. When failing to conclude that the two suspect samples do not share a common source, the two-stage approach requires the use of evidence from a specified background population to assess the evidential support for the proposition that the suspect samples do share a common source. For an overview on the comparison of likelihood-based and two-stage approaches, see [22]; for a discussion as it applies to handwriting examination, see [21]; and for a rigorous statistical discussion on two-stage approaches for Bayesian model selection as it pertains to specific-source propositions, see [23].

The present study contributes to the body of research on handwriting evidence interpretation in two ways. First, we successfully repurposed FLASH ID® from a feature extraction program designed to estimate the likelihood that a questioned handwriting sample was written by each writer in a list of candidate writers residing in a reference database to population distributions for a precise numeric estimation of the rarity of feature dissimilarities. In this way, the output of the automated system characterized handwriting feature dissimilarities and their distributions rather than writer identification. This is an important step toward automating the two-stage process, particularly with respect to the discrimination stage. Conventionally, an FDE might observe similarities among questioned documents and conclude that they were written by a single individual. Such an opinion would be strengthened based on the classical premise and estimating the rarity combinations of features and characteristics or their variability, leading to the conclusion that the chance of observing this combination of features and characteristics from samples in the background population is extremely low. Unfortunately, examiners reach conclusions about the relevant population based on experience alone with little or no support from actual prevalence data. With few exceptions [8,9], estimates of population variance in handwriting features or estimates of the dissimilarity in features between two writers are unavailable.

Moreover, there is growing support cautioning against reaching a conclusion based on handwriting evidence that identifies an individual writer [7]. In its current form, FLASH ID® outputs a ranking of the reference set of candidate writers based on feature extraction. Our approach generates a score of the rarity of feature dissimilarities without any identification inference.

The second contribution of this study to the body of research on handwriting evidence interpretation stems from the results of the FDE survey. While it was not possible to know whether FDEs applied a two-stage or likelihood approach when reaching decisions in support for or against a particular proposition, their scores and decision patterns were consistent with what would be expected in a difficult-case scenario. For example, in the discrimination stage of the two-stage evaluative process, if there are many similarities in the writing pairs, then that pair may be considered typical of the within-writer population and one would expect the FDE to respond with strong to extremely strong support for the same-writer proposition. On the other hand, if the opposite were true, with many differences in the writing pairs, then that pair may be considered typical of the between-writer population and one would expect the FDE to respond in strong to extremely strong support for the different-writer proposition. Neither of these outcomes was evident from the survey results. However, if there are both similarities and differences in the writing pair, then that pair is somewhere in the overlap of the two populations and you would expect the FDE to respond with only moderate support of the propositions. Indeed, this was the most frequent response pattern observed from the survey results. The average strength of support for the same-writer proposition when presented with cursive sample pairs from the same writer was 4.25 (3.89 for print), while the average strength of support for the different-writer proposition when presented with cursive sample pairs from the different writers was 3.90 (3.78 for print). This is not surprising considering the survey was designed to include item pairs drawn from the tails of their respective population distributions where the magnitude of the feature dissimilarity would be uncommon considering the ground truth of that pair.

The value of an automated feature-based program is under-scored further by the results from our predictive regression analyses. Two feature-based scores derived from the repurposed FLASH ID® program representing each stage of the two-stage. process combined to account for 57% of the variability in FDE strength of support for the proposition that a pair of handprinted samples were produced by different writers are a blinded test of ground truth. While these findings are limited to handprinting, they support our overall hypothesis that a feature-based automated system could be deployed to perform a two-stage evaluation of handwriting evidence.

It is important to recognize that the results from the FDE survey cannot be used to evaluate proficiency. The questions posed to the examiners focused on strength of support or confidence that a specific writership proposition is correct. These scores cannot be converted to proficiency with respect to ground truth. A further limitation of this study is that it is unclear which interpretation paradigm the examiner used (e.g., likelihood-based or two-stage). We observed a number of examiners whose measures of support for the prosecution hypothesis were near perfectly negatively correlated with their measures of support for the defense hypothesis, as well as a number of examiners whose measures of support were minimally or positively correlated. This indicates that different examiners are using different methods of evidence interpretation. A natural extension of this research is to focus on alternative survey designs to shed light on this issue.

In conclusion, the present study demonstrated that with careful reprogramming and armed with a large database of features and feature differences, automated systems can be repurposed to generate population density curves for a numeric estimation of the rarity of feature dissimilarity scores from within a large reference set. In this way, the automated system would inform the discrimination stage of the two-stage framework and support FDE examination process.

2.4.1

Acknowledgments

The authors acknowledge Dr. Linton Mohammed, Dr. Carolyne Bird, and Jonathan Morris for their contributions in the design and piloting of the survey and in recruiting volunteers who participated in the survey and Brenda Lanners and Gina Hunter for their assistance in collecting the handwriting samples. We wish to thank the volunteers who participated providing the handwriting samples or participated in the survey.

References

1. Kirk PL, Thornton JI. Crime investigation, 2nd edn. New York, NY: John Wiley and Sons Ltd., 1974;9-17.

2. Parker J. A statistical treatment of identification problems. J Forensic Sci Soc 1966;6:33-9.

3. Parker J. The mathematical evaluation of numerical evidence. J Forensic Sci Soc 1967;7(3):134-44.

4. Parker J, Holford A. Optimum test statistics with particular reference to a forensic science problem. J R Stat Soc Ser C Appl Stat 1968;17(3):237-51. https://doi.org/10.2307/2985461.

5. Evett IW, Berger CEH, Buckleton JS, Champod C, Jackson G. Finding the way forward for forensic science in the US - a commentary on the PCAST report. Forensic Sci Int 2017;273:16-23. https://doi.org/10.1016/j.forsciint.2017.06.018.

6. Huber RA, Headrick AM. Handwriting identification: facts and fundamentals. Boca Raton, FL: CRC Press LLC., 1999;33-59.

7. Found B, Bird C. The modular forensic handwriting method. J Forensic Doc Exam 2016;26:7-83. https://doi.org/10.31974/jfde26-7-83.

8. Kam M, Wetstein J, Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39(1):5-14. https://doi.org/10.1520/JFS13565J.

9. Kam M, Gummadidala Fielding G, Conn R. Signature authentication by forensic document examiners. J Forensic Sci 2001;46(4):884-8.

10. Kam M, Lin E. Writer identification using hand-printed and non-handprinted questioned documents. J Forensic Sci 2003;48:1391-5. https://doi.org/10.1520/JFS15062J.

11. Galbraith O, Galbraith CS, Galbraith NG. The principle of the "Drunkards" search as a proxy for scientific analysis: the misuse of handwriting test data in a law journal article. Int J Forensic Doc Exam 1995;1:7-17.

12. Sita J, Found B, Rogers DK. Forensic handwriting examiners' expertise for signature comparison. J Forensic Sci 2002;47(5):1117-24. https://doi.org/10.1520/JFS15521J.

13. Bird C, Found B, Rogers D. Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors. J Forensic

Sci 2010;55(5):1291-5.
https://doi.org/10.1111/j.1556-4029.2010.01456.x.

14. Johnson ME, Vastrick TW, Schuetzner E. Measuring the frequency of occurrence of handwriting and handprinting characteristics.
J Forensic Sci 2017;62(1):142-63. https://doi.org/10.1111/1556-4029.13248.

15. Vastrick TW, Schuetzner E, Osborn K. Measuring the frequency occurrence of handwritten numeral characteristics. J Forensic Sci 2018;63(4):1215-20. https://doi.org/10.1111/1556-4029.13678.

16. U.S. v Lewis, 220 F. Supp. 2d 548 (S.D. W. Va. 2002).

17. U.S. v Johnsted, 30 F. Supp. 3d 814 (W.D. Wis. 2003).

18. U.S. v Saelee, 549 U.S. 1147, 127 S. Ct. 1016 L. Ed. 2d 766 (2007).

19. Miller JJ, Patterson RB, Gantz DT, Saunders CP, Walch MA, Buscaglia J. A set of handwriting features for use in automated writer identification. J Forensic Sci 2017;62(3):722-34.
https://doi.org/10.1111/1556-4029.13345

20. Core R, Team R. a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.

21. Expert Working Group for Human Factors in Handwriting Examination. Forensic handwriting examination and human factors: improving the practice through a systems approach. NISTIR 8282. Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology, 2020.
https://doi.org/10.6028/NIST.IR.8282.

22. The Statistical and Applied Mathematical Sciences Institute. Statistics and applied mathematical science aspects of forensic science - Part 1. 2015.

https://www.samsi.info/news-and-media/29-sep-christopher-saunders-samsi/ (accessed June 27, 2020).

23. Ausdemore MA, Neumann C, Saunders CP, Armstrong D, Muehlethaler C. Two-stage approach for the inference of the source of high-dimensional and complex chemical data in forensic science. J Chemom 2020:1-16. https://doi.org/10.1002/cem.3247

CHAPTER 3

Wasserstein Distance Score for Handwriting Feature Measurements

The following paper by Ommen et al. [14], published in Forensic Science International, is rewritten to reflect the current notational conventions. Ommen et al. introduces the Wasserstein distance score (WDS) metric used on the output of the MovAlyzeR® software to compare to the same survey of handwriting examiners (as discussed in Fuglsby et al. [9]. The MovAlyzeR® software outputs many kinematic measurements on each stroke of writing, measuring 100 strokes a second. The samples collected from the MovAlyzeR® software may have differing numbers of strokes (observations) measured on each writing sample. The WDS metric is designed to measure the difference between two samples with potentially differing numbers of observations. The WDS is also adaptable to univariate and multivariate input. My contributions to this work was on the development of the Wasserstein distance score and on writing the paper.

Advances toward validating examiner writership opinion based on handwriting kinematics

Abstract

A National Research Council report on strengthening forensic science raised concern over the lack of scientific studies supporting the validity of examining and interpreting forensic evidence. However, establishing the foundational validity of subjective methods can be challenging. The present study aimed to establish the scientific validity of expert writership opinions and the two-stage approach to evidence interpretation using measures derived from research on handwriting motor control. Regression-based procedures were used to address two experimental questions: 1) what are the relative contributions of kinematic and pressure features in predicting examiner support for alternate writership propositions when examining pairs of questioned handwriting samples; and 2) to what extent does information about the rarity of the kinematic feature dissimilarity scores improve the accuracy of a predictive model based on dissimilarity alone. Regarding the first question, we identified a multi-factor model consisting of feature dissimilarity scores and their population distributions having correlation coefficients (R^2) of 0.84 and 0.88 for the same-writer and different-writers propositions, respectively. Temporal features contributed up to 21% to the predictive value of the model, whereas spatial features contributed only 9% and pen pressure contributed up to 17%. When we compared models reflecting a single-stage process (based on feature dissimilarities) of forming opinions with models reflecting a two-stage process

Rewritten for notational clarity from Forensic Science International, 318, D. M. Ommen, C. Fuglsby, and M. P. Caligiuri, Advances toward validating examiner writership opinion based on handwriting kinematics, (based on feature dissimilarities and rarity) we found that the two-stage models had an average of 15.25% greater predictive value than single-stage models. These findings support the scientific validity of FDE writership determinations and underscore the importance of the two-stage approach for evidence interpretation.

3.1

Introduction

A National Research Council report on strengthening forensic science in the United States raised concern over the lack of scientific studies supporting the validity of examining and interpreting forensic evidence [1]. The report pointed to the general requirements under ISO/IEC 17025:2005 for competence testing and laboratory calibration based on established principles to substantiate validity. These requirements should include: (1) calibration of laboratory or examination practices using a standard reference, (2) ensuring agreement between two independent methods in reaching the same result, (3) inter-laboratory comparisons, (4) assessing confounding factors, and (5) recognition of the uncertainty based on scientific and theoretical principles underlying the method.

Forensic document examiners (FDE) are faced with an additional challenge in meeting the ISO/IEC criteria. In handwriting comparisons, the FDE constitutes a significant part of the measurement instrument subject to the same "laboratory calibration" as would be standard practice in laboratory-based disciplines. While attention has been paid to human sources of error including cognitive bias in addressing the problem of calibration and reliability in handwriting examination [2,3], cross-validation studies employing independent scientifically established methods have not been conducted or subject to peer review.

The PCAST report [2] identified essential criteria in establishing foundational validity of a measure in forensic science. First, the methods used to evaluate, for example pattern evidence, must undergo empirical testing by multiple groups under case-relevant conditions. Second, these studies must demonstrate that the method is reproducible and provide valid estimates of the accuracy of the method. While various research strategies can be deployed to test a method's accuracy, in subjective feature comparisons such as handwriting, studies of error rates (also known as black box studies) are common and offer general support for the comparative processes followed by experienced FDEs. However, as noted above in handwriting examination, the examiner is the measurement instrument and ground truth is not generally known as a means to establish accuracy outside the laboratory setting. Given the uncertainty of whether the examination process meets criteria for foundational validity in actual practice, experienced examiners will avoid making claims about the source of a handwritten sample that cannot be firmly established. Rather, the safe approach would be to offer an opinion on how strongly the examination supports the proposition that a questioned sample was or was not written by the suspect rather than reaching an explicit attribution. This approach to the interpretation of handwriting evidence was echoed in the recent NIST report on Forensic Handwriting Examination and Human Factors [3] noting that "Uniqueness and individualization in forensic science no longer correspond to the conventional, strict interpretation of these terms and can lead to an exaggeration of the strength of the evidence. Indeed, empirical research and statistical reasoning do not support source attribution to the exclusion of all others. In practice, examiners often (but not always) explain in reports and testimony that an identification to the exclusion of all others cannot be proven" (p. 47).

As the discipline of forensic handwriting examination moves away from claims of individualization, studies of error rates are less relevant in legal challenges of foundational validity. Establishing the foundational validity of subjective feature comparison methods can be difficult under the PCAST criteria. Rather a different approach to measurement validation is needed. The present study is based on a reframing of the concept of validity as is traditionally understood in forensic science to one that is widely accepted in the social sciences [4]. We consider a measurement to be valid if it accurately reflects or assesses the specific concept that examiner is attempting to measure. Several terms have been used to capture this notion including face validity and concurrent or convergent validity. The key advantages to this approach are that measurement validity can be established in the absence of ground truth and relies upon independent reference standards derived from reliable scientific methods.

Laboratory research on handwriting kinematics published over the past 35 years has contributed to the development of a reliable quantitative method for extracting specific features from handwriting samples [5-9]. The dynamic methodology yields numerous independent features characterizing the spatial and temporal characteristics of pen strokes. These features can be compared between two handwriting samples since it is expected that samples written by the same writer will have smaller feature dissimilarities than samples written by different writers. The present study aims to establish the scientific validity of FDE writership opinions and a common approach to evidence interpretation using methods derived from research on handwriting motor control. We utilized multiple regression procedures to quantify the relative importance of differences in spatial, temporal, and pressure features in predicting FDE strength of support for same-writer or different-writers propositions when examining pairs of unknown handwriting samples and the extent to which the rarity of the feature dissimilarities improves the kinematic predictive models. More generally, by examining the extent to which the rarity of the feature dissimilarities improves the predictive models, the present study is an empirical examination of the contribution of handwriting kinematics to the two-stage approach for evidence interpretation [10,11].

3.2

Methods

3.2.1

Handwriting samples

Thirty-three subjects were recruited from the San Diego Sheriff's Crime Laboratory to participate in the handwriting collection portion of the study. Subjects agreeing to participate were seated at a table and asked to write six phrases from the London Letter using an inking pen on lined paper placed atop a Wacom (Intuos Pro, model PTH-660) digitizing tablet. Subjects wrote each of six phrases five times in cursive and five times in print styles for a total of 60 samples per writer. To increase the yield of within-writer samples, subjects were asked to provide a second set of samples (consisting of five repetitions of each of the same six phrases from the first session in both cursive and print styles) two weeks later. Seven subjects provided two sets of samples. The digitized handwriting samples were recorded at a capture rate of 100 measurements/second in x, y, and z dimensions using MovAlyzeR® software. The digitized samples were subjected to analyses of kinematic features to be used in the predictive modeling component of this research, while the hard copy ink samples were used to create an FDE opinion survey to capture strength of support opinions for alternate writership propositions [12]. Further details of this FDE opinion survey are provided in the FDE Writership Opinions section below. Use of human subjects in the handwriting sample procurement portion of this study was reviewed and approved by the University of California San Diego Institutional Review Board.

3.2.2

Handwriting kinematic analyses

Samples were automatically segmented into upstrokes and downstrokes using MovAlyzeR®. Pen stroke segmentation points were determined based on the zero-axis crossing of the vertical velocity curve throughout time. Zero velocity along the curve reflected no vertical pen movement, thus marking a change in stroke direction. Multiple spatial, temporal and pressure features were then automatically extracted from each upward and downward pen stroke. For the purpose of this study, spatial features consisted of vertical and horizontal stroke amplitude, slant, loop surface, and trace length. Temporal features consisted of stroke duration, peak velocity and average velocity. Pen pressure was treated as a third feature set with a single feature. Throughout this paper, we use the term "kinematics" to refer to both movement and pressure features. Table 3.1 lists the kinematic features and definitions used in the present study to model FDE writership opinions. The measurements of all these features are recorded for each stroke segment within each sample and will be used to calculate kinematic feature dissimilarity scores between any two handwriting samples (within the same style and phrase).

Stroke Feature	Definition
Duration	Time interval (in milliseconds) between the first and last recorded measurements in a stroke.
Vertical Amplitude	Vertical vector difference between beginning and end of a stroke in centimeters.
Horizontal Amplitude	Horizontal vector difference between beginning and end of a stroke in centimeters.
Peak Vertical Velocity	First derivative of vertical displacement (centimeters/second). Also referred to as peak instantaneous velocity and is independent of segment duration.
Average Absolute Velocity	The absolute difference in segment amplitude divided by the difference in segment duration (centimeters/second). This measure does not distinguish upstrokes from downstrokes.
Slant	The angle or inclination of the axes of letters relative to the perpendicular to the baseline of the writing (in radians).
Loop Surface	Surface or the area of the loop enclosed by the previous and present stroke in square centimeters. The surface is not normalized. If the crossing does not occur within the previous stroke, although a loop has been formed, the loop area will be zero.
Trace Length	The length of a segment from beginning to end following its trajectory. It is calculated by summing the distances (in centimeters) between all consecutive recorded measurements or pixels.
Pen Pressure	Relative axial pressure on the pen tip when the pen is on the paper (ranging from 0 to 2047).

Table 3.1: Handwriting stroke features used in this study and their operational definitions.[1]

[1] http://www.neuroscript.net/help/viewingtrials.html

3.2.3

Handwriting feature dissimilarity scores and empirical cumulative distribution functions

MovAlyzeR® was used to extract multiple kinematic features from each pen stroke of the digitally acquired handwriting samples. These features characterize the handwriting in spatial, temporal and pressure dimensions, each having different units of measure. For example, stroke size is measured in centimeters, velocity in centimeters/second, stroke duration and road length in milliseconds, and pen pressure in digital units. Several transformations were necessary in order to reduce the multidimensional kinematic features into a single score representing the dissimilarity between two handwriting samples. We developed a new dissimilarity score to measure the difference between two writing samples for each style of writing (print or cursive) and for each selected phrase of the London Letter.

This dissimilarity score is constructed by first identifying the measurements for all upstrokes then using Linear Discriminant Analysis (LDA) to find the direction of maximum separation between the feature sets of the two handwriting samples (which does not require ground truth about the writer, only labels for which strokes belong to which sample in the pair). Once this direction is determined by LDA, then it is used (without sample labels) to classify each stroke as either belonging to the first or second sample in the pair. This results in an estimated posterior probability of each upstroke belonging to the first handwriting sample. In the situation where the handwriting samples were produced by two different writers, ideally, we would see that every segment from the first sample would have a very large posterior probability (near 1) and all segments from the second sample would have very small posterior probability (near 0) of belonging to the first writing sample. In the situation where the handwriting samples were written by the same writer, we would expect to see posterior probabilities spanning the entire range between 0 and 1. Finally, we compare the estimated posterior probabilities for all segments between the first and the second handwriting samples by computing the integrated squared error difference of the corresponding quantile functions. This procedure results in a measure of the dissimilarity between two quantile functions also known as the Wasserstein distance score (WDS) [13,14]. The WDS values range from 0 to 1 where values near 0 indicate the two samples are similar and values near 1 indicate that they are dissimilar. An analogous set of steps are then repeated for the downstrokes.

With 33 writers, each writing six phrases five times each in both cursive and print there were 13,530 sample pairs available per phrase and writing style to generate distribution functions and calculate population-level dissimilarity scores for multidimensional handwriting feature sets. The population-level dissimilarity scores were calculated from the cumulative distribution of the WDS relative to the writing style and phrase. These scores reflect the "rarity" of a given WDS score in two different populations, one representing within-writer pairs (sample pairs were written by the same person) and the other between-writer pairs (sample pairs written by two different people). These distributions were estimated from all remaining between- or within-writer WDS from pairs not used in the survey. The resulting covariates are probability values (ranging from 0 to 1) that a given WDS is less than or equal to the WDS of the survey sample pair according to the estimated between-writer and within-writer populations, and is referred to as the empirical cumulative distribution function or ECDF$_b$ and ECDF$_w$, respectively. Higher ECDF values indicate that a WDS is larger than many other scores in the population. That is, the calculated WDS indicates that the writing pair is more dissimilar than other pairs from the population. Conversely, an ECDF value near zero indicates that the WDS is smaller than many other scores in the population. That is, the calculated WDS indicates that the writing pair is more similar than other pairs from the population. An ECDF near 1.0 indicates that the WDS for the given pair is in the upper tail of the population distribution, whereas ECDF values near 0 indicates that the WDS for the given pair is in the lower tail of the population distribution. ECDF values near the 0 or 1 extremes indicate the pair is rare within the population whereas ECDF values near 0.5 indicate the pair is common within the population. Fig. 3.1 shows examples of population density curves and their respective ECDF associated with one sample pair for the phrase "but Vienna and Berlin are quiet."

As shown in Fig. 3.1A, the population density of WDS for between-writer pairs range from 0 to 1 and average around 0.25, whereas Fig. 3.1B shows the population density of WDS for within-writer pairs has a narrower range from 0 to 0.1 with an average around 0.025. The sample pair written by different writers shown in Fig. 3.1 (top) has a WDS of 0.086 which is more common among between-writer pairs versus within-writer pairs due to its respective position in the body (A) versus tail (B) of the densities, respectively. This trend was observed for other between-writer pairs, and the reverse trend is observed for within-writer pairs. For the given sample pair, the corresponding ECDF$_b$ value is 0.230 and the ECDF$_w$ value is 0.973 (denoted by the shaded area in Fig. 3.1A and B, respectively). This indicates that the vast majority of the population-level WDS for within-writer pairs (B) fall at or below 0.086; whereas relatively few WDS from the population fall at or below 0.086 among between-writer pairs (A).

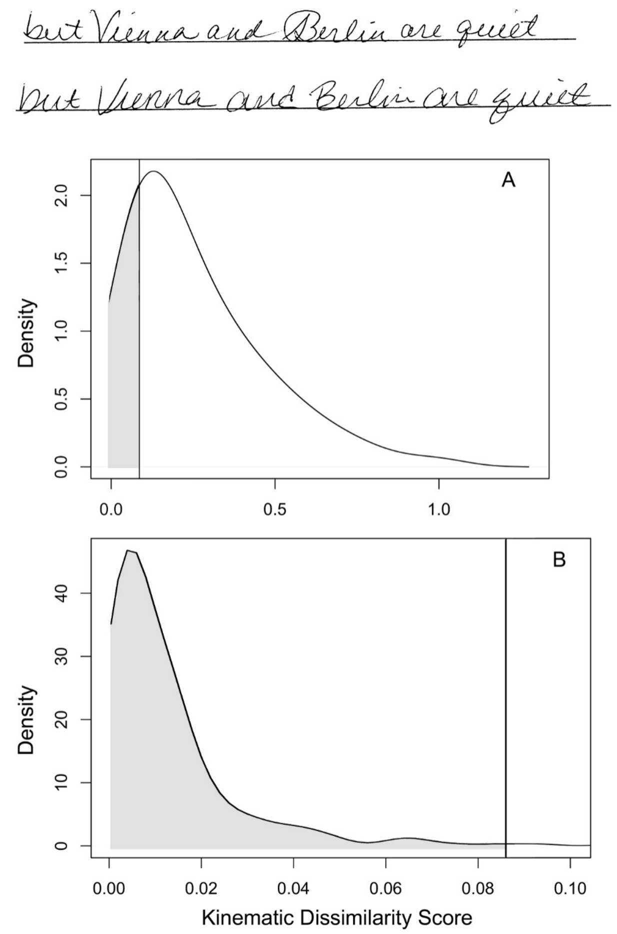


Figure 3.1: Examples of population density curves for between-writer (A) and within-writer (B) pairs. For this sample, survey item no. 33, the kinematic dissimilarity score (WDS) was 0.086 and is denoted by the vertical line. Shaded areas represent the ECDF value associated with this pair from a large pool of pairs of the same phrase written in cursive. See text for further explanation.

3.2.4

FDE writership opinions

Each page of the hard copy ink samples was scanned at 600 dpi, cropped into individual samples, and saved as separate 16-bit TIFF files. Individual samples were then paired with a sample of the same phrase from either another writer or a different sample from the same writer. An automated feature recognition program (FLASH ID®) was used to calculate dissimilarity scores for all possible sample pairs. These dissimilarity scores were then rank ordered for pairs within a given phrase and writing style. Forty of these pairs were included in a survey [12] to encompass two difficult case scenarios: 1) 30 between-writer pairs with low dissimilarity among the population of samples from different writers and 2) 10 within-writer pairs with high dissimilarity among the population of samples from the same writer. Of the 33 writers, 20 contributed handwriting samples to the survey based on having met the difficult case scenario criteria under the automated selection algorithm. Further details concerning the FLASH ID® dissimilarity score and the design of the survey can be found in Fuglsby et al. [12].

The online survey consisted of 20 print pairs (15 between-writer and five within-writer) and 20 cursive pairs (15 between-writer and five within-writer). Participants were blinded to the writer source for each pair. To assess repeatability, five pairs were repeated, however for these items, the order of Sample 1 and Sample 2 was reversed. Participants were asked to examine each sample pair and score their strength of support for each of two propositions. The first proposition pertained to the examiner's strength of support for the hypothesis that the two samples were written by the same writer (i.e. the prosecution hypothesis). The second proposition pertained to the examiner's strength of support for the hypothesis that the two samples were written by different writers (i.e. the defense hypothesis). Participants indicated their strength of support using a 7-point scale ranging from extremely strong support (7) to extremely low support (1). The survey yielded 90 strength of support scores for analysis from each participating FDE.

Invitations to participate in the writership survey were sent to 60 FDEs spanning three regions: North America, Europe, and Australia/New Zealand. Forty-one FDEs completed the survey (68.3% response rate) with the majority coming from European countries (26 FDEs), followed by Australia or New Zealand (9

FDEs) and North America (6 FDEs). Participants were asked to provide minimal de-identified demographic information pertaining to their work environment and experience. 90.2% of the examiners worked in government laboratories; 80.5% reported that a majority of their casework involved handwriting; and 75.6% reported having been in practice for at least 10 years. Use of human subjects for the survey was reviewed and approved by the University of California San Diego Institutional Review Board.

Next, we will summarize the results of the FDE opinion survey previously provided in Fuglsby et al. [12]. When addressing the prosecution hypothesis, FDEs strength of support for pairs known to come from the same writer averaged a 4.25 for cursive and 3.89 for print. Alternatively, FDEs strength of support for pairs known to come from different writers averaged a 3.17 for cursive and 3.16 for print. Therefore, when addressing the prosecution hypothesis, FDEs scored pairs known to come from the same writer 1.08 points higher than pairs known to come from different writers, on average, for cursive writing and 0.73 points higher for print writing. Similarly, when addressing the defense hypothesis, FDEs strength of support for pairs known to come from different writers averaged 3.90 and 3.78 for cursive and print, respectively, whereas the averages for pairs known to come from the same writer were 2.91 for cursive and 3.14 for print. So, when addressing the defense hypothesis, FDEs scored pairs known to come from different writers 1 point higher for cursive and 0.64 points higher for print writing than pairs known to come from the same writer, on average. All differences were statistically significant with a p-value of less than 0.001. This indicates that, by and large, FDEs tend to provide opinions that correspond to ground truth.

3.2.5

Statistical analyses

Two sets of analyses involving multiple regression were performed on the FDE strength of support scores. The goal of the first set of analyses is to quantify how important each of the handwriting feature sets are to FDEs when determining their strength of support opinions. This was done by comparing a full model with all feature sets included to a reduced model in which one feature set is excluded. If the feature set is important, the full model will show a stronger relationship to the FDE strength of support than the reduced model. The purpose of the second set of analyses is to explore the benefit of incorporating measures of rarity for handwriting features when FDEs determine their strength of support opinions. This was done by comparing a single-stage model that includes only dissimilarity scores to a two-stage model that also includes rarity as measured by ECDF values. If the second stage is beneficial, the two-stage model will show a stronger relationship to the FDE strength of support than the single-stage model.

In the first set, a model was designed consisting of the following explanatory variables: source of the sample pair (whether the pair was from the same or different writers), the WDS encompassing all relevant kinematic feature sets (spatial, temporal, pressure), the ECDF for between-writer and within-writer population distributions, and the interaction between the two ECDFs. This was considered the full model. Under our approach, we removed one feature set at a time from the calculation of the WDS and ECDF variables, while keeping the number of explanatory variables the same. These were considered reduced models. Thus, for each proposition and style of handwriting four models were run: 1) a full model where WDS and ECDF values were computed using all three feature sets; 2) a reduced model where WDS and ECDF values were computed after excluding the spatial features;

3) a reduced model where WDS and ECDF values were computed after excluding the temporal features; and 4) a reduced model where WDS and ECDF values were computed after excluding the pressure feature set. The difference in R^2 between the full model and each of the reduced models served to scale the importance of a given handwriting feature set in explaining the variability in FDE strength of support for either the same-writer or different-writers proposition. Separate models were run for each writership proposition and each style of handwriting (print or cursive) allowing broader interpretation. Therefore, each model was fit to 20 observations corresponding to 15 known between-writer and 5 known within-writer survey pairs. Based on prior research showing differences across several kinematic features and their variability between upstrokes from downstrokes [10,15,16], we ran separate regression models for each stroke direction.

In the second, we compared the strengths of association (R^2) between models of the feature dissimilarity scores (referred to as a single-stage model) and models of both feature dissimilarity scores and their ECDF scores (referred to as a two-stage model). This permitted an assessment of the added value of the rarity of the features dissimilarities to multivariate models predicting FDE responses to alternative writership propositions. More generally, this analysis enabled a cross-validation between handwriting kinematics and opinions rendered under a two-stage approach in evidence interpretation. Contrasts were examined for each proposition and writing style separately.

The reasoning for our statistical approach is because it is unclear whether traditional statistical inference would apply due to a variety of limitations with our data. First, the kinematic variables are highly correlated with each other making it difficult to interpret the individual contribution of any feature set using a traditional stepwise regression approach. For example, vertical amplitude (from the spatial feature set) is highly correlated with peak vertical velocity (from the temporal feature set) [17]. If, for example, vertical amplitude is removed during stepwise regression, it can still be accounted for in the model since it is confounded with peak vertical velocity. However, there is no way to tell the magnitude of the contribution of vertical amplitude within the peak vertical velocity covariate using traditional methods. Further details on the complications with interpreting stepwise regression procedures in the presence of multicollinearity can be found in [18]. Next, an exploratory analysis of the data revealed that several of the explanatory variables are highly correlated (for example, the WDS is highly correlated with the ECDF value for the between-writer population). Due to the interpretation issues in the presence of correlated explanatory variables, we make no attempt to interpret the significance of the coefficients for individual regressors. This is relevant to both sets of analyses. Finally, an important limitation of our data is the construction of survey items as pairs of samples from the same set of writers. While it is necessary to assume that the observations are independent for the regression analyses, it is unclear whether or not this is actually true. For example, survey items that contain a sample from writers with more natural variation may be associated with less strength of support than writers with less natural variation (causing the responses to be dependent on writer combination). To our knowledge, there is no easy way to check for this type of dependence caused by the pairwise structure. As a result, we have chosen to use the regression analyses in an exploratory manner, and refrain from making the usual statistical inferences associated with this type of analysis (e.g. statistical tests for significance and standard interpretations of regression coefficients). Further research is required to determine how to properly perform statistical inference for these pairwise data.

3.3

Results

The results are presented in two sections. In the first, we present findings on the relative importance of three kinematic feature sets in predicting FDE strength of support for alternate writership propositions in both print and cursive handwriting. In the second, we present the contrasts between the single-stage (feature dissimilarities) and two-stage (feature dissimilarities and their population density distributions) predictive models.

3.3.1

The relative importance of kinematic feature sets in predicting FDE responses

The full regression model having the highest correlation between the handwriting kinematic variables and FDE strength of support for alternate propositions consisted of five covariates: 1) the sample type as either a within-writer pair or between-writer pair; 2) the logistic transform of the WDS using the natural logarithm for a given pair; 3) the logistic transform of the ECDF using the natural logarithm for within-writer distributions; 4) the logistic transform of the ECDF using the natural logarithm for between-writer distributions; and 5) the interaction between the two logistic-transformed ECDF scores. The 40 sample pairs from the survey (disregarding the five survey items to assess reliability) were available to model FDE strength of support scores and generate the R^2 coefficients. This 5-factor model resulted in multiple correlation coefficients of 0.92 (R^2 = 0.84) and 0.94 (R^2 = 0.88) for downstroke kinematics from cursive samples for the same-writer and different-writers propositions, respectively. Correlation coefficients were lower for print samples (with R^2 values of 0.51 and 0.61 for the two propositions respectively) and for upstroke kinematics in general.

Table 3.2 shows the coefficients of determination (R^2) for the full and reduced models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting. Decreasing R^2 when removing a feature set from the model indicates that the feature set was important in the overall prediction. Table 3.3 shows the differences in R^2 between the full and three reduced models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting. Here, negative values indicate that removing a feature set from the full model reduces R^2 and weakens the strength of the relationship between kinematic feature dissimilarity and their rarity and FDE strength of support. Therefore, feature sets with negative values in Table 3.3 strengthen the overall relationship between examiner writership opinion and the full set of kinematic features, whereas feature sets with positive values weaken this relationship.

For cursive handwriting, the contribution of temporal features for downstrokes to the model predicting FDE strength of support ranged from $15\% - 21\%$ (depending on the proposition) of the predictive value, whereas for printed samples, including temporal features for downstrokes in the full model reduced the predictive value by $5\% - 7\%$ (depending on the proposition). Temporal features for upstrokes imparted minimal effects on the predictive modeling of FDE responses with changes in R^2 between -1% and $+3\%$. Overall, we found that temporal features had the largest contribution when assessing cursive downstrokes.

Spatial features extracted from downstrokes imparted a modest effect on the correlation between handwriting kinematic dissimilarity scores (and distributions) and FDE support for alternate writership propositions. Contributions ranged from $7\% - 9\%$ for cursive samples and $3\% - 7\%$ for print samples. The contribution of spatial features from upstrokes was mixed. Minimal effects were observed for cursive upstrokes ($1\% - 5\%$); whereas for print samples, spatial features weakened the relationship to examiner opinion by as much as 16% for FDE responses to the different-writers proposition. Therefore, we found that spatial features contribute to assessing all types of comparisons with the exception of upstrokes in print writing.

Pen pressure for downstrokes imparted a larger contribution for cursive writing ($15\% - 17\%$) than printed writing ($0\% - 1\%$) in predicting FDE support for the writership propositions. The opposite pattern was observed for upstrokes with pen pressure weakening the relationship to examiner opinion by $5\% - 8\%$ for cursive and strengthening the relationship by $5\% - 6\%$ for printed samples. Accordingly, pressure has the largest contribution to assessing cursive downstrokes, followed by print upstrokes.

	Cursive	
	Same Writer	Different Writers
Downstrokes		
Full	0.84	0.88
-Temporal	0.69	0.67
-Spatial	0.77	0.79
-Pressure	0.69	0.71
Upstrokes		
Full	0.61	0.65
-Temporal	0.63	0.68
-Spatial	0.60	0.60
-Pressure	0.69	0.70
	Print	
	Same Writer	Different Writers
Downstrokes		
Full	0.51	0.61
-Temporal	0.58	0.66
-Spatial	0.48	0.54
-Pressure	0.51	0.60
Upstrokes		
Full	0.54	0.56
-Temporal	0.53	0.59
-Spatial	0.67	0.72
-Pressure	0.49	0.50

Table 3.2: R^2 values for full and reduced multiple regression models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting.

	Cursive	
	Same Writer	Different Writers
Downstrokes		
-Temporal	-15%	-21%
-Spatial	-7%	-9%
-Pressure	-15%	-17%
Upstrokes		
-Temporal	$+2\%$	$+3\%$
-Spatial	-1%	-5%
-Pressure	$+8\%$	$+5\%$
	Print	
	Same Writer	Different Writers
Downstrokes		
-Temporal	$+7\%$	$+5\%$
-Spatial	-3%	-7%
-Pressure	0%	-1%
Upstrokes		
-Temporal	-1%	$+3\%$
-Spatial	$+13\%$	$+16\%$
-Pressure	-5%	-6%

Table 3.3: Differences in R^2 between full and three reduced models in the prediction of FDE strength of support for two writership propositions for cursive and print handwriting.

3.3.2

Contrasts between the single-stage and two-stage predictive models

Table 3.4 shows the results contrasting the single-stage and two-stage regression models. Overall, the two-stage models increased the predictive value (R^2) an average of 15.25% (range 4% – 35%). Effects of including the rarity measures along with the feature dissimilarity scores in the predictive models were greater for cursive than printed handwriting.

	Cursive	
	Same Writer	Different Writers
Downstrokes		
R^2 (Single-Stage Model)	0.53	0.53
R^2 (Two-Stage Model)	0.84	0.88
R^2 Difference	0.31	0.35
Upstrokes		
R^2 (Single-Stage Model)	0.44	0.51
R^2 (Two-Stage Model)	0.61	0.65
R^2 Difference	0.17	0.14
	Print	
	Same Writer	Different Writers
Downstrokes		
R^2 (Single-Stage Model)	0.46	0.52
R^2 (Two-Stage Model)	0.51	0.61
R^2 Difference	0.05	0.09
Upstrokes		
R^2 (Single-Stage Model)	0.47	0.52
R^2 (Two-Stage Model)	0.54	0.56
R^2 Difference	0.07	0.04

Table 3.4: Correlation coefficient (R^2) for single-stage and two-stage models and their differences for cursive and print handwriting, upstrokes and downstrokes for FDE strength of support for the prosecution or same-writer hypothesis (H1) and defense or different-writers hypothesis (H2).

3.4

Discussion

The present study aimed to establish the scientific validity of expert writership opinions and the two-stage approach by correlating expert opinions with handwriting kinematics and pen pressure. Handwriting samples from 33 writers were subjected to digital analyses to extract spatial, temporal, and pen pressure measures from each pen stroke. These measures were further analyzed to produce feature dissimilarity scores and the rarity of the dissimilarity scores within a population of over 13,000 possible pairs of print and cursive handwriting for each phrase. Regression-based techniques were applied to test the correlation between expert opinions of writership for pairs of unknown handwriting samples and their corresponding feature dissimilarity and population distribution scores.

Two key findings emerged from this study. First, we examined the relative contribution of a specific set of kinematic features within a larger explanatory model predicting FDE writership opinion by selectively removing spatial, temporal, and pen pressure feature sets from a full model. The full model consisted of feature dissimilarity scores from all three feature sets as well as their population distribution scores (i.e. the rarity of the dissimilarity score). This model resulted in multiple correlation coefficients (R^2) of 0.84 and 0.88 from cursive samples for the same-writer and different-writers propositions, respectively. Results from our "step-down" approach indicated that temporal features contributed up to 21% to the overall accuracy; spatial features contributed only 9%; while pen pressure contributed up to 17% to the correlation between handwriting kinematics and FDE writership opinion.

Secondly, we compared the strengths of association (R^2) between models of the feature dissimilarity scores (referred to as a single-stage model) and models of both feature dissimilarity scores and their empirical distribution functions scores (referred to as a two-stage model). This permitted an assessment of the added value of the rarity of the features dissimilarities to models predicting FDE responses to alternative writership propositions. Results comparing the single-stage and two-stage models revealed that the two-stage models increased the predictive value over the single-stage models an average of 15.25%. This finding underscores the importance of the discrimination stage of the two-stage evaluative process when forming opinions about whether pairs of unknown handwriting samples were written by a single writer or different writers.

The results of this study have three important implications. First, kinematic downstroke feature dissimilarity scores from cursive samples and their distributions were highly correlated with FDE strength of support opinions for alternative propositions. As such, FDE training or continuing education programs could benefit from curriculum that include exposure to handwriting motor control and handwriting kinematics. Second, FDE responses to the defense (different writer) proposition had higher correlation coefficients with kinematic variables than FDE responses to the prosecution (same writer) proposition. This suggests that FDEs are more likely to rely on the perceived rarity or population distribution of feature dissimilarities when evaluating samples from unknown but different writers than when evaluating samples from the same writer. In the absence of available data on the population density at the time of the examination, this information likely comes from experience. Further studies are necessary to confirm this hypothesis. Third, our models showed that the relationships between FDE support for specific writership propositions and kinematic feature scores and distributions were consistently stronger for writership opinions involving cursive handwriting than printing. The average multiple correlation coefficients (R^2) for cursive writing was 0.75 and 0.56 for printed writing for the two-stage model (see Table 3.3). The reasons for this pattern are unclear. However, we can speculate that cursive samples contain more relevant kinematic features than printed samples to inform expert writership opinions.

In the online survey designed for this study, FDEs examined digitized versions of the original hardcopy samples. We chose not to make paper copies of the original samples to distribute to all participating FDEs due to known limitations of examining copies. It is hypothesized that the same type of study conducted on copy samples would lead to smaller strength of support values from the FDEs for the comparisons included in the survey. This would potentially lead to weaker associations between the FDE opinions and the kinematic feature sets. However, further research is needed to explore this hypothesis for other survey designs that use copy documents.

Since Daubert, several challenges to the reliability, validity, and subsequent admissibility of handwriting evidence have been raised in Federal court. Of particular relevance to the problem addressed by this research are the conclusions reached by Judge Rakoff in Almeciga v Centers for Investigative Reporting. In his 2015 ruling, Judge Rakoff excluded expert testimony on handprinting following a Daubert hearing concluding that without a refined methodology, forensic document examination "is virtually untestable, rendering it an unscientific endeavor". The results from the present study suggests that trained forensic document examiners deliver opinions in support for a single-writer proposition and conversely, for a different-writers proposition that were associated with spatial, temporal, and pen pressure features for handprinting. Our results offer support for a scientific basis underlying FDE opinions and for the two-stage approach to evidence interpretation. It is important to place this interpretation in proper context given the aforementioned limitations in the ability of this study to address the statistical significance of the findings. The limited nature of our conclusions is due to the lack of rigorous methodology to evaluate the statistical significance of this type of experimental data.

Finally, it is important to place the present findings in the context of the 2016 PCAST report [2]. In addressing the foundational validity of feature-comparison methods, the report noted: " . . . that neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of confidence based on personal professional experience or expressions of consensus among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a sine qua non. Nothing can substitute for it." (p. 6) The overarching goal of the present study was to establish a causal relationship between examiner opinion and basic science. PCAST considered validation studies within the feature-comparison disciplines to be black-box studies. Black-box studies are those in which examiners express conclusions about questioned and known samples in order to establish the proficiency of the examiner though analysis of error rates. White-box studies on the other hand are designed to shed light on factors that contribute to the examiners' conclusions. While white-box studies are prevalent in the latent print literature [19, 20], the present study is one of just a few white-box studies from within the questioned document discipline designed to identify factors contributing to examiners' conclusions.

3.5

Conclusions

The present study aimed to establish the scientific validity of expert writership opinions and the two-stage approach using methods derived from research on handwriting motor control. Multiple regression models of feature dissimilarities and their population distributions were used to predict writership opinions by forensic document examiners. The observed correlation coefficients for the same-writer and different-writers propositions suggest strong associations between FDE opinions and kinematic feature dissimilarities. Temporal features (stroke duration and velocity) contributed more to the predictive value of the kinematic models, followed by pen pressure and spatial features (vertical and horizontal stroke amplitude, slant, loop surface, and trace length). Further examination revealed that the two-stage models (based on feature dissimilarities and distribution functions) produced, on average, 15.25% greater predictive value than single-stage models (based only on feature dissimilarities). These findings support the scientific validity of FDE writership determinations and under-score the importance of the two-stage process for evidence interpretation.

References

1. National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, 2009, doi:http://dx.doi.org/10.17226/12589.

2. President's Council of Advisors on Science and Technology (PCAST), Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, (2016).

3. Expert Working Group for Human Factors in Handwriting Examination. Forensic Handwriting Examination and Human Factors: Improving the Practice through a Systems Approach, U.S. Department of Commerce, National Institute of Standards and Technology. NISTIR 8282, 2020, doi:http://dx.doi.org/10.6028/NIST.IR.8282.

4. D. Borsboom, G.J. Mellenbergh, J. van Heerden, The concept of validity, Psychol. Rev. 111 (2004) 1061-1071, doi:http://dx.doi.org/10.1037/0033-295X.111.4.1061.

5. G. van Galen, Structural complexity of motor patterns: a study on reaction times of handwritten letters, Psychol. Res. 46 (1984) 49-57.

6. A. van Gemmert, G. van Galen, Dynamical features of mimicking another person's writing and signature, in: M.L. Simner, C.G. Leedham, A.J.W.M. Thomassen (Eds.), Handwriting and Drawing Research: Basic and Applied Issues, IOP Press, Amsterdam, 1996, pp. 459-471.

7. B. Ostrum, T. Tanaka, Another look at handwriting movement, J. Am. Soc. Quest. Doc. Exam. 9 (2006) 57-67.

8. L. Mohammed, B. Found, M.P. Caligiuri, D. Rogers, Dynamic characteristics of signatures: effects of writer style on genuine and simulated signatures, J. Forensic Sci. 60 (2015) 89-94, doi:http://dx.doi.org/10.1111/1556-4029.12605.

9. M.P. Caligiuri, L.A. Mohammed, The Neuroscience of Handwriting: Applications for Forensic Document Examination, CRC Press, Boca Raton; FL, 2012, pp. 95-111.

10. P.L. Kirk, J.I. Thornton, Crime Investigation, second ed., John Wiley and Sons Ltd., New York, NY, 1974.

11. J.B. Parker, A statistical treatment of identification problems, J. Forensic Sci. Soc. 6 (1966) 33-39.

12. C. Fuglsby, C. Saunders, D.M. Ommen, M.P. Caligiuri, Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage evaluative process, J. Forensic Sci. 65 (6) (2020) 2080-2086, doi:http://dx.doi.org/10.1111/1556-4029.14547.

13. E. del Barrio, J.A. Cuesta-Albertos, C. Matrn, J.M. Rodriguez-Rodriguez, Tests of goodness of fit based on the L2-Wasserstein distance, Ann. Stat. 27 (1999) 1230-1239.

14. E. del Barrio, J.A. Cuesta-Albertos, C. Matrn, S. Csrg, C.M. Cuadras, T. de Wet, et al., Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests, Test 9 (2000) 1-96, doi:http://dx.doi.org/10.1007/BF02595852.

15. G.P. van Galen, J.F. Weber, On-line size control in handwriting demonstrates the continuous nature of motor programs, Acta Psychol. 100 (1998) 195-216.

16. F. Maarse, A. Thomassen, Produced and perceived writing slant: difference between up and down strokes, Acta Psychol. 54 (1983) 131-147.

17. P. Viviani, C. Terzoulo, Space-time invariance in learned motor patterns, in: G.A. Stelmach, J. Requin (Eds.), Tutorials in Motor Behavior, North-Holland, Amsterdam, 1980, pp. 525-533.

18. A.J. Izenman, Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, New York, 2013.

19. R.A. Hicklin, J. Buscaglia, M.A. Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (2013) 106-117, doi:http://dx.doi.org/10.1016/j.forsciint.2012.12.015.

20. B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Changes in latent fingerprint examiners’ markup between analysis and comparison, Forensic Sci. Int. 247 (2014) 54-61, doi:http://dx.doi.org/10.1016/j.forsciint.2014.11.021.

CHAPTER 4

Regression Modeling for the Comparison of the Euclidean Distance Score and the Wasserstein Distance Score

The following paper by Fuglsby et al. [8] has been rewritten to reflect the current notational conventions. Fuglsby et al. does not include the survey of FDEs and instead introduces the use of regression models to compare the scores of an opaque system (FLASH ID®, using the Euclidean distance scoring method on the VOS output) and a transparent system (MovAlyzeR®, using the WDS method.) The appendix of this publication is included to incorporate the model development. My main contributions to this paper were the FLASH ID® VOS calculations and graphical design.

Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons

Abstract

Recent advances in complex automated handwriting identification systems have led to a lack of understandability of these systems' computational processes and features by the forensic handwriting examiners that they are designed to support. To mitigate this issue, this research studied the relationship between two systems: FLASH ID®, an automated handwriting/black box system that uses measurements extracted from a static image of handwriting, and MovAlyzeR®, a system that captures kinematic features from pen strokes. For this study, 33 writers each wrote 60 phrases from the London Letter using cursive writing and handprinting, which led to thousands of sample pairs for analysis. The dissimilarities between pairs of samples were calculated using two score functions (one for each system). The observed results indicate that dissimilarity scores based on kinematic spatial-geometric pen stroke features (e.g., amplitude and slant) have a statistically significant relationship with dissimilarity scores obtained using static, graph-based features used by the FLASH ID® system. Similar relationships were observed for temporal features (e.g., duration and velocity) but not pen pressure, and for both handprinting and cursive samples. These results strongly imply that both the current implementation of FLASH ID® and MovAlyzeR® rely on similar features sets when measuring differences in pairs of handwritten samples. These results suggest that studies of biometric discrimination using MovAlyzeR®, specifically those based

Rewritten for notational clarity from the Journal of Forensic Sciences, 57, C. Fuglsby, C. Saunders, D. M. Ommen, J. Buscaglia, and M. P. Caligiuri, Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons, pages 642-650, on the spatial-geometric feature set, support the validity of biometric matching algorithms based on FLASH ID® output.

4.1

Introduction

In forensic science, examiner-based black box studies "evaluat[e] the examiners' accuracy and consensus in making decisions, rather than attempting to determine or dictate how those decisions are made." [1] More broadly, an examiner-based black box study is "an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples" ([2]; p. 48). Typically, the examiner is viewed as a black box, and the aim of the research is to measure the degree to which the output or response from the black box examiner conforms with ground truth. Conversely, white box studies "are detailed assessments of the bases of examiners' decisions, focused not just on the end decisions but the features and attributes used by the examiners in rendering conclusions" [3]. Although the concepts of black box and white box methods of examiner testing in forensic science have become well-known in recent years, black box and white box methods have their roots in computer systems testing. With advances in automated feature recognition systems for forensic science applications, the forensic focus on black box methods should include both machine-based decision systems and human examiners, with increasing emphasis on interpretable artificial intelligence.

Approaches to automated handwriting identification and verification have been developed since the mid-1980s [4]. Several systems have emerged over the years including CEDAR-FOX [5], Forensic Information System for Handwriting (FISH), WANDA [6], and FLASH ID® (Sciometrics, LLC). FLASH ID® is an automated handwriting feature extraction program designed for closed-set identification of writers [7]. FLASH ID® relies on complex algorithms using graph theory to skeletonize and segment handwriting from a scanned document into graphemes (or subgraphs) having nodes and edges. Each grapheme is assigned an "isomorphism class" based on the connectivity structure and a "shape class" based on a set of rules centered on each grapheme's geometry. Each grapheme also has a feature vector of physical measurements within the geometric-spatial domain. Similar to FLASH ID®, other automated systems segment handwriting into smaller pieces in order to extract meaningful measurements from a larger handwriting sample. The responses produced by FLASH ID® involve multiple decisions for segmenting and classifying features based on graphemes, but the precise methods of doing so are not disclosed to the system's users. In this sense, FLASH ID® may be considered a black box evaluative system because the transfer function between input and output response is not transparent.

In contrast, MovAlyzeR® (Neuroscript, LLC) is a program that records and analyzes dynamic pen movements. MovAlyzeR® cap-tures the digitized writing sample and then segments the writing sample into individual strokes based on change in stroke direction; it encodes the on-line pen strokes to generate spatial-geometric and temporal metrics (i.e., kinematics) and pen pressure to characterize the handwritten features. The on-line decoding of pen strokes and reduction of feature metrics by the MovAlyzeR® system is fully transparent to the user and, as such, we considered it to be a white-box evaluative system. The process of disentangling the inner workings of an automated black box system may not be trivial and, in some cases, the user may only have access to the input objects and their outputs but not complete access to the black box system. Using the inputs, a white box system can deconstruct each object and gain a broader/deeper understanding of the closed black box system. These details may be used to model the black box system and determine if the features measured are significant in predicting the outputs of the black box system. The black box and white box systems chosen for modeling are FLASH ID® and MovAlyzeR®, respectively. The first goal of this study is to use MovAlyzeR® to elucidate the informative characteristics of a black box automated handwriting feature recognition system used in forensic handwriting comparisons (i.e., FLASH ID®). The second goal is to determine the strength of associations (if any) of feature differences between the two systems for handprinting and cursive styles of handwriting across different features. Finally, the third goal is to provide empirical support for the validity of the two automated handwriting feature analysis systems used.

To accomplish the first study goal, both systems are deployed on the same handwriting sample pairs, and feature dissimilarity scores are calculated and used to evaluate the relationship between these two systems. Specifically, we are interested in determining whether feature differences between two samples of handwriting obtained from a black box automated system are associated with feature differences obtained from a white box automated system.

The second goal is to determine the strength of these associations (if any) for handprinting and cursive styles of handwriting across multiple feature sets. Based on preliminary power studies (see the Appendix) and some knowledge about each system's capabilities, we formed four expectations. First, knowing that FLASH ID® uses a static image, we expect to observe a relationship between FLASH ID® dissimilarity scores and the scores for static spatial-geometric MovAlyzeR® features. Second, as FLASH ID® does not accept dynamic pen features as input, we did not expect to observe a relationship between FLASH ID® dissimilarity scores and scores for the dynamic temporal MovAlyzeR® features. Third, because FLASH ID® uses static images, we did not expect to observe a relationship between FLASH ID® dissimilarity scores and scores for the dynamic pen pressure features from

MovAlyzeR®. Fourth, we expected these relationships to hold for both writing styles (i.e., cursive writing vs. handprinting).

There is evidence that both MovAlyzeR® and FLASH ID® are considered valid instruments when applied to their designed purpose. Regarding MovAlyzeR®, support comes from controlled validation studies designed to assess the accuracy of spatial-geometric and temporal kinematic features and pen pressure in distinguishing genuine from simulated signatures [8,9], measuring signature complexity [10] and for distinguishing handwriting samples from two unknown writers [11]. Several studies summarized in Miller et al. [7] support the validity of several versions of FLASH ID®. Walch and colleagues [12] reported performance rates from two experiments of FLASH ID® deployed in a pairwise comparison of topological and geometric classes extracted from handwritten samples. They found 100% correct classification from 194 test documents (100 writers) in the first experiment and 100% correct classification from 590 test documents (300 writers) in the second. Another study by Walch et al. [13] used grapheme-based shape codes processed from 200 test documents to test the performance of FLASH ID®. They reported 99.5% accuracy in correctly identifying same-source documents. These studies motivated the third goal of this study, namely, to provide further empirical support for the validity of MovAlyzeR® and FLASH ID® as measures of handwriting feature and pattern analysis systems. A fundamental principle in scientific measurement validation is that one of the instruments under study exhibits performance characteristics that are consistent with the expected response pattern of the behavior being measured [14]. The third aim extends this principle to forensic measurement validation, as recommended in the PCAST Report ([2]; p. 14) as applied to handwriting feature and pattern analysis systems.

4.2

Methods

4.2.1

Study participants and handwriting sample collection

The study recruited 33 volunteer writers from the San Diego Sheriff's Crime Laboratory; each subject was asked to write six phrases from the London Letter [15] and to repeat each phrase five times using both handprinting and cursive writing styles (for a total of 60 writing samples per subject). Handwriting data from these subjects were used in two prior studies aimed at further understanding the decision-making process of forensic document examiners [16,17]. Subjects were asked to write the handwriting sample phrases with an inking pen on lined papers placed on top of a Wacom (Intuos Pro, model PTH-660) digitizing tablet. The stimulus phrase was shown on the top of each page, and repetitions were written vertically, five per page. A total of 1980 separate handwriting samples were collected on both paper (for processing in FLASH ID®) and digital forms (for processing in MovAlyzeR®) from 33 writers. The 60 handwritten samples from each subject collected on paper were scanned to digital format and underwent feature extraction via FLASH ID®, whereas the 60 digital samples collected on the Wacom tablet underwent direct feature extraction via MovAlyzeR®. Then, for any given stimulus phrase and style of writing, the comparison of the features between all pairs of samples resulted in a large set of dissimilarity scores, as described later.

4.2.2

FLASH ID® feature dissimilarity scores

For this study, we modified the scoring output (but not the feature extraction) of FLASH ID®, as previously described in Fuglsby et al. [16]. The output of FLASH ID® encodes all the graphemes in a document relative to a reference set of writers (in this case, 50 writers from the “FBI100” data set, described in Saunders et al. [18]; the reference set is a term used in FLASH ID® to denote a list of possible writers of interest for the original recommendation system). The graphemes used for this encoding were derived from a base set of 50 different writers (in this case, the remaining 50 writers from the “FBI100” data set). The FLASH ID® system uses the idea of reward functions to construct an omnibus score for the corresponding recommendation system. We use the idea of a reward function to construct our Vector of Scores (VOS); that is, each grapheme receives a set of rewards based on the recommender algorithm built by the reference set documents (one reward per grapheme for each reference set writer). Although the specific mechanism for assigning rewards is not revealed to the user, it is known that a larger reward indicates a greater similarity of that grapheme to the reference writer’s samples (M. Walch, D. Gantz, J. Miller, J. Buscaglia, personal communication, September 8-11, 2009). For each reference writer, these rewards are then summed over all the graphemes in a document, resulting in an omnibus VOS (comparable with the vector of counts method in Gantz et al. [19], for which the rewards are split among a reference set of writers) for each document. Calculating the Euclidean distance between the two VOSs (one per writing sample in a pair) yielded the dissimilarity score between the pair of writing samples. Larger Euclidean distance scores between two VOSs reflect larger feature dissimilarities. This was repeated for all possible sample pairs within a given phrase (from the London letter) and writing style. With 33 writers and five repeats for each of six phrases, this procedure yielded dissimilarity scores for 81,180 possible pairs for each writing style. The structure of this class of score functions leaves much to be desired in terms of how to interpret and explain the resulting dissimilarity.

To the best of our knowledge, the 33 writers who participated in this study are not part of the "FBI100" data set, given that they were collected approximately 15 years apart in different collections. However, as part of our ethical obligation to protect the privacy of study subjects, we could not cross-compare identity between the two groups.

4.2.3

MovAlyzeR® kinematic feature dissimilarity scores

Handwriting samples were automatically segmented into upstrokes and downstrokes using MovAlyzeR®. Pen stroke segmentation points were determined based on the zero-axis crossing of the vertical velocity curve over time. The zero velocity points along the curve reflect a momentary absence of vertical pen movement just prior to a direction change. The segmentation criterion is a user-defined property that was applied to all samples consistently. Several spatial-geometric, temporal, and pressure features were then automatically extracted from each upward and downward pen stroke. The set of spatial-geometric features included vertical and horizontal stroke amplitude, slant, loop surface, and trace length. The set of temporal features included stroke duration, peak velocity, and average velocity. Pen pressure was treated as a third feature set with only a single feature: the average pen pressure during the stroke.

These features characterize handwriting movement in multiple dimensions. The multidimensional kinematic features were transformed into a single score representing the dissimilarity between two handwriting samples, as in Ommen et al.

[17]. First, using the kinematic features for all upstrokes in a pair of handwriting samples, a dissimilarity score is constructed by determining the direction of maximum separation by applying linear discriminant analysis (LDA). The LDA method uses this direction to classify each upstroke to either the first or second sample in the pair by providing an estimated posterior probability of belonging to the first sample. For handwriting pairs produced by two different writers, every upstroke from the first sample should have posterior probabilities near one, and all upstrokes from the second sample should have posterior probabilities near zero. For sample pairs produced by the same writer, both samples should have posterior probabilities anywhere between zero and one (depending on the range of natural within-writer variation). Then, the integrated squared error difference of the two quantile functions for estimated posterior probabilities of upstrokes between the pair of handwriting samples is computed. This calculation is a measure of the dissimilarity between two quantile functions and is known as the Wasserstein distance score (WDS) [20,21]. The WDS values range from zero to one, where values near zero indicate more overlap in the posterior probabilities for the two samples, and values near one indicate less overlap. The level of dissimilarity between the measured features of each pairwise comparison is therefore determined by the corresponding WDS value. An analogous set of steps are repeated to obtain kinematic dissimilarity scores for the downstrokes.

4.2.4

Regression models of pairwise comparisons

A total of 1980 separate handwriting samples were collected on both paper and in digital forms from 33 writers. Hard copy samples were digitally scanned at 600 pixels per inch (ppi). The MovAlyzeR® feature dissimilarity scores for each pair were used to model the FLASH ID® feature dissimilarity score as follows. Separate simple linear regression models were run for each kinematic feature set (spatial-geometric, temporal, and pressure), for stroke direction (upstrokes and downstrokes), for each writing style (handprinting and cursive), and for six phrases for a total of 72 regression models ($3 \times 2 \times 2 \times 6 = 72$). We established that the large number of potential co-dependences across multi-writer input samples can inflate the Type I error (see Appendix). To minimize the threat stemming from multiple comparisons involving the same writer, we developed a robust statistical approach that takes the comparison/ dependence structure into account.

We assume that the collections of writing samples (with one collection per writer) are independent and identically distributed random elements; in effect, we have a simple random sample of writers, and from each writer, we have observed one collection of writing samples. For each of the writing samples, we have measured two sets of features: one corresponding to the FLASH ID® VOS dissimilarity score and a second set of features extracted from the MovAlyzeR® system. We further reduced the features from the MovAlyzeR® system into six sets of subfeatures: spatial-geometric, temporal, and pen pressure feature sets for both upstrokes and downstrokes.

For each of these seven sets of measurements (one FLASH ID® score and six kinematic feature scores), we developed a pairwise dissimilarity score to represent a document-level comparison. Following Ommen et al. [17], the pairwise dissimilarity is computed using a modification of the WDS (see the Appendix for further details). The goal was to create six different regression models to assess the marginal relationship between the MovAlyzeR® features and the FLASH ID® features, where the WDS for one of the six kinematic feature sets is used as the explanatory variable and the FLASH ID® dissimilarity score is used as the response variable. However, this became difficult because the observations (i.e., document-level dissimilarity scores) are not independent, although the assumption of independence is required to perform regression.

When the original set of samples are assumed to be a simple random sample (as in this case), the act of performing pairwise comparisons to produce a score introduces a dependence structure that must be accounted for before any statistical tests can be performed at the desired nominal level. If the full dependency structure (i.e., covariance matrix) is known up to a constant, then the generalized least-squares (GLS) approach can be used. Unfortunately, in this setting, there are three distinct terms that are needed before we can perform a GLS-based analysis. We do have the advantage of being able to solve out for the eigenvectors, but not the eigenvalues, of the pairwise dissimilarity scores covariance matrix. These issues are explored in greater detail in Appendix.

To address the issue of independence, the regression approach was modified. A summary measure was obtained for each pair of writers by averaging their 25 between-writer document-level dissimilarity scores. This resulted in a reduction of the 13,530 document-level dissimilarity scores for each phrase and style of writing to 528 writer-level dissimilarity scores. (See the Appendix for further details.) We performed the modified regression analyses for each of the six phrases, handprinting and cursive separately, and only considered one of the kinematic feature sets at a time. This resulted in a total of 72 tests and corresponding p-values.

4.3

Results

Scatterplots with regression lines-of-best fit are shown in Figure 4.1 for the phrase “Our London business is good” for the set of upstrokes for cursive (top row) and handprinting (bottom row) styles, respectively. The points on the scatterplots represent the average dissimilarity scores across all pairwise comparisons between a pair of writers. Each plot contains 528 averaged dissimilarity scores; for a detailed description of these averaged dissimilarity scores, see Appendix. The red regression lines are fit using the averaged pairwise scores, and the black line is the average of the red lines in each plot. Each plot shows the relationships between individual FLASH ID® VOS dissimilarity scores (y-axis) and MovAlyzeR® spatial-geometric, temporal, and pressure feature dissimilarity scores (x-axis) for the set of upstrokes for cursive (top row) and handprinting (bottom row) for all possible pairs for this phrase. The more negative the kinematic feature dissimilarity score (along the x-axis) is, the less dissimilarity there is in that feature between a given pair of writers.

Inspection of the scatterplots reveals a strong positive relationship for spatial-geometric feature dissimilarities between the two systems. Surprisingly, a modest positive relationship between the FLASH ID® VOS dissimilarity score and temporal feature dissimilarity score from MovAlyzeR® was observed. Lower FLASH ID® VOS dissimilarity scores were associated with lower kinematic spatial-geometric and temporal feature-based dissimilarity scores for cursive samples, whereas only spatial-geometric feature-based dissimilarity scores were significantly associated with FLASH ID® VOS dissimilarity scores for handprinted samples. Similar plots were obtained for downstrokes and for all phrases other than phrase 3.

Results from the regression models for average dissimilarity scores for the relationships between FLASH ID® VOS dissimilarity scores and MovAlyzeR® spatial-geometric, temporal, and pen pressure feature dissimilarities across all pairs of writers for cursive writing and handprinting are shown in Tables 4.1-4.3, respectively. Results show that spatial-geometric dissimilarity scores were significant ($p\text{-}value < 0.05$) in predicting FLASH ID® VOS dissimilarity scores for both handprinting and cursive sample pairs as well as upstrokes and downstrokes. The relationships between temporal feature dissimilarity scores and FLASH ID® VOS dissimilarity scores were significant ($p\text{-}value < 0.05$) for cursive sample pairs only, whereas the average pen pressure dissimilarity scores across samples between two writers was not a significant factor ($p\text{-}value > 0.05$) in predicting FLASH ID® VOS dissimilarity scores. With the exception of phrase 3, these patterns were consistent across stroke direction and across the different phrases from the London Letter. Phrase 3 differed from the other five phrases as it contains unfamiliar words such as "Mr. Lloyd" and "Switzerland," which may have contributed to greater dysfluencies and subsequently more variability in feature sets across writers as writers self-checked spelling and punctuation of this phrase.

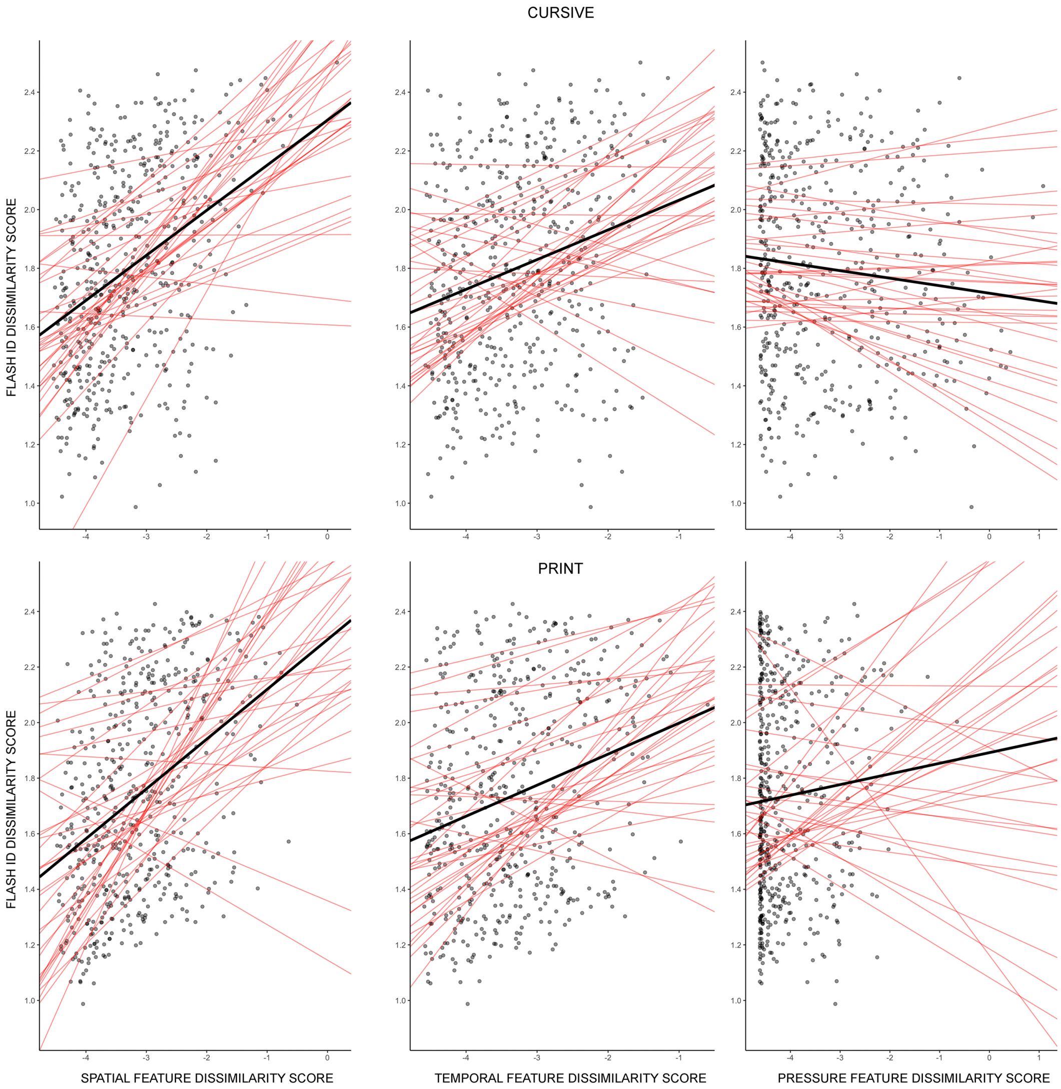


Figure 4.1: Scatterplots with individual (red) and average (black) lines of best fit for cursive (top row) and handprinting (bottom row) handwriting showing the relationship between FLASH ID® dissimilarity score (y-axis) and the dissimilarity scores for spatial-geometric (left), temporal (center), and pressure (right) features for upstrokes for the phrase "Our London business is good." The red regression lines are fit using the averaged pairwise scores-one score per pair of writers, each line representing the 33 scores with one fixed writer for a total of 33 red lines. The thick black line is the average of the red lines in each plot.

	Downstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	0.259	0.001	0.204	0.003
2	0.315	<0.001	0.309	<0.001
3	0.306	<0.001	0.076	0.174
4	0.233	<0.001	0.212	<0.001
5	0.185	0.005	0.250	<0.001
6	0.263	<0.001	0.187	0.001

	Upstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	0.197	0.043	0.155	0.016
2	0.315	0.001	0.283	<0.001
3	0.243	0.012	0.062	0.167
4	0.215	0.017	0.172	0.001
5	0.030	0.740	0.220	<0.001
6	0.250	0.001	0.180	<0.001

Table 4.1: Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® spatial-geometric dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes

	Downstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	0.051	0.528	0.106	0.232
2	0.094	0.373	0.197	0.043
3	0.020	0.857	−0.006	0.910
4	0.005	0.948	0.166	0.016
5	−0.063	0.464	0.277	0.002
6	0.138	0.174	0.176	0.023

	Upstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	0.113	0.218	0.167	0.057
2	0.173	0.083	0.332	<0.001
3	0.084	0.482	0.039	0.449
4	0.032	0.762	0.179	0.004
5	−0.074	0.381	0.183	0.004
6	0.134	0.127	0.177	0.004

Table 4.2: Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® temporal dissimilarity scores for cursive writing and handprinting sample pairs for upstrokes and downstrokes

	Downstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	−0.032	0.666	−0.064	0.193
2	−0.034	0.645	−0.078	0.138
3	−0.053	0.508	−0.022	0.636
4	−0.053	0.498	−0.015	0.738
5	−0.041	0.631	−0.079	0.119
6	0.003	0.972	−0.078	0.133
	Upstrokes			
	Print		Cursive	
Phrase	Slope Coefficient	p-value	Slope Coefficient	p-value
1	0.087	0.501	−0.062	0.270
2	−0.008	0.952	−0.016	0.813
3	0.092	0.568	0.032	0.530
4	−0.066	0.610	0.012	0.823
5	0.071	0.635	−0.011	0.835
6	0.017	0.897	−0.024	0.637

Table 4.3: Results from regression models predicting FLASH ID® dissimilarity scores based on MovAlyzeR® pen pressure dissimilarity scores for cursive writing and hand-printing sample pairs for upstrokes and downstrokes

4.4

Discussion

In the present study, we expected to observe three patterns. First, we expected that we would observe a relationship between these instruments for spatial-geometric features. We found that dissimilarity scores calculated from spatial-geometric stroke kinematics were significantly associated with dissimilarity scores calculated from an independent, automated feature recognition system in support of our hypothesis. As expected, the relationships between FLASH ID® VOS and MovAlyzeR® dissimilarity scores for spatial-geometric features were generally consistent, regardless of handwriting style. This finding implies that the spatial-geometric features detected and used by the FLASH ID® algorithm in its feature quantification may be robust to writing style.

For our second expectation, we did not expect to observe a relationship between dissimilarity scores produced by FLASH ID® VOS and those produced by kinematic analyses of temporal features. For handprinting, we did not find statistically significant relationships. However, contrary to this, we found significant relationships in the temporal domain for cursive handwriting. This is likely due to the well-established relationship between stroke velocity and stroke amplitude for limb movement in general [22] and handwriting specifically [14]. FLASH ID® relies upon complex algorithms to skeletonize and segment writing into graphemes, classify these graphemes using the resulting nodes and edges, and calculate the physical measurements exclusively within the spatial-geometric domain. Although it is a black-box system, the static input (i.e., digital scan of a document) contains no temporal components for the algorithms to utilize. Because two of the three parameters that make up the temporal feature set are velocity measures, it is possible that the temporal features were correlated with at least one of the spatial-geometric features driving the FLASH ID®-kinematic relationship. Thus, at least for cursive handwriting, velocity and amplitude are probably not independent features.

For the third expectation, we did not expect to observe a relationship between dissimilarity scores produced by FLASH ID® VOS and those associated with pen pressure. This expectation holds as we did not find any statistically significant relationships. As a static feature encoding system, FLASH ID® was not designed to encode pressure features in handwriting. However, considering that pen pressure often affects line thickness in the static handwriting sample, it is possible that pressure variation could affect the skeletonization and attribution of some grapheme structures in FLASH ID® (e.g., lower case "e" and "i"). Although line thickness can also be impacted by writing instrument (e.g., ballpoint pen vs marker), in the present study, all writers used the same writing instrument.

The kinematic feature dissimilarity scores for upstrokes behaved similarly to downstrokes with regard to their correlations with FLASH ID® VOS dissimilarity scores. This observation is not surprising, given that some of the graphemes used in the FLASH ID® system will contain both upstrokes and downstrokes. Further research may disentangle a stroke-direction effect that this study did not capture. There are strong correlations between the upstroke and downstroke dissimilarity scores (for both spatial-geometric and temporal); thus, seeing the significant p-values of these models with respect to the FLASH ID® VOS dissimilarity scores is not surprising.

The third goal of the present study was to provide empirical support for the validity of two automated handwriting feature analysis systems, MovAlyzeR® and FLASH ID®. Our results support both the construct and convergent validity of MovAlyzeR® and FLASH ID® as instruments capable of detecting differences in handwriting features between two samples written by different writers. The construct itself is a "process or characteristic believed to account for individual or group differences in behavior" ([23]; p. 1) where construct validity refers to how well an instrument measures that behavior or characteristic [24,25]. Handwriting consists of a series of individual pen movements or strokes, each characterized by multiple features in the spatial-geometric, temporal, and pressure domains. These characteristics form the construct used by examiners to understand variability within and across writers. Based on the robust statistical relationships between dissimilarity scores measured by our two instruments, especially in the spatial-geometric domain, we may conclude that both instruments are valid as measures of the construct that handwriting is a series of spatial-geometric parameters or patterns.

Convergent validity reflects the relationship among different measures of the same construct [23]. The present study demonstrated empirically that different measures of the same construct were statistically related. Dissimilarity scores derived from two different approaches to measuring handwriting converged along with some (but not all) features. Specifically, we observed convergence for spatial-geometric features such as vertical and horizontal stroke amplitude, slant, and trace length; however, such convergence was not observed for pen pressure. Where present, convergent validity held for both handprinting and cursive writing styles.

Within a statistical framework, validity can be defined as the absence of both random and systematic measurement error [14]. Although it is unreasonable to expect the complete absence of random or unexplained error between two independent measurement systems, minimizing systematic error is an attainable goal. Results from the present study demonstrate that there is at least a linear relationship between the FLASH ID® VOS dissimilarity scores and the previously noted subsets of the kinematic dissimilarity scores. In the present study, individual regression models for each of the kinematic feature scores were used, which ignores any possible interactions between the kinematic features. In the future, a single model that incorporates all the kinematic features could be developed using more sophisticated statistical tools. However, before these methods can be applied, they must be fully developed for pairwise comparison data [26].

Last, the guidance document published by the Presidents' Council of Advisors on Science and Technology [2] on ensuring scientific validity of forensic feature comparison methods recognizes a valid scientific instrument as one that "has shown, based on empirical studies, to be reliable with levels of repeatability, reproducibility, and accuracy that are appropriate to the intended application." (p. 48). The PCAST position on scientific validity is that if a measurement of a feature (or in this case, feature-based dissimilarity scores) produced accurate results (based on some accepted standard) and these results can be reproduced, then one can claim that the measurement system is valid within the context of legal discourse. Results from the present study demonstrate the scientific validity that is accepted in legal discourse for our intended application of both MovAlyzeR® and FLASH ID® as biometric verification systems.

Computational algorithms used in proprietary automated forensic biometric identification systems are considered black box systems and, therefore, pose a challenge for proper discovery in the U.S. judicial system. To increase their transparency and interpretability, many have called for the release of algorithm source code, potentially infringing the intellectual property of the algorithm developers. Our approach offers an alternative to the access to intellectual property while addressing the need for transparency and interpretability of such algorithms by developing techniques to characterize the performance of a black box algorithm in terms of a transparent system.

The present research focused on two systems, and any extension of the results of this research to other systems is not warranted at this time. Further research is needed to test whether the correlations observed in the present study between a black box system designed for writer verification and an open handwriting kinematic feature analysis system generalize to other automated systems such as CEDAR-FOX [5] or WANDA [6]. Such studies would strengthen the construct and convergent validity of these and other automated handwriting feature recognition systems.

In conclusion, the present study demonstrated that a white box system has the potential to inform the user of, and to validate, a black box system. Using handwriting data, the results of the testing showed a significant relationship between the FLASH ID® system and the spatial-geometric kinematic features measured by MovAlyzeR®, robust to writing content and writing styles.

Acknowledgements

The authors acknowledge Sierra Lutz for her contribution to the handwriting sample formatting, and Brenda Lanners and Gina Hunger for their assistance in collecting the handwriting samples. We wish to thank the volunteers who participated in providing the handwriting sample. The authors would like to acknowledge Sciometrics for providing access to the FLASH ID® system to use for this research.

Disclaimer

This is publication number 21-21 of the Laboratory Division of the Federal Bureau of Investigation (FBI). The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. government. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI.

References

1. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA. 2011;108(19):7733-8. https://doi.org/10.1073/pnas.1018707108

2. President's Council of Advisors on Science and Technology (PCAST). Report to the President. Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Washington, DC: PCAST; 2016. Accessed 27 Sept 2021 https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

3. Hicklin RA, Ulery B, Roberts MA, Buscaglia J. Black box and white box forensic examiner evaluations: Understanding the details. In: Proceedings of the 69th Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2017. p. 480.

4. Plamondon R, Lorette G. Automatic signature verification and writer identification - The state of the art. Pattern Recognit. 1989;22(2):107-31. https://doi.org/10.1016/0031-3203(89)90059-9

5. Srihari SN, Srinivasan H, Desai K. Questioned document examination using CEDAR-FOX. J Forensic Doc Exam. 2007;18:1-20. https://doi.org/10.31974/jfde28-15-26

6. Franke K, Schomaker L, Vuurpijl L, Giesler S. FISH-New: a common ground for computer-based forensic writer identification. In: Proceedings of the Third European Academy of Forensic Science Triennial Meeting; 2003 Sept 22-27; Istanbul, Turkey. Rome, Italy: Eur Acad Forensic Sci. 2003;136(S1-S432):84.

7. Miller JJ, Patterson RB, Gantz DT, Saunders CP, Walch MA, Buscaglia J. A set of handwriting features for use in automated writer identification. J Forensic Sci. 2017;62(3):722-34. https://doi.org/10.1111/1556-4029.13345

8. Caligiuri MP, Mohammed LA, Found B, Rogers D. Nonadherence to the Isochrony Principle in forged signatures. Forensic Sci Int. 2012;223:228-32. https://doi.org/10.1016/j.forsciint.2012.09.008

9. Mohammed L, Found B, Caligiuri M, Rogers D. The dynamic character of disguise behavior for text-based, mixed, and stylized signatures. J Forensic

Sci. 2011;56(Suppl 1):S136-41.
https://doi.org/10.1111/j.1556-4029.2010.01584.x

10. Angel M, Cavanaugh M, Caligiuri MP. Kinematic models of subjective complexity in handwritten signatures. J Am Soc Quest Doc Exam. 2017;20(2):3-10.

11. Caligiuri M, Mohammed L, Lanners B, Hunter G. Kinematic validation of FDE determinations about writership in handwriting examination: a preliminary study. J Am Soc Quest Doc Exam. 2018;21(1):3-12.

12. Walch M, Gantz D, Miller J, Saunders C, Lancaster M, Buscaglia J. Evaluation of the individuality of handwriting using FLASH ID - A totally automated language-independent system for handwriting identification. In: Proceedings of the 60th Annual Scientific Meeting of the American Academy of Forensic Sciences; 2008 Feb 18-23; Washington, DC. Colorado Springs, CO: American Academy of Forensic Sciences. 2008. p. 388.

13. Walch M, Gantz D, Miller J, Buscaglia J. Evaluation of the language-independent process in the FLASH ID system for handwriting identification. In: Proceedings of the 61st Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2009. p. 381-2.

14. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. Psychol Rev. 2004;111(4):1061-71. https://doi.org/10.1037/0033-295X.111.4.1061

15. Osborn AS. Questioned documents. 2nd edn. New York, NY: Boyd Printing Co.; 1929. p. 34.

16. Fuglsby C, Saunders C, Ommen DM, Caligiuri MP. Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage eval-

uative process. J Forensic Sci. 2020;65(6):2080-6. https://doi.org/10.1111/1556-4029.14547

17. Ommen D, Fuglsby C, Caligiuri MP. Advances toward validating examiner writership opinion based on handwriting kinematics. Forensic Sci Int. 2021;318:110644. https://doi.org/10.1016/j.forsciint.2020.110644

18. Saunders C, Davis L, Lamas A, Miller J, Gantz D. Construction and evaluation of classifiers for forensic document analysis. Ann Appl Stat. 2011;5(1):381-99. https://doi.org/10.1214/10-AOAS379

19. Gantz DT, Miller JJ, Saunders CP, Walch MA, Buscaglia J. New results for addressing the open set problem in automated handwriting identification. In: Proceedings of the 62nd Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2010. p. 431-2.

20. del Barrio E, Cuesta-Albertos JA, Matrn C, Rodriguez-Rodriguez JM. Tests of goodness of fit based on the L2-Wasserstein distance. Ann Stat. 1999;27(4):1230-9.

21. del Barrio E, Cuesta-Albertos JA, Matrn C, Csrg S, Cuadras CM, de Wet T, et al. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. Test. 2000;9:1-96. https://doi.org/10.1007/BF02595852

22. Viviani P, Terzoulo C. Space-time invariance in learned motor patterns. In: Stelmach GA, Requin J, editors. Tutorials in motor behavior. Amsterdam, Netherlands: North-Holland Publishing Company; 1980. p. 525-33.

23. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. Ann Rev Clin Psychol. 2009;27:1-25.

https://doi.org/10.1146/annurev.clinpsy.032408.153639

24. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302. https://doi.org/10.1037/h0040957

25. Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas: Issues Pract. 1995;14(4):5-8. https://doi.org/10.1111/j.1745-3992.1995.tb00881

26. Rosen SL, Saunders CP, Guharay SK. A structured approach for rapidly mapping multilevel system measures via simulation metamodeling. Syst Engin. 2015;18:87-101. https://doi.org/10.1002/sys.21290

Appendix

4.4.1

Data formatting and dissimilarity scores

For each phrase within each writing style, there are five documents from each writer. For the purpose of comparing the i^{th} and j^{th} writers (for $i = 1, 2, \ldots, n-1$, $j = i+1, \ldots, n$, and $i \neq j$) under the given phrase and writing style conditions, there are 25 unique ways to pair a document from the i^{th} writer to a document from the j^{th} writer, excluding pairs where both documents come from the same writer. For each of these document pairs, we computed seven dissimilarity scores:

1. Euclidean distance between FLASH ID® Vector-of-Scores (VOS);

2. Wasserstein distance between MovAlyzeR® spatial measurements on upstrokes;

3. Wasserstein distance between MovAlyzeR® spatial measurements on downstrokes;

4. Wasserstein distance between MovAlyzeR® temporal measurements on upstrokes;

5. Wasserstein distance between MovAlyzeR® temporal measurements on downstrokes;

6. Wasserstein distance between MovAlyzeR® pressure measurements on upstrokes;

7. Wasserstein distance between MovAlyzeR® pressure measurements on downstrokes.

Since the dissimilarity scores from the kinematics features (scores 2-7) were constrained to be between zero and one, we performed a logistic transform with a correction (adding 0.01 in the numerator of the fraction and subtracting from 1.01 in the denominator) to avoid taking the log of zero. Then, for each type of dissimilarity score, we took the average of all 25 of the resulting pairwise scores. This results in seven average dissimilarity scores for each writer pair.

4.4.2

Regression models of pairwise comparisons

Let Y_{ij} denote the average FLASH ID® VOS dissimilarity score between the i^{th} and j^{th} writers (for $i = 1, 2, \ldots, n-1$, $j = i+1, \ldots, n$, and $i \neq j$); let x_{ij} denote the average MovAlyzeR® dissimilarity score corresponding to one of six feature sets, described above, between the i^{th} and j^{th} writers. Then consider the model:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + a_i + a_j + \epsilon_{ij}$$

where the a_is and ϵ_{ij}s are independent random variables with variance terms σ_a^2 and σ_ϵ^2, respectively. The above linear model captures the natural dependencies from pairwise comparisons, and we will use it as a simple approximate model for analysis. Let

$$\begin{aligned} \mathbf{Y} &= [Y_{12}\, Y_{13}\ \ldots Y_{n-1n}]^t \\ \mathbf{X} &= \left(\mathbf{1}, [x_{12}\ x_{13}\ \ldots x_{n-1n}]^t\right) \end{aligned} \tag{4.1}$$

with $\mathbf{1}$ being a column vector of ones, $\mathbf{P}$ is a design matrix that has the form of:

$$\mathbf{P} = [P_{12}\ P_{13}\ \ldots P_{n-1n}]^t$$

where each P_{ij} row is a vector of zeroes of length n, that has ones in the i and j indices of that row,

$$\begin{aligned} \mathbf{a} &= [a_1 \; a_2 \; \ldots \; a_n]^t \\ \epsilon &= [\epsilon_{12} \; \epsilon_{13} \; \ldots \epsilon_{n-1n}]^t , \end{aligned} \tag{4.2}$$

and

$$\boldsymbol{\beta} = [\beta_0 \; \beta_1]^t$$

We can then rewrite the model in the corresponding matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Pa} + \epsilon.$$

This model is closely related to the models proposed in Gantz and Saunders (A1), O'Brien (A2), Armstrong et al. (A3), and Ausdemore et al. (A4). The models are also related to the correlation structures used by Schuckers (A5). The aforementioned authors' models apply to $Z_{ij} = Y_{ij} - \beta_0 - \beta_1 x_{ij}$, where the covariance models in the aforementioned papers would apply to the Z_{ij}. Additionally, we limit our focus to symmetric dissimilarity score (meaning the ordering of the pairs does not affect the final score), which is not the case for Schuckers work. In the context of the proposed model, we are interested in the following hypotheses:

H_0**:** $\beta_1 = 0$

H_1**:** $\beta_1 \neq 0$

specifically, we are interested in testing H_1 with a least squares-based approach. If we were to use a simple least squares approach, it would ignore the correlation structure, resulting in highly inflated type one error (see Figure 4.2 below for results from a basic simulation). In light of this concern, we have chosen to use a generalized least squares approach taking advantage of dependencies between the $Y_{ij}s$. As noted in Gantz and Saunders (A1), the dependency structure (in terms of the eigenvectors of the covariance matrix of the error terms is completely determined by the matrix $\mathbf{PP}^t$. Following Gantz and Saunders, we will consider three subspace vectors of $\mathbf{PP}^t$; the first corresponds to a projection space that is proportional to a vector of ones, the second is a projection onto a space spanned by the eigenvectors of $\mathbf{PP}^t$ that are orthogonal to the vector of ones, and the final space corresponds to the Null-space of $\mathbf{PP}^t$. Note that the projections for these spaces are completely determined by the known matrix $\mathbf{PP}^t$. (This is analogous to a generalized least squares approach, in each subspace, to regression modeling. See Christensen (A6) or similar texts for details.) We will refer to the swecond set of projections as the Writer-space and the corresponding set of estimates as the riter-based estimates and p-values. The third space will be referred to as the Null-space and the corresponding set of estimates as the Null-based estimates and p-values.

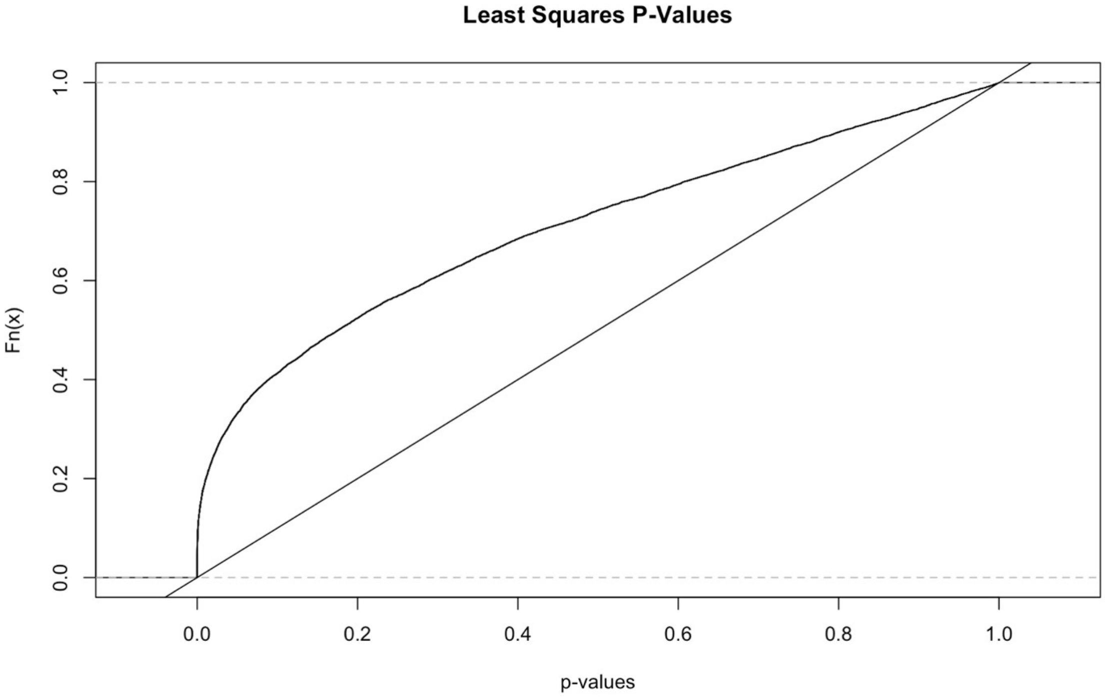


Figure 4.2: Simple least squares p-values for testing if $\beta_1 = 0$

4.4.3

Concerns about the Null-space estimation

Null distributions

To test the performance of the proposed linear models when there is no relationship between the x covariates and the response Y, we performed a basic Monte Carlo simulation. We considered a set of 33 independent and identically distributed (i.i.d.) multivariate normal random vectors in fourth dimension Euclidean space ($\mathbb{R}^4$) with a mean vector of zeroes and an identity co-variance matrix. The simulated response variable, Y_{ij}, is the pairwise Euclidean distance between the first two elements of the i^{th} and j^{th} normal vectors. The simulated covariate, x_{ij}, is the pairwise Euclidean distance between the third and fourth elements of the i^{th} and j^{th} normal vectors. Then, we performed the simple least squares regression in addition to regression in both the Writer- and Null-spaces, as described above, to obtain the corresponding regression estimates and p-values. Since each element of these four-dimensional vectors are uncorrelated, i.e., simulated using an identity covariance matrix, then $\beta_1 = 0$ for this simulation. We repeated this process 10,000 times (for a total of 10,000 Monte Carlo simulations), recorded the p-value for each iteration of the simulation, and summarized the distribution of the p-values with the corresponding empirical cumulative distribution function (ECDF). It is a well-known fact that the distribution of p-values under the null hypothesis of $\beta_1 = 0$ should be uniform. Therefore, we plotted the Uniform cumulative distribution function (CDF) on the same plot for reference. If the p-values for our test are performing reasonably, the ECDF of the p-values should overlay the uniform CDF.

From this simulation, the null distribution of p-values derived from the simple least squares approach behaved poorly, exhibiting strong departures from the Uniform CDF (see Figure 4.2). In contrast, the null distribution of p-values based on the Writer-space estimates perform exactly as we would expect (see Figure 4.3). Unfortunately, the null distribution of p-values in the Null-space seems to be conservative for lower ranges and liberal for the larger values (see Figure 4.4). This result did not "washout" as we increased sample size. We expect that the approximate distribution for the test statistic is related to the Chi-squared family and not the more commonly encountered normal family. Due to our lack of understanding of the null distribution for the p-values in the Null-space and the nice behavior of the null distribution of p-values in the Writer-space, we chose to focus on results from the Writer-space in this paper.

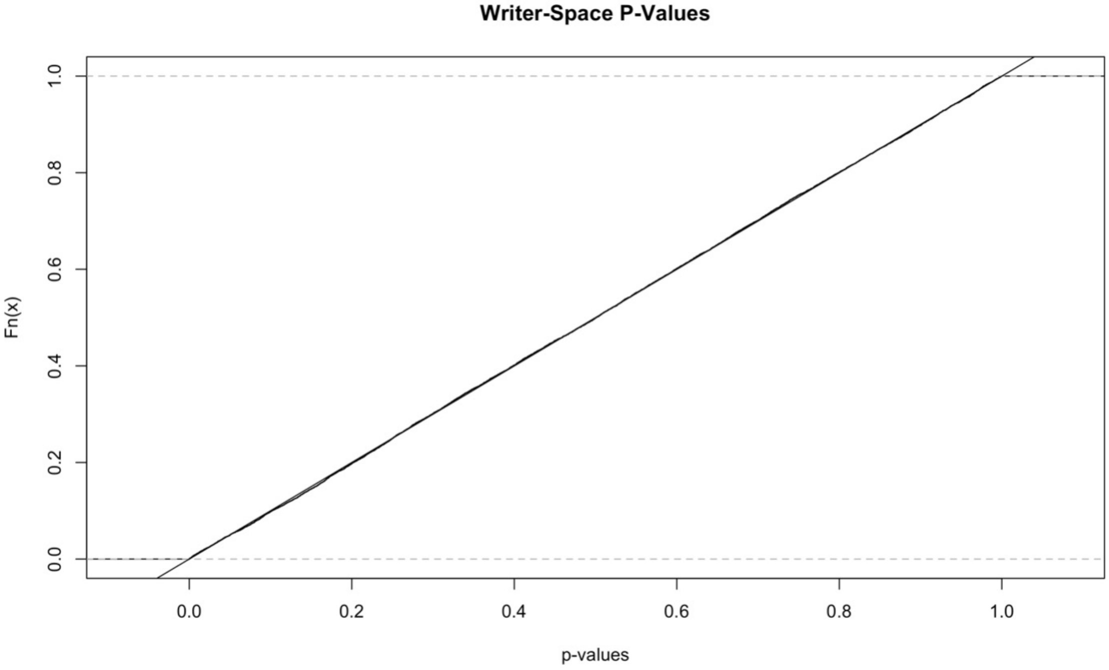


Figure 4.3: Writer-space p-values for testing if $\beta_1 = 0$

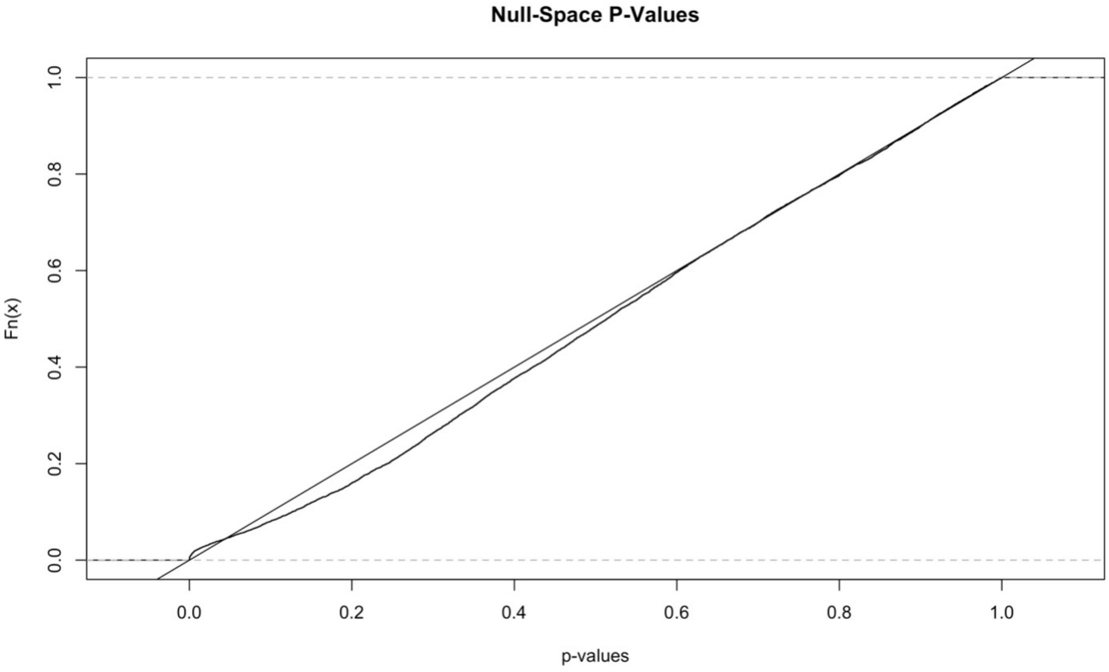


Figure 4.4: Null-space p-Values for testing if $\beta_1 = 0$

Power

As before, to test the power of the proposed linear models when there is a relationship between the covariates, x, and the response, Y, we performed a sequence of Monte Carlo simulations. However, for this simulation, we considered a set of 33 i.i.d. multivariate normal random vectors in $\mathbb{R}^4$ with a mean vector of zeroes and with a covariance matrix Σ. The diagonal elements of Σ are one with the off-diagonal elements being equal to ρ. The simulated response variable, Y_{ij}, is the pairwise Euclidean distance between the first two elements of the i^{th} and j^{th} normal vectors. The simulated covariate, x_{ij}, is the pairwise Euclidean distance between the third and fourth elements of the i^{th} and j^{th} normal vectors. Then, we performed the regression in the Writer-space, as described above, to obtain the Writer-based regression estimates and p-values. Because each element of these four-dimensional vectors is correlated, i.e., the off-diagonal elements of Σ are non-zero, then $\beta_1 \neq 0$ for this simulation. We generated 1,000 simulations, recorded the p-value for each iteration of the simulation, and summarized the power at a significance level of 0.05. We performed the power studies for $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and the results are summarized in Table 4.4 below. As shown in Table 4.4, in this study, by the time $\rho = 0.6$, the test has approximately 80% power. These results confirmed our decision to restrict testing to the Writer-space only.

Rho	Simulated Power
0.1	0.05
0.2	0.08
0.3	0.15
0.4	0.3
0.5	0.5
0.6	0.79
0.7	0.95
0.8	0.99
0.9	1

Table 4.4: Power study results

REFERENCES

A1. Gantz DT, Saunders S. Quantifying the effects of database size and sample quality on measures of individualization validity and accuracy in forensics. Washington, DC: U.S. Department of Justice; 2014 Mar. Report No.: 2009-DN-BX-K234 248670.

A2. O'Brien A. A kernel based approach to determine atypicality [dissertation]. Brookings, SD: South Dakota State University; 2017.

A3. Armstrong DE, Neumann C, Saunders CP, Gantz DT, Miller JJ, Stoney DA. Kernel-based methods for source identification using very small particles from carpet fibers. Chemom Intel Lab Syst. 2017;160:99-209.
doi: 10.1016/j.chemolab.2016.10.004.

A4. Ausdemore MA, Neumann C, Saunders CP, Armstrong D, Muehlethaler C. Two-stage approach for the inference of the source of high-dimensional and complex chemical data in forensic science. J Chemom. 2021;35(1):e3247. doi: 10.1002/cem.3247.

A5. Schuckers ME. Computational methods in biometric authentication: Statistical methods for performance evaluation. London, U.K.: Springer-Verlag London; 2010. p. 48-51.

A6. Christensen R. Plane answers to complex question, 4th edn. New York, NY: Springer-Verlag; 2011. p. 43.

CHAPTER 5

Regression Models for Comparing Metrics

This chapter contains the work completed on this research line since the publication of Fuglsby et al. [8]. Parts of the work in this chapter was presented at the South Dakota State University Data Science Symposium in 2022 as a poster (see Appendix 2), however is updated to reflect advancements on the mathematical development. One such update is how the regression modeling of the Euclidean distance scores behaves for different numbers of elements of a vector (EOV) that make up the VOS. For larger EOV, this allows us to use the Null-space mentioned in the Appendix of Fuglsby et al. [8] and Chapter 4, allowing us to treat the data sets with a sample size of $N = \binom{n}{2}$ instead of n, where n represents the number of writers.

For the Monte Carlo simulations in [8] and in Chapter 4, we assumed there were two EOV in the vector of scores (VOS) for each object. We increased the number of EOV that make up the input VOS, and included a step-wise increase in our simulations. As we increased the number of EOV that make up the VOS, the Null-space p-value distribution (under the null hypothesis) converges to a Uniform distribution. The distribution of the simulated Writer-space p-values showed no change, as well as the distribution of the p-values for the simple linear regression simulation.

Increasing the number of EOV in the VOS was also applied to the power study in [8] and in Chapter 4. Under the alternate hypothesis that there is a relationship between the scores of the two features being tested, the power study showed that the Null-space will catch this relationship a large portion of the time, and the more

EOV in the score, the quicker it gains power. The power in the Writer-space increases at a slower rate, however it does gain power as the number of EOV in the score increase. These results are summarized in Tables 5.1 - 5.3.

The next step after the discovery of increasing the intrinsic dimension is to generalize the process. The EOV used can be any subset of the intrinsic dimensionality.

5.1

Setting Up the Models

In this set of experiments, we have n potential sources. Let O_i denote the sample from the i^{th} source, for $i = 1, 2, \ldots, n$. For each of the samples, suppose we have a VOS for each feature in the output of a feature extraction system. To compare the same feature in the i^{th} and j^{th} samples, take the Euclidean distance between the corresponding VOS. Our goal is to measure the marginal relationship between each of the feature Euclidean distances.

Let $O_i, i = 1, \ldots, n$ be a set of $i.i.d.$ random elements in O. Let $Y^{\{m\}} : O \times O \to \mathbb{R}$ be a symmetric map such that m is the intrinsic dimension of O_i and O_j. Let $x^{\{q\}} : O \times O \to \mathbb{R}$ be a symmetric map such that q is the intrinsic dimension of O_i and O_j. For writers i and j, $1 \leq i < j \leq n$, define the following:

$$\begin{aligned}
Y_{ij}^{\{m\}} &= Y^{\{m\}}(O_i, O_j) = Y^{\{m\}}(O_j, O_i), \\
\mathbf{Y}^{\{m\}} &= [Y_{12}^{\{m\}}, Y_{13}^{\{m\}}, \ldots, Y_{n-1n}^{\{m\}}]^t, \\
x_{ij}^{\{q\}} &= x^{\{q\}}(O_i, O_j) = x^{\{q\}}(O_j, O_i), \\
\mathbf{x}^{\{q\}} &= (x_{12}^{\{q\}}, x_{13}^{\{q\}}, \ldots, x_{n-1n}^{\{q\}})^t, \\
\mathbf{X}_{ij}^{\{q\}} &= [1\ x_{ij}^{\{q\}}]^t \text{ a } 2 \times 1 \text{ column vector}, \\
\mathbf{X}^{\{q\}} &= [\mathbf{X}_{12}^{\{q\}}, \mathbf{X}_{13}^{\{q\}}, \ldots, \mathbf{X}_{n-1n}^{\{q\}}]^t = (\mathbf{1}, \left[x_{12}^{\{q\}}\ x_{13}^{\{q\}} \ldots x_{n-1n}^{\{q\}}\right]^t).
\end{aligned}$$

Let $Y_{ij}^{\{m\}}$ be a VOS for a given feature, and $x_{ij}^{\{q\}}$ be a VOS for a different feature. We want to use a least-squares based approach to test the Null hypothesis:

$$
\begin{aligned}
H_0 &: \beta_1^{\{m,q\}} = 0 \\
H_1 &: \beta_1^{\{m,q\}} \neq 0
\end{aligned}
$$

A naive simple least squares model is

$$E(Y_{ij}^{\{m\}}|\mathbf{X}_{ij}^{\{q\}}) = \beta_0^{\{m,q\}} + \beta_1^{\{m,q\}} x_{ij}^{\{q\}}, \tag{5.1}$$

in matrix form:

$$E(\mathbf{Y}^{\{m\}}|\mathbf{X}^{\{q\}}) = \mathbf{X}^{\{q\}}\boldsymbol{\beta}^{\{m,q\}}. \tag{5.2}$$

A simple least squares approach ignores the correlation structure of pairwise comparisons in the data set, which leads to an inflated type 1 error for increasing number of EOV in the response score (see Figure 5.1). To account for this, we explore spectral decomposition of the pairwise score structure.

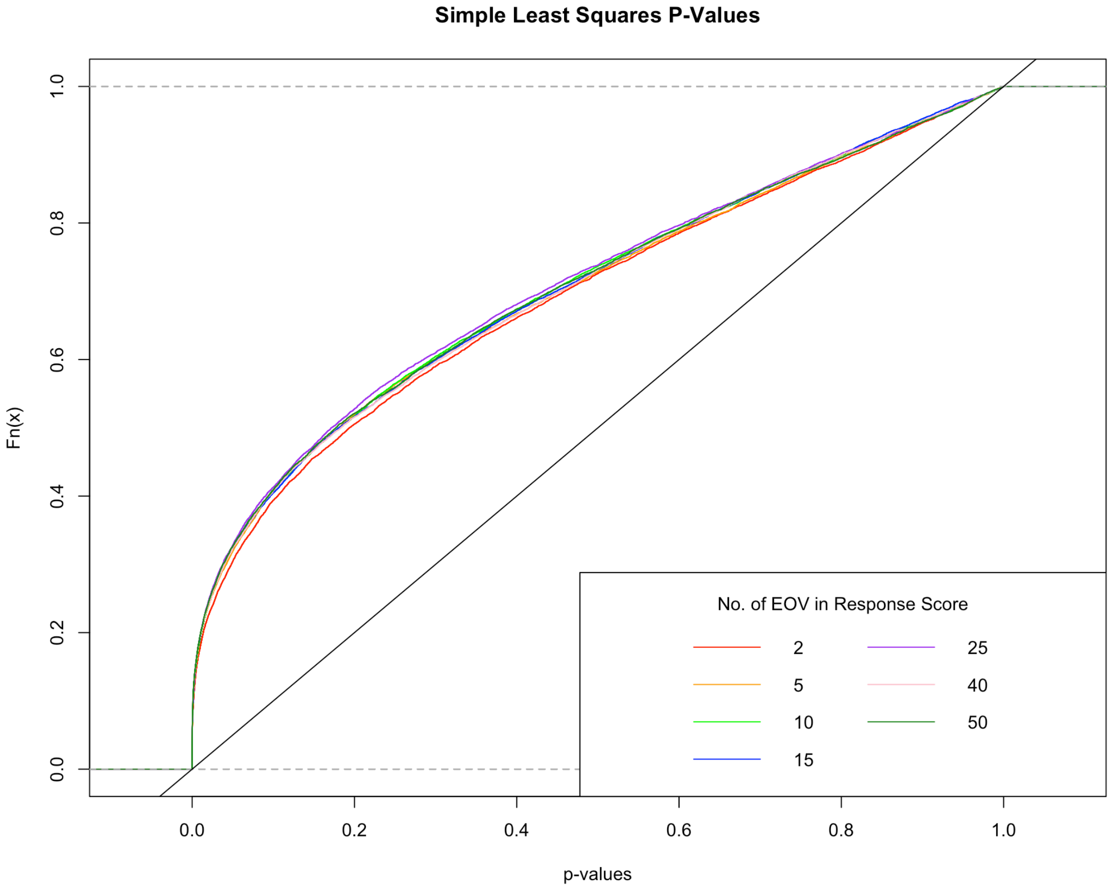


Figure 5.1: Monte Carlo simulation results of simple least-squares regression performed on sets of m-dimensional multivariate Normals.

Define a design matrix

$$\mathbf{P} = \left[\mathbf{P}_{12}\,\mathbf{P}_{13}\ldots\mathbf{P}_{n-1n}\right]^t, \tag{5.3}$$

where $\mathbf{P}_{ij}$ is a vector of zeroes with ones in places i and j.

$$\begin{aligned}\mathbf{a}^{\{m,q\}} &= [a_1^{\{m,q\}}, a_2^{\{m,q\}}, \ldots, a_n^{\{m,q\}}]^t \\ \boldsymbol{\epsilon}^{\{m,q\}} &= [\epsilon_{12}^{\{m,q\}}, \epsilon_{13}^{\{m,q\}}, \ldots, \epsilon_{n-1n}^{\{m,q\}}]^t \\ \boldsymbol{\beta}^{\{m,q\}} &= \left[\beta_0^{\{m,q\}}\ \beta_1^{\{m,q\}}\right]^t\end{aligned} \tag{5.4}$$

Note that $Cov(\mathbf{Y}^{\{m\}} - \mathbf{X}^{\{q\}}\boldsymbol{\beta}^{\{m,q\}}) = \mathbf{PP}^t\sigma^2_{a\{m,q\}} + \sigma^2_{\epsilon\{m,q\}}\mathbf{I}$. A model that shares the same first two moments is

$$\mathbf{Y}^{\{m\}} = \mathbf{X}^{\{q\}}\boldsymbol{\beta}^{\{m,q\}} + \mathbf{Pa}^{\{m,q\}} + \boldsymbol{\epsilon}^{\{m,q\}}, \tag{5.5}$$

where $a_i^{\{m,q\}} \overset{i.i.d}{\sim} N(0, \sigma^2_{a\{m,q\}}),\ i = 1, \ldots, n$ and $\epsilon_{ij}^{\{m,q\}} \overset{i.i.d}{\sim} N(0, \sigma^2_{\epsilon\{m,q\}}),\ 1 \leq i < j \leq n$. Note that $\sigma_{a\{m,q\}}$ and $\sigma_{\epsilon\{m,q\}}$ are functions of $\beta^{\{m,q\}}$. Using the design matrix $\mathbf{P}$, we can use a least squares approach that takes into account the pairwise structure in our data set.

5.2

Dependency Structures and Error Terms

Instead of using a simple least squares approach, we use a generalized least squares (GLS) approach taking advantage of dependencies between the $Y_{ij}^{\{m\}}$s. The dependency structure (in terms of the eigenvectors of the covariance matrix) of the error terms is completely determined by the matrix **PP**^{t} (see Appendix 1, rewritten from Gantz and Saunders [10] to reflect current notational conventions.) Following Gantz and Saunders, we will consider three subspace vectors of **PP**^{t}:

1. A projection space proportional to a vector of ones.
2. A projection onto a space spanned by the eigenvectors of **PP**^{t}, orthogonal to the vector of ones. Referred to as the Writer-space and any corresponding estimates as Writer-based.
    a. Define this set of eigenvectors to be $\mathbf{E}_{\mathrm{W}}$.
3. The Null-space of **PP**^{t}. Corresponding estimates are referred to as Null-based estimates and p-values.

a. Define this set of eigenvectors to be $\mathbf{E}_{\mathrm{N}}$.

The projections for these spaces are completely determined by the known matrix $\mathbf{PP}^t$. (Analogous to a GLS approach, in each subspace, to regression modeling. See Christensen [4] or similar texts for details.) Assuming the linear model used is true, the eigenvalues of the three distinct spaces are difficult to estimate.

Using the eigenvectors of each space, we can fit regression models in each space. Define the projection of scores into the Writer-space as

$$\begin{aligned}\mathbf{Y}_{\mathrm{W}}^{\{m\}} &= \mathbf{E}_{\mathrm{W}}^{t}\mathbf{Y}^{\{m\}} \\ \mathbf{X}_{\mathrm{W}}^{\{q\}} &= \mathbf{E}_{\mathrm{W}}^{t}\mathbf{X}^{\{q\}},\end{aligned} \tag{5.6}$$

and the projection of scores into the Null-space as

$$\begin{aligned}\mathbf{Y}_{\mathrm{N}}^{\{m\}} &= \mathbf{E}_{\mathrm{N}}^{t}\mathbf{Y}^{\{m\}} \\ \mathbf{X}_{\mathrm{N}}^{\{q\}} &= \mathbf{E}_{\mathrm{N}}^{t}\mathbf{X}^{\{q\}}.\end{aligned} \tag{5.7}$$

Then we can fit the regression models for the Writer- and Null-space.

$$\begin{aligned}E_{\mathrm{W}}Y^{\{m\}} &= E_{\mathrm{W}}X^{\{q\}} + \epsilon_{\mathrm{W}} \\ E_{\mathrm{N}}Y^{\{m\}} &= E_{\mathrm{N}}X^{\{q\}} + \epsilon_{\mathrm{N}}\end{aligned} \tag{5.8}$$

5.3

Simulation Under the Null Hypothesis

Following O'Brien [13], a simulation study has been developed to test the distributions of the p-values when the number of EOV that go into the score increases. This is tested for the simple least-squares model and for models using the design matrix $\mathbf{P}$. As the intrinsic dimension of the EOV in the response score, m, increases, $Y_{ij}^{\{m\}} - \beta_0^{\{m,q\}}\mathbf{X}_{ij}^{\{q\}}$ appears to follow a multivariate Normal distribution, characterized by

$$Y_{ij}^{\{m\}} - \beta_0^{\{m,q\}}\mathbf{X}^{\{\mathbf{q}\}}{}_{ij} = a_i^{\{m,q\}} + a_j^{\{m,q\}} + \epsilon_{ij}^{\{m,q\}}, \tag{5.9}$$

where $a_i^{\{m,q\}}$ are i.i.d. Normal random variables with mean 0 and variance $\sigma_a^{\{m,q\}}$, and $\epsilon_{ij}^{\{m,q\}}$ are i.i.d. Normal random variables with mean 0 and variance $\sigma_\epsilon^{\{m,q\}}$ and where $\beta_0^{\{m,q\}}$ represents the true coefficient and not the intercept. If this model holds under the null hypothesis, then the least squares fits in the Writer-space and Null-space will be exact and the corresponding p-values should be Uniform.

The results of Algorithm 1 are shown in Figures 5.2 - 5.4. Figure 5.2 shows the ECDFs of the p-values from the Writer-space. These ECDFs overlay a Uniform CDF for every k. Since we are testing the null hypothesis, we feel comfortable using the Writer-space for modeling the pairwise VOS scores. In Figures 5.3 and 5.4, the p-values from the Null-space converge approximately to a Uniform CDF as k increases.

Algorithm 1: Monte Carlo Algorithm for Testing the Null Hypothesis - Increasing Features

```
for $k$ in $\{4, 7, 12, 17, 27, 42, 52\}$ do
    for 10,000 times do
        1. Generate 30 i.i.d. $MVN(\mathbf{0}, \mathbf{I}_{kxk})$ random variables in $\mathbb{R}^k$.
        2. $x_{ij}^{\{1\}}$ - the pairwise Euclidean distance between the first two EOV
           of the $i^{th}$ and $j^{th}$ random variables.
        3. $Y_{ij}^{\{k\}}$ - the pairwise Euclidean distance between the $3^{rd}$ through $k^{th}$
           EOV of the $i^{th}$ and $j^{th}$ random variables.
        4. Fit a simple least squares regression in the Writer-space:
           $\mathbf{Y}_{\mathrm{W}}^{\{k\}} = \mathbf{X}_{\mathrm{W}}^{\{1\}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{W}}$.
        5. Fit a simple least squares regression in the Null-space:
           $\mathbf{Y}_{\mathrm{N}}^{\{k\}} = \mathbf{X}_{\mathrm{N}}^{\{1\}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{N}}$.
        6. Record $k$ and the resulting $p$-value for $\hat{\beta}_1$ from the Writer-space
           regression model.
        7. Record $k$ and the resulting $p$-value for $\hat{\beta}_1$ from the Null-space
           regression model.
```

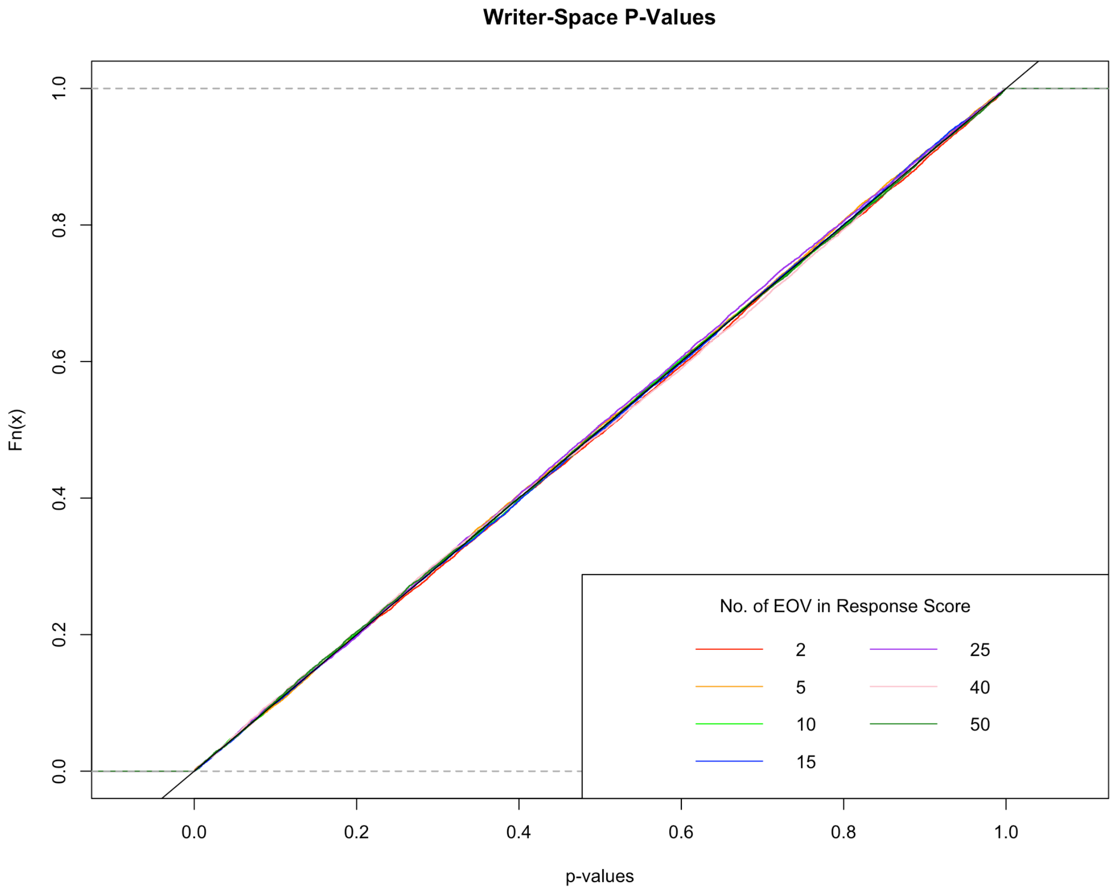


Figure 5.2: ECDFs of p-values in the Writer-space as k increases, resulting from Algorithm 1.

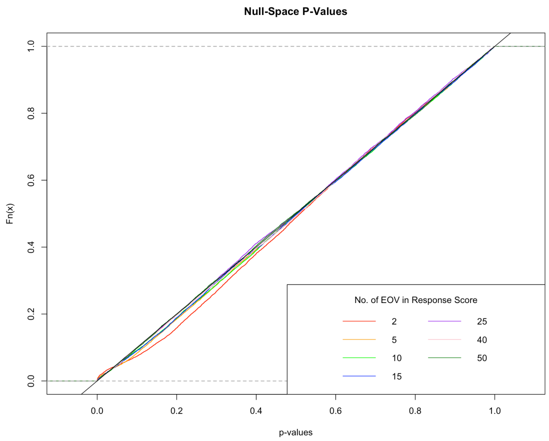


Figure 5.3: ECDFs of p-values in the Null-space as k increases, resulting from Algorithm 1.

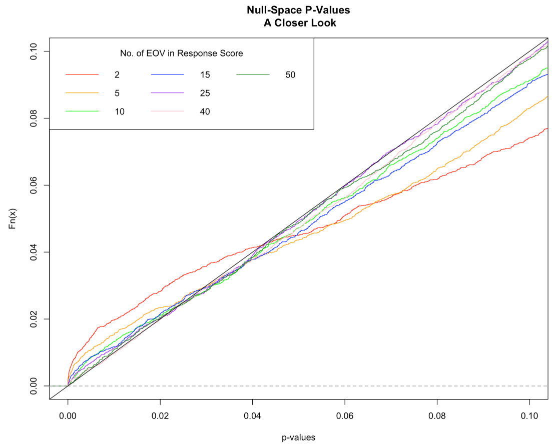


Figure 5.4: Close-up of the lower tail of Null-space p-values CDF. Note the colors red, orange, and green are furthest from the Uniform CDF. These correspond to a lower number of features, k that make up the response score.

5.3.1

Power Study

Before we apply the modeling in the Writer-space and Null-space to the handwriting data set, we want to test the performance of the proposed linear models when there is a relationship between the $\mathbf{X}^{\{q\}}$ covariates and the response $\mathbf{Y}^{\{m\}}$ ($\beta_1^{\{m,q\}} \neq 0$). Due to the nature of the Null-space p-values converging to a Uniform CDF as the number of features in the response score increase, this power study will consider both the Writer-space and the Null-space as the number of features in the response score increase (following O'Brien [13].) See Algorithm 2 for details.

Algorithm 2: Power Study

```
for $k$ in $\{4, 7, 12, 17, 27, 42, 52\}$ do
    for $\rho = \{0, 0.1, \ldots, 0.9\}$ do
        for 10,000 times do
            1. Generate 30 i.i.d. $MVN(\mathbf{0}, \mathbf{\Sigma}_{k \times k})$ random variables in $\mathbb{R}^k$,
               $\mathbf{\Sigma}_{k \times k} = (1 - \rho)\mathbf{I}_{k \times k} + \rho \mathbf{1}\mathbf{1}^t$.
            2. $x_{ij}^{\{1\}}$ - the pairwise Euclidean distance between the first two
               EOV of the $i^{th}$ and $j^{th}$ random variables.
            3. $Y_{ij}^{\{k\}}$ - the pairwise Euclidean distance between the $3^{rd}$
               through $k^{th}$ EOV of the $i^{th}$ and $j^{th}$ random variables.
            4. Fit a simple least squares regression in the Writer-space:
               $\mathbf{Y}_{\mathrm{W}}^{\{k\}} = \mathbf{X}_{\mathrm{W}}^{\{1\}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{W}}$.
            5. Fit a simple least squares regression in the Null-space:
               $\mathbf{Y}_{\mathrm{N}}^{\{k\}} = \mathbf{X}_{\mathrm{N}}^{\{1\}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{N}}$.
            6. Record $k$ and the resulting $p$-value for $\hat{\beta}_1$ from the
               Writer-space regression model.
            7. Record $k$ and the resulting $p$-value for $\hat{\beta}_1$ from the Null-space
               regression model.
```

For each feature set k and value of ρ, the power was calculated at a significance level of 0.05. Results of the power study in the Writer-space are in Tables 5.1, 5.2, and 5.3. These results indicate that under the alternative hypothesis, the Writer-space does not have much power for a few number of EOV contributing to the score except for larger values of ρ (i.e. a stronger relationship exists between the two features being compared). The Writer-space does gain power as the number of EOV increase, however it is again for larger values of ρ. The Null-space contains more power, even for a few number of EOV contributing to the score and for smaller values of ρ.

Rho	Writer-Space Power	Null-Space Power
0.0	0.049	0.047
0.1	0.057	0.129
0.2	0.104	0.421
0.3	0.213	0.767
0.4	0.415	0.939
0.5	0.661	0.992
0.6	0.849	0.999
0.7	0.969	1.000
0.8	0.998	1.000
0.9	1.000	1.000

Table 5.1: Power Study results for **5** EOV in the response score.

Rho	Writer-Space Power	Null-Space Power
0.0	0.051	0.048
0.1	0.070	0.244
0.2	0.181	0.698
0.3	0.382	0.931
0.4	0.619	0.991
0.5	0.827	0.999
0.6	0.943	1.000
0.7	0.991	1.000
0.8	0.999	1.000
0.9	1.000	1.000

Table 5.2: Power Study results for **25** EOV in the response score.

Rho	Writer-Space Power	Null-Space Power
0.0	0.049	0.048
0.1	0.082	0.320
0.2	0.222	0.769
0.3	0.430	0.953
0.4	0.670	0.994
0.5	0.849	0.999
0.6	0.952	1.000
0.7	0.993	1.000
0.8	0.999	1.000
0.9	1.000	1.000

Table 5.3: Power Study results for **50** EOV in the response score.

5.4

Application to Handwriting Samples

5.4.1

FLASH ID® Shape Codes

The work presented in this section was originally presented at the 2022 SDSU Data Science Symposium[1] as a poster. This further discusses the FLASH ID® system, including another output that has not been previously mentioned.

The FLASH ID® system is considered a "black-box" writer identification system due to the hidden set of algorithms that extract feature sets from a given handwriting sample. The FLASH ID® system segments the handwriting on the page using these hidden algorithms, and since it is a language-independent system, the segmentation is not based on the individual alpha-numeric characters. One of the outputs is the shape code and provides the orientation of the segmented writing. This orientation is a " numeric encoding based on the compass direction between the prime vertex and all other vertices." (Quote from the FLASH ID® User Manual [16].) The assignment of the vertices are hidden to the user, however the shape-code encoding is known (see Figure 5.5). From the prime vertex, the next vertex receives a number based on the compass direction in Figure 5.5. Figure 5.6 (from the FLASH ID® User Manual [16].) contains examples of shape code encoding.

[1] https://openprairie.sdstate.edu/datascience_symposium/

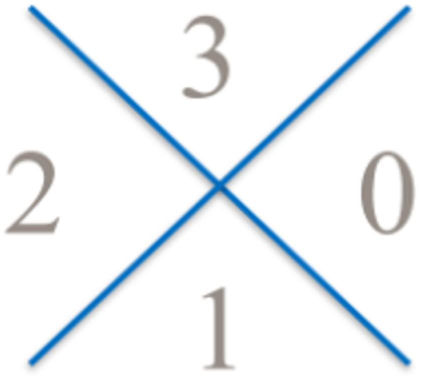


Figure 5.5: The orientation of the shape code encoding directions. A vertex going "north" would be assigned the number "3". (Image from the FLASH ID® User Manual [16].)

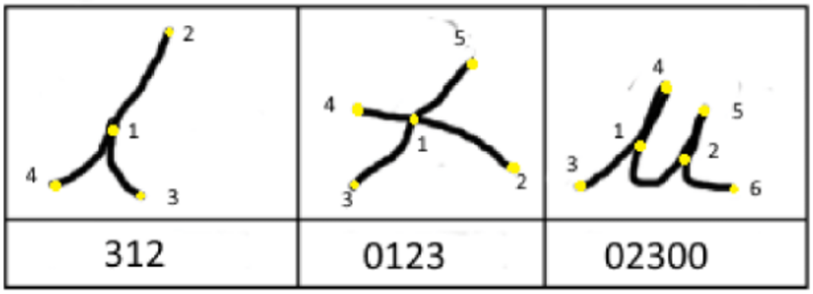


Figure 5.6: Example of three shape codes. The images show the order of the vertices (as determined by the FLASH ID®) and the number below each image is the shape code encoding. (Image from the FLASH ID® User Manual [16].)

For this experiment, we used a different set of data collected by West Virginia University (WVU) on behalf of the FBI Laboratory via a convenience sample. The subset of 30 writing samples used were selected based on their image quality. Each writing sample had a different writer ($n = 30$); the writers were given suggested writing prompts, their word choice was their own.

These writing samples were scanned in to FLASH ID® and the resulting shape codes for the writing segments were recorded. Only four shape codes were found in each writing sample, which will be the four we focus on. Due to the nature of the data collection, we did not have access to other outputs of the FLASH ID® system. The author uploaded pages of letters and numbers to the FLASH ID® system to parse out what the resulting four shape codes may look like. The four found are shown in Figures 5.7 - 5.10. These images are provided by the FLASH ID® system. Each image contains two letters, one with the FLASH ID® segmentation and vertices, and one without to show the writing. Note that the vertices are not labelled.

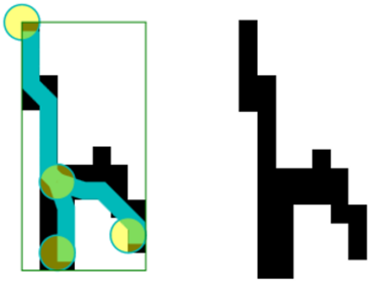


Figure 5.7: Shape code 013.

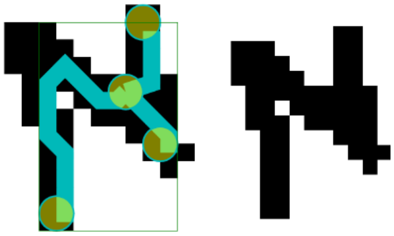


Figure 5.8: Shape code 113.

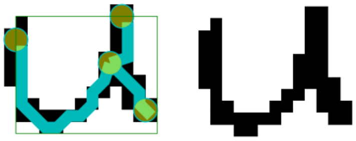


Figure 5.9: Shape code 123.

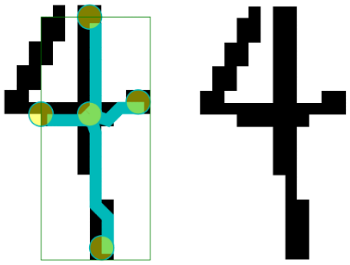


Figure 5.10: Shape code 0123.

The FLASH ID® system provides the VOS output, discussed in Chapter 2, for each shape code as well. For a given shape code, the Euclidean distance was calculated on each pair of the 30 writing samples, resulting in 435 Euclidean distances per shape code. Each shape code was then modeled against another shape code for both the Writer-space and Null-space transformations. The results of these models are found in Table 5.4.

Response	Explanatory	Writer-space p-values	Null-space p-values
123	013	0.031*	< 0.001*
123	113	0.027*	< 0.001*
123	0123	0.138	< 0.001*
013	113	0.019*	< 0.001*
013	0123	0.030*	< 0.001*
113	0123	<0.001*	< 0.001*

Table 5.4: Results of modeling the FLASH ID Shape Code VOS Euclidean distance using the Writer-space and Null-space transformations. The starred p-values are all less than $\alpha = 0.05$, indicating a rejection of the null hypothesis that $\beta_1 = 0$.

Table 5.4 shows that the majority of shape code pairings in the Writer-space and in the Null-space have significant p-values, indicating that the way the 30 writers write these shape codes are related. Due to the limited number of shape codes, our conclusions about the shape codes and the writing are also limited.

5.4.2

Handwriting Kinematics

The kinematics of handwriting refers to how a person uses the pen to write. The speed at which they write, the velocity changing as they curve a stroke, the pressure at which they push their pen down, and so on. Software, such as the MovAlyzeR® software, is used to measure the various features of how someone writes. Due to the type of measurements, the MovAlyzeR® software is considered a “white-box” algorithm.

The data for the handwriting kinematics is previously discussed in Chapters 3 and 4. Thirty-three writers were asked to write on a piece of paper, placed on top of a tablet with the MovAlyzeR® software installed, all using the same pen. They were asked to write six phrases from the London letter, each repeated five times, in their natural print and cursive style of handwriting, for a total of 60 lines.

5.5

Wasserstein Distance Score Development

The Wasserstein distance score (WDS) development is shown in Ommen et al. [14] and rewritten in Chapter 3. Here is a more in-depth development.

The WDS is a method developed by del Barrio et al. [3] that considers the L_2 distance between the quantile functions of two objects. Here we developed it to consider the L_2 distance between the posterior probabilities of two objects from performing Linear Discriminant Analysis (LDA). This process produces a univariate score between zero and one, with scores close to one implying little to no overlap between the two quantiles (i.e. the two objects are more dissimilar.)

This score computes the L_2-distance between a pair of quantiles, Q. In an analogous manner to del Barrio et al. [3] we define the L_2 Wasserstein distance using the following equation:

$$w(O_i, O_j) = \left(\int_0^1 \left(Q_i(z) - Q_j(z) \right)^2 dx \right)^{1/2}. \tag{5.10}$$

Algorithm 3: Wasserstein Distance Score Development

```
for $i$ in $\{1, 2, \ldots, n-1\}$ do
    for $j$ in $\{j+1, \ldots, n\}$ do
        1. For the feature measurements of $O_i$ and $O_j$ from the MovAlyzeR® software output, perform LDA using the proportions of observations for each object as the prior.
        2. Take the posterior classification probability of each object belonging to the source (i.e. model class) of $O_i$. Each observation in each object will be given a posterior classification probability.
        3. Let $Q_i$ be the quantile of the posterior classification probabilities of the observations from $O_i$ belonging to the source of $O_i$.
        4. Let $Q_j$ be the quantile of the posterior classification probabilities of the observations from $O_j$ belonging to the source of $O_i$.
        5. Compute $w(O_i, O_j)$ from Eq. 5.10 to produce the score.
```

5.5.1

Results

For each pair of writers, the kinematics of the lines of writing samples with the same writing style and phrase were applied to the WDS process, producing one score per pair of lines per pair of writers. For a given writing style and phrase, the 25 scores per pair of writers were then averaged to produce one score for each pair of writers.

For a given phrase, writing style, stroke direction, and for each pair of MovAlyzeR® features, the pairwise scores were then projected into the Writer-space (Equation 5.6) and into the Null-space (Equation 5.7), and then modeled following Equation 5.8 where Y is now the vector for one set of pairwise scores for a given MovAlyzeR®

feature, and X is the vector for one set of pairwise scores for another MovAlyzeR® feature. Note that the superscripts of m and q do not apply here as we are not considering the intrinsic dimension of the handwritten objects.

There are a total of 24 writing categories composed of phrase, writing style, and stroke direction ($6 \times 2 \times 2$). For downstrokes we considered seven features, plus a score comprised of all features, resulting in $\binom{8}{2} = 28$ regression models. For upstrokes, we considered nine features, plus a score comprised of all features, resulting in $\binom{10}{2} = 45$ regression models. This resulted in a total of $(12 * 28) + (12 * 45) = 876$ regression models. For ease of understanding, Table 5.5 displays the proportion of p-values that are less than a chosen α-level of 0.05 for the 24 different writing categories. Under the null hypothesis that $\beta_1 = 0$, or that there is no relationship between two features, we would expect that this proportion would be about 0.05.

While Table 5.5 shows valuable information about the p-values, it leaves out information about the relationships between the pairs of features. The graphics in Figure 6.2 in Appendix 2 displays tanglegrams of all 24 writing categories. These plots compare a transformed p-value for a pair of MovAlyzeR® features from the Writer-space to the Null-space. The transformed p-value, $-2 * log(1 - p\text{-value})$ is interpreted in nearly the same way as the p-value, except the range of the transformation is now $(0, \infty)$. Note that, for our chosen α-level of 0.05, $-2 * log(1 - 0.05) \approx 0.1$.

	Downstrokes			
	Print		Cursive	
Phrase	Writer-Space	Null-space	Writer-Space	Null-Space
1	0.32	0.46	0.39	0.43
2	0.36	0.54	0.29	0.54
3	0.39	0.54	0.54	0.50
4	0.32	0.47	0.46	0.54
5	0.36	0.50	0.29	0.50
6	0.39	0.54	0.39	0.43

	Upstrokes			
	Print		Cursive	
Phrase	Writer-Space	Null-space	Writer-Space	Null-Space
1	0.27	0.49	0.24	0.42
2	0.31	0.51	0.36	0.36
3	0.31	0.53	0.60	0.64
4	0.24	0.35	0.44	0.49
5	0.42	0.40	0.33	0.53
6	0.56	0.44	0.42	0.51

Table 5.5: Proportion of the slope p-values < 0.05 from the regression models predicting pairs of MovAlyzeR® feature WDS for cursive writing and handprinting sample pairs for upstrokes and downstrokes in the Writer-space and Null-space. Note that the Downstrokes have an $n = 28$ and the Upstrokes have an $n = 45$.

It is not clear, however, if the regression modeling using the WDS metric will behave appropriately under the null and alternative hypotheses. Since the WDS metric does not use a similar set of input vectors as Algorithms 1 and 2, there needs to be a new Monte Carlo simulation developed that reflects the style of input the

MovAlyzeR® software provides.

Algorithm 4: Monte Carlo algorithm for testing the null hypothesis using the WDS

```
for a subset $n^*$ from $n$ writers do
    for $1 \leq i < j \leq n^*$ do
        1. From the mclust library in R, use Mclust function to estimate
           the mean and covariance parameters of the estimated mixture
           models for the Object $O_i$ from the $i^{th}$ writer.
        2. From the mvtnorm library in R, use the rmvnorm function to
           generate random observations from multivariate normal
           distributions with the estimated mean and covariance parameters
           from Step 1. The number of observations is based on $100 \times$
           proportion of observations Mclust assigned to each model. Call
           this new set of generated data $O_i^*$.
        3. Repeat Steps 1 and 2 for the Object $O_j$ from the $j^{th}$ writer, creating
           $O_j^*$.
        4. Calculate the WDS between the simulated $O_i^*$ and $O_j^*$ for each
           feature.
        for all pairs of MovAlyzeR® features do
            5. $x_{ij}$ - the WDS between $O_i^*$ and $O_j^*$ for the first feature.
            6. $Y_{ij}$ - the WDS between $O_i^*$ and $O_j^*$ for the second feature.
            7. Fit a simple least squares regression in the Writer space:
               $\mathbf{Y}_{\mathrm{W}} = \mathbf{X}_{\mathrm{W}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{W}}$.
            8. Fit a simple least squares regression in the Null space:
               $\mathbf{Y}_{\mathrm{N}} = \mathbf{X}_{\mathrm{N}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\mathrm{N}}$.
            9. Record the resulting p-value for $\hat{\beta}_1$ from the Writer-space
               regression model.
            10. Record the resulting p-value for $\hat{\beta}_1$ from the Null-space
                regression model.
```

Algorithm 4 was ran for Cursive Phrase 1 Upstrokes. The results for this simulation are shown in a tanglegram in Appendix 3. The interpretation of the tanglegram follows that of the Appendix 1 tanglegrams.

CHAPTER 6

Discussion and Future Research

6.1

Discussion

Black-box systems such as AIS for handwriting samples often provide the user with accurate information, however its decision-making process is typically unknown to the user. Pairwise comparison metrics, often used for biometric verification, introduce another level of complexity to be accounted for in the algorithms. The metrics chosen should be adapted to the style of output the system generates while also behaving properly under the null hypothesis. I have demonstrated how to construct metrics based off the output of the system at hand.

The FLASH ID® VOS Euclidean distance score from Chapter 2 showed it is useful for the development of a handwriting survey, and for the comparison to FDE opinions. We displayed that, in general, FDEs note stronger support for the correct proposition for pairs of handwriting samples. This suggests that FDEs display knowledge of how dissimilar two handwriting samples may be.

In Chapter 3, we showed the development and usefulness of the WDS from MovAlyzeR®. We ran models to observe how the R^2 values change when we remove sets of features based on their general categorization (spatial-geometric, temporal, and pressure) when predicting the strength of support from the survey results of FDEs. These results provide some information on what category of features do FDEs draw from.

Chapter 4 covers the comparison of the Euclidean distance score from the FLASH

ID® VOS and the WDS from MovAlyzeR®. Regression models were fit using the WDS from one of the three feature categories from MovAlyzeR®, spatial-geometric, temporal, or pressure as the explanatory variable, and the FLASH ID® VOS Euclidean distance scores as the response variable. Monte Carlo simulations were developed to test these models, and when the behavior under the null hypothesis (that there is no relationship between the FLASH ID® VOS Euclidean distance scores and the MovAlyzeR® WDS) proved to not behave (i.e. the ECDF of the p-values did not follow a Uniform CDF), then the work by Gantz and Saunders [10] for pairwise metrics were implemented, but only in the Writer-space. This allowed us to write another simulation to show that with their method, the p-values behaves under the null hypothesis. Thus, we modeled the FLASH ID® VOS Euclidean distance scores against the WDS of each of the MovAlyzeR® categories. The results showed us that the spatial-geometric category was correlated with the FLASH ID® VOS Euclidean distance scores for a majority of the phrase/stroke direction/writing style combinations. The temporal category had a few correlations, however the pressure category had zero.

Chapter 5 is a continuation of Chapter 4. The work by Gantz and Saunders [10] for these pairwise distance metrics is extended to the Null-space by taking into account the intrinsic dimension of the handwritten objects. This discovery led to adjusting the simulations from Chapter 4 to account for the intrinsic dimension. The higher the dimension, the more the ECDFs of the Null-space p-values converge towards the Uniform CDF.

I have also discussed another application where I looked at the FLASH ID® VOS scores for a different set of handwriting samples where I explored the shape code output of FLASH ID®. I also explored modeling the WDS of pairs of MovAlyzeR® features, and how to set up a simulation to spot relationships using the WDS.

6.2

Future Research

There are many areas of this research line I would like to explore, including theoretical properties dependent on the score method in use, a variety of simulations, algorithmic estimation for formal generalized least squares, and a number of applications unrelated to handwriting identification systems.

First, the finding from Section 5.3 concerning the Null-space suggests that there is a normality assumption that is becoming more reasonable as the intrinsic dimension of the response score increases.

Conjecture 1. *As the intrinsic dimension of the response score increases, the normality of the least squares model becomes more reasonable.*

The exact form of Conjecture 1 will depend on the the score function(s) used. See Armstrong [2] (p. 94-104) for a summary of the kernels used and simulations about the normality assumptions.

Here are the next set of simulations related to this research line that I would like to implement.

1. Test how increasing the sample size, n, affects the p-values of the Writer-space and Null-space when also increasing the number of features that go into the score.

2. Repeat the simulations with the Writer-space and Null-space where the features contributing to each score overlap:

   a) $Y_{ij}^{\{k\}}$ - the pairwise Euclidean distance between the 3^{rd} through k^{th} EOV of the i^{th} and j^{th} random variables.

b) $x_{ij}^{\{1\}}$ - the pairwise Euclidean distance between the 1^{st} through r^{th} EOV of the i^{th} and j^{th} random variables, with r increasing from 3 to k.

3. Repeat the simulations on both the Writer-space and Null-space while the number of features that go into the predictor score, x, also increases.

4. Repeat the simulations on both the Writer-space and Null-space where a single feature contributes to each score $Y_{ij} = |O_i^{\{1\}} - O_j^{\{1\}}|,\ x_{ij} = |O_i^{\{2\}} - O_j^{\{2\}}|$.

   a) Y_{ij} - the pairwise absolute difference between the first element of the i^{th} and j^{th} random variables.

   b) x_{ij} - the pairwise absolute difference between the second element of the i^{th} and j^{th} random variables.

5. Repeat the simulations where we model $Y^{\{k\}}$ using $x^{\{1\}}$, then model $x^{\{1\}}$ using $Y^{\{k\}}$.

I am also working on the algorithmic estimation for the formal generalized least squares. The following algorithm iteratively estimates the eigenvalues of the covariance matrix using REML estimates analogous to Gantz and Saunders [10]. The convergence criteria for this problem is still unknown and will be further explored with a collaboration with Dr. Jung-Han Kimn and his high performance computing group at South Dakota State University.

```
Algorithm 5: Iterative Estimation of Eigenvalues
  Data: A set of response scores, $\mathbf{Y}$, and a set of predictor scores, $\mathbf{x}$.
  Initialize $\hat{\boldsymbol{\beta}}^{(1)} = [\beta_0^1 \; \beta_1^1]^t$
  Initialize $k = 2$
  repeat
      1. Calculate $\mathbf{Y} - \hat{\boldsymbol{\beta}}^{(k-1)}\mathbf{X} = \mathbf{Pa} + \boldsymbol{\epsilon}$
      2. Calculate $Cov(\mathbf{Y} - \hat{\boldsymbol{\beta}}^{(k-1)}\mathbf{X}) = \mathbf{PP}^t\sigma_{\mathbf{a}}^2 + \sigma_{\boldsymbol{\epsilon}}^2\mathbf{I}_{N\times N}$
      3. Define $d_{ij}^{(k)} = Y_{ij} - \hat{\boldsymbol{\beta}}^{(k-1)}\mathbf{x}_{ij}$, and the vector of $d_{ij}^{(k)}$s as $\mathbf{d}^{(k)}$
      4. Estimate the variance of the $a_i$ terms $\hat{\sigma}_{\mathbf{a}}^{2(k)} = E_{\mathbf{W}}^t\mathbf{d}^{(k)}$
      5. Estimate the variance of the error terms $\hat{\sigma}_{\boldsymbol{\epsilon}}^{2(k)} = E_{\mathbf{N}}^t\mathbf{d}^{(k)}$
      6. Estimate the covariance $\hat{\boldsymbol{\Psi}}^{(k)} = \mathbf{PP}^t\hat{\sigma}_{\mathbf{a}}^{2(k)} + \hat{\sigma}_{\boldsymbol{\epsilon}}^{2(k)}\mathbf{I}_{N\times N}$
      7. Estimate $\hat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}}{argmin}(\mathbf{Y} - \hat{\boldsymbol{\beta}}^{(k-1)}\mathbf{X})\hat{\boldsymbol{\Psi}}^{(k)}(\mathbf{Y} - \hat{\boldsymbol{\beta}}^{(k-1)}\mathbf{X})$
      8. Set $k = k + 1$
  until convergence criteria is met;
```

Finally, this research can be applied to many different black- and white-box algorithms addressing different types of biometrics (and other uses beyond biometrics). I am working on applying this to data concerning blockchain transactions. The current methods employ different metrics that attempt to track trades between sets of wallets (see Akcora et al. [1]). One such metric may employ knowing users' accounts and wallets, thus knowing all of their transactions. Another metric may only know the transactions that occur between wallets. The methods proposed in this dissertation can be applied to these hypothetical metrics to measure the degree of "connectedness", or how well the second, less intrusive metric can predict the more intrusive metric. This line of research grew from the workshops associated with an NSF grant on blockchain data analysis that I served as co-principle investigator on[1].

[1] National Science Foundation Grant No. 2139349 Graph Theoretical Methods for Blockchain

Data Analysis. Division of Mathematical Sciences. Algorithms for Threat Detection.

APPENDIX 1

This Appendix is a rewrite of Gantz and Saunders [10] (p. 54-57), rewritten with the language and notational conventions used throughout this dissertation and included for clarity.

A Parametric Model for Comparing Pairs of Pairwise Scoring Methods:

Development of the Parametric Model:

Using the score development from Section 5.1, the parametric model from Equation 5.5 accounts for two scores between pairs of the same objects. For a single pair, that model is

$$Y_{ij} = x_{ij} + a_i + a_j + \epsilon_{ij},$$

where Y_{ij}, x_{ij}, a_i, ϵ_{ij} follow the definitions from Section 5.1, and the superscripts representing the EOV were removed for brevity. We can rewrite this model to follow Equation 5.5. The design matrix $\mathbf{P}$ from Equation 5.3 which has $N = \binom{n}{2}$ rows and n columns, and is incorporated into the matrix form of the parametric model.

Thus our model becomes

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{x} + \mathbf{P}\mathbf{a} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$, $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$ and $\mathbf{a}$ are as defined in Section 5.1.

We are interested in the $N \times N$ covariance matrix of $\mathbf{Y}$,

$$\boldsymbol{\Sigma} = \sigma_e^2 \mathbf{I}_N + \sigma_a^2 \mathbf{P}\mathbf{P}^t.$$

Note that $\mathbf{\Sigma}$ does not rely on the normality assumptions of the a_is and the ϵ_{ij}.

The two matrices in Equation A.1, are an example where $n = 6$ and $N = 15$. The left matrix represents the vector of Y_{ij}s, and the right matrix represents the resulting $\mathbf{P}$ matrix. For example, the fourth line of $\mathbf{P}$ has ones in columns 1 and 5 for selecting objects 1 and 5. Each column has five ones because each object is compared to the other five objects.

$$\mathbf{Y} = \begin{bmatrix} Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{23} \\ Y_{24} \\ Y_{25} \\ Y_{26} \\ Y_{34} \\ Y_{35} \\ Y_{36} \\ Y_{45} \\ Y_{46} \\ Y_{56} \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{A.1}$$

Eigenstructure of $\mathbf{\Sigma}$:

To find the eigenstructure of $\mathbf{\Sigma}$, we can first focus on the eigenstructure of $\mathbf{P}\mathbf{P}^t$. An easier task is to compute the eigenstructure of $\mathbf{P}^t\mathbf{P}$, which is shown in Equation A.2.

$$\begin{aligned}\mathbf{P}^t\mathbf{P} &= \begin{bmatrix} n-1 & 1 & \dots & 1 \\ 1 & n-1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & n-1 \end{bmatrix} \\ &= (n-2)\mathbf{I}_n + \mathbf{1}_n\mathbf{1}_n^t \end{aligned} \tag{A.2}$$

$\mathbf{1}_n\mathbf{1}_n^t$ has one eigenvalue of n with eigenvector of $\mathbf{1}_n/(\mathbf{1}_n^t\mathbf{1}_n)^{1/2}$ and $n-1$ eigenvalues of zero with every eigenvector orthogonal to $\mathbf{1}_n$. For $\mathbf{P}^t\mathbf{P}$, there is one eigenvalue of $2(n-1)$ and $n-1$ eigenvalues of $(n-2)$. Note that the non-zero eigenvalues of $\mathbf{P}^t\mathbf{P}$ are the same as the non-zero eigenvalues of $\mathbf{P}\mathbf{P}^t$. (See Roy, 1954.) This leads to the following set of eigenvalues of $\mathbf{P}\mathbf{P}^t$: one eigenvalue of $2(n-1)$, $n-1$ eigenvalues of $(n-2)$, and $N-n$ eigenvalues of zero.

We can use this information to obtain the eigenvectors and eigenvalues of $\mathbf{\Sigma} = \sigma_e^2\mathbf{I}_N + \sigma_a^2\mathbf{P}\mathbf{P}^t$,

1. 1 eigenvector ($\mathbf{e}_1 = \mathbf{1}_N/\sqrt{N}$) with eigenvalue $\lambda_1 = \sigma_\epsilon^2 + 2(n-1)\sigma_a^2$
2. $n-1$ eigenvectors ($\mathbf{e}_2$ to $\mathbf{e}_n$) with eigenvalue $\lambda_2 = \sigma_\epsilon^2 + (n-2)\sigma_a^2$
3. $N-n$ eigenvectors ($\mathbf{e}_{n+1}$ to $\mathbf{e}_N$) with eigenvalue $\lambda_3 = \sigma_\epsilon^2$

Since σ_a^2 and σ_ϵ^2 are both greater than zero, implying each eigenvalue is greater than zero, $\mathbf{\Sigma}$ has full rank. Also note that $\mathbf{e}_v^t\mathbf{1}_N = 0$ for all $v > 1$.

U-Process Development

There is a U-process structure involving β that will serve as a motivation to use the sets of eigenvectors (the Writer-Space and Null-Space) separately.

The following U-process development is a generalization, and so the EOV sets m and q will not be considered. A simplified version of the variables from Section 5.1 are considered.

For writers i and j, $1 \leq i < j \leq n$, define the following:

$$\begin{aligned} Y_{ij} &= Y(O_i, O_j) = Y(O_j, O_i), \\ \mathbf{Y} &= [Y_{12}, Y_{13}, \ldots, Y_{n-1n}]^t, \\ x_{ij} &= x(O_i, O_j) = x(O_j, O_i), \\ \mathbf{x} &= (x_{12}, x_{13}, \ldots, x_{n-1n})^t, \\ \mathbf{X}_{ij} &= [1\ x_{ij}]^t \text{ a } 2 \times 1 \text{ column vector}, \\ \mathbf{X} &= [\mathbf{X}_{12}, \mathbf{X}_{13}, \ldots, \mathbf{X}_{n-1n}]^t = (\mathbf{1}, [x_{12}\ x_{13} \ldots x_{n-1n}]^t). \end{aligned}$$

We are interested in the behavior of $\hat{\beta}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \underset{\beta \epsilon \mathbb{R}^{p \times 1}}{argmin} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}. \end{aligned} \tag{A.3}$$

Define a U-process of β as

$$U_n(\beta) = \binom{n}{2}^{-1} \sum\sum_{1 \leq i < j \leq n} (Y_{ij} - \boldsymbol{\beta}\mathbf{X}_{ij})^2. \tag{A.4}$$

Define the difference in the sum as

$$(Y_{ij} - \boldsymbol{\beta}\mathbf{X}_{ij}) = D_{ij}(\boldsymbol{\beta}), \tag{A.5}$$

$$D(\beta) = \begin{bmatrix} D_{12}(\boldsymbol{\beta}) & D_{13}(\boldsymbol{\beta}) & ... & D_{n-1n}(\boldsymbol{\beta}) \end{bmatrix}^t. \tag{A.6}$$

Then the U-process can be written as

$$U_n(\boldsymbol{\beta}) = \binom{n}{2}^{-1} D^t(\boldsymbol{\beta}) D(\boldsymbol{\beta}). \tag{A.7}$$

The identity matrix $\mathbf{I}_{N\times N}$ can be rewritten in terms of an orthonormal set of vectors in $\mathbb{R}^N$. For clarity, I will show the proof for the equality $\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t = \mathbf{I}$ for use in Equation A.6.

Let $\mathbf{E} = \{\mathbf{e}_v\}_{v=1}^N$ be a set of orthonormal vectors in $\mathbb{R}^N$. Then $\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t = \mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix. By definition of $\mathbf{E}$, it is an orthonormal set of vectors of size N and is a basis for $\mathbb{R}^N$. Let $\mathbf{c}$ be a vector in $\mathbb{R}^N$, we can write $\mathbf{c}$ as

$$\mathbf{c} = \sum_{v=1}^{N} \mathbf{e}_v^t \mathbf{c} \mathbf{e}_v.$$

Consider

$$\begin{aligned} \left(\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t\right) \mathbf{c} &= \left(\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t\right) \sum_{v=1}^{N} \mathbf{e}_v^t \mathbf{c} \mathbf{e}_v \\ &= \left(\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t\right) \sum_{v'=1}^{N} \mathbf{e}_{v'}^t \mathbf{c} \mathbf{e}_{v'} \\ &= \sum_{v=1}^{N} \sum_{v'=1}^{N} \mathbf{e}_v \mathbf{e}_v^t \mathbf{e}_{v'}^t \mathbf{c} \mathbf{e}_{v'} \\ &= \sum_{v=1}^{N} \sum_{v'=1}^{N} \mathbf{e}_{v'}^t \mathbf{c} \mathbf{e}_v \mathbf{e}_v^t \mathbf{e}_{v'} \end{aligned} \tag{A.8}$$

Where the final line of Equation A.8 is done by relocating the constant $\mathbf{e}_{v'}^t\mathbf{c}$.

Recall the definition of orthonormal vectors,

$$\mathbf{e}_v^t \mathbf{e}_{v'} = \begin{cases} 0 & \text{if } v \neq v' \\ \\ 1 & \text{if } v = v' \end{cases}$$

Thus, the final line of Equation A.8 can be rewritten as

$$\sum_{v=1}^{N} \mathbf{e}_{v'}^t \mathbf{c} \mathbf{e}_v = \mathbf{c}.$$

Since $\left(\sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t \right) \mathbf{c} = \mathbf{c}$, then

$$\begin{aligned} \mathbf{I} &= \sum_{v=1}^{N} \mathbf{e}_v \mathbf{e}_v^t \\ &= \mathbf{e}_1 \mathbf{e}_1^t + \sum_{v=2}^{n} \mathbf{e}_v \mathbf{e}_v^t + \sum_{v=n+1}^{N} \mathbf{e}_v \mathbf{e}_v^t \\ &= \mathbf{E}_1 \mathbf{E_1}^t + \mathbf{E}_{\mathrm{W}} \mathbf{E}_{\mathrm{W}}^t + \mathbf{E}_{\mathrm{N}} \mathbf{E}_{\mathrm{N}}^t, \end{aligned} \tag{A.9}$$

We can take advantage of the identity matrix equality within the U-process from Equation A.7,

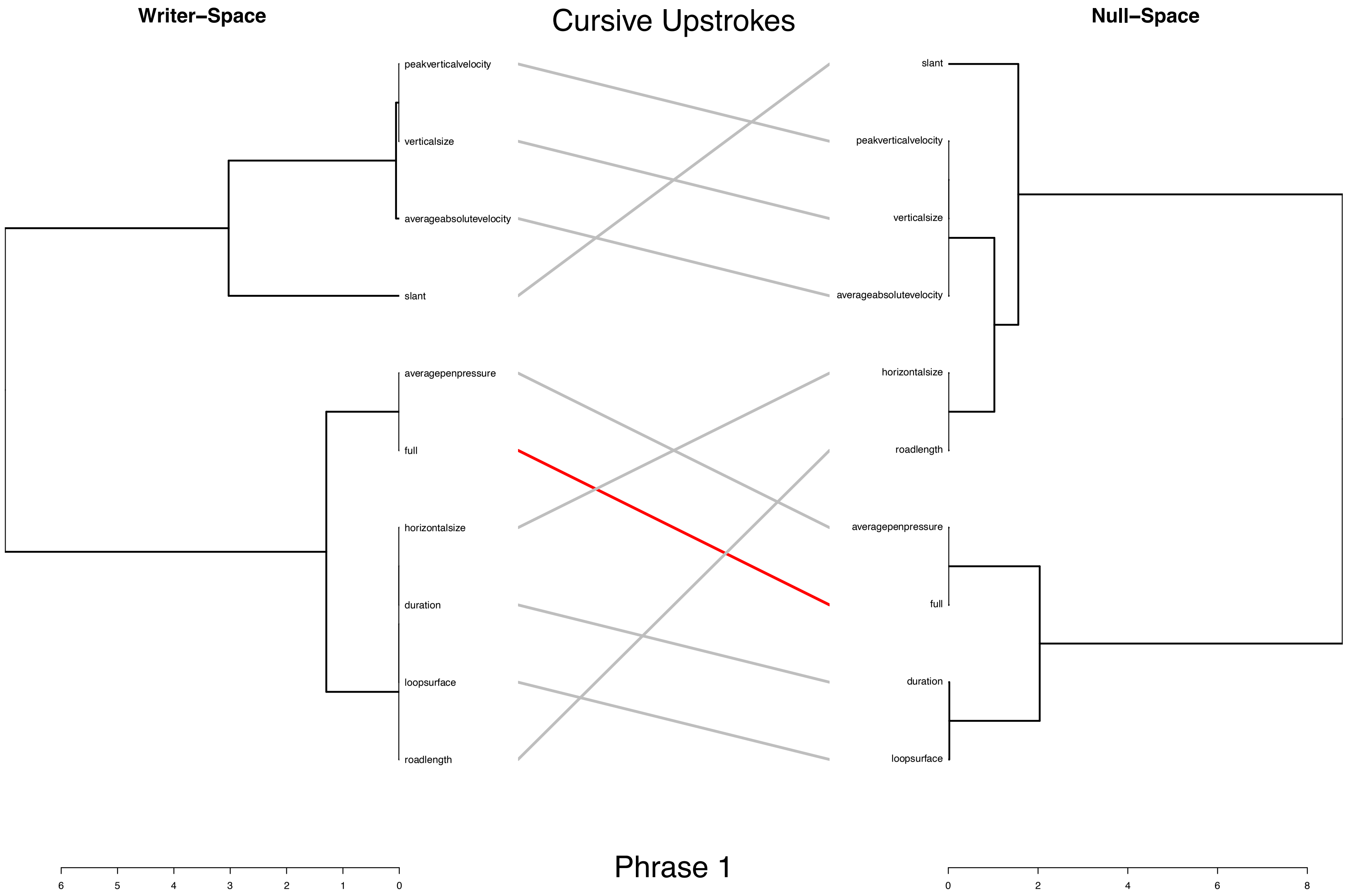
$$\begin{aligned} U_n(\boldsymbol{\beta}) &= \binom{n}{2}^{-1} D^t(\boldsymbol{\beta}) D(\boldsymbol{\beta}) \\ &= \binom{n}{2}^{-1} \left(D^t(\boldsymbol{\beta}) \left[\mathbf{E}_1 \mathbf{E_1}^t + \mathbf{E}_{\mathrm{W}} \mathbf{E}_{\mathrm{W}}^t + \mathbf{E}_{\mathrm{N}} \mathbf{E}_{\mathrm{N}}^t \right] D(\boldsymbol{\beta}) \right) \\ &= \binom{n}{2}^{-1} D^t(\mathbf{E}_1 \mathbf{E_1}^t) D(\boldsymbol{\beta}) + \binom{n}{2}^{-1} D^t(\boldsymbol{\beta}) \mathbf{E}_{\mathrm{W}} \mathbf{E}_{\mathrm{W}}^t D(\boldsymbol{\beta}) \\ &\quad + \binom{n}{2}^{-1} D^t(\boldsymbol{\beta}) \mathbf{E}_{\mathrm{N}} \mathbf{E}_{\mathrm{N}}^t D(\boldsymbol{\beta}) \end{aligned} \tag{A.10}$$

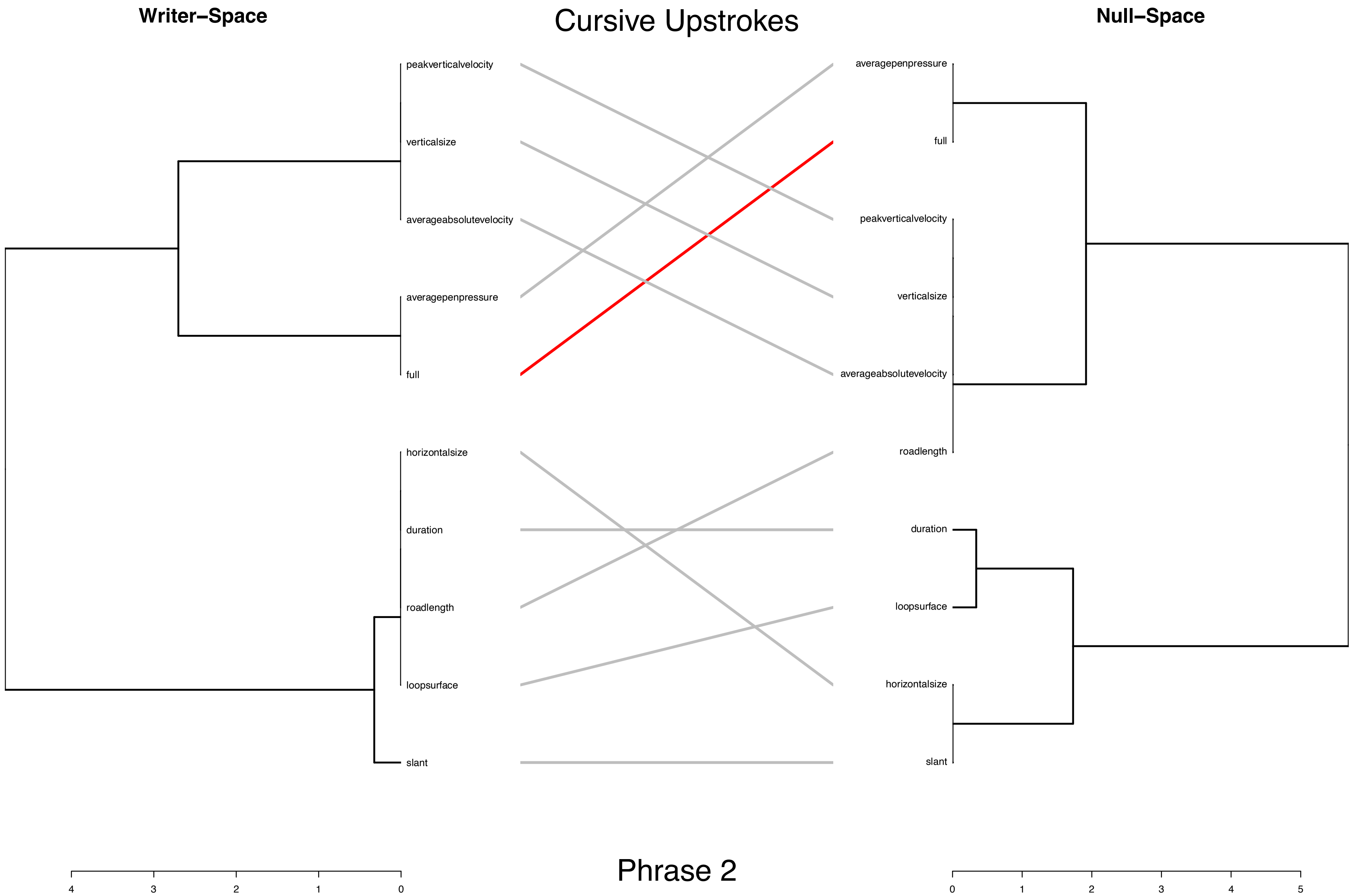
The first term in the final line of Equation A.10, $\binom{n}{2}^{-1} D^t(\mathbf{E}_1 \mathbf{E_1}^t) D(\boldsymbol{\beta})$ represents the first eigenvector. The second term, $\binom{n}{2}^{-1}D^t(\boldsymbol{\beta})\mathbf{E}_{\mathrm{W}}\mathbf{E}_{\mathrm{W}}^t D(\boldsymbol{\beta})$ represents the Writer-space scores in the U-process. Finally, the the term, $\binom{n}{2}^{-1}D^t(\boldsymbol{\beta})\mathbf{E}_{\mathrm{N}}\mathbf{E}_{\mathrm{N}}^t D(\boldsymbol{\beta})$ represents the Null-space scores in the U-process. For further U-process developments, see Deborah Nolan [12].

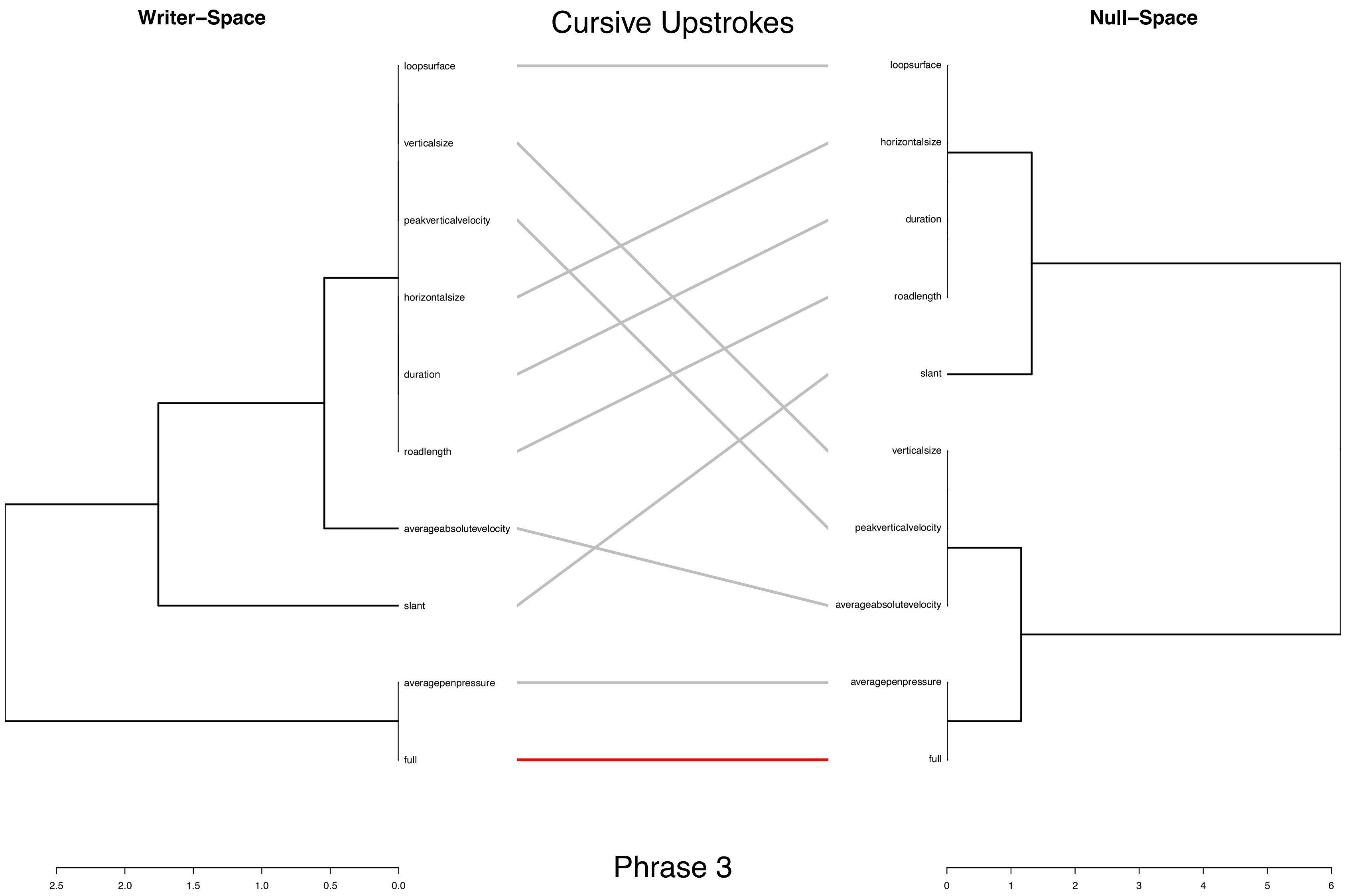
APPENDIX 2

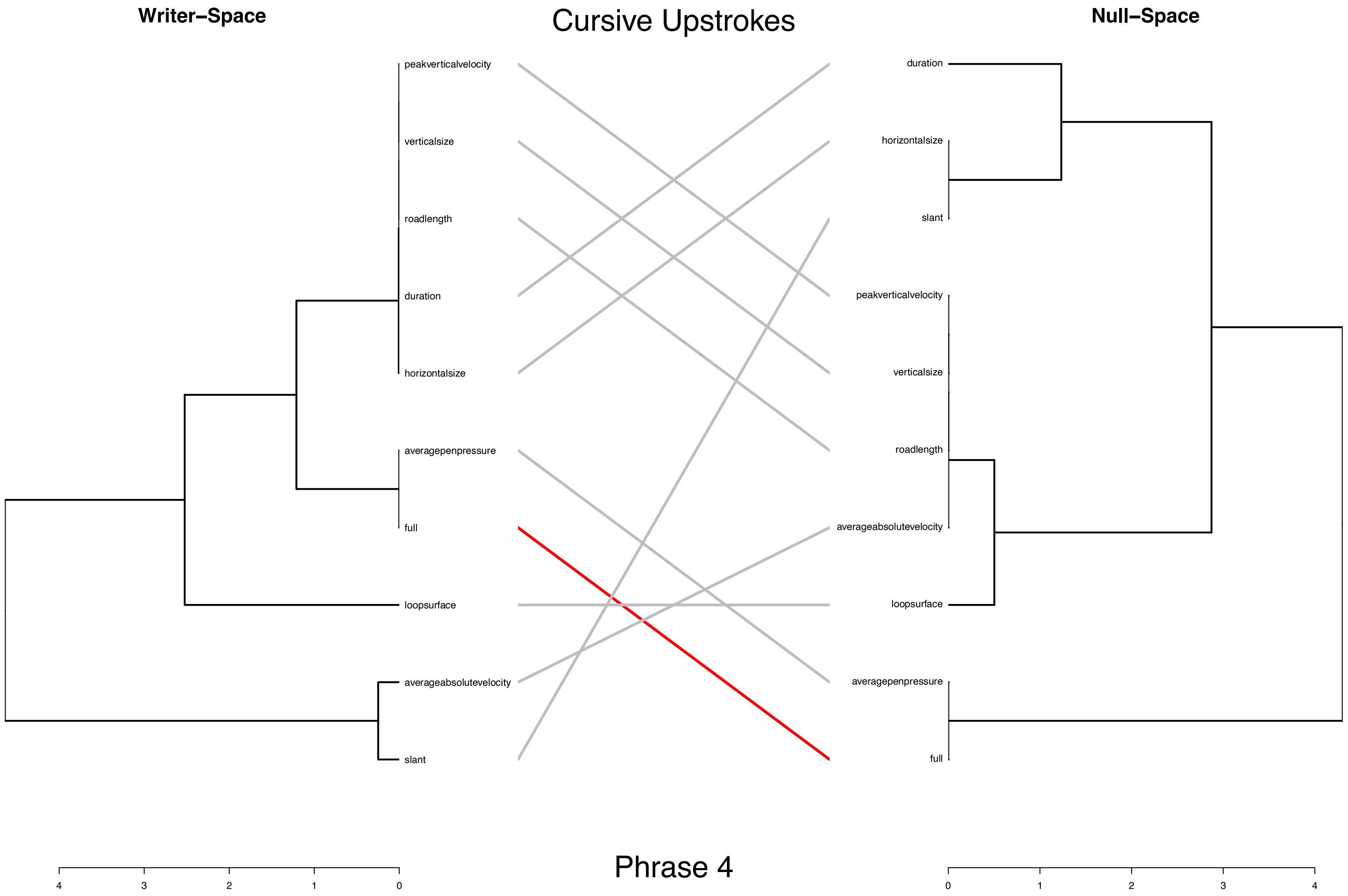
The following plots are a visual representation of the relationships between pairwise features from regression modeling of the MovAlyzeR® WDS. The left side of each plot represents the models fit in the Writer-space, and the right side of each plot represents the models fit in the Null-space. The x-axes represent a transformation of the p-values associated with $\hat{\beta}_1$ of $-2 * log(1 - p\text{-value})$. Note that $-2 * log(1 - 0.05) \approx 0.1$. The solid red line in each plot represents the 'full' score, i.e. the WDS calculated with all of the other features pooled together.

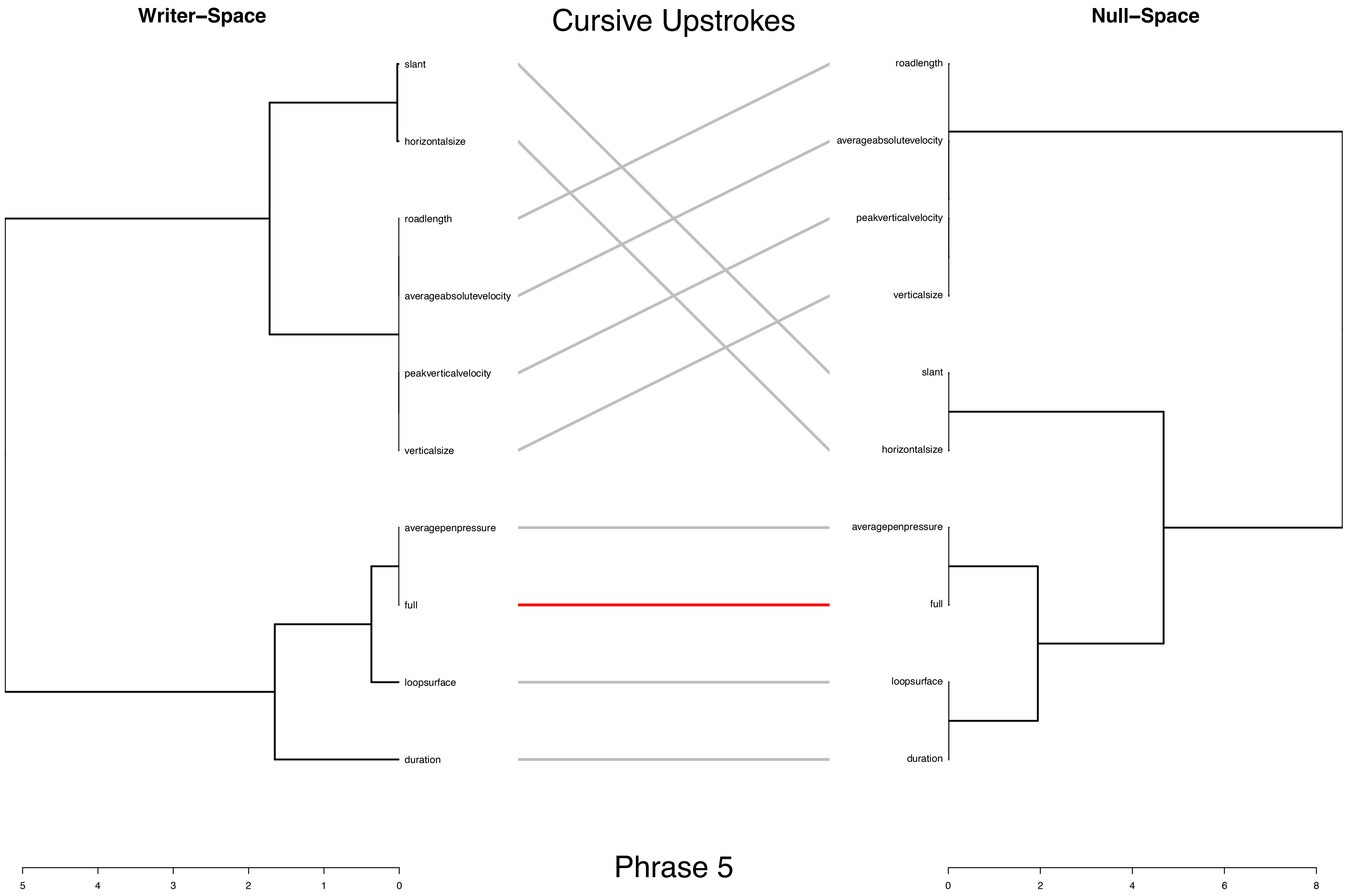
For a majority of these plots, there appears to be a relationship between the Full WDS and the Average Pen Pressure WDS for both the Writer-space and the Null-space.


Writer–Space
Cursive Upstrokes
Null–Space
peakverticalvelocity
verticalsize
averageabsolutevelocity
slant
averagepenpressure
full
horizontalsize
duration
loopsurface
roadlength
slant
peakverticalvelocity
verticalsize
averageabsolutevelocity
horizontalsize
roadlength
averagepenpressure
full
duration
loopsurface
6
5
4
3
2
1
0
Phrase 1
0
2
4
6
8


Writer–Space
Cursive Upstrokes
Null–Space
peakverticalvelocity
verticalsize
averageabsolutevelocity
averagepenpressure
full
horizontalsize
duration
roadlength
loopsurface
slant
averagepenpressure
full
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
duration
loopsurface
horizontalsize
slant
4
3
2
1
0
Phrase 2
0
1
2
3
4
5

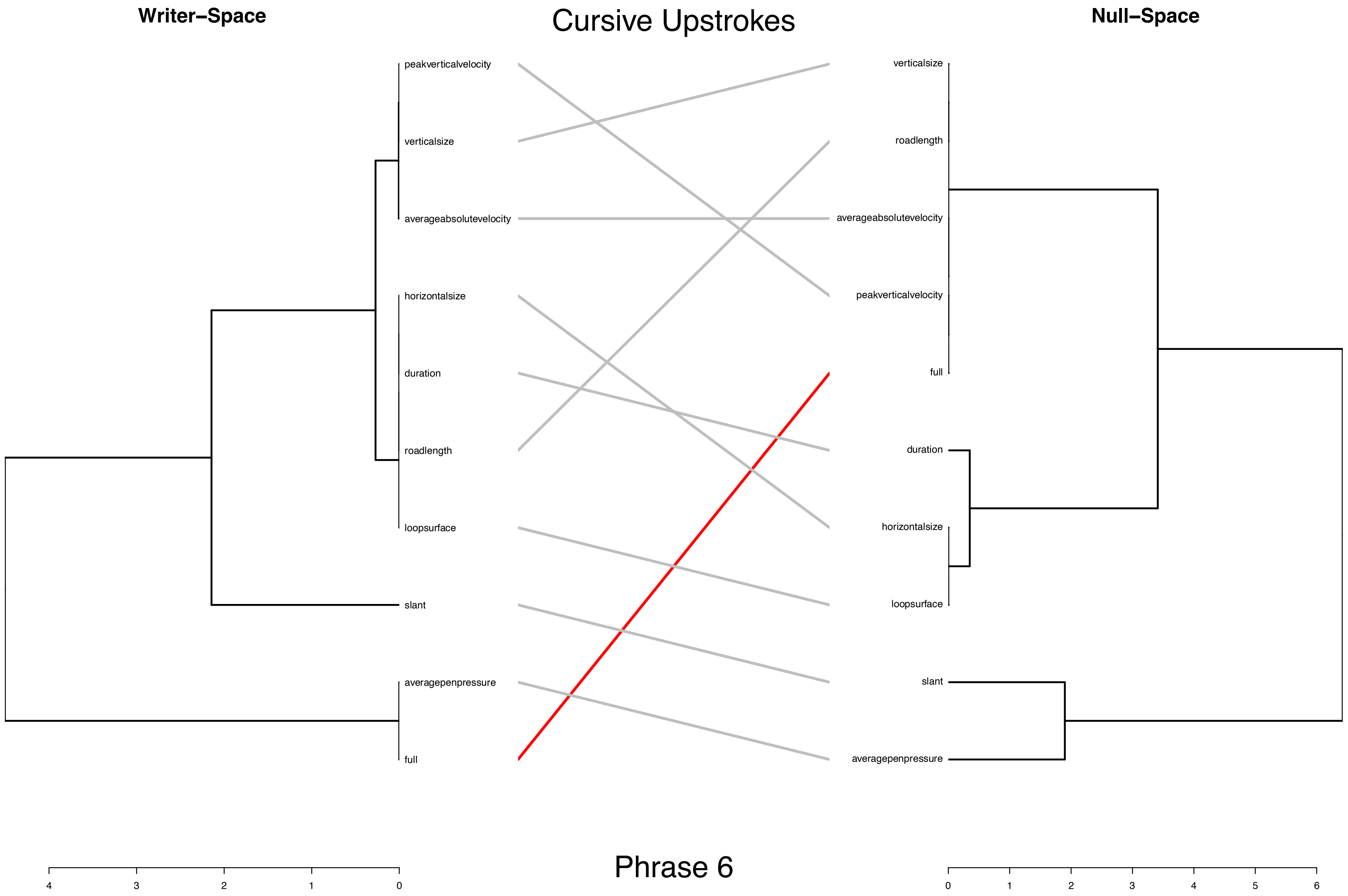

Writer–Space
Cursive Upstrokes
Null–Space
loopsurface
verticalsize
peakverticalvelocity
horizontalsize
duration
roadlength
averageabsolutevelocity
slant
averagepenpressure
full
loopsurface
horizontalsize
duration
roadlength
slant
verticalsize
peakverticalvelocity
averageabsolutevelocity
averagepenpressure
full
2.5
2.0
1.5
1.0
0.5
0.0
Phrase 3
0
1
2
3
4
5
6

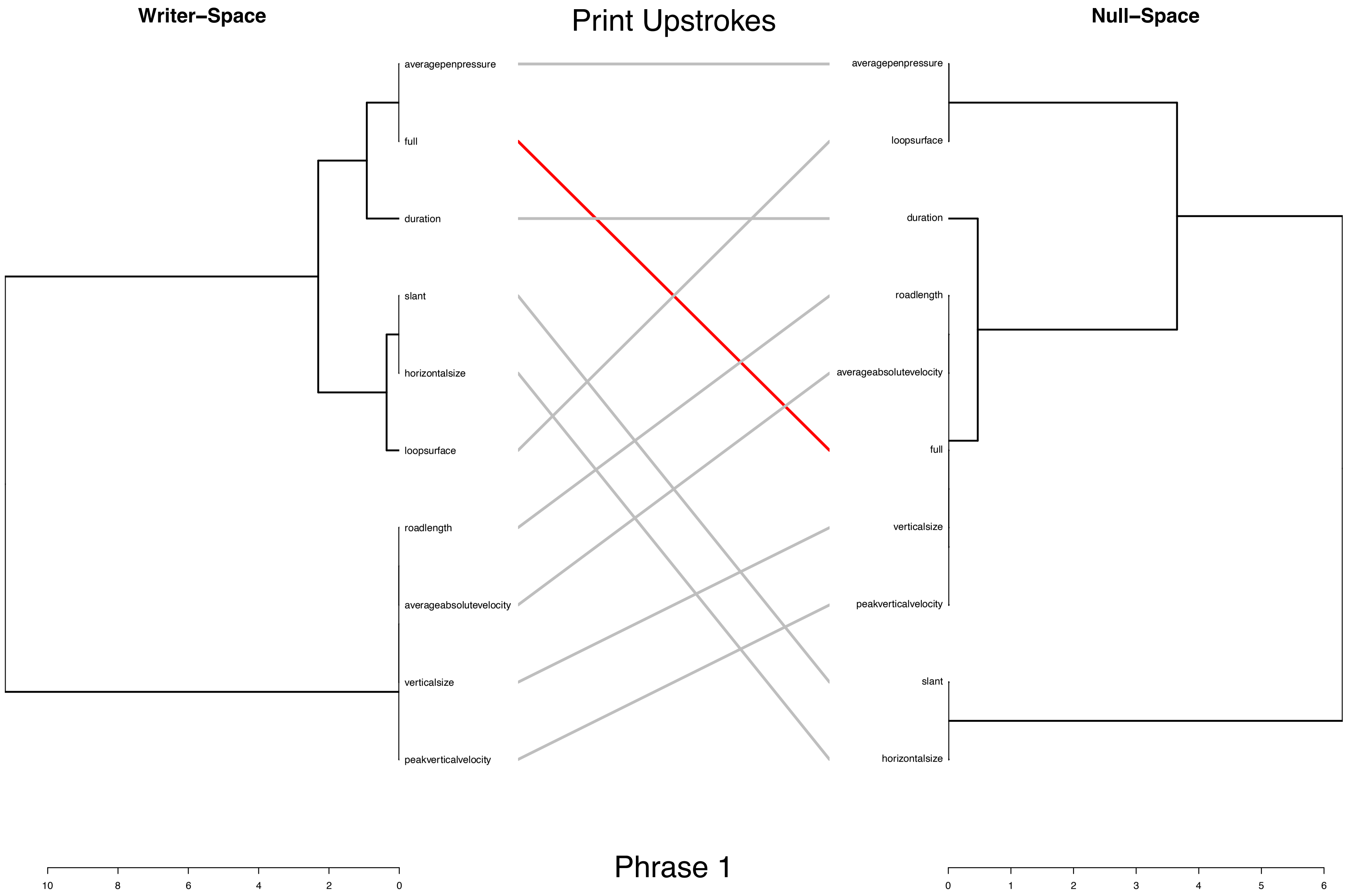

Writer–Space
Cursive Upstrokes
Null–Space
peakverticalvelocity
verticalsize
roadlength
duration
horizontalsize
averagepenpressure
full
loopsurface
averageabsolutevelocity
slant
duration
horizontalsize
slant
peakverticalvelocity
verticalsize
roadlength
averageabsolutevelocity
loopsurface
averagepenpressure
full
4
3
2
1
0
Phrase 4
0
1
2
3
4

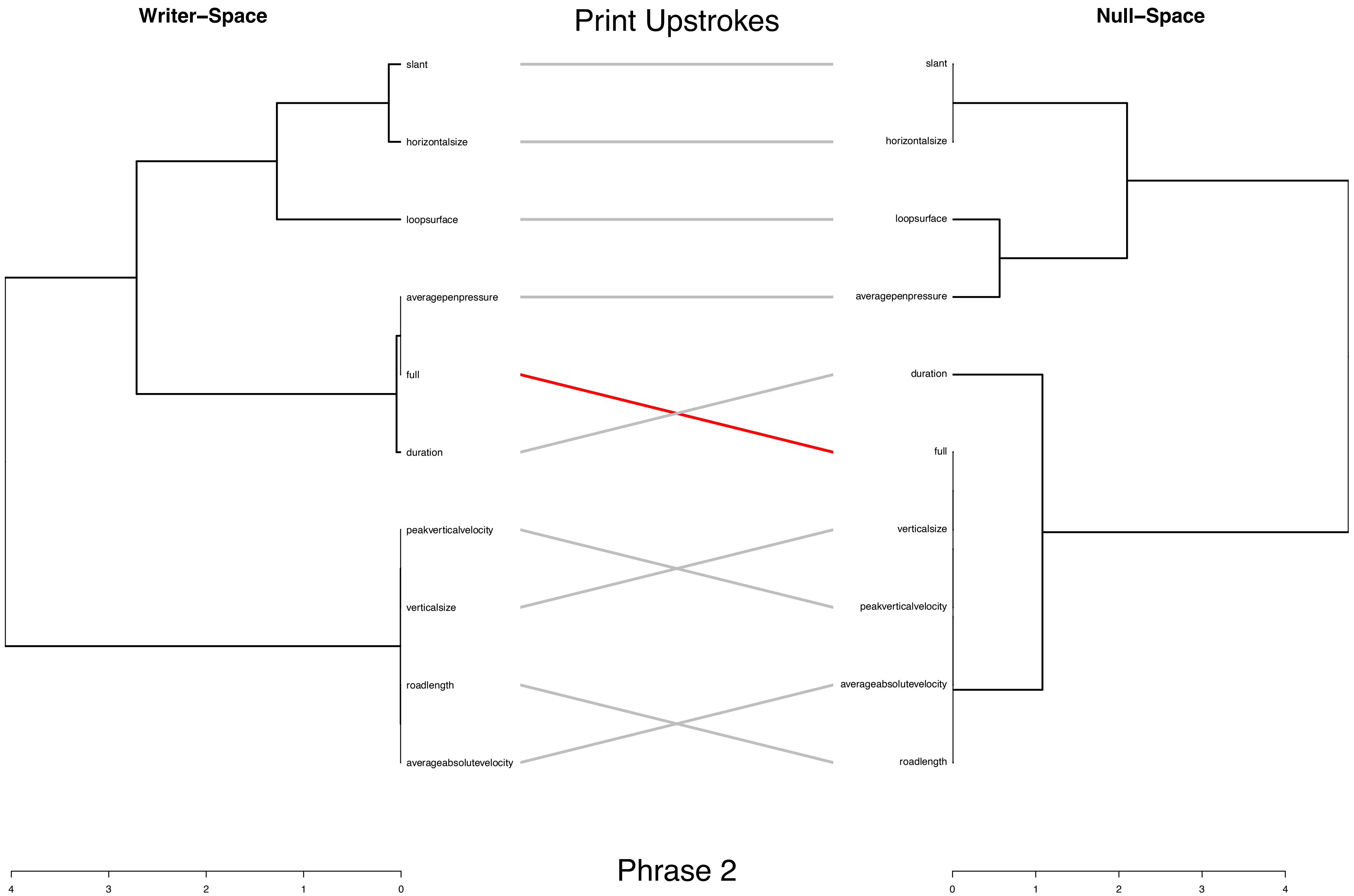


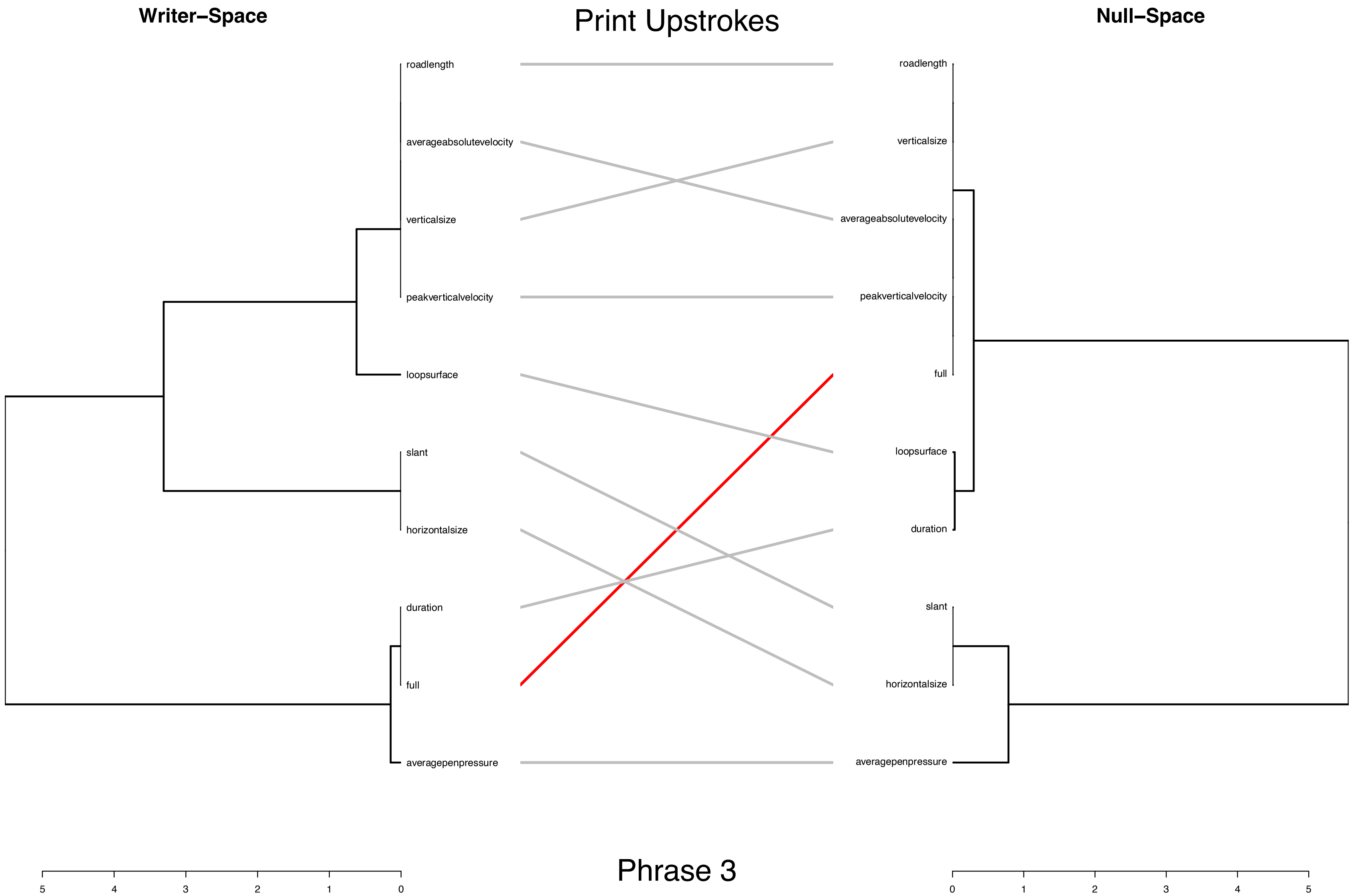
Cursive Upstrokes

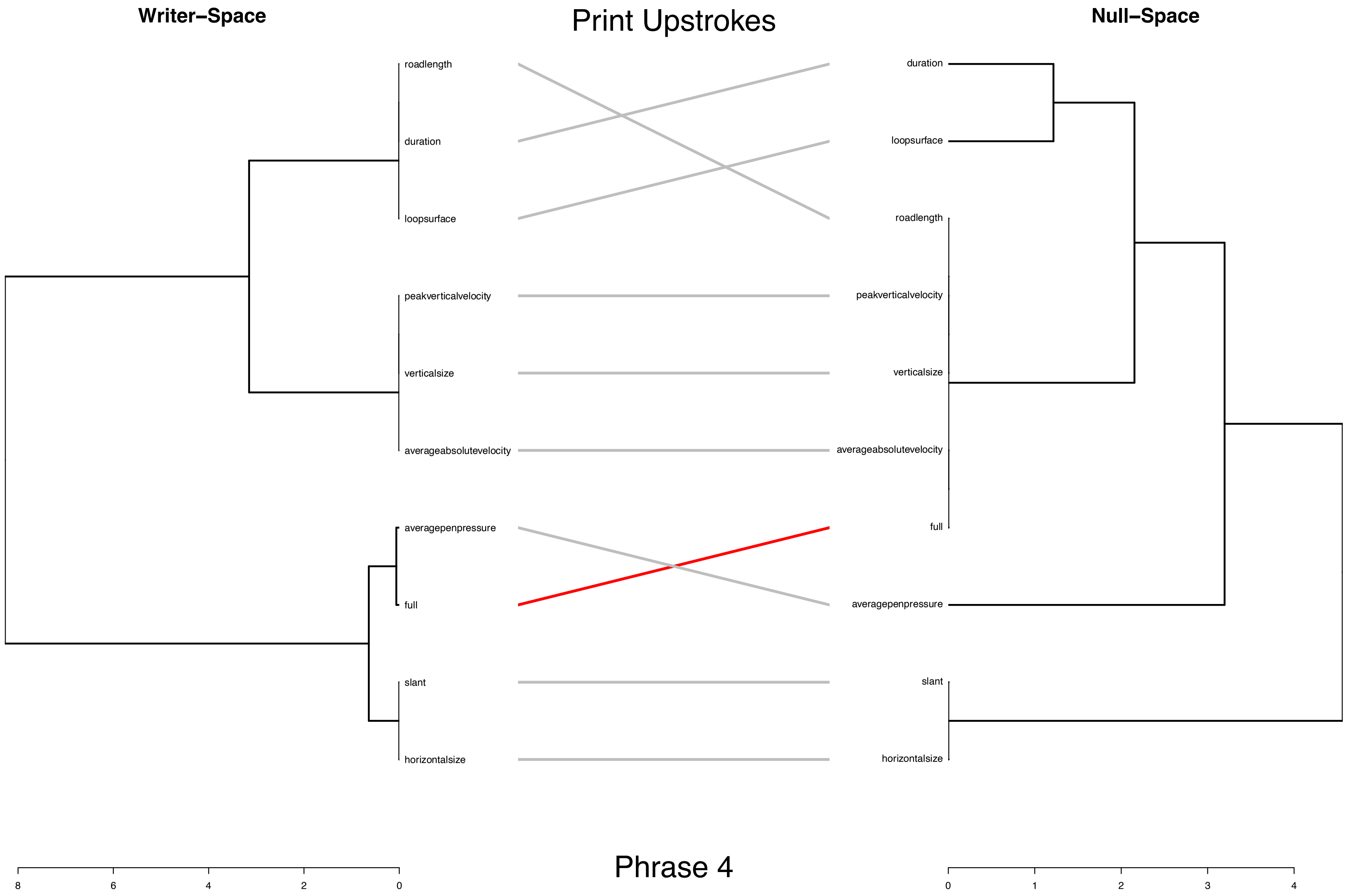
Phrase 5

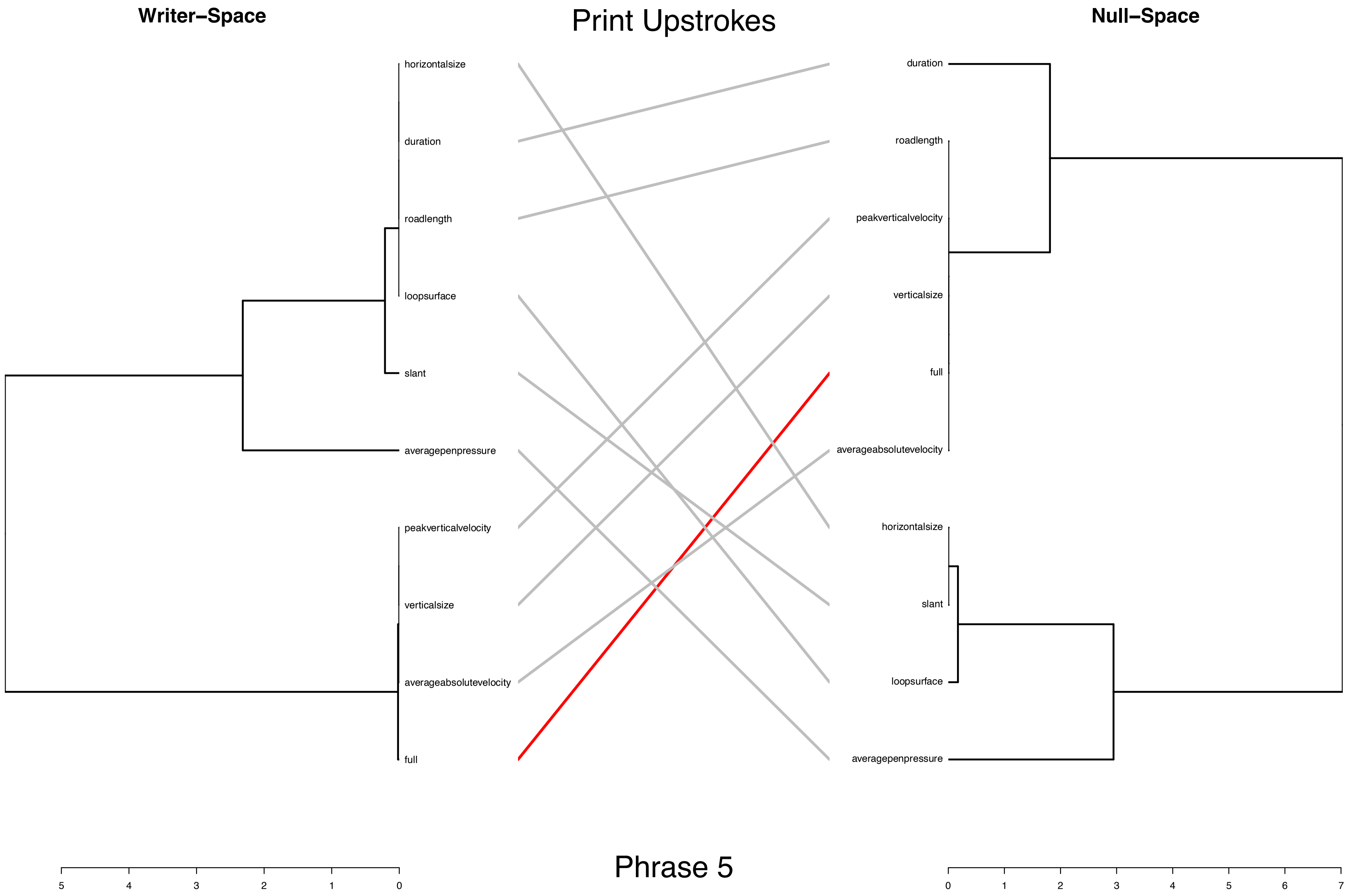

Writer–Space
Cursive Upstrokes
Null–Space
peakverticalvelocity
verticalsize
averageabsolutevelocity
horizontalsize
duration
roadlength
loopsurface
slant
averagepenpressure
full
verticalsize
roadlength
averageabsolutevelocity
peakverticalvelocity
full
duration
horizontalsize
loopsurface
slant
averagepenpressure
4
3
2
1
0
Phrase 6
0
1
2
3
4
5
6

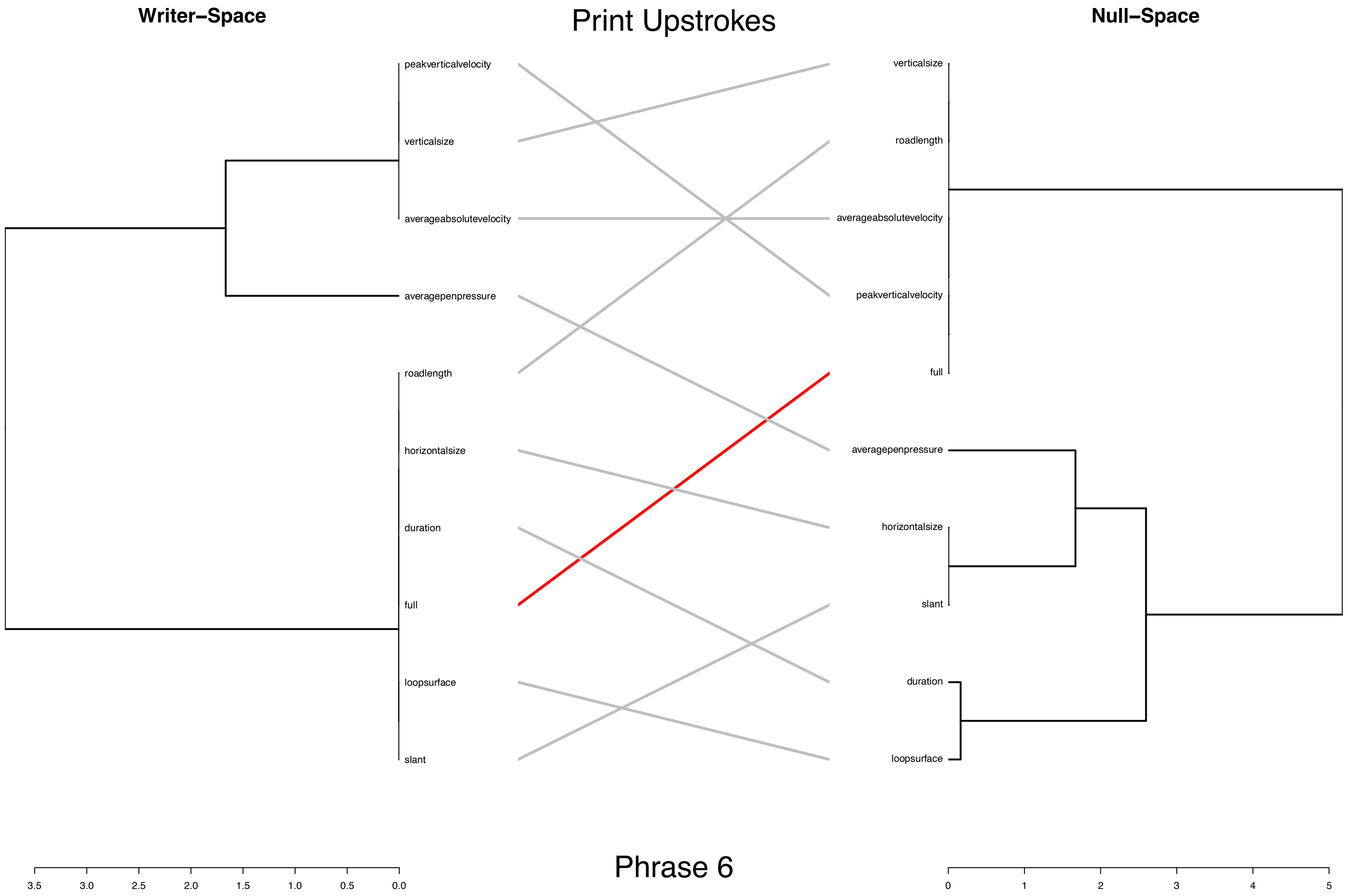

Writer–Space
Print Upstrokes
Null–Space
averagepenpressure
full
duration
slant
horizontalsize
loopsurface
roadlength
averageabsolutevelocity
verticalsize
peakverticalvelocity
averagepenpressure
loopsurface
duration
roadlength
averageabsolutevelocity
full
verticalsize
peakverticalvelocity
slant
horizontalsize
10
8
6
4
2
0
Phrase 1
0
1
2
3
4
5
6

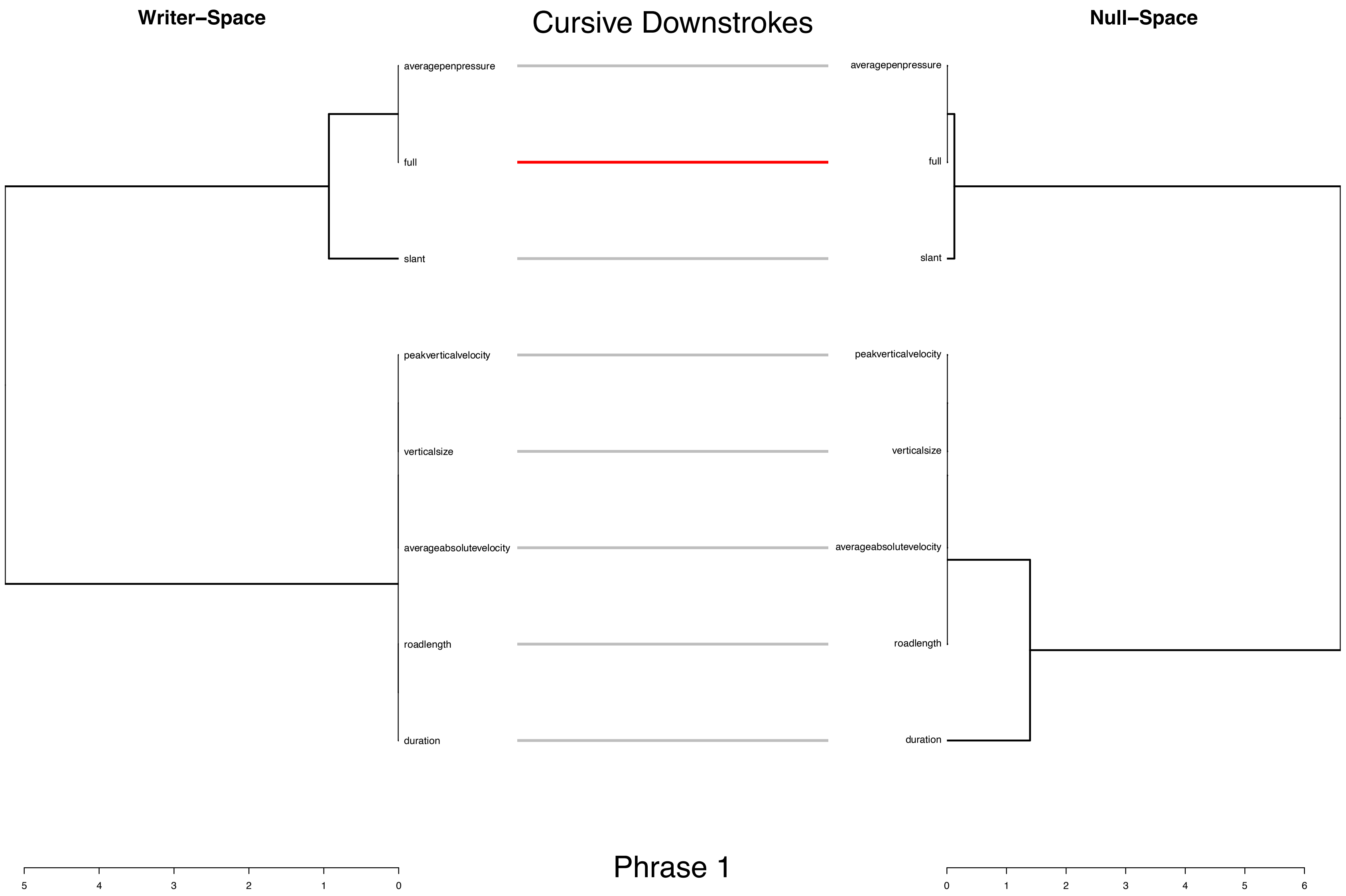

Writer–Space
Print Upstrokes
Null–Space
slant
horizontalsize
loopsurface
averagepenpressure
full
duration
peakverticalvelocity
verticalsize
roadlength
averageabsolutevelocity
slant
horizontalsize
loopsurface
averagepenpressure
duration
full
verticalsize
peakverticalvelocity
averageabsolutevelocity
roadlength
4
3
2
1
0
Phrase 2
0
1
2
3
4

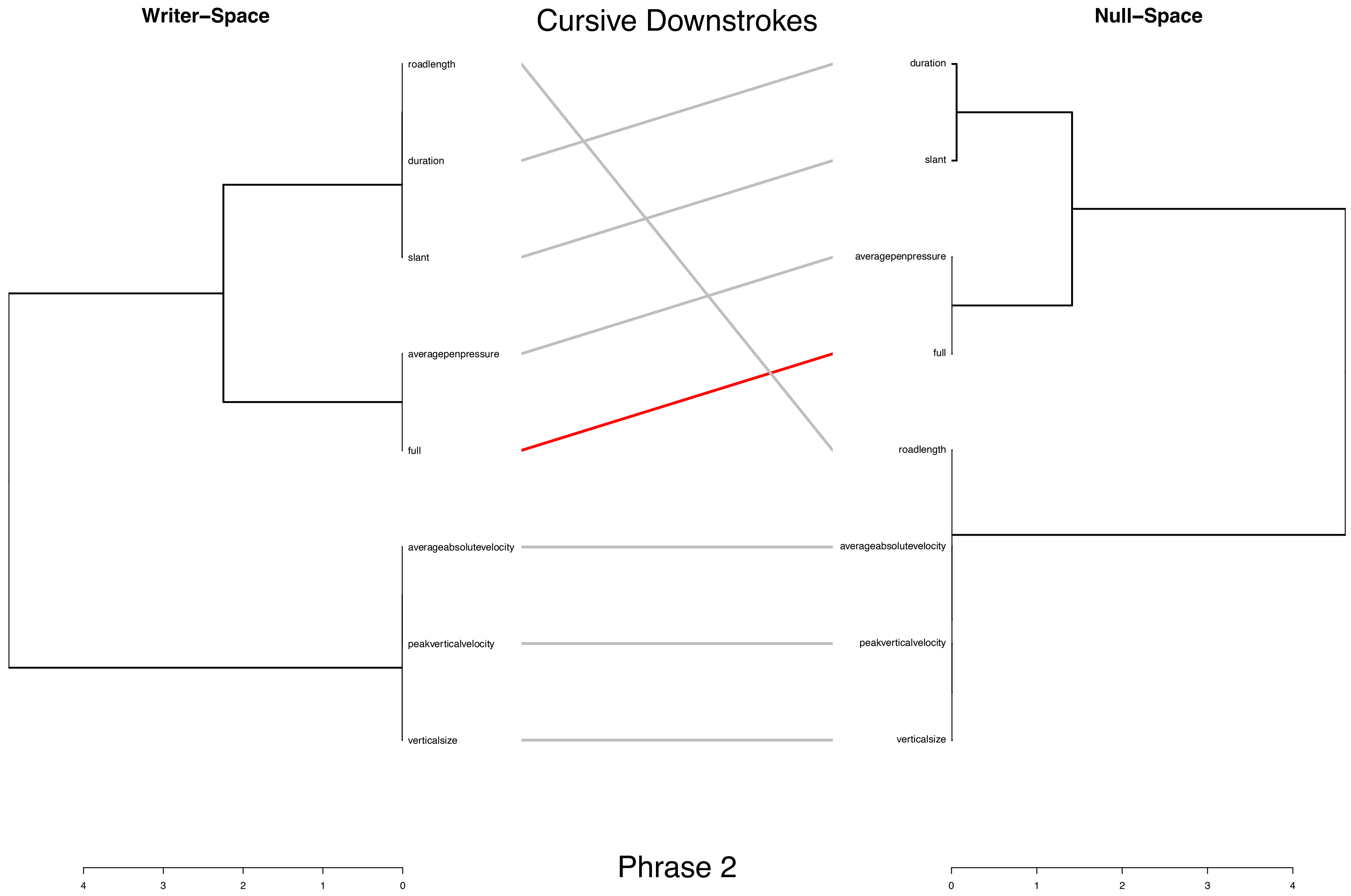

Writer–Space
Print Upstrokes
Null–Space
roadlength
averageabsolutevelocity
verticalsize
peakverticalvelocity
loopsurface
slant
horizontalsize
duration
full
averagepenpressure
roadlength
verticalsize
averageabsolutevelocity
peakverticalvelocity
full
loopsurface
duration
slant
horizontalsize
averagepenpressure
5
4
3
2
1
0
Phrase 3
0
1
2
3
4
5


Writer–Space
Print Upstrokes
Null–Space
roadlength
duration
loopsurface
peakverticalvelocity
verticalsize
averageabsolutevelocity
averagepenpressure
full
slant
horizontalsize
duration
loopsurface
roadlength
peakverticalvelocity
verticalsize
averageabsolutevelocity
full
averagepenpressure
slant
horizontalsize
Phrase 4
8
6
4
2
0
0
1
2
3
4


Writer–Space
Print Upstrokes
Null–Space
horizontalsize
duration
roadlength
loopsurface
slant
averagepenpressure
peakverticalvelocity
verticalsize
averageabsolutevelocity
full
duration
roadlength
peakverticalvelocity
verticalsize
full
averageabsolutevelocity
horizontalsize
slant
loopsurface
averagepenpressure
5
4
3
2
1
0
Phrase 5
0
1
2
3
4
5
6
7

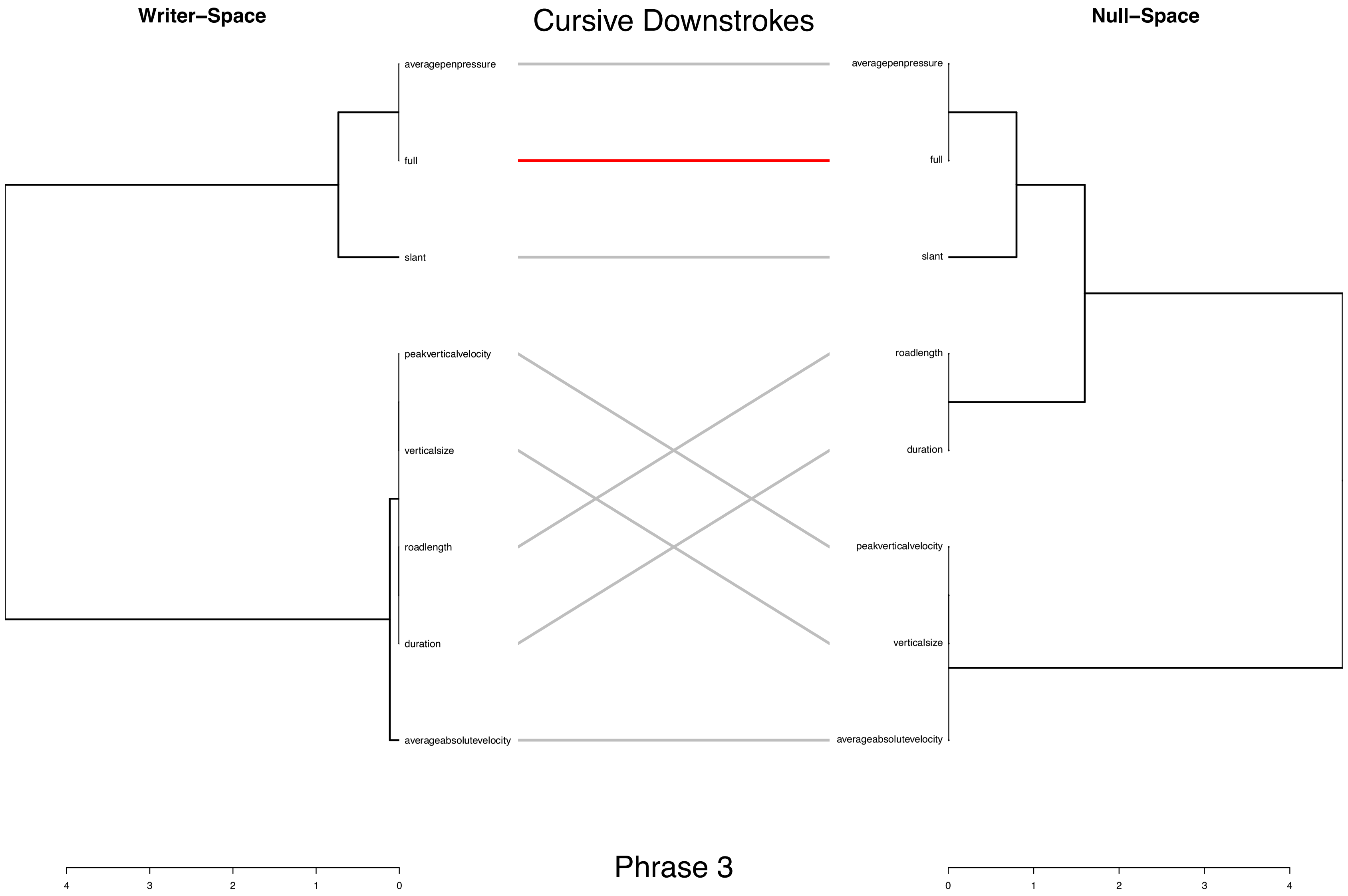

Print Upstrokes
Writer–Space
Null–Space
peakverticalvelocity
verticalsize
averageabsolutevelocity
averagepenpressure
roadlength
horizontalsize
duration
full
loopsurface
slant
verticalsize
roadlength
averageabsolutevelocity
peakverticalvelocity
full
averagepenpressure
horizontalsize
slant
duration
loopsurface
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
Phrase 6

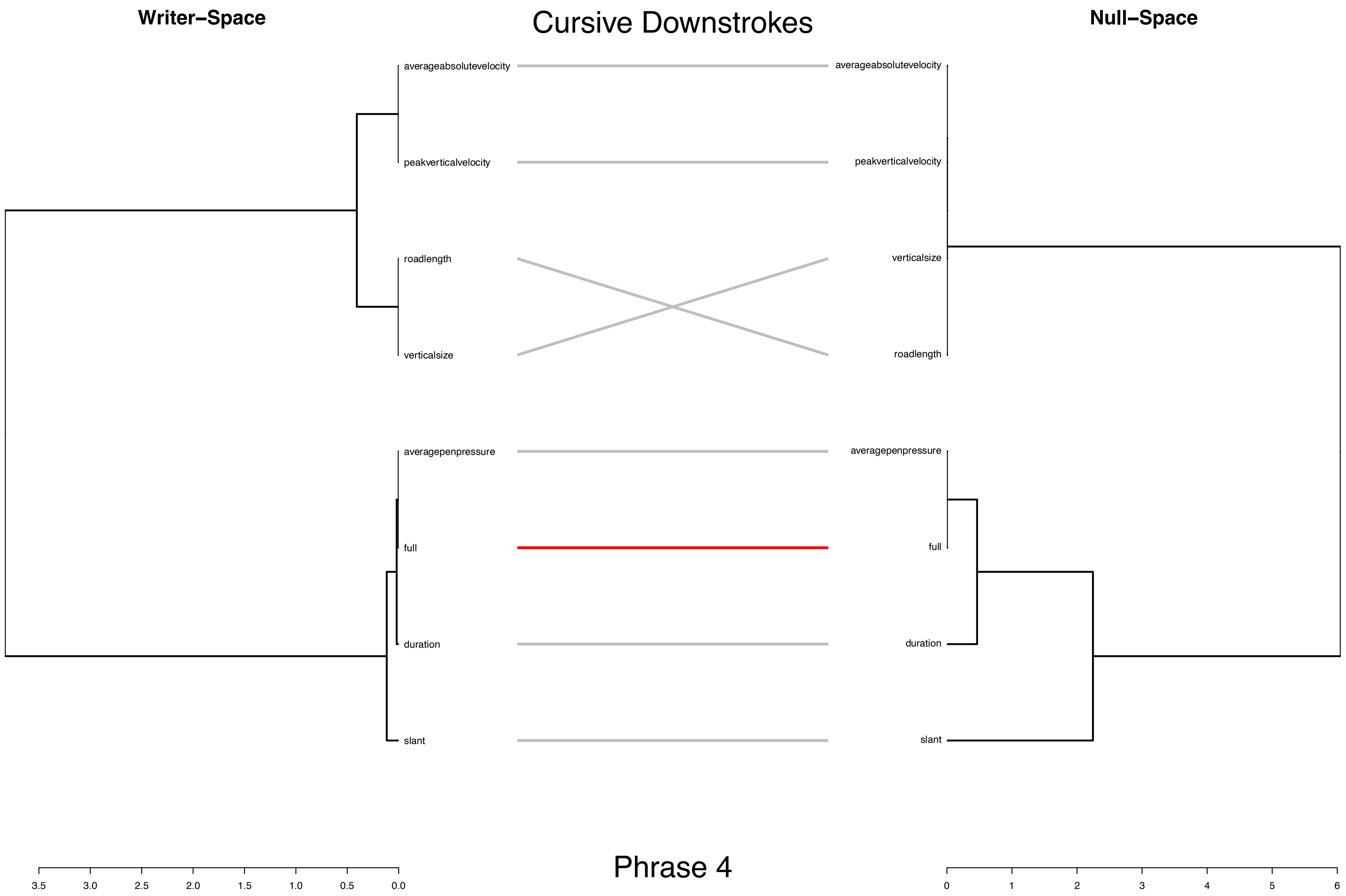

Writer–Space
Cursive Downstrokes
Null–Space
averagepenpressure
full
slant
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
duration
averagepenpressure
full
slant
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
duration
5
4
3
2
1
0
Phrase 1
0
1
2
3
4
5
6



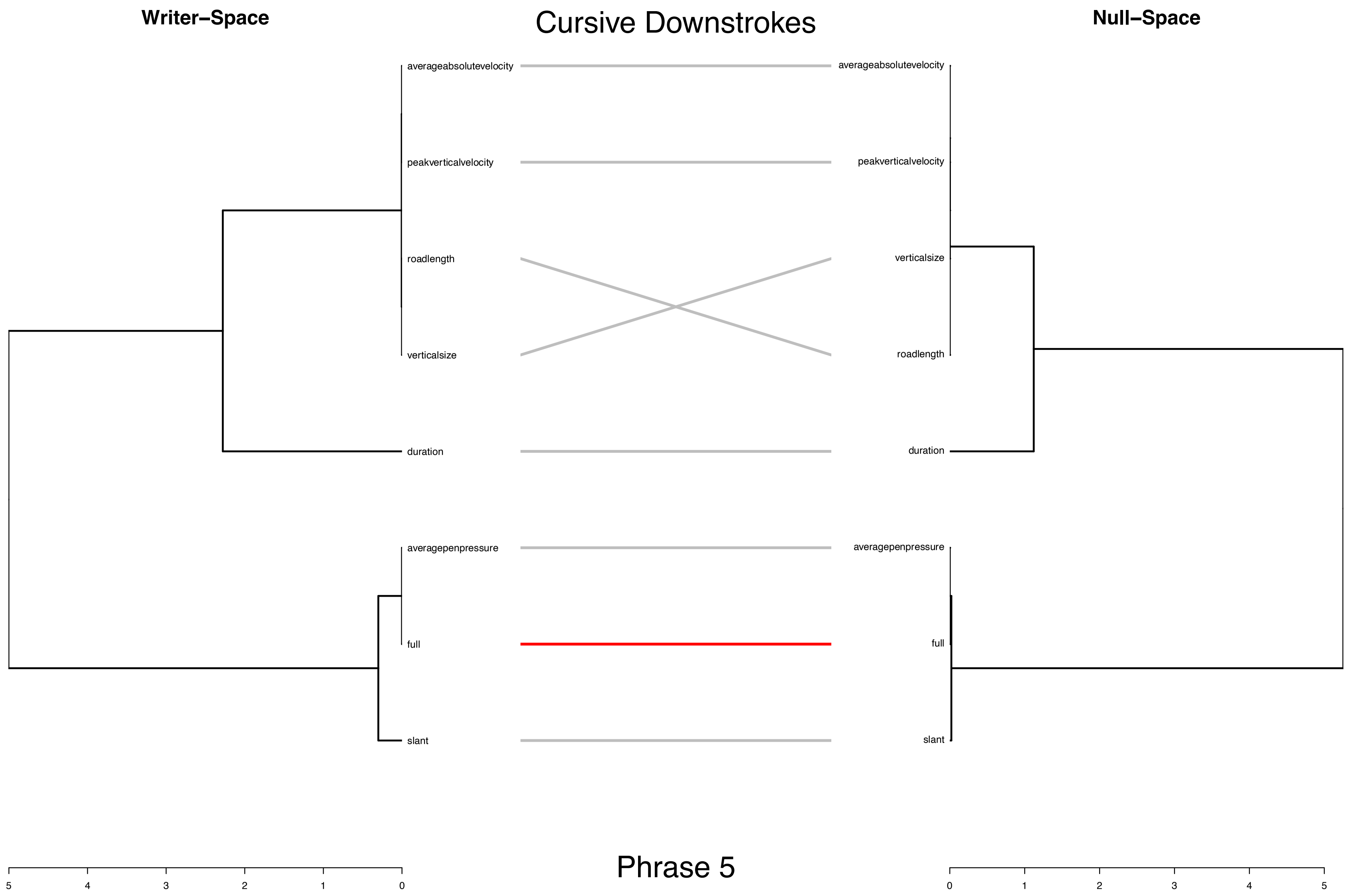
Cursive Downstrokes

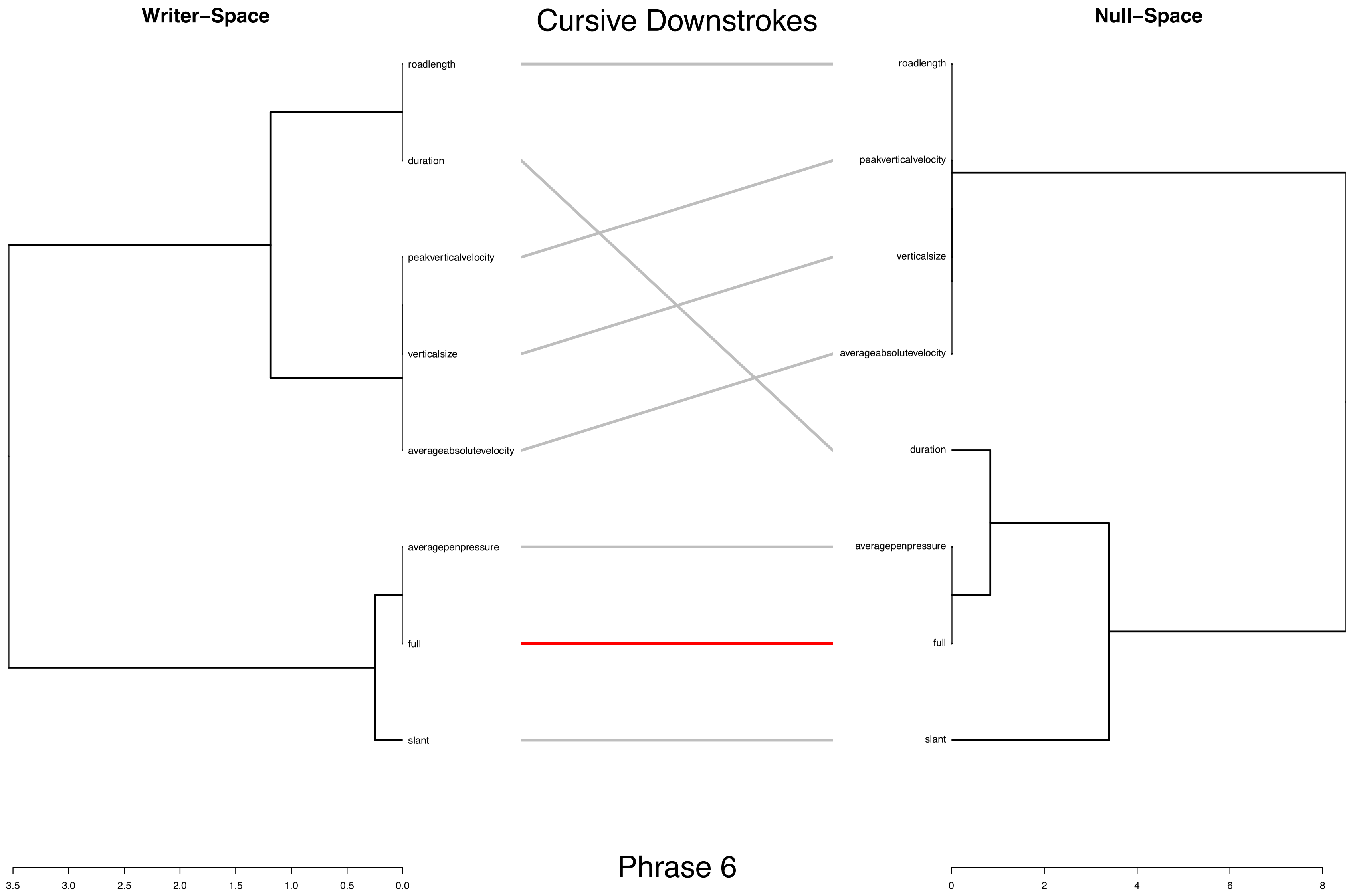
Phrase 2


Writer–Space
Cursive Downstrokes
Null–Space
averagepenpressure
full
slant
peakverticalvelocity
verticalsize
roadlength
duration
averageabsolutevelocity
averagepenpressure
full
slant
roadlength
duration
peakverticalvelocity
verticalsize
averageabsolutevelocity
4
3
2
1
0
Phrase 3
0
1
2
3
4


Cursive Downstrokes

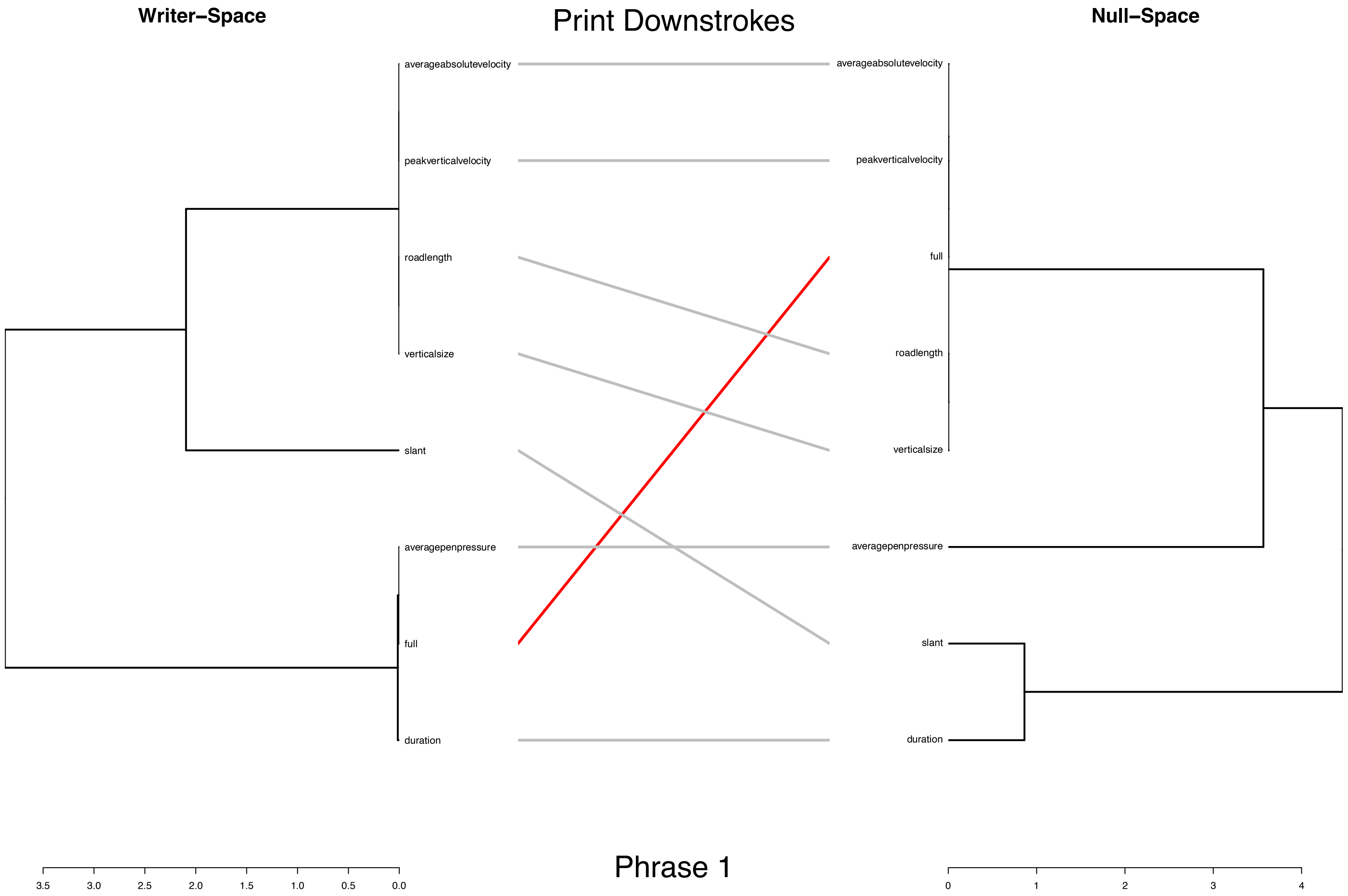
Phrase 4

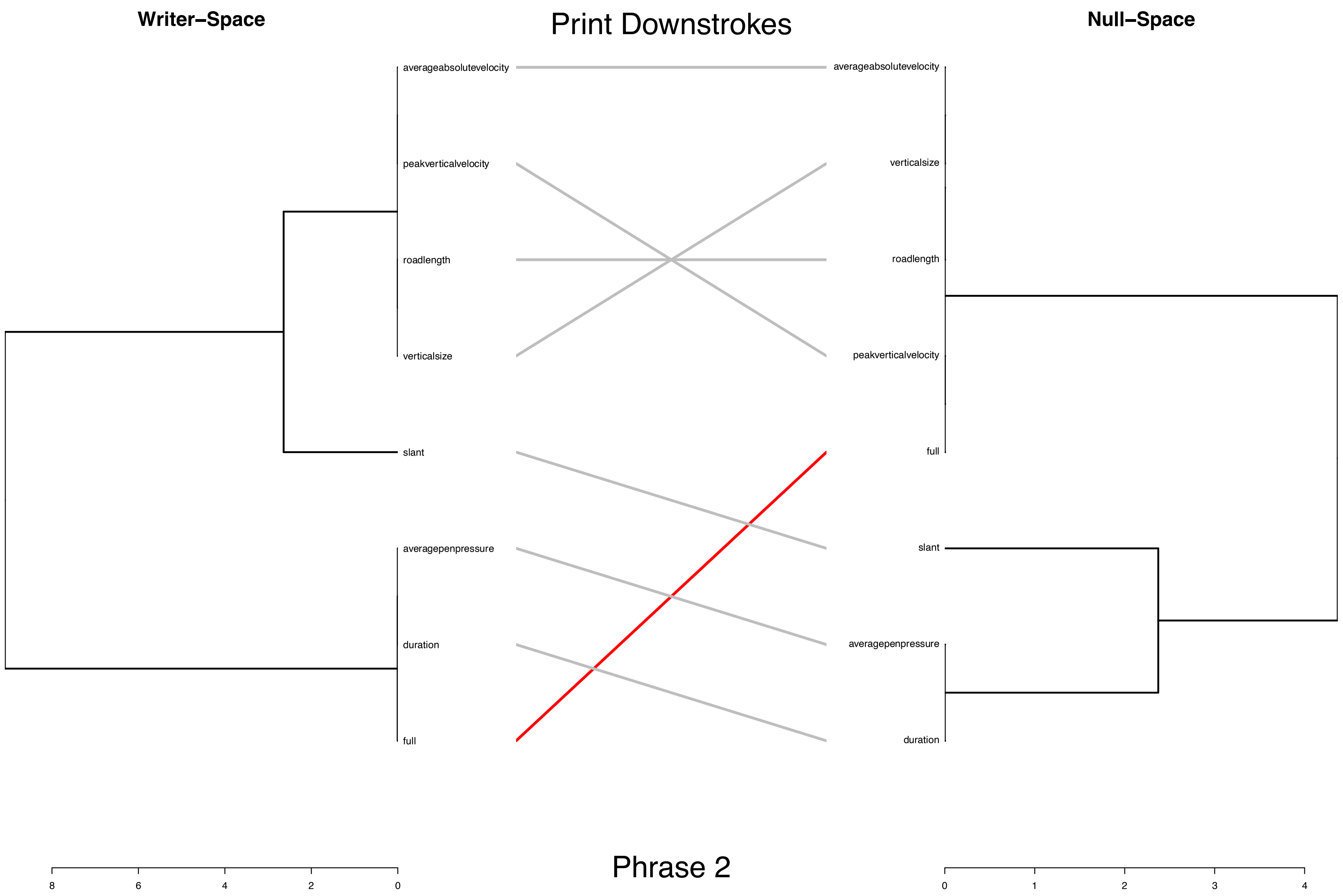

Writer–Space
Cursive Downstrokes
Null–Space
averageabsolutevelocity
peakverticalvelocity
roadlength
verticalsize
duration
averagepenpressure
full
slant
averageabsolutevelocity
peakverticalvelocity
verticalsize
roadlength
duration
averagepenpressure
full
slant
5
4
3
2
1
0
Phrase 5
0
1
2
3
4
5

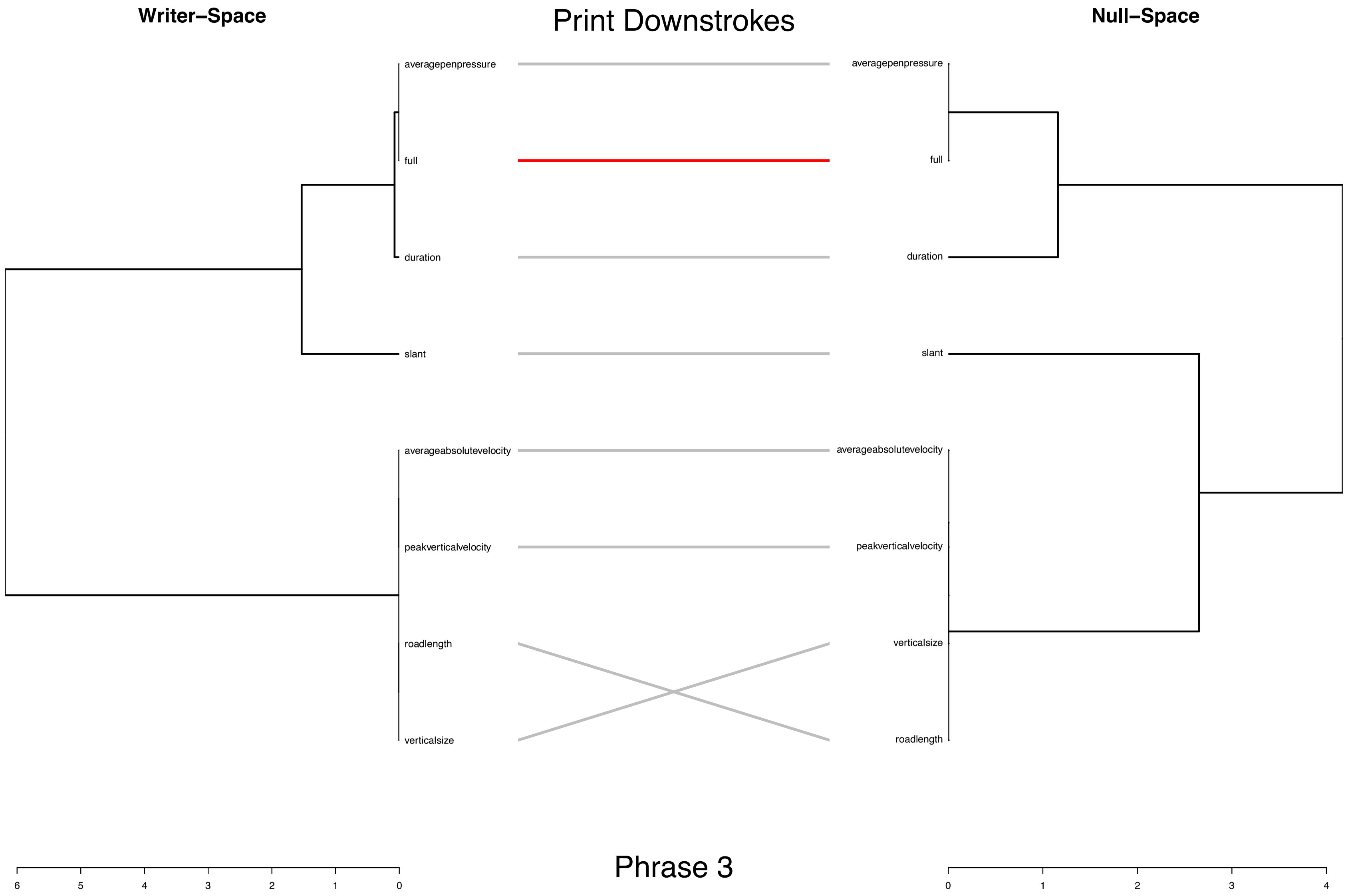


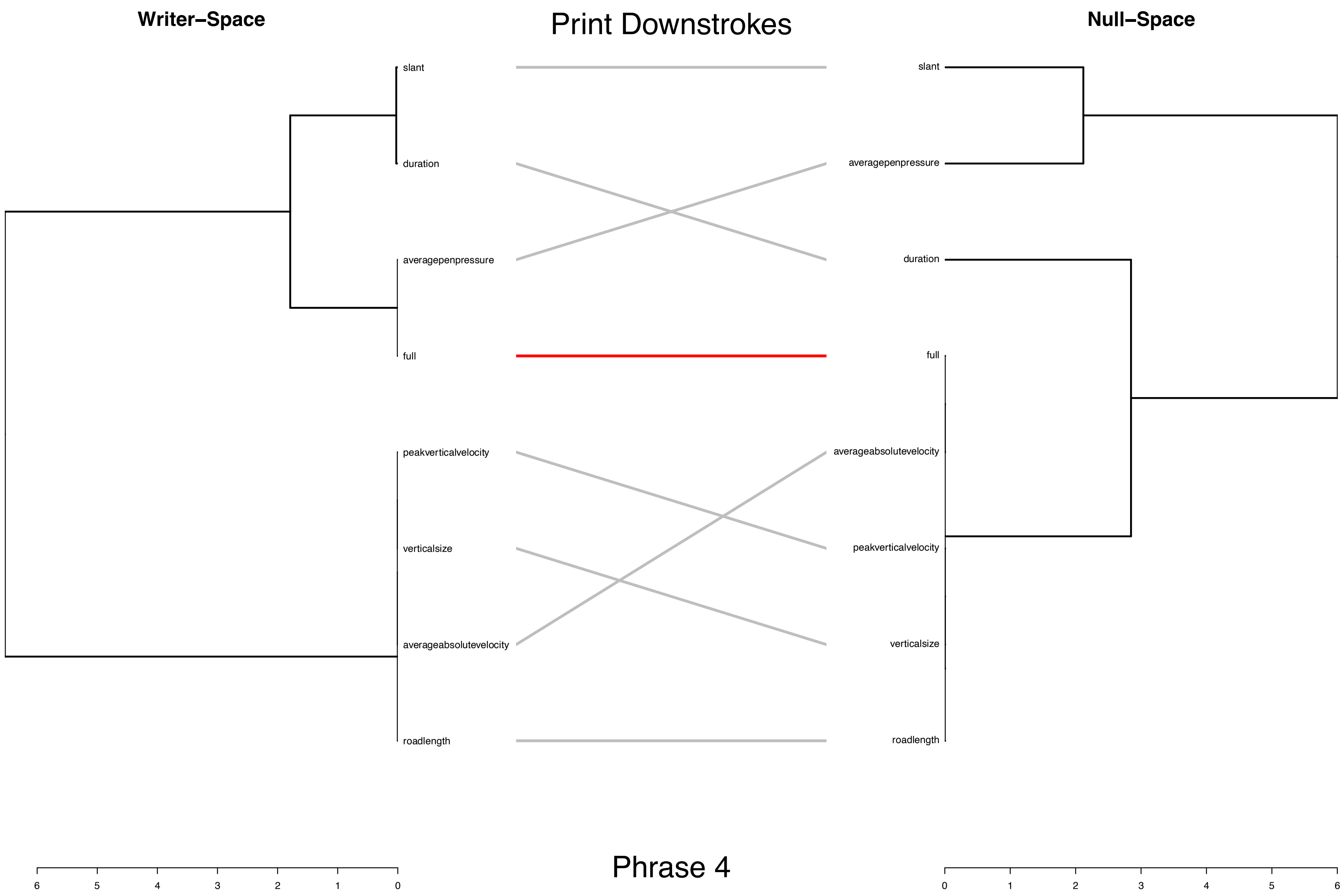
Cursive Downstrokes

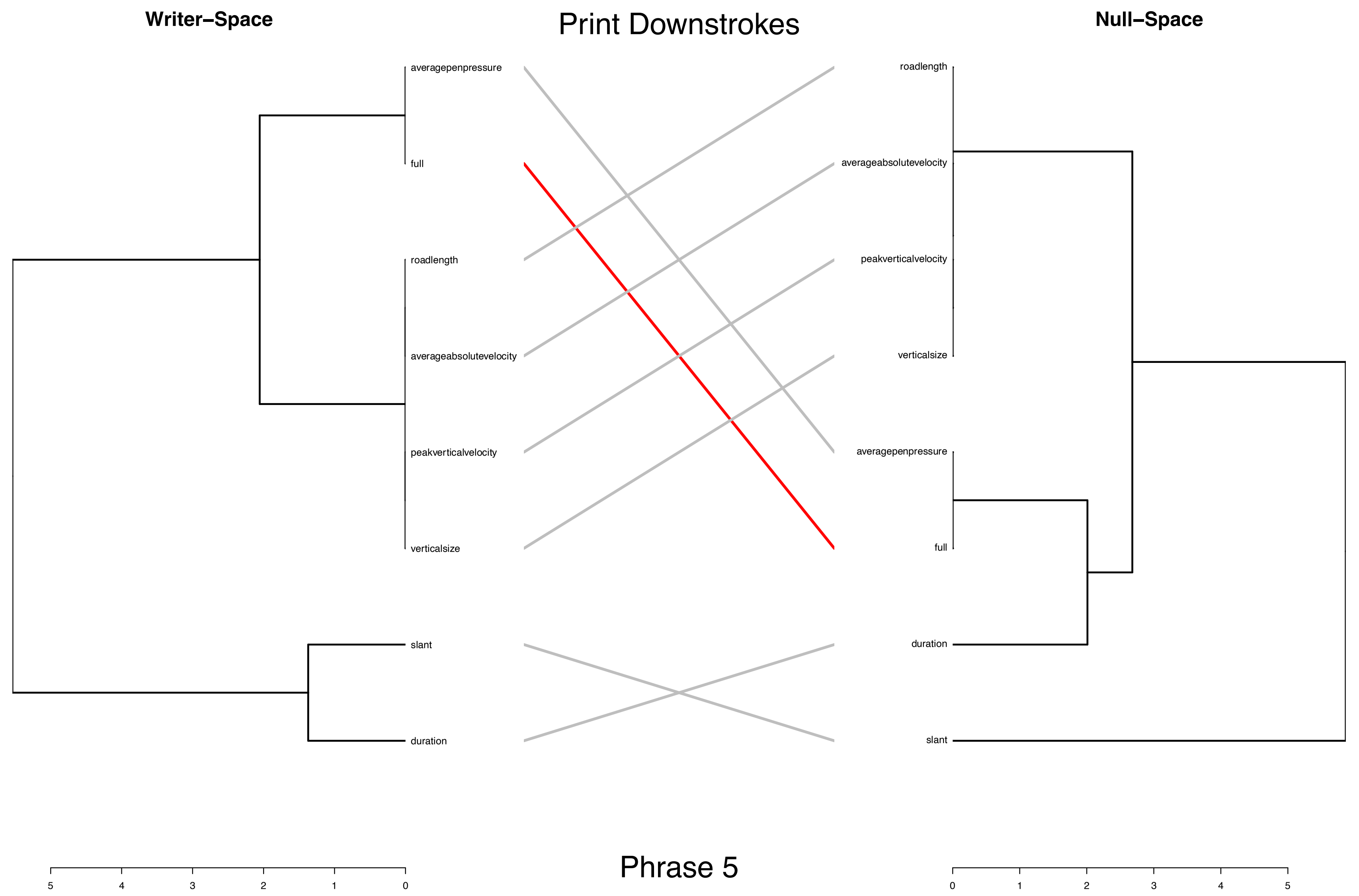
Phrase 6

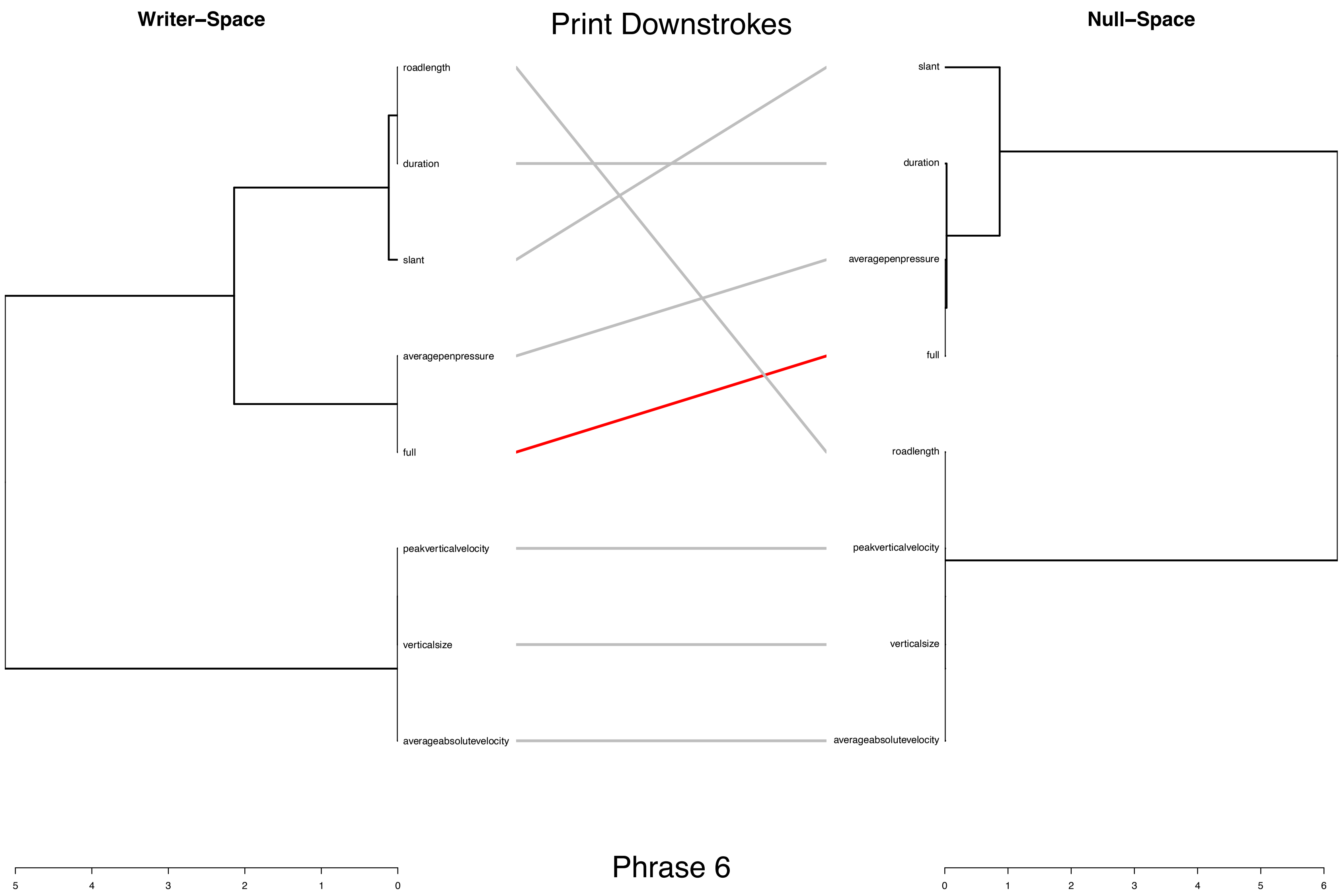

Writer–Space
Print Downstrokes
Null–Space
averageabsolutevelocity
peakverticalvelocity
roadlength
verticalsize
slant
averagepenpressure
full
duration
averageabsolutevelocity
peakverticalvelocity
full
roadlength
verticalsize
averagepenpressure
slant
duration
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Phrase 1
0
1
2
3
4


Writer–Space
Print Downstrokes
Null–Space
averageabsolutevelocity
peakverticalvelocity
roadlength
verticalsize
slant
averagepenpressure
duration
full
averageabsolutevelocity
verticalsize
roadlength
peakverticalvelocity
full
slant
averagepenpressure
duration
8
6
4
2
0
Phrase 2
0
1
2
3
4


Writer–Space
Print Downstrokes
Null–Space
averagepenpressure
full
duration
slant
averageabsolutevelocity
peakverticalvelocity
roadlength
verticalsize
averagepenpressure
full
duration
slant
averageabsolutevelocity
peakverticalvelocity
verticalsize
roadlength
6
5
4
3
2
1
0
Phrase 3
0
1
2
3
4


Writer–Space
Print Downstrokes
Null–Space
slant
duration
averagepenpressure
full
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
slant
averagepenpressure
duration
full
averageabsolutevelocity
peakverticalvelocity
verticalsize
roadlength
6
5
4
3
2
1
0
Phrase 4
0
1
2
3
4
5
6


Writer–Space
Print Downstrokes
Null–Space
averagepenpressure
full
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
slant
duration
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
averagepenpressure
full
duration
slant
5
4
3
2
1
0
Phrase 5
0
1
2
3
4
5

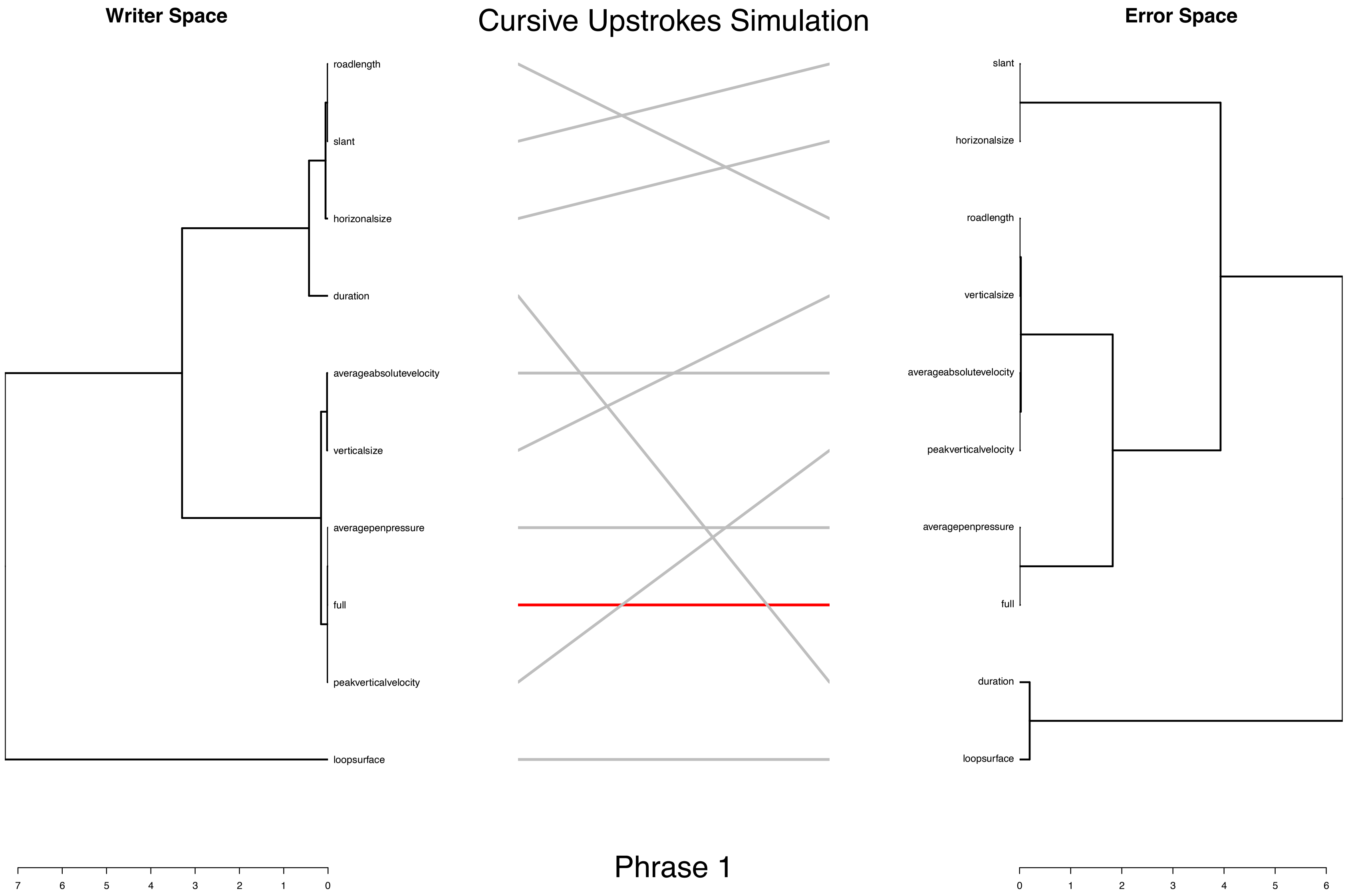

Writer–Space
Print Downstrokes
Null–Space
roadlength
duration
slant
averagepenpressure
full
peakverticalvelocity
verticalsize
averageabsolutevelocity
slant
duration
averagepenpressure
full
roadlength
peakverticalvelocity
verticalsize
averageabsolutevelocity
5
4
3
2
1
0
Phrase 6
0
1
2
3
4
5
6

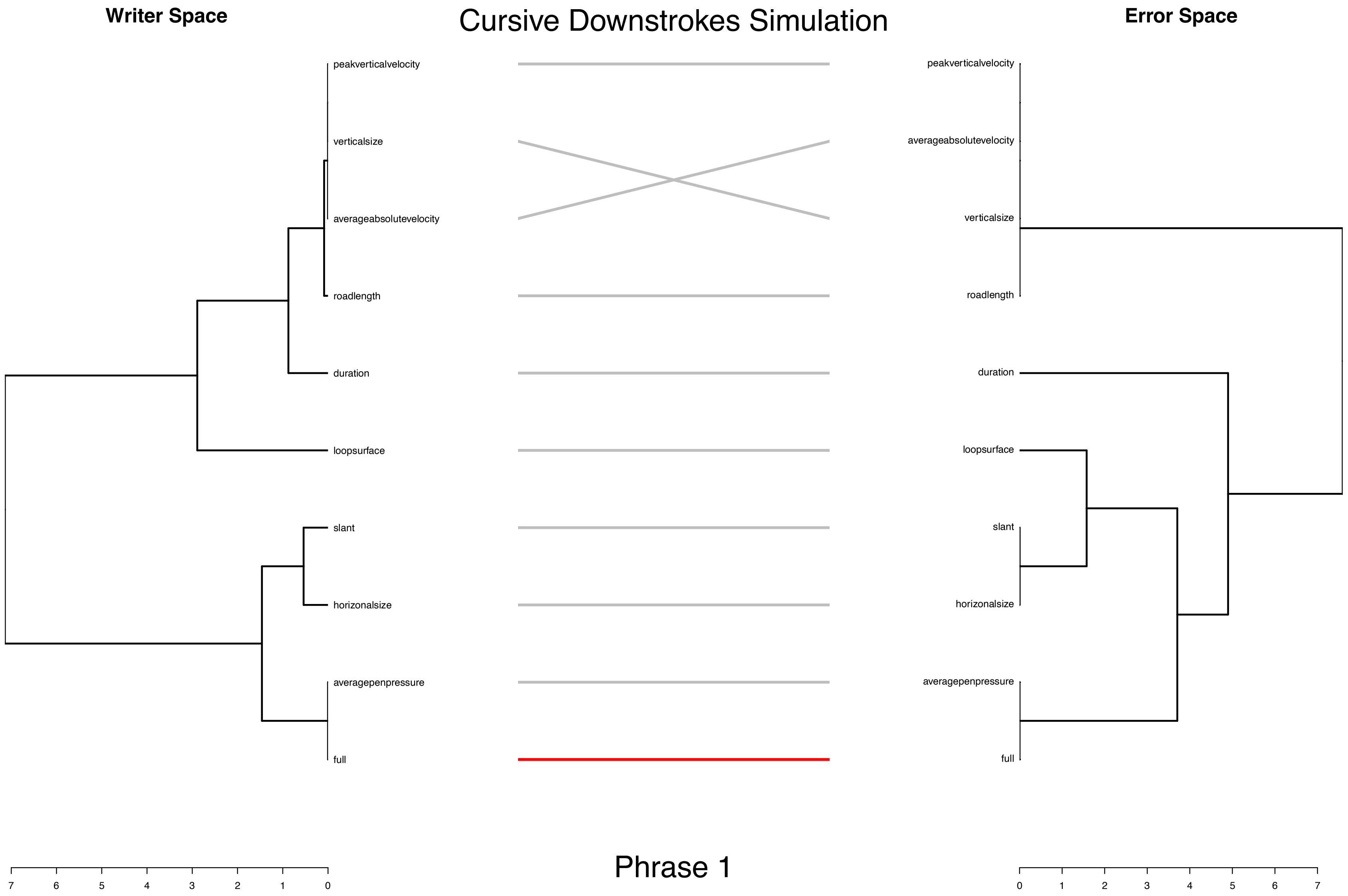
APPENDIX 3

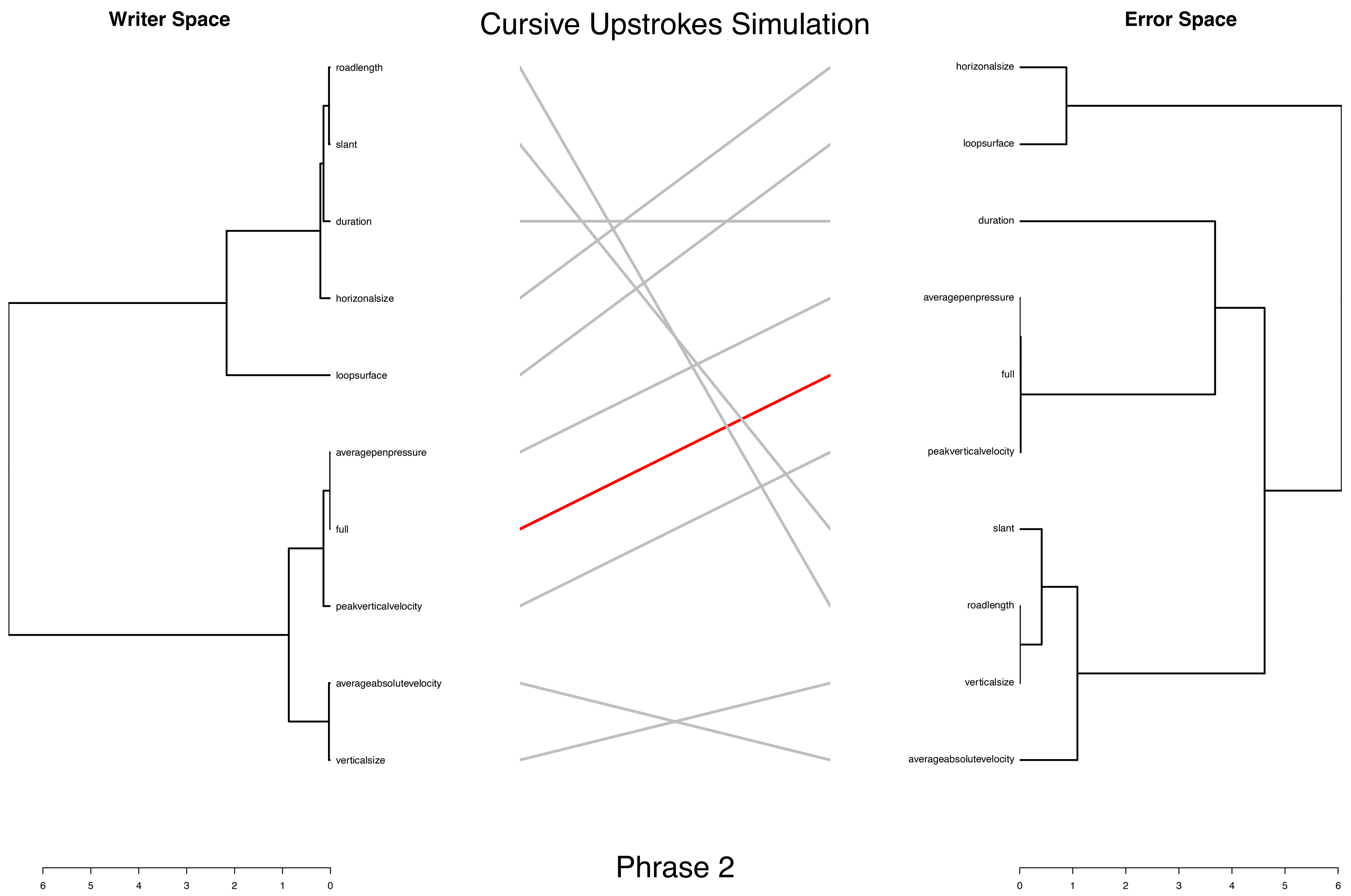
The following plot is a visual representation of the relationships between pairwise features from regression modeling of the simulated MovAlyzeR® WDS. The left side of each plot represents the models fit in the Writer-space, and the right side of each plot represents the models fit in the Null-space. The x-axes represent a transformation of the p-values associated with $\hat{\beta}_1$ of $-2 * log(1 - p\text{-value})$. Note that $-2 * log(1 - 0.05) \approx 0.1$. The solid red line in each plot represents the 'full' score, i.e. the WDS calculated with all of the other features pooled together.

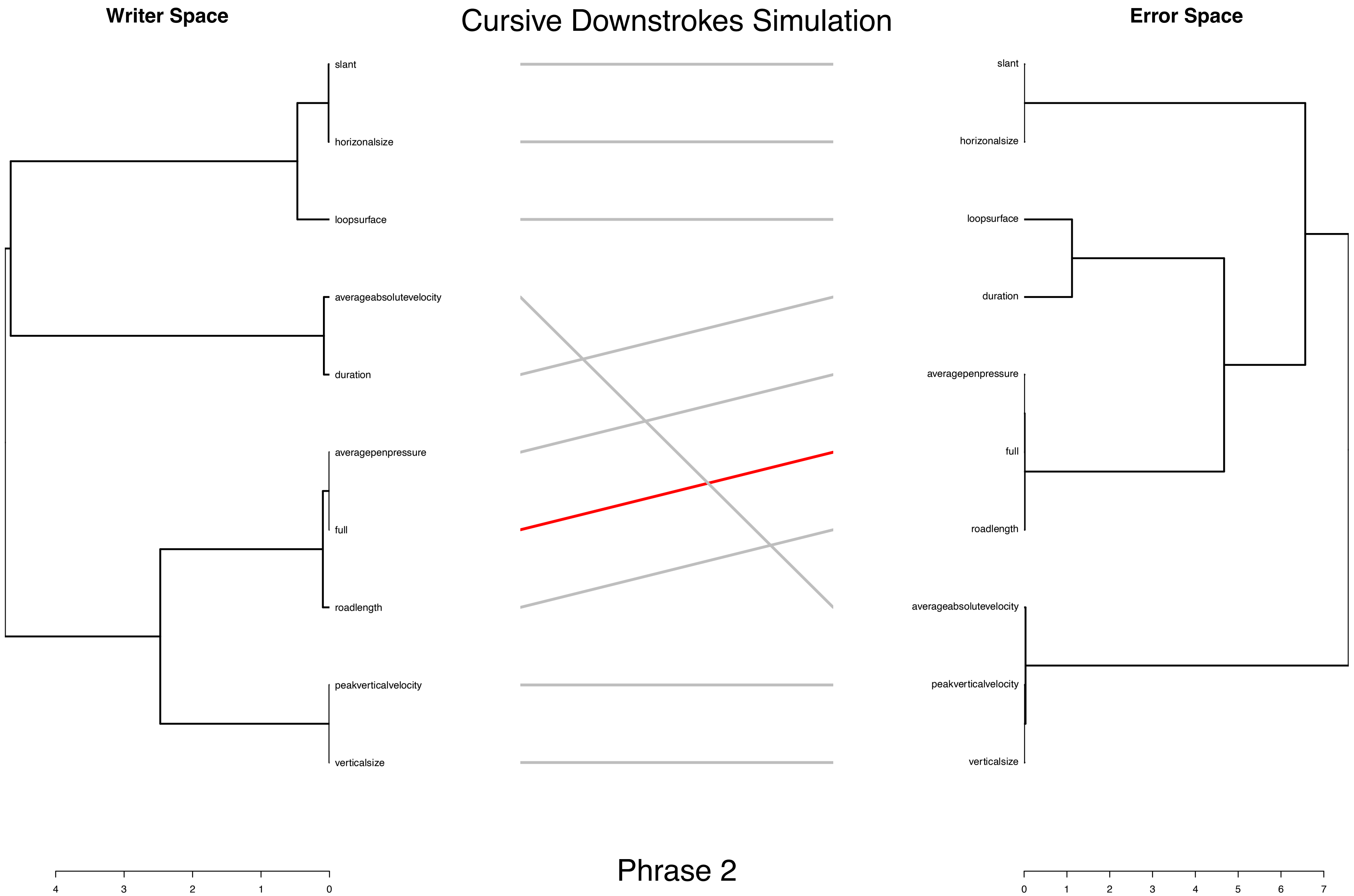
One of the main takeaways is that the Full WDS still maintains a relationship with the Average Pen Pressure WDS in both the Writer-space and Null-space.

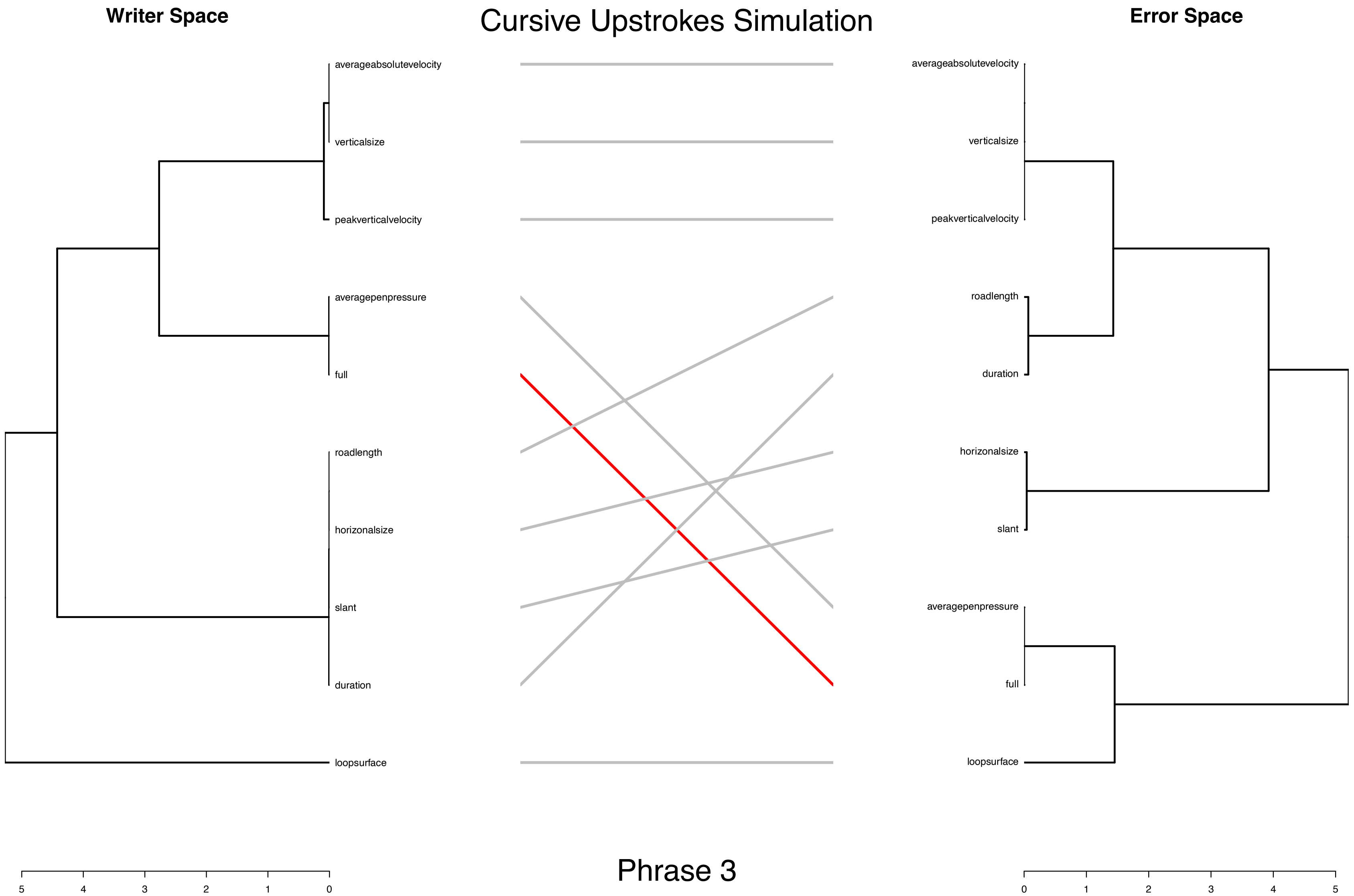
Relationships found in the simulation tanglegram that are not found in the Appendix 2 tanglegrams, or vice versa, may be indicative that the naive assumption that each feature follows a Normal distribution with sample mean and sample variance may not be entirely correct.


Writer Space
Cursive Upstrokes Simulation
Error Space
roadlength
slant
horizonalsize
duration
averageabsolutevelocity
verticalsize
averagepenpressure
full
peakverticalvelocity
loopsurface
slant
horizonalsize
roadlength
verticalsize
averageabsolutevelocity
peakverticalvelocity
averagepenpressure
full
duration
loopsurface
7
6
5
4
3
2
1
0
Phrase 1
0
1
2
3
4
5
6

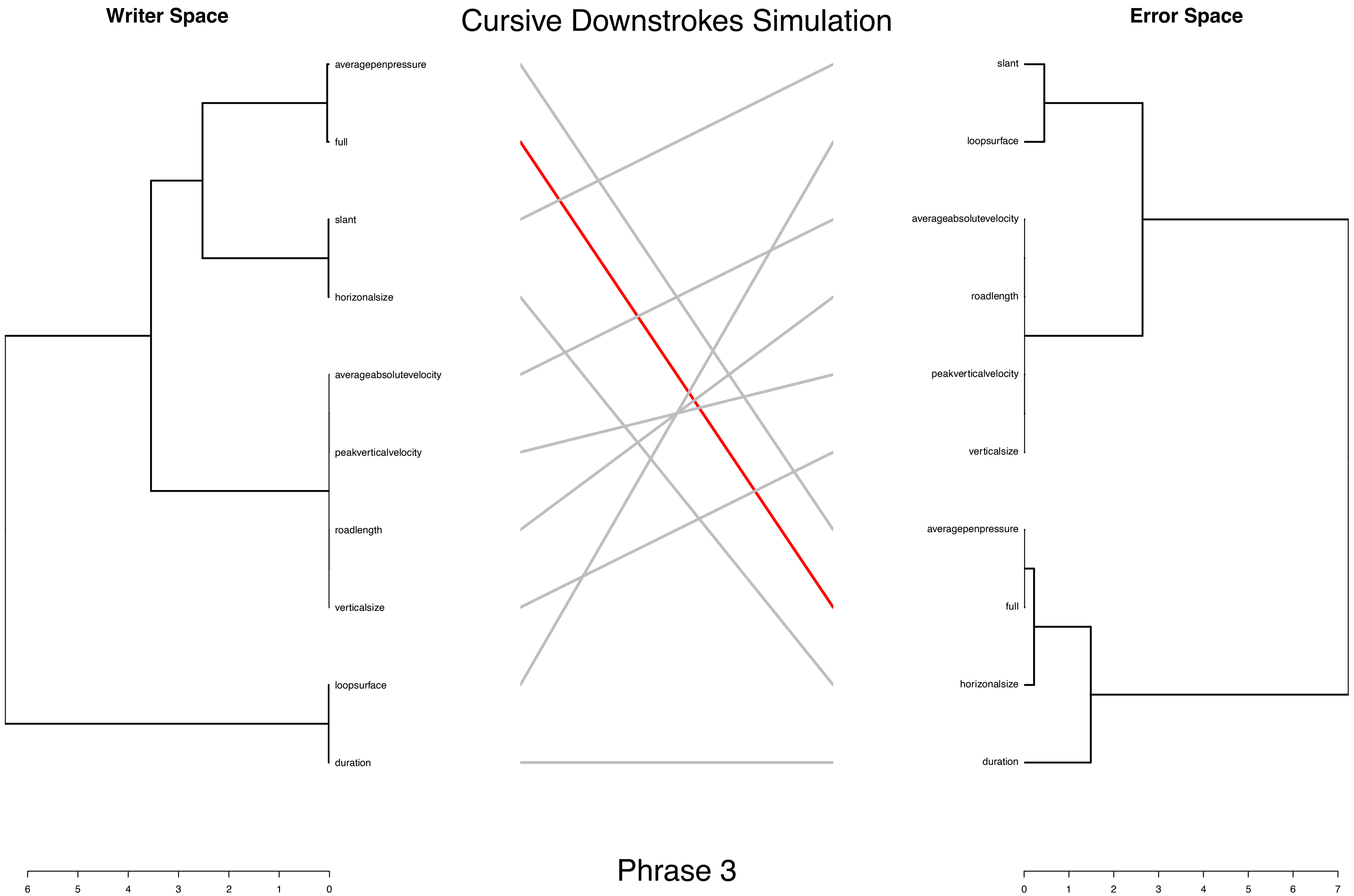

Writer Space
Cursive Downstrokes Simulation
Error Space
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
duration
loopsurface
slant
horizonalsize
averagepenpressure
full
peakverticalvelocity
averageabsolutevelocity
verticalsize
roadlength
duration
loopsurface
slant
horizonalsize
averagepenpressure
full
7
6
5
4
3
2
1
0
Phrase 1
0
1
2
3
4
5
6
7


Writer Space
Cursive Upstrokes Simulation
Error Space
roadlength
slant
duration
horizonalsize
loopsurface
averagepenpressure
full
peakverticalvelocity
averageabsolutevelocity
verticalsize
6
5
4
3
2
1
0
horizonalsize
loopsurface
duration
averagepenpressure
full
peakverticalvelocity
slant
roadlength
verticalsize
averageabsolutevelocity
0
1
2
3
4
5
6
Phrase 2

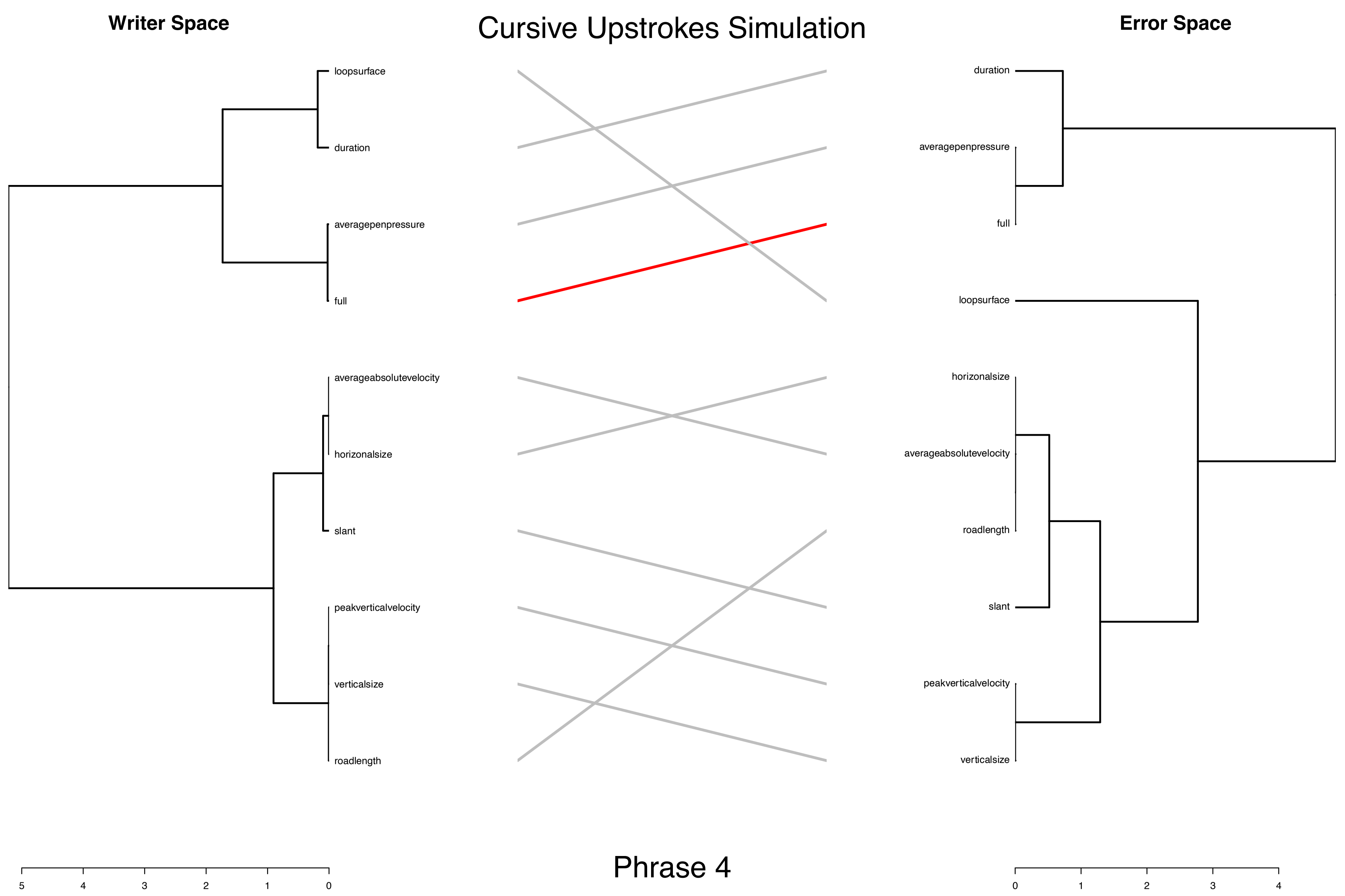

Writer Space
Cursive Downstrokes Simulation
Error Space
slant
horizonalsize
loopsurface
averageabsolutevelocity
duration
averagepenpressure
full
roadlength
peakverticalvelocity
verticalsize
slant
horizonalsize
loopsurface
duration
averagepenpressure
full
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
4
3
2
1
0
Phrase 2
0
1
2
3
4
5
6
7

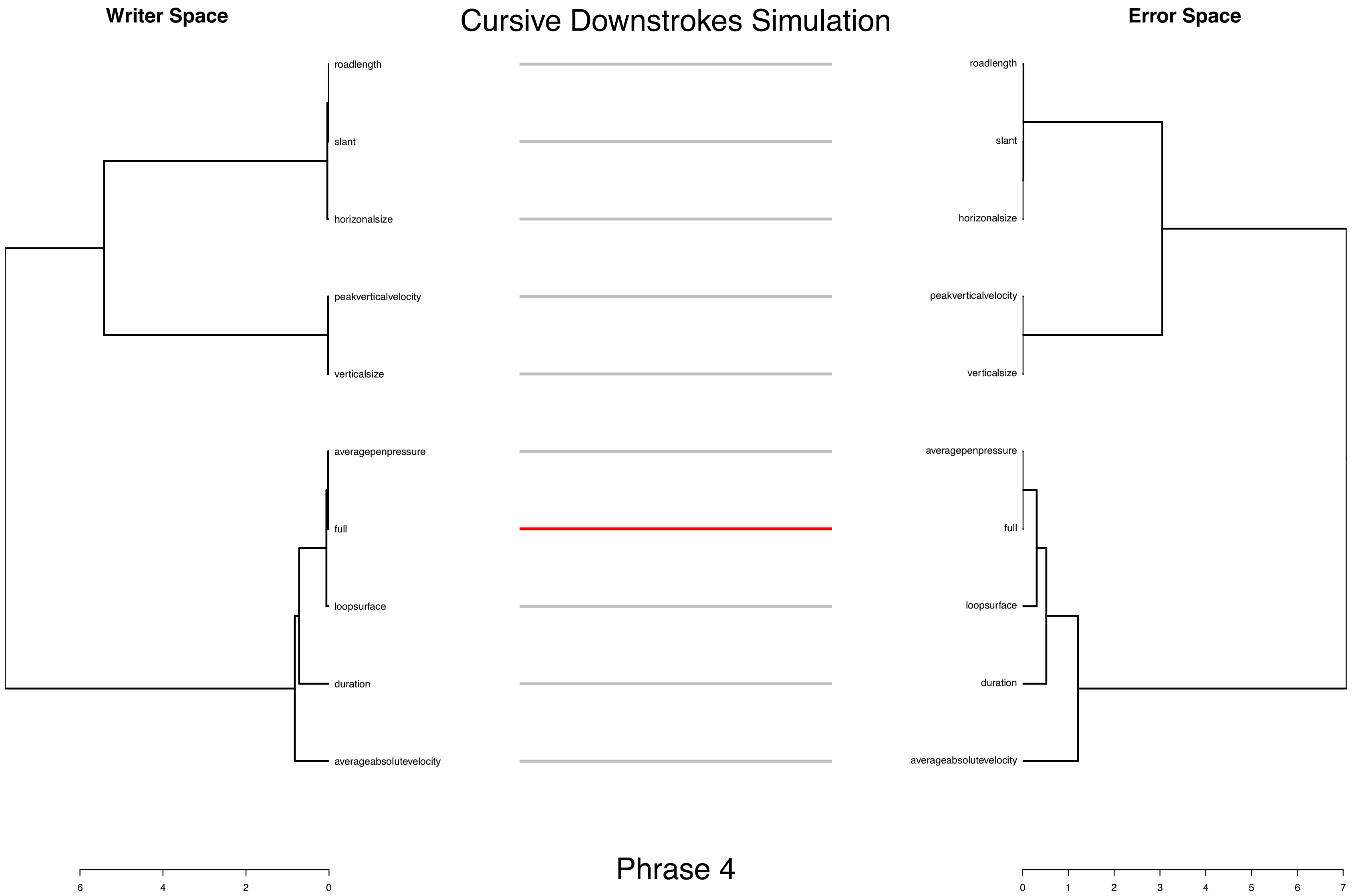

Cursive Upstrokes Simulation
Writer Space
Error Space
averageabsolutevelocity
verticalsize
peakverticalvelocity
averagepenpressure
full
roadlength
horizonalsize
slant
duration
loopsurface
averageabsolutevelocity
verticalsize
peakverticalvelocity
roadlength
duration
horizonalsize
slant
averagepenpressure
full
loopsurface
5
4
3
2
1
0
0
1
2
3
4
5
Phrase 3

Cursive Downstrokes Simulation

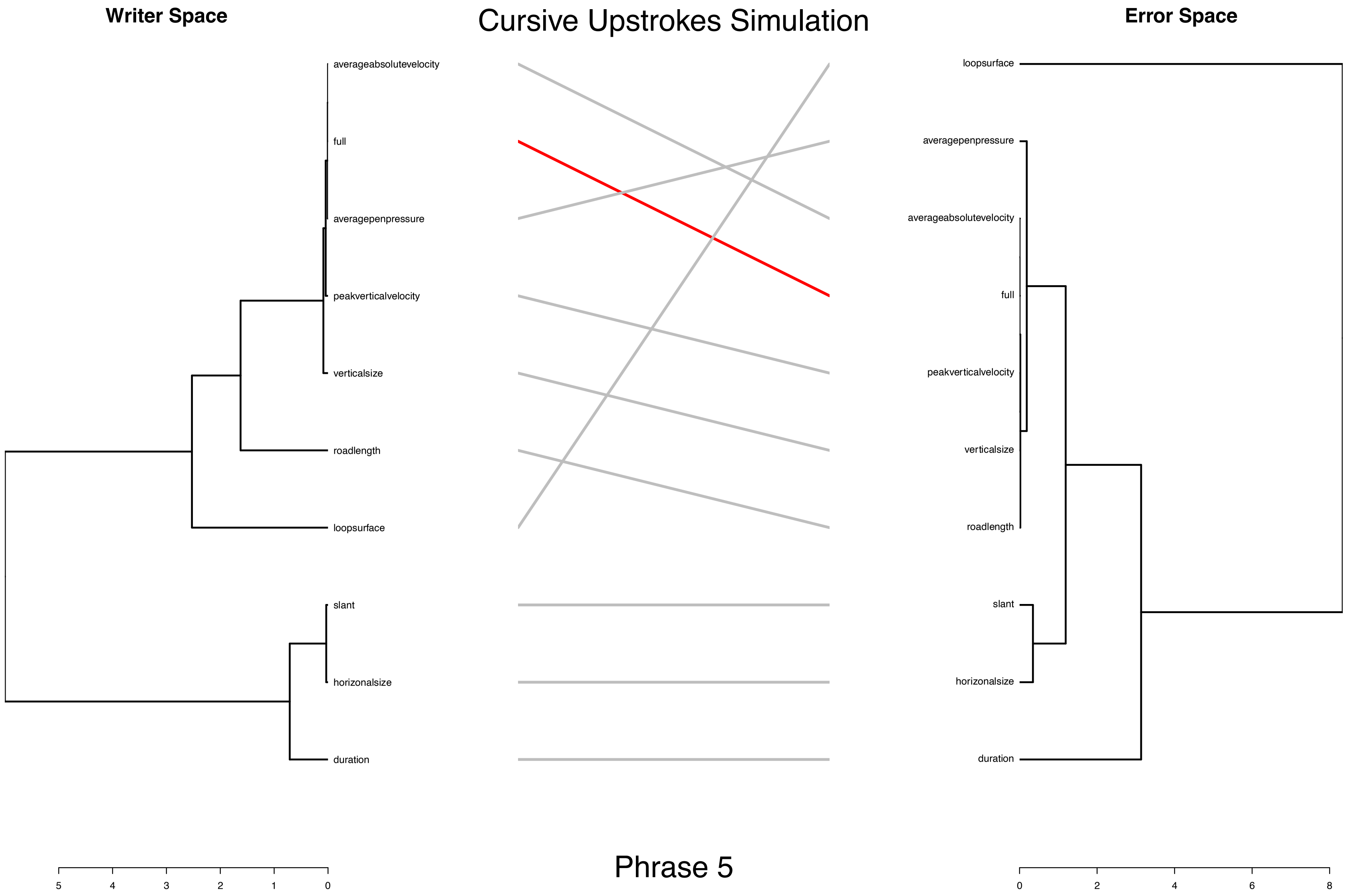


Phrase 3

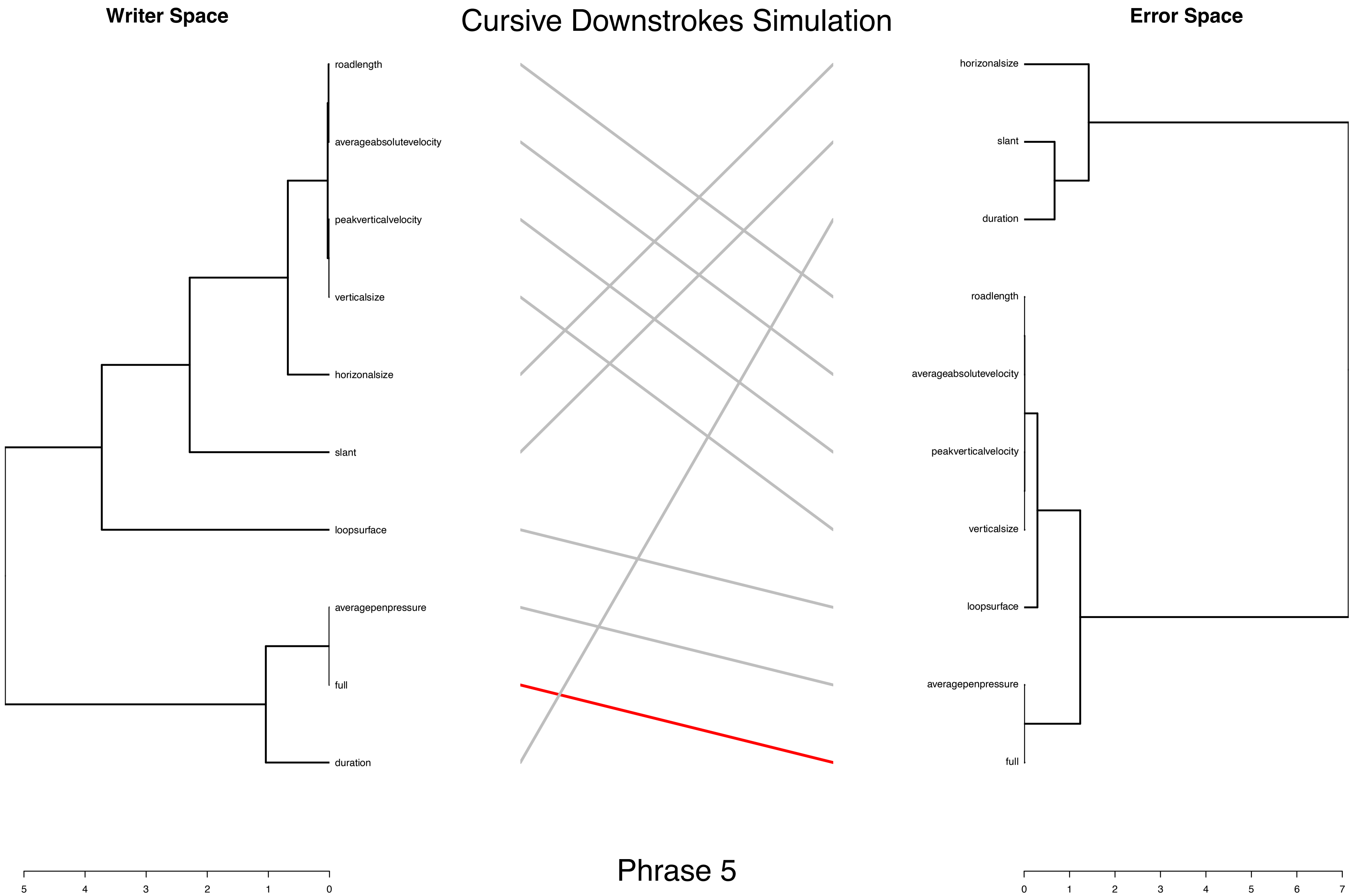

Writer Space
Cursive Upstrokes Simulation
Error Space
loopsurface
duration
averagepenpressure
full
averageabsolutevelocity
horizonalsize
slant
peakverticalvelocity
verticalsize
roadlength
duration
averagepenpressure
full
loopsurface
horizonalsize
averageabsolutevelocity
roadlength
slant
peakverticalvelocity
verticalsize
5
4
3
2
1
0
Phrase 4
0
1
2
3
4

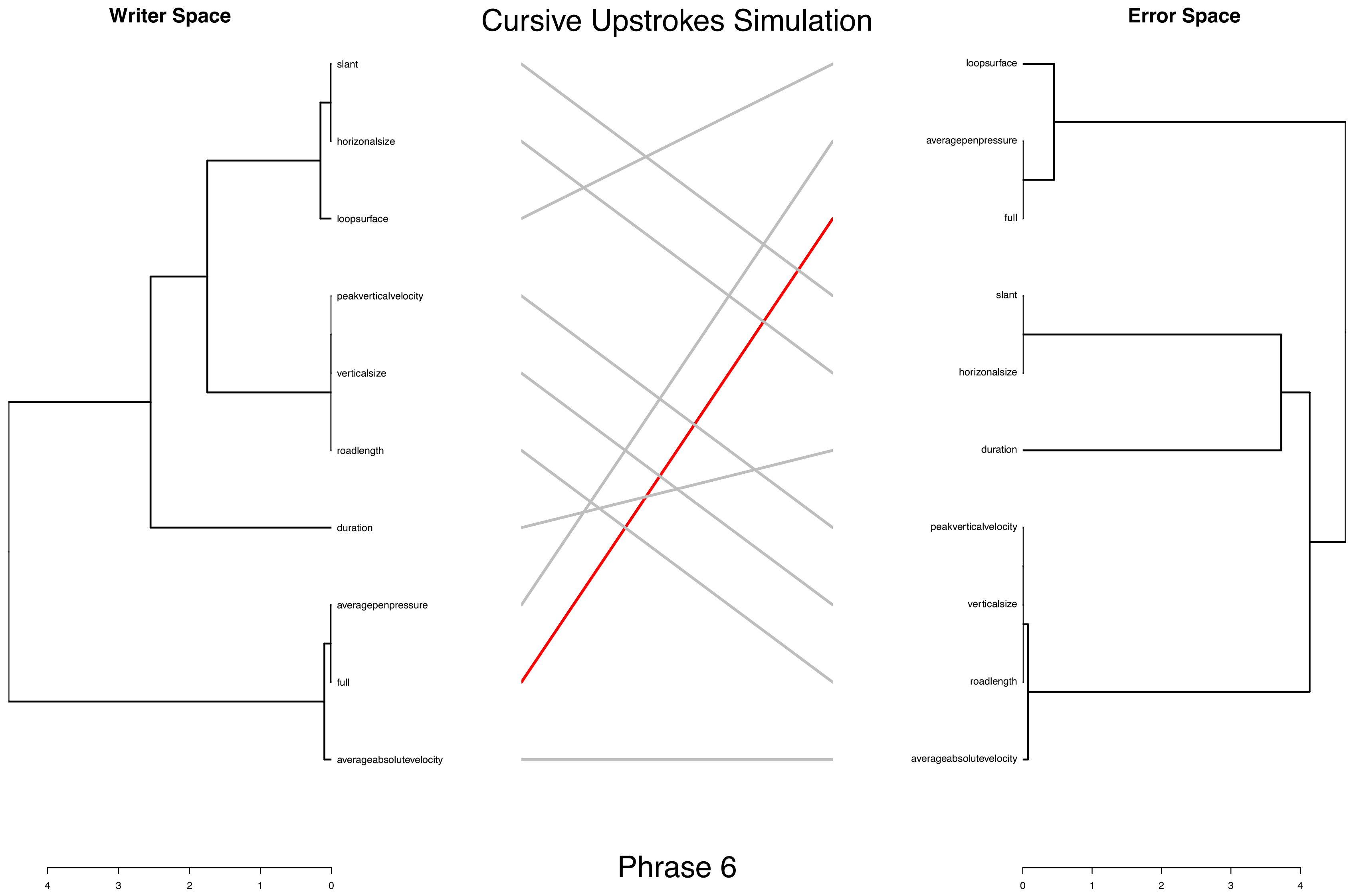

Cursive Downstrokes Simulation
Writer Space
Error Space
roadlength
slant
horizonalsize
peakverticalvelocity
verticalsize
averagepenpressure
full
loopsurface
duration
averageabsolutevelocity
6
4
2
0
0
1
2
3
4
5
6
7
Phrase 4

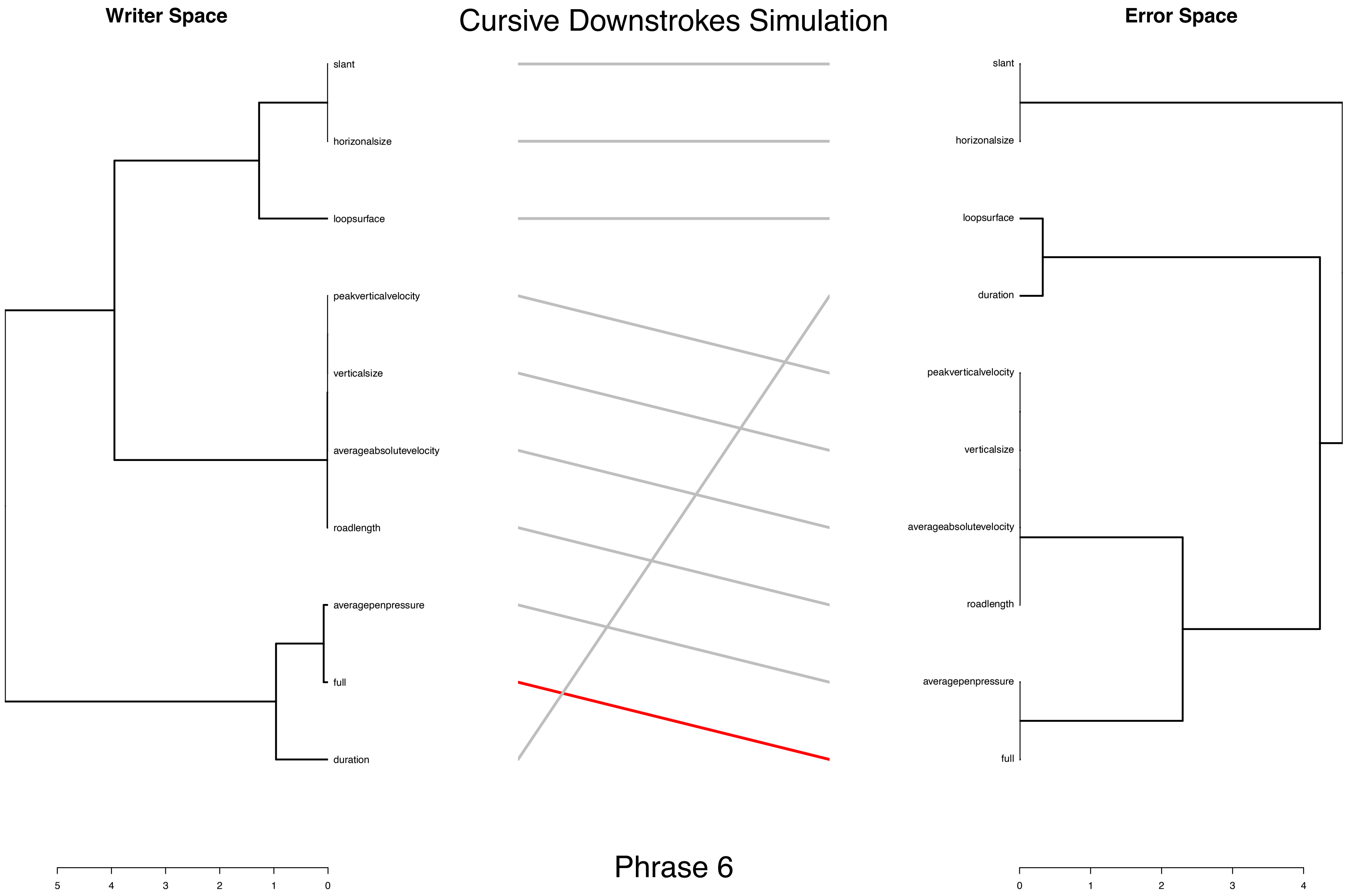
Cursive Upstrokes Simulation



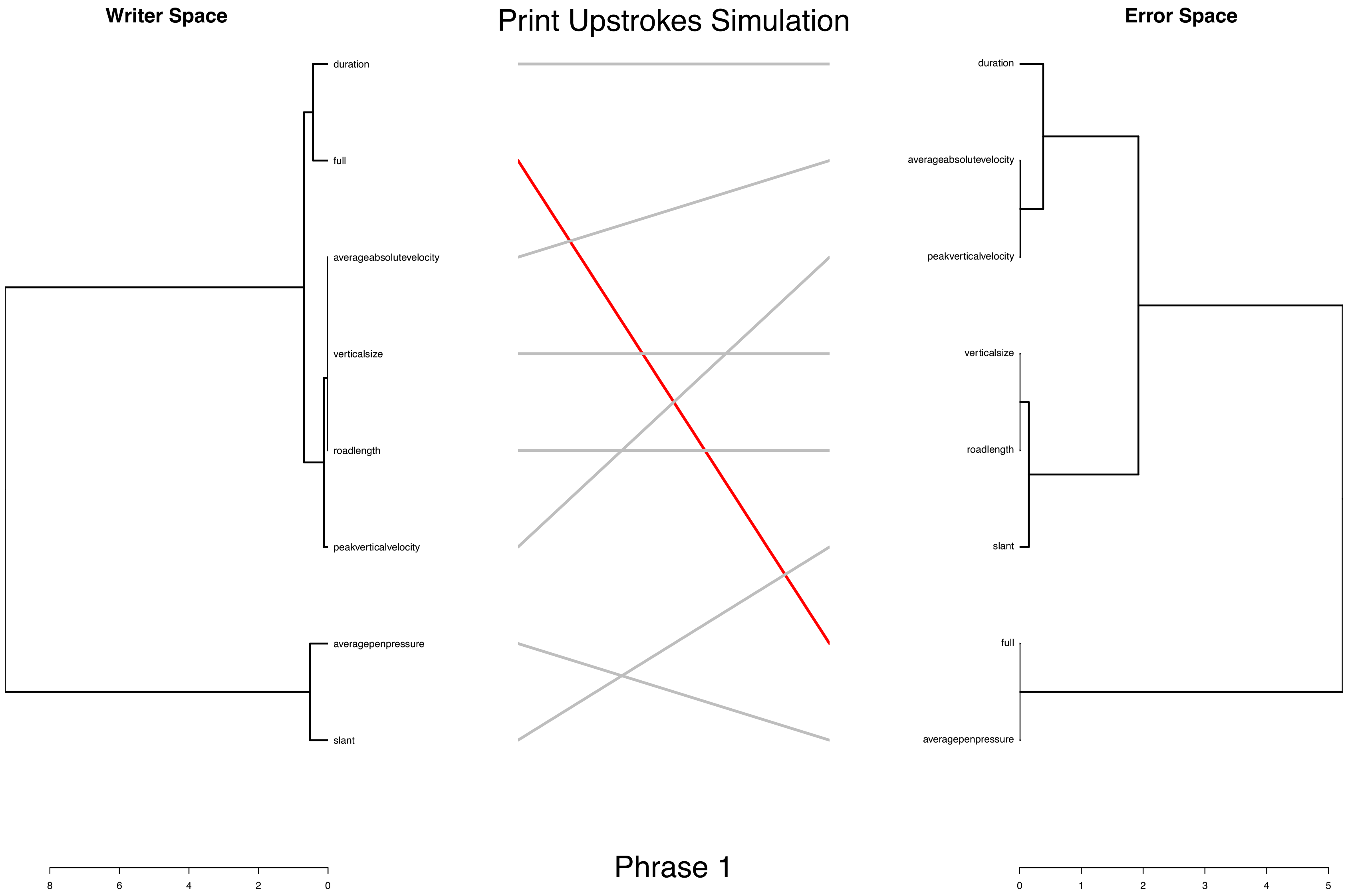
Phrase 5


Writer Space
Cursive Downstrokes Simulation
Error Space
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
horizonalsize
slant
loopsurface
averagepenpressure
full
duration
5
4
3
2
1
0
horizonalsize
slant
duration
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
loopsurface
averagepenpressure
full
0
1
2
3
4
5
6
7
Phrase 5


Writer Space
Cursive Upstrokes Simulation
Error Space
slant
horizonalsize
loopsurface
peakverticalvelocity
verticalsize
roadlength
duration
averagepenpressure
full
averageabsolutevelocity
loopsurface
averagepenpressure
full
slant
horizonalsize
duration
peakverticalvelocity
verticalsize
roadlength
averageabsolutevelocity
4
3
2
1
0
Phrase 6
0
1
2
3
4


Cursive Downstrokes Simulation
Writer Space
slant
horizonalsize
loopsurface
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
averagepenpressure
full
duration
5
4
3
2
1
0
Error Space
slant
horizonalsize
loopsurface
duration
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
averagepenpressure
full
0
1
2
3
4
Phrase 6

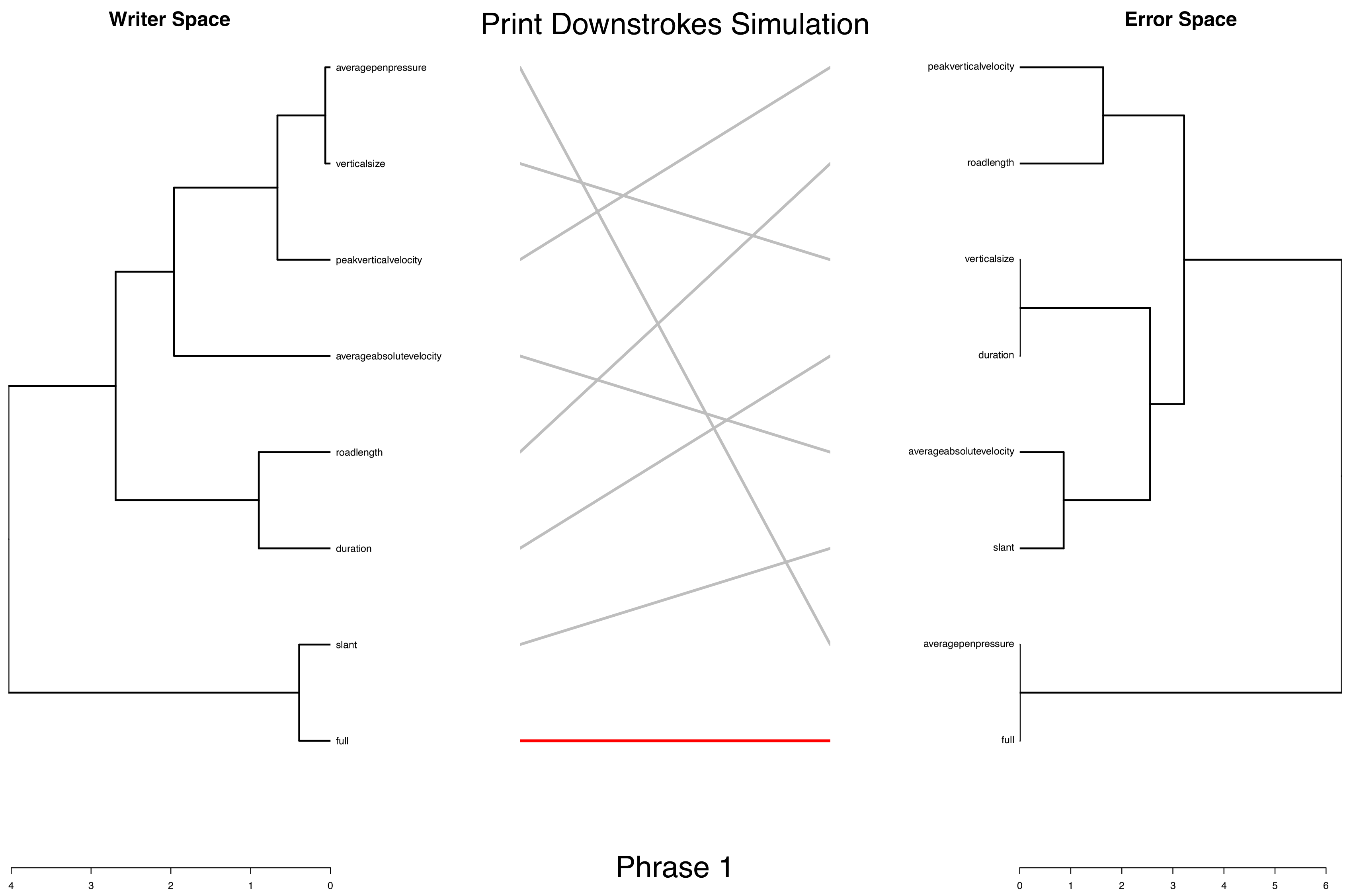
Print Upstrokes Simulation



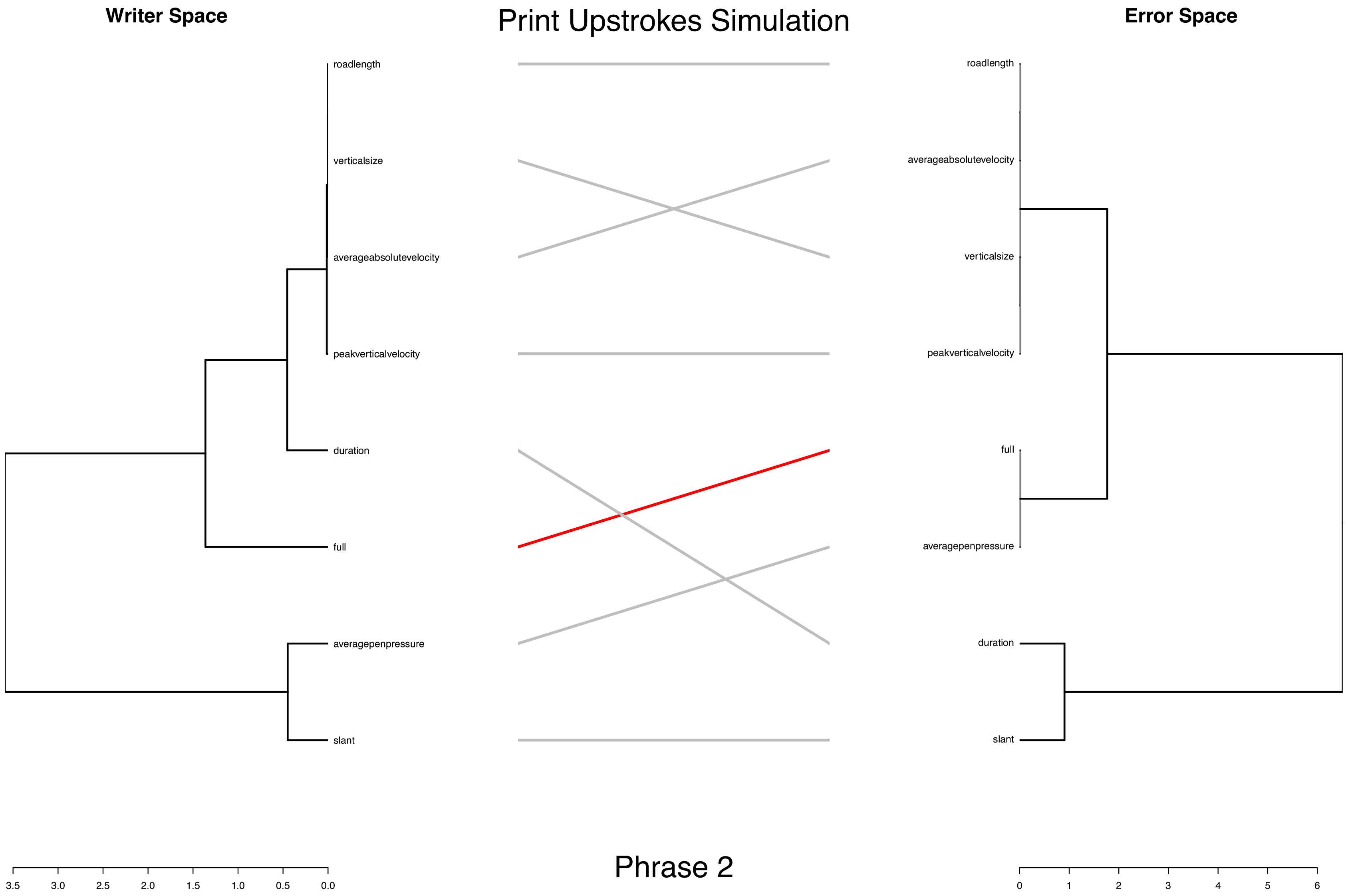
Writer Space

Error Space

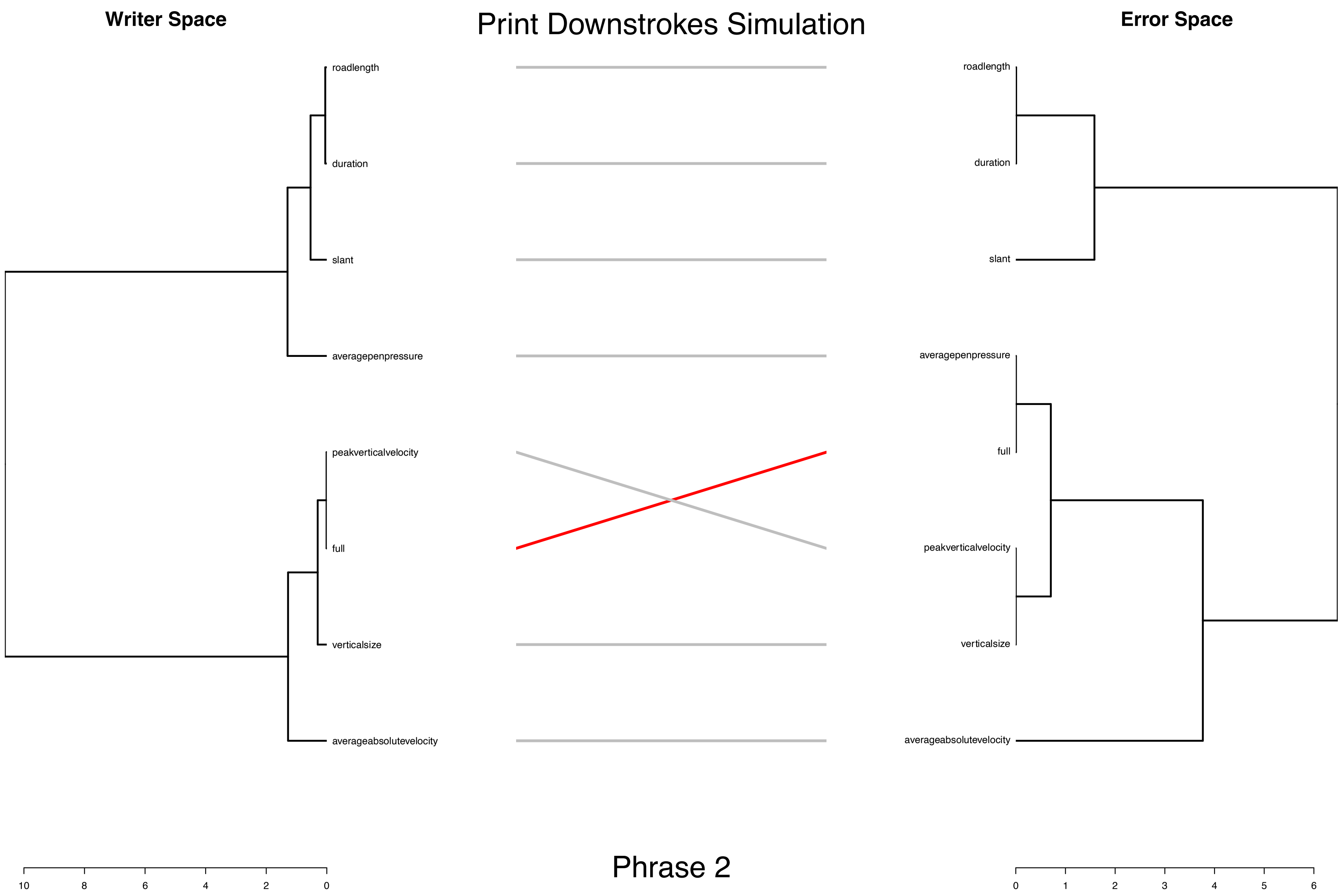
Phrase 1

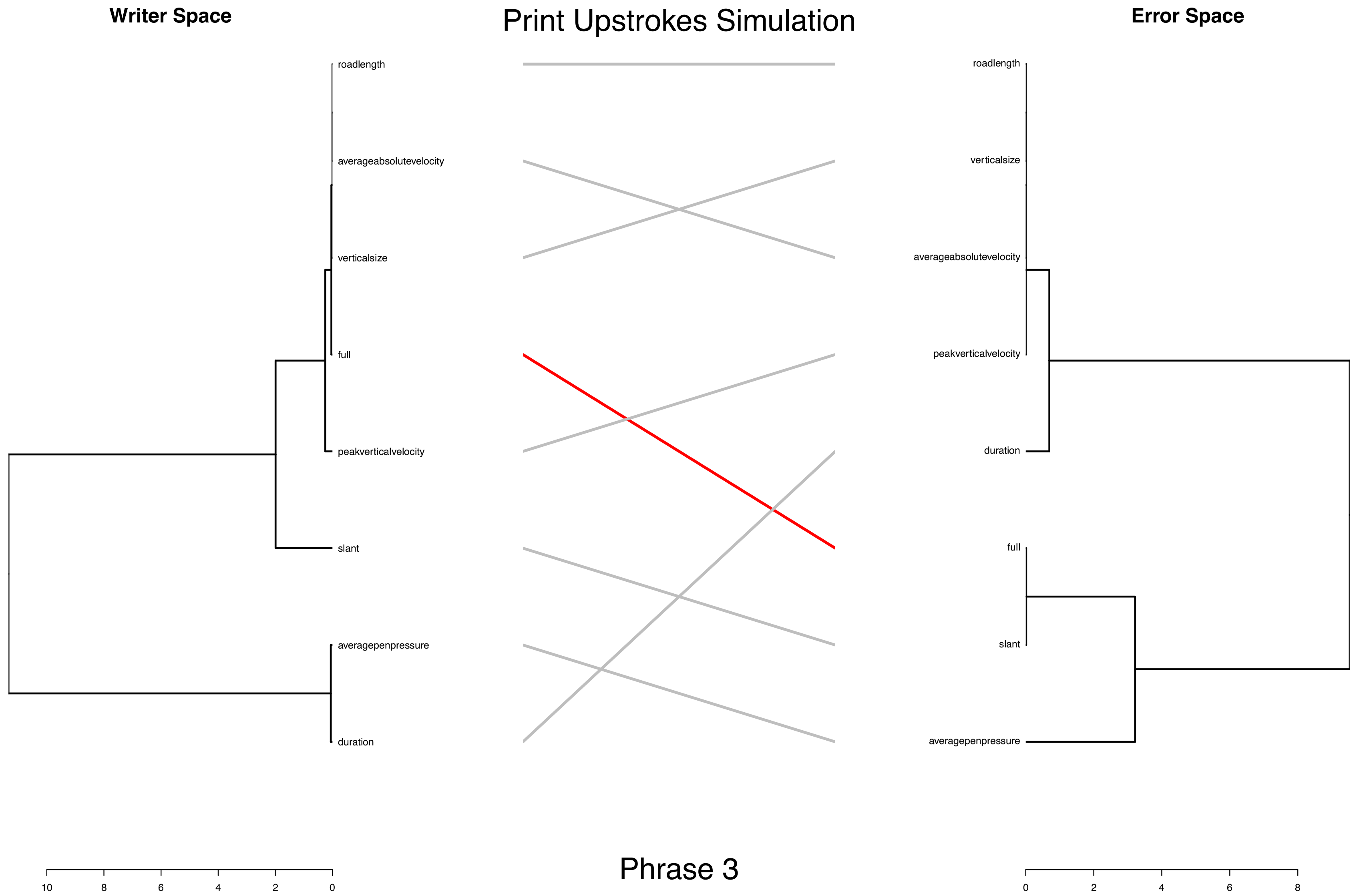

Print Downstrokes Simulation
Writer Space
averagepenpressure
verticalsize
peakverticalvelocity
averageabsolutevelocity
roadlength
duration
slant
full
4
3
2
1
0
Error Space
peakverticalvelocity
roadlength
verticalsize
duration
averageabsolutevelocity
slant
averagepenpressure
full
0
1
2
3
4
5
6
Phrase 1

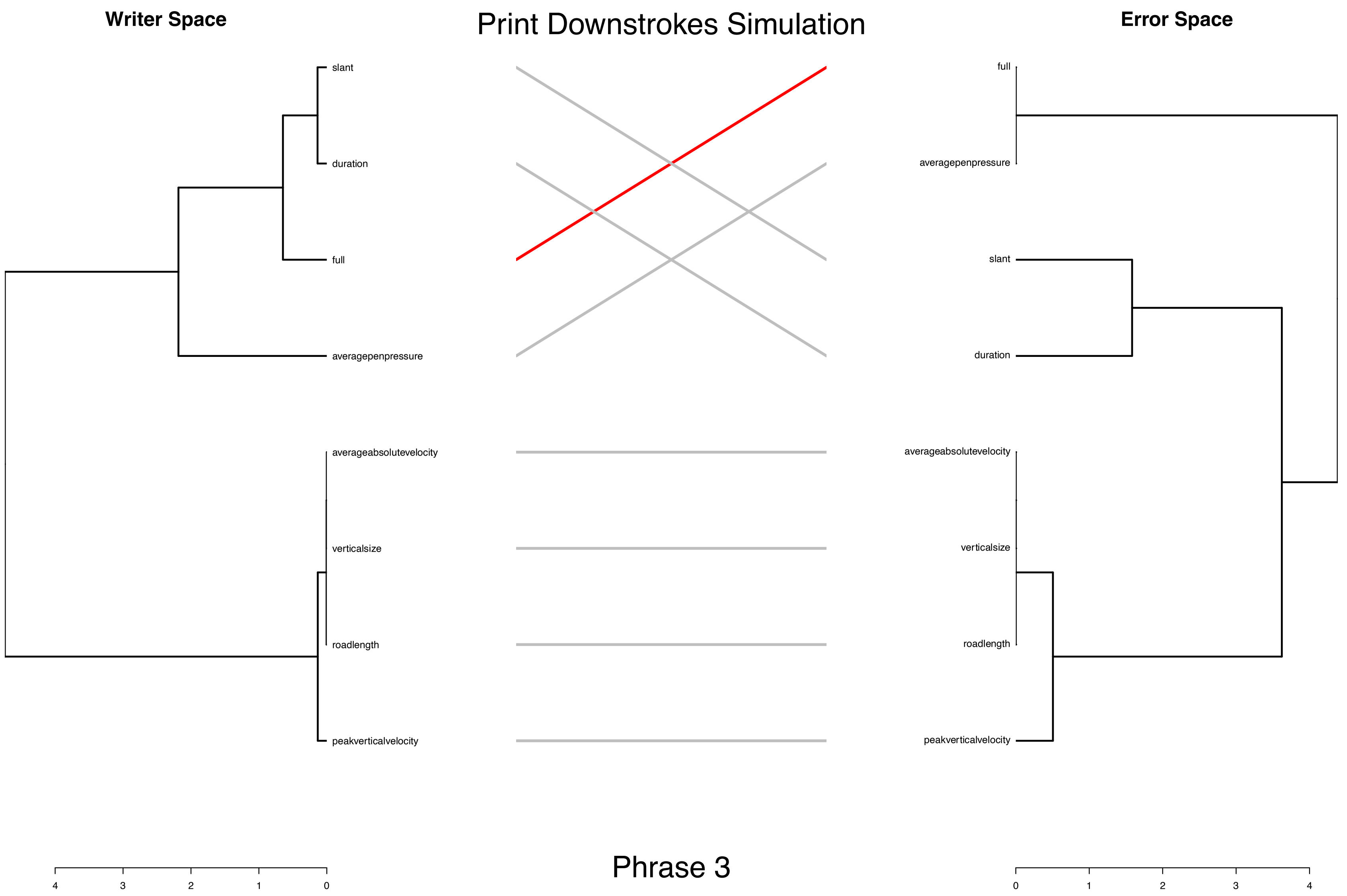
Print Upstrokes Simulation

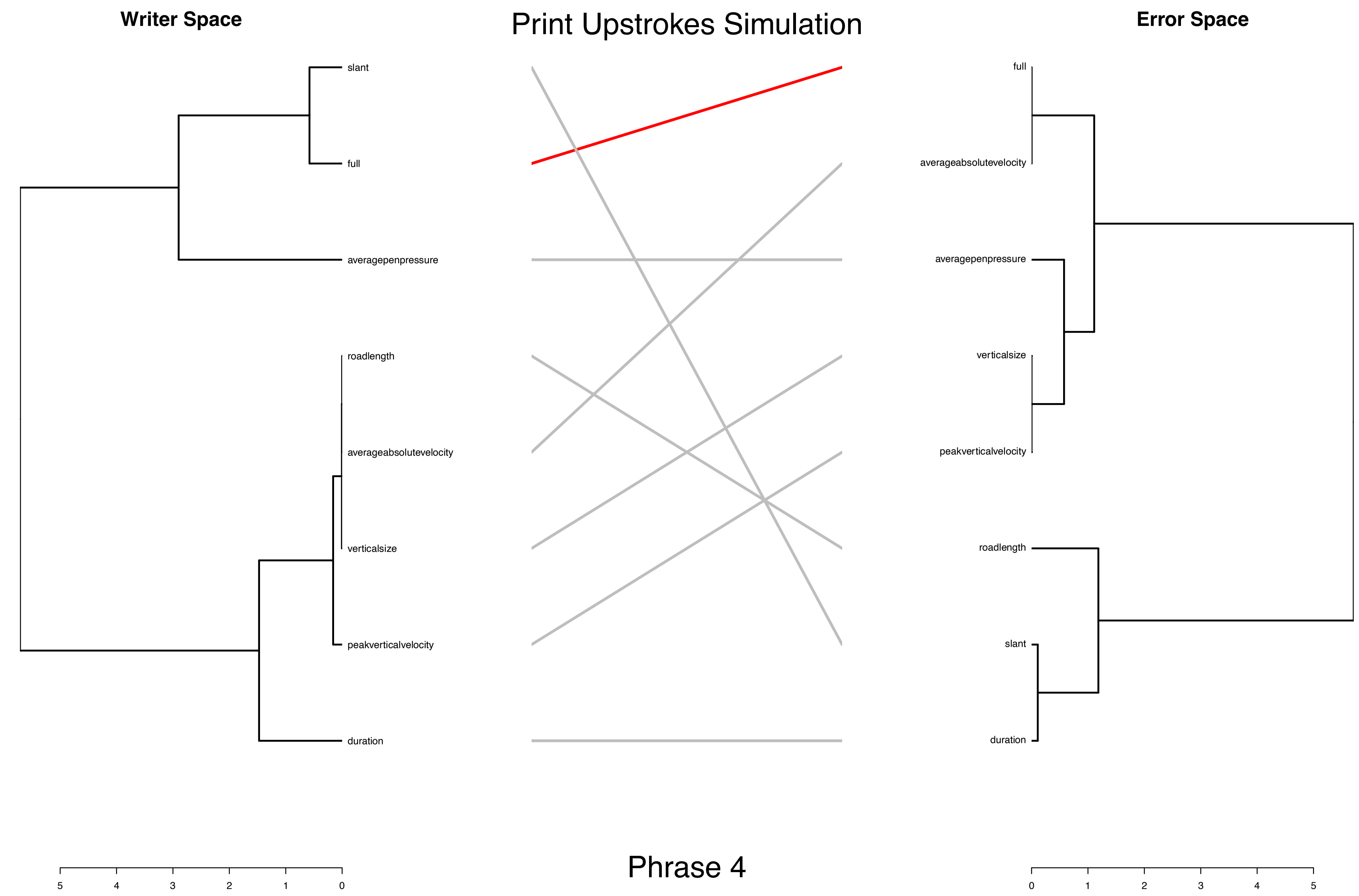


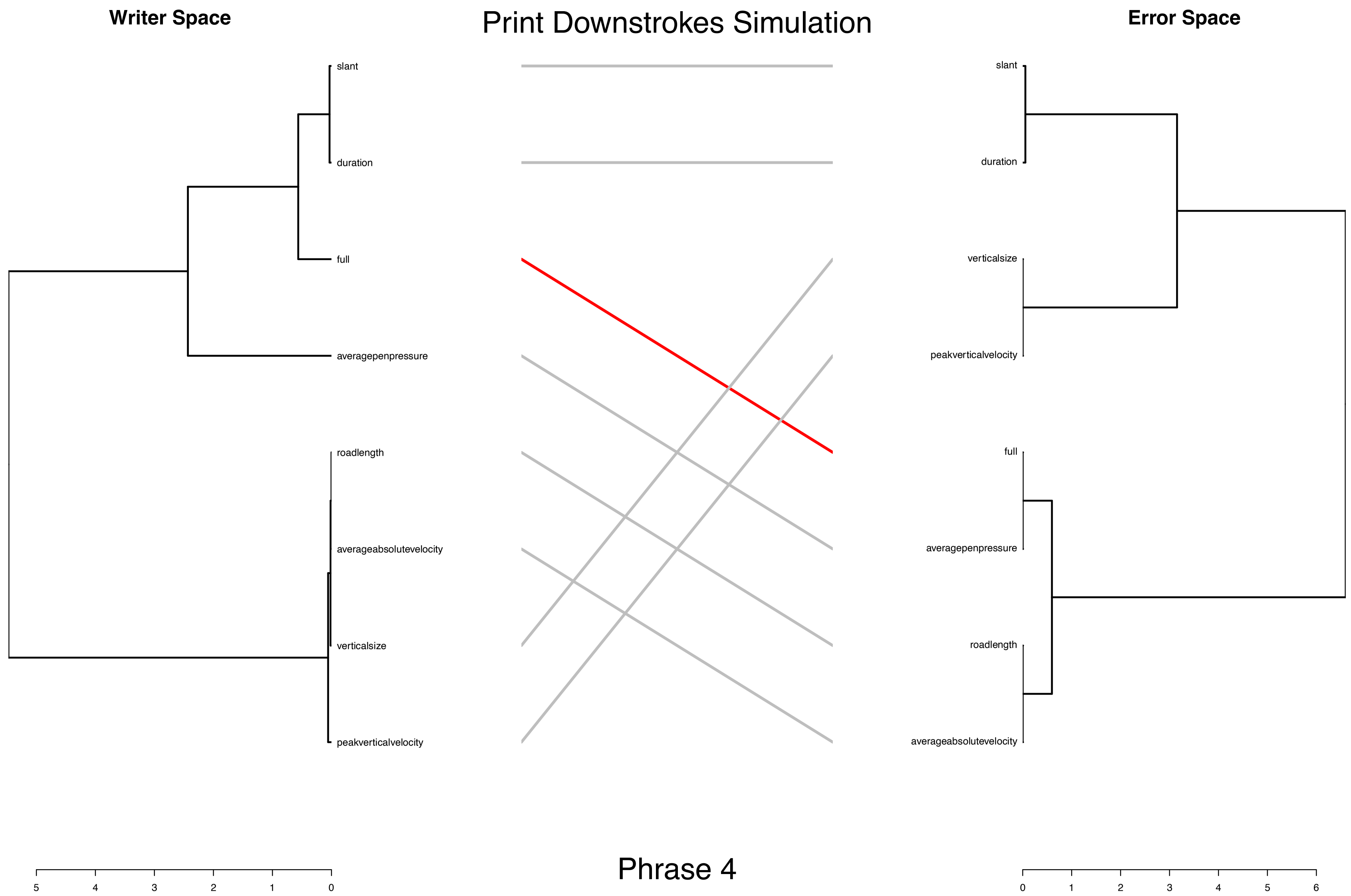
Phrase 2

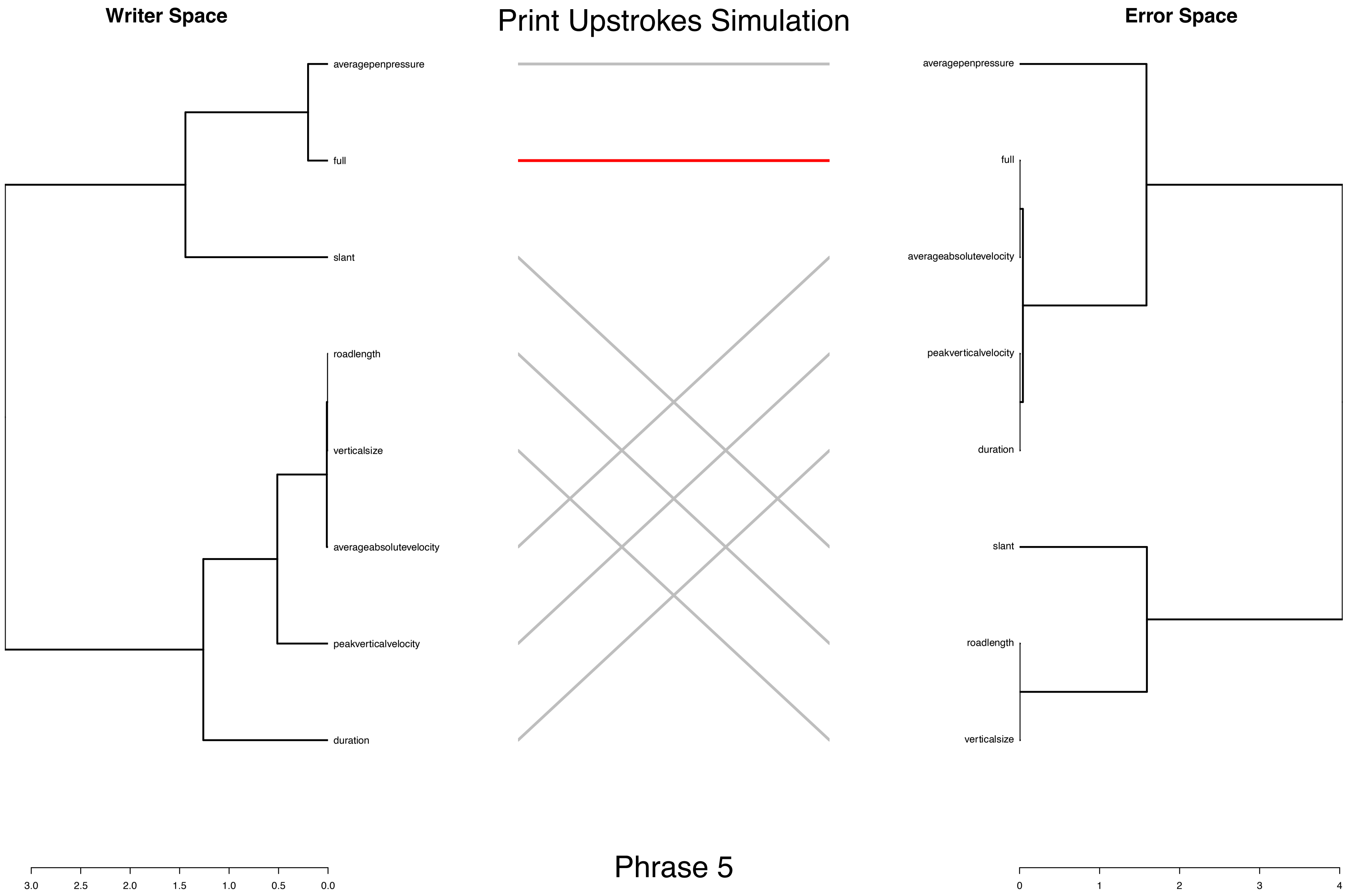

Writer Space
Print Downstrokes Simulation
Error Space
roadlength
duration
slant
averagepenpressure
peakverticalvelocity
full
verticalsize
averageabsolutevelocity
roadlength
duration
slant
averagepenpressure
full
peakverticalvelocity
verticalsize
averageabsolutevelocity
10
8
6
4
2
0
Phrase 2
0
1
2
3
4
5
6


Writer Space
Print Upstrokes Simulation
Error Space
roadlength
averageabsolutevelocity
verticalsize
full
peakverticalvelocity
slant
averagepenpressure
duration
roadlength
verticalsize
averageabsolutevelocity
peakverticalvelocity
duration
full
slant
averagepenpressure
10
8
6
4
2
0
Phrase 3
0
2
4
6
8


Writer Space
Print Downstrokes Simulation
Error Space
slant
duration
full
averagepenpressure
averageabsolutevelocity
verticalsize
roadlength
peakverticalvelocity
full
averagepenpressure
slant
duration
averageabsolutevelocity
verticalsize
roadlength
peakverticalvelocity
4
3
2
1
0
Phrase 3
0
1
2
3
4

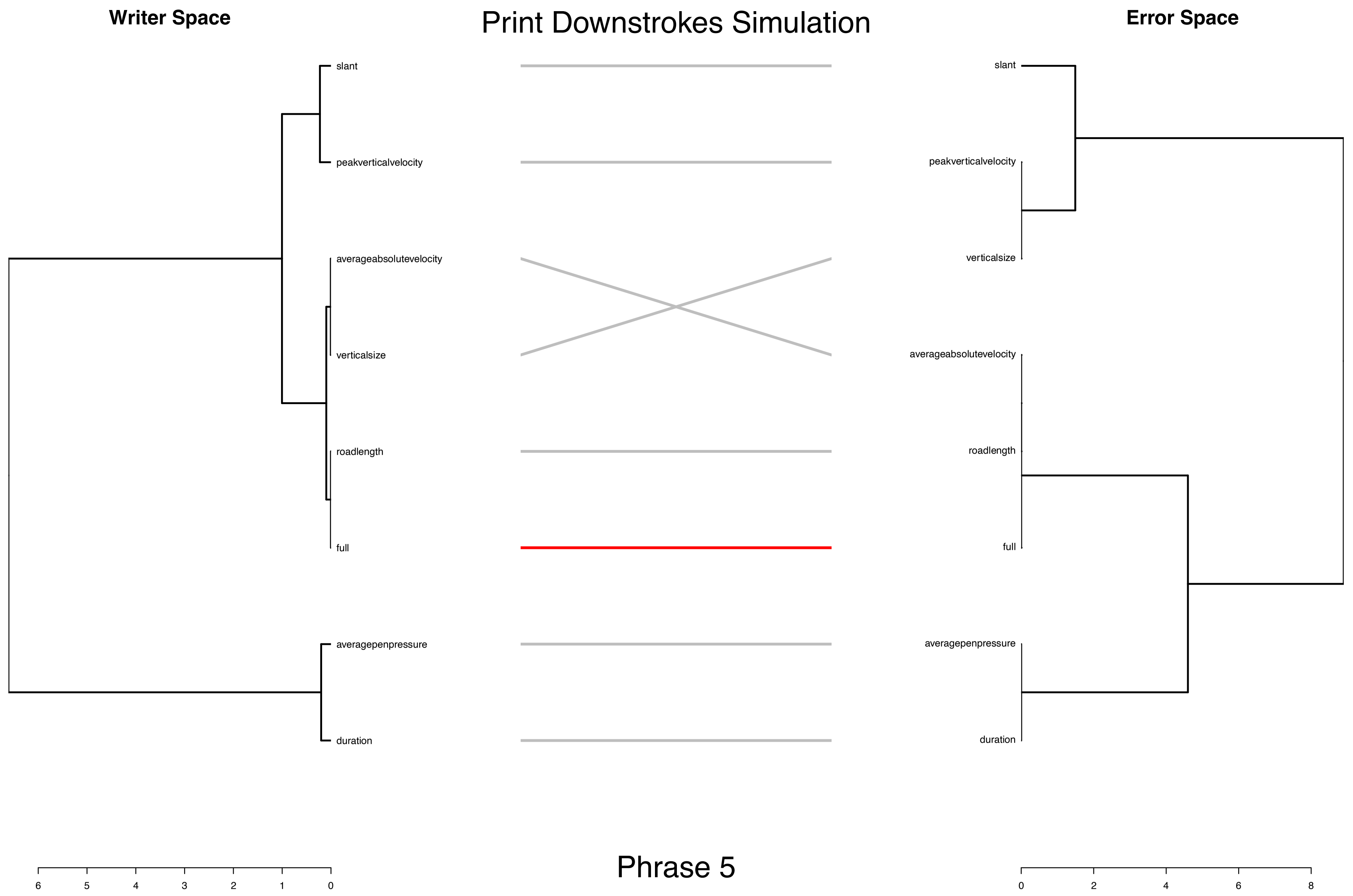

Writer Space
Print Upstrokes Simulation
Error Space
slant
full
averagepenpressure
roadlength
averageabsolutevelocity
verticalsize
peakverticalvelocity
duration
full
averageabsolutevelocity
averagepenpressure
verticalsize
peakverticalvelocity
roadlength
slant
duration
5
4
3
2
1
0
Phrase 4
0
1
2
3
4
5

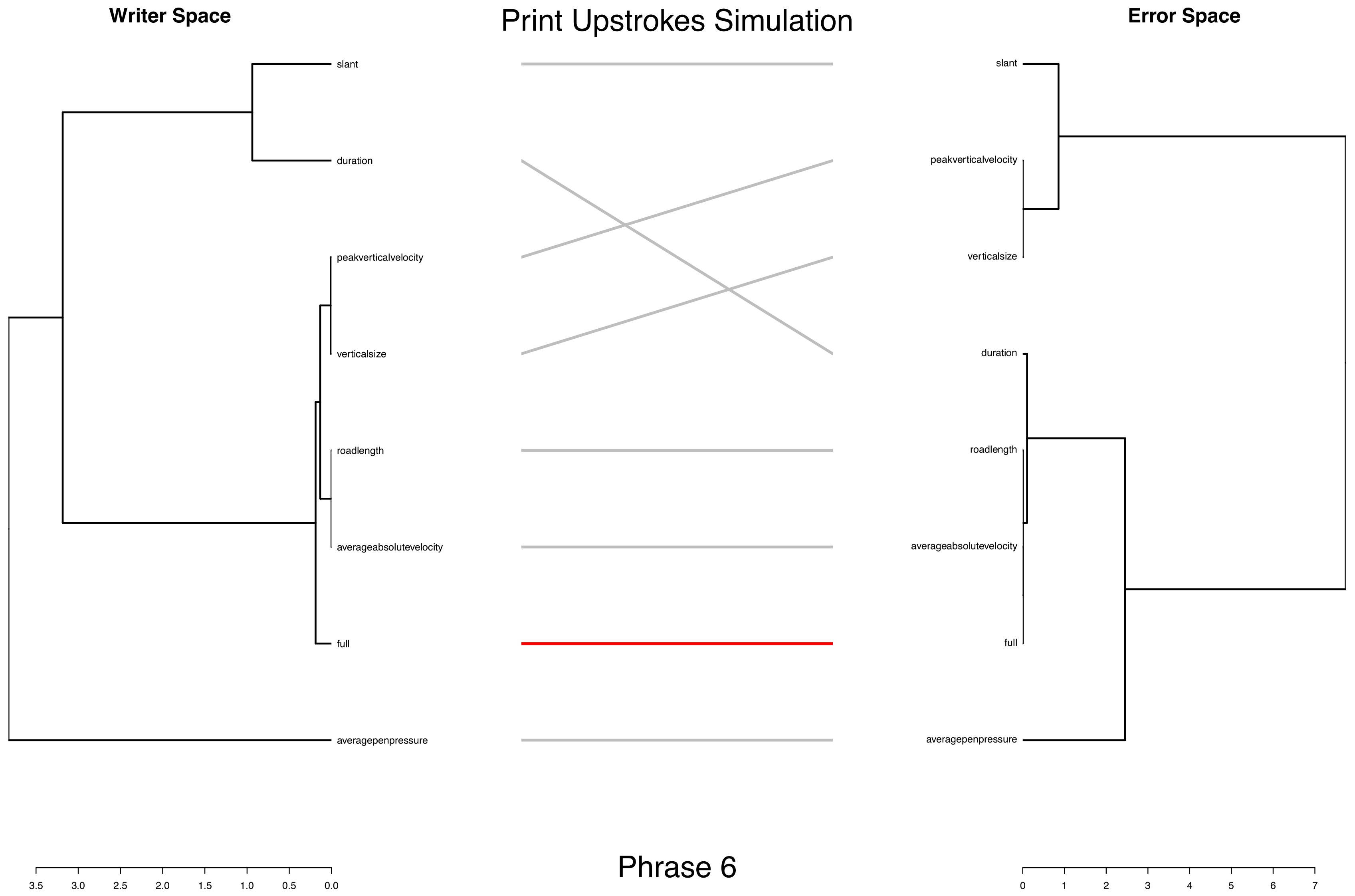

Writer Space
Print Downstrokes Simulation
Error Space
slant
duration
full
averagepenpressure
roadlength
averageabsolutevelocity
verticalsize
peakverticalvelocity
slant
duration
verticalsize
peakverticalvelocity
full
averagepenpressure
roadlength
averageabsolutevelocity
5
4
3
2
1
0
Phrase 4
0
1
2
3
4
5
6

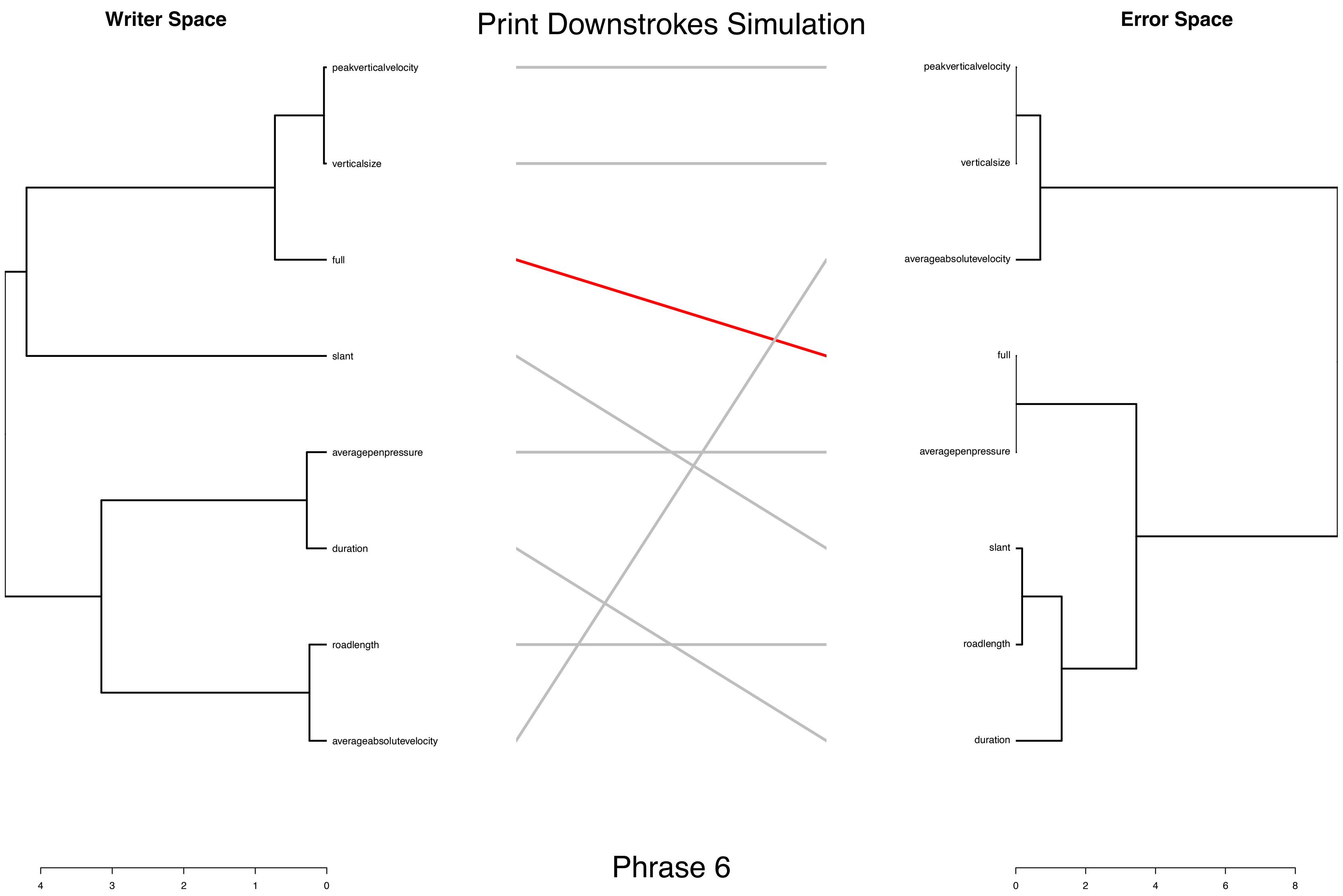


Print Upstrokes Simulation

Phrase 5


Writer Space
Print Downstrokes Simulation
Error Space
slant
peakverticalvelocity
averageabsolutevelocity
verticalsize
roadlength
full
averagepenpressure
duration
slant
peakverticalvelocity
verticalsize
averageabsolutevelocity
roadlength
full
averagepenpressure
duration
6
5
4
3
2
1
0
Phrase 5
0
2
4
6
8


Writer Space
Print Upstrokes Simulation
Error Space
slant
duration
peakverticalvelocity
verticalsize
roadlength
averageabsolutevelocity
full
averagepenpressure
slant
peakverticalvelocity
verticalsize
duration
roadlength
averageabsolutevelocity
full
averagepenpressure
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Phrase 6
0
1
2
3
4
5
6
7


Writer Space
Print Downstrokes Simulation
Error Space
peakverticalvelocity
verticalsize
full
slant
averagepenpressure
duration
roadlength
averageabsolutevelocity
peakverticalvelocity
verticalsize
averageabsolutevelocity
full
averagepenpressure
slant
roadlength
duration
4
3
2
1
0
Phrase 6
0
2
4
6
8

Bibliography

[1] Cuneyt Gurcan Akcora, Yulia R. Gel, and Murat Kantarcioglu. Blockchain: A graph primer. arXiv:2212.1708.08749v2 [cs.CY], 2022.

[2] Douglas Armstrong. *Development and properties of kernel-based methods for the interpretation and presentation of forensic evidence*. Electronic theses and dissertations, South Dakota State University, 2017.

[3] E. Del Barrio, E. Gine, and C. Matran. Central limit theorems for the wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2):1009:1071, 1999.

[4] Ronald Christensen. *Plane answers to complex questions: The Theory of Linear Models*, page 43. Springer Texts in Statistics. Springer - Verlag, 4 edition, 2011.

[5] H. W. Eiserman and M. R. Hecker. Fish-computers in handwriting examinations. 44th Annual Meeting of the American Society of Questioned Document Examiners, 1986.

[6] Katrin Franke, Lambert Schomaker, Christian Veenhuis, Louis Vuurpijl, Merijn van Erp, and Isabelle Guyon. A common ground for forensic handwriting examination and writer identification. *ENFHEX News- Bulletin of the European Network of Forensic Handwriting Experts*, 1(4), 2001.

[7] Katrin Franke, Lambert Schomaker, Louis Vuurpijl, and Stefan Giesler. Fishnew: A common ground for computer-based forensic writer identification. *Forensic Science International*, 136(S1-S432):84, December 2003.

[8] Cami Fuglsby, Christopher Saunders, Danica M. Ommen, JoAnn Buscaglia, and Michael Caligiuri. Elucidating the relationships between two automated

handwriting feature quantification systems for multiple pairwise comparisons. *Journal of Forensic Sciences*, 67:642:650, March 2022.

[9] Cami Fuglsby, Christopher Saunders, Danica M. Ommen, and Michael Caligiuri. Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage evaluative process. *Journal of Forensic Sciences*, 65(6):2080–2086, November 2020.

[10] Donald T Gantz and Christopher Saunders. Quantifying the effects of database size and sample quality on measures of individualization validity and accuracy in forensics. Grant Report 2009 - DN - BX - K234, U.S. Department of Justice, March 2014.

[11] Christoph Molnar. Interpretable machine learning: A guide for making black box models explainable (2nd ed.), 2022.

[12] Debora Nolan and David Pollard. Functional limit theorems for u-processes. *The Annals of Probability*, 16(3):1291:1298, 1988.

[13] Austin O'Brien. *A Kernel Based Approach to Determine Atypicality*. Electronic theses and dissertations, South Dakota State University, 2017.

[14] Danica M. Ommen, Cami Fuglsby, and Michael P. Caligiuri. Advances toward validating examiner writership opinion based on handwriting kinematics. *Forensic Science International*, 318, December 2021.

[15] Arun A. Ross, Karthik Nandakumar, and Anil K. Jain. *Handbook of Multibiometrics*. International Series on Biometrics. Springer Science + Business Media, LLC, 233 Spring Street, New York, New York 10013, 2006.

[16] LLC Sciometrics and Global Strategies Group. *FLASH ID in-depth*. Sciometrics, LLC and Global Strategies Group., Chantilly, VA, 2017.

[17] Sargur Srihari, Yong-Chul Shin, Sangjik Lee, Venugoal Govindaraju, Sung-Hyuk Cha, Catalin I. Tomai, Bin Zhang, Ajay Shekhawat, Dave Bartnik, Wen-Jann Yang, Srirangaraj Setlur, Phil Kilinskas, Fred Kunderman, Xia Liu, Zhixin Shi, and Vemulapati Ramanaprasad. Method and apparatus for analyzing and/or comparing handwritten and/or biometric samples. U.S. Patent no. 7,580,551, August 2009.

CURRICULUM VITAE

Education

Fall 2017 - *present*	Ph.D. Candidate Computational Science and Statistics South Dakota State University *Dissertation title:* Functionals of U-Processes for the Characterization of Operating Characteristics for Forensic Handwriting Comparison Systems Advisor and Chair: Christopher Saunders *Expected Graduation Summer 2023*
Fall 2015 - Fall 2017	M.S. Mathematics with emphasis in Statistics South Dakota State University *Thesis*: U-Statistics for Characterizing Forensic Sufficiency Studies Advisor and Chair: Christopher Saunders
Fall 2011 - Spring 2015	B.S. Mathematics and Statistics South Dakota State University

Professional Experience

starting August 2023	*Assistant Professor of Computer Science* Augustana University
May 2023 - August 2023	*Graduate Research Assistant* South Dakota State University
May 2022 - May 2023	*Instructor of Record* South Dakota State University
May 2021 - May 2022	*NSF NRT Fellow* National Science Foundation's National Research Trainee Program
August 2020 - May 2021	*Graduate Teaching Assistant* South Dakota State University
May 2018 - August 2020	*Graduate Research Assistant* South Dakota State University
August 2017 - May 2018	*Graduate Teaching Assistant* South Dakota State University
May 2017 - August 2017	*Graduate Research Assistant* South Dakota State University
August 2015 - May 2017	*Graduate Teaching Assistant* South Dakota State University

Courses Taught

May 2022 - August 2022

Undergraduate	Introduction to Probability & Statistics
Undergraduate	Basic R Programming
Graduate	Basic R Programming

August 2022 - December 2023

Undergraduate	Time Series Analysis
Graduate	Time Series Analysis

January 2023 - May 2023

Undergraduate	SAS Programming
Graduate	SAS Programming

Awards and Honors

Stephen E. Fienberg CSAFE Young Investigator Travel Award [$1500] 2017

ISBA 2018 World Meeting Travel Awards [$500] 2018

NSF support to attend Preparing to Teach Workshop at JSM [$800] 2018

American Statistical Association Poster Prize 10^{th} International Workshop on Simulation and Statistics. [$440] 2019

Forensic Sciences Foundation, Inc. Student Affiliate Scholarship. American Academy of Forensic Science 75th Annual Scientific Conference, complimentary registration. [$155] 2023.

Funded Research and Support

National Institute of Justice Grant 2017-DN-BX-0148 Kinematic Validation of FDE Determinations about Writership in Questioned Handprinting and Handwriting. Research and Development in Forensic Science for Criminal Justice Purposes. *Graduate Research Assistant, 2018-2019.*

National Institute of Justice Grants 2018-DU-BX-0192 and 2018-DU-BX-0193 Post-Blast Explosives Attribution and Persistence of Touch DNA for Forensic Analysis. Research and Development in Forensic Science for Criminal Justice Purposes. *Graduate Research Assistant, 2020.*

National Science Foundation. The Center for Security Printing and Anti-Counterfeiting Technology. *National Research Trainee Fellow, May 2021-May 2022.*

National Science Foundation Grant No. 2139349 Graph Theoretical Methods for Blockchain Data Analysis. Division of Mathematical Sciences. Algorithms for Threat Detection. $7,562. *Co-Principal Investigator.*

Peer-Reviewed Publications

Fuglsby, C., Saunders, C., Ommen, D.M. and Caligiuri, M.P. (2020), Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage evaluative process. J Forensic Sci, 65: 2080-2086. DOI 10.1111/1556-4029.14547 *funding was provided by NIJ Grant* 2017-DN-BX-0148

Fuglsby, C., Saunders, C.P., Buscaglia, J. (2020), U-statistics for estimating performance metrics in forensic handwriting analysis. J Statistical Computation and Simulation, 90:6, 1082-1117. DOI 10.1080/00949655.2020.1715406 *funding was provided by NIJ Grants* 2009-DN-BX-K234 *and* 2014-IJ-CX-K088

Ommen, D.M., **Fuglsby, C.**, Caligiuri, M.P. (2021), Advances toward validating examiner writership opinion based on handwriting kinematics. Forensic Sci Int, 318:110644.
DOI 10.1016/j.forsciint.2020.110644 *funding was provided by NIJ Grant* 2017-DN-BX-0148

Fuglsby, C. Saunders, C. Ommen, D.M. Buscaglia, J. Caligiuri, M.P. (2022), Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons. J Forensic Sci, 67: 642-650. DOI 10.1111/1556-4029.14914 *funding was provided by NIJ Grant* 2017-DN-BX-0148

Ippoliti, P. Werlich, J. **Fuglsby, C.** Yarnes, C. Saunders, C. Dettman, J. (2023), Linking Ammonium Nitrate Aluminum (AN-AL) Post-Blast Residues to Pre-Blast Explosive Materials Using Isotope Ratio and Trace Elemental Analysis for Source Attribution. J Forensic Sci, 68: 407-415. DOI 10.1111/1556-4029.15190

Presentations

Danica Ommen (presenter), Larry Tang, **Cami Fuglsby**. To differentiate or not to differentiate: An unholy love-affair between the Two-Stage and Bayesian approaches. CSAFE sponsored Symposium on Error Rates for Evidence Interpretation, Arlington, VA. Invited Presentation. *Travel support provided by CSAFE*. January

2018.

Cami Fuglsby (presenter), Linton Mohammed (presenting), JoAnn Buscaglia, Christopher Saunders. Sufficiency and Complexity Factors in Handwriting Examination. Impression, Pattern, and Trace Evidence Symposium, Arlington, VA. Accepted Oral Presentation. *Travel support provided by NIJ and RTI.* January 2018.

Cami Fuglsby. An Introduction to the Forensic Identification of Source Problem and Related Problems in Statistical Model Selection. Computational Science and Statistics Seminar, Brookings, SD. Oral Presentation.

Cami Fuglsby (presenter), Christopher P. Saunders, JoAnn Buscaglia, Danica Ommen. A Modified Two-Stage Approach to the Interpretation of Forensic Evidence. Joint Statistical Meetings, Vancouver, BC. Oral Presentation. August 2018.

Linton Mohammed (presenter), **Cami Fuglsby** (presenter), Christopher P. Saunders, Danica Ommen, Michael Caligiuri, JoAnn Buscaglia. FDE Conclusion Scales: Rev. Bayes or Prof. Kirk? (Part I). The 76th Annual General Meeting of the American Society of Questioned Document Examiners, Park City, UT. Conference Attendance by Invitation Only. August 2018.

Cami Fuglsby (presenter), Linton Mohammed (presenter), Christopher P. Saunders, Danica Ommen, Michael Caligiuri, JoAnn Buscaglia. FDE Conclusion Scales: Rev. Bayes or Prof. Kirk? (Part II). The 76th Annual General Meeting of the American Society of Questioned Document Examiners, Park City, UT. Conference Attendance by Invitation Only. August 2018.

Danica Ommen (presenter), **Cami Fuglsby**, Christopher P. Saunders, Michael Caligiuri, Linton Mohammed, JoAnn Buscaglia. Pairwise Comparison Scores for Handwritten Questioned Documents. American Academy of Forensic Sciences 71^{st} Annual Scientific Meeting. Baltimore, MD. February 2019.

Danica Ommen (presenter), JenaMarie Baldaino, **Cami Fuglsby** , Christopher P. Saunders (presenting), Jack Hietpas, JoAnn Buscaglia. On the Development of Score Rules for the Pairwise Sample Comparison of Particle Micromorphometry of Aluminum (Al) Powders. American Academy of Forensic Sciences 71^{st} Annual Scientific Meeting. Baltimore, MD. February 2019.

Cami Fuglsby (presenter), Christopher P. Saunders, Danica Ommen, JenaMarie Baldaino, JoAnn Buscaglia, Jack Hietpas. New Developments in the Interpretation of Pairwise Comparison Procedures for a Class of Forensic Applications Related to Improvised Explosive Devices. University of Kentucky Department of Statistics Seminar. Lexington, KY. February 2019. *Invited presentation*.

Cami Fuglsby, Christopher P. Saunders (presenter), Danica Ommen, JoAnn Buscaglia. Bayesian Characterizations Of U-processes Used In Pattern Recognition With Application To Forensic Source Identification. 10^{th} International Workshop on Simulation and Statistics. Salzburg, Austria. September 2019. *Invited Talk*.

Cami Fuglsby, Danica Ommen, JoAnn Buscaglia, Christopher P. Saunders (presenter). Bayesian Characterizations of U-processes Used In Pattern Recognition with Application To Forensic Source Identification. University of Nebraska Medical Center. December 2019. *Invited Talk.*

Cami Fuglsby (presenter), Michael Caligiuri, Danica Ommen, Christopher P. Saunders, JoAnn Buscaglia. The Interaction of Writing Profiles and Automated Scoring Rules. American Academy of Forensic Sciences 72^{nd} Annual Scientific Meeting. Anaheim, CA. February 2020.

Cami Fuglsby (presenter), Christopher P. Saunders, Danica Ommen, JoAnn Buscaglia, Michael Caligiuri. Advancements in Black-box metamodeling of pairs of features from pairwise scores of handwritten items. Virginia Commonwealth University Forensic Science seminar. September 20^{th} 2022. *Invited Talk.*

Cami Fuglsby (presenter), Christopher P. Saunders, Danica Ommen, JoAnn Buscaglia. White-box metamodeling for feature pairs for handwriting data. Joint Statistical Meetings, Toronto, Ontario, CA. Topic-Contributed Oral Presentation. August 2023.

Posters

Cami Fuglsby (presenter), JoAnn Buscaglia, Christopher Saunders. Incomplete U-Processes for Forensic Sufficiency Studies in Questioned Document Examination. Joint Statistical Meetings, Baltimore, ML. Contributed Poster Presenter. August 2017.

Cami Fuglsby (presenter), Christopher P. Saunders, JoAnn Buscaglia. Incomplete U-Processes for Forensic Sufficiency Studies in Questioned Document Examination. International Conference on Forensic Inference and Statistics, Minneapolis, MN. Contributed Poster Presenter. *Travel Support provided by Stephen E. Fienberg CSAFE Young Investigator Award.* September 2017.

JenaMarie Baldaino (presenter), **Cami Fuglsby**, Danica M. Ommen, Christopher P. Saunders, Jack Hietpas, and JoAnn Buscaglia. Characterization of Aluminum Powders in Explosives Utilizing Particle Micromorphometry. International Conference on Forensic Inference and Statistics, Minneapolis, MN. September 2017.

JenaMarie Baldaino (presenter), Danica M. Ommen, **Cami Fuglsby**, Christopher P. Saunders, Jack Hietpas, and JoAnn Buscaglia. Statistical Characterization of Commercial and Home-made Aluminum Powders in Explosives Using Automated Particle Micromorphometry. Impression, Pattern, and Trace Evidence Symposium, Arlington, VA. January 2018.

JenaMarie Baldaino (presenter), Danica M. Ommen, **Cami Fuglsby**, Christopher P. Saunders, Jack Hietpas, and JoAnn Buscaglia. Statistical Characterization of Aluminum (Al) Powders in Explosives Using Automated Particle Micromorphometry. American Academy of Forensic Sciences 70^{th} Annual Scientific Meeting, Seattle, WA. February 2018.

Cami Fuglsby (presenter), Christopher P. Saunders, Danica M. Ommen, JoAnn Buscaglia. On the Use of Bayesian P-Values for Forensic Identification of Source Problems. 2018 International Society for Bayesian Analysis World Meeting, Edinburgh, UK. Travel Support provided by ISBA 2018 World Meeting Travel Awards. June 2018.

JenaMarie Baldaino (presenter), Danica M. Ommen,**Cami Fuglsby**, Christopher P. Saunders, Jack Hietpas, and JoAnn Buscaglia (presenter). Characterization of commercial and home-made aluminum powders via micromorphometric analysis. 8^{th}

European Academy of Forensic Science Conference, Lyon, France. August 2018.

Danica Ommen (presenter), **Cami Fuglsby**, Christopher Saunders, Michael Caligiuri, Linton Mohammed, JoAnn Buscaglia. Pairwise Scores for Designing Handwritten Document Comparisons. Forensics @NIST, Gaithersburg, MD. November 2018.

Danica Ommen (presenter), Larry Tang, **Cami Fuglsby**, Christopher P. Saunders, Susan Vanderplas. Statistical Infrastructure for the Use of Error Rate Studies in the Interpretation of Forensic Evidence. CSAFE All-Hands Meeting. Ames, IA. May 2019

Danica Ommen (presenter), Michael Caligiuri, **Cami Fuglsby**, Christopher P. Saunders, Linton Mohammed. Kinematic Validation of FDE Determinations about Writership for Questioned Handprinting and Handwriting. CSAFE All-Hands Meeting. Ames, IA. May 2019.

Cami Fuglsby (presenter), Christopher P. Saunders, Danica Ommen, Michael Caligiuri. A Class of Score Functions for the Analysis of Kinematic Handwriting Data.(Won the 1^{st} American Statistical Association Poster Prize) 10^{th} International Workshop on Simulation and Statistics. Salzburg, Austria. September 2019.

Kayla Moquin (presenter), **Cami Fuglsby** (presenter), JenaMarie Baldaino, Danica Ommen, Christopher P. Saunders, Jack Hietpas, JoAnn Buscaglia. Further Development of Scoring Rules for Sample Comparisons Using Automated Particle Micromorphometry of Aluminum (Al) Powders. American Academy of Forensic Sciences 72^{nd} Annual Scientific Meeting. Anaheim CA. February 2020.

Cami Fuglsby (presenter), Christopher Saunders, Danica M Ommen, JoAnn Buscaglia, Michael Caligiuri. Increasing the Transparency of Black-Box Systems. 2021 Joint Statistical Meetings, virtual conference. Contributed Poster Presenter. August 2021.

Cami Fuglsby (presenter), Kayla M. Moquin, Christopher P. Saunders, Danica Ommen, Michael Caliguiri, JoAnn Buscaglia. Assessing the Dependency Structure Between Shape Codes for Forensic Handwriting Data. American Academy of Forensic Sciences 74^{th} Annual Scientific Meeting. Seattle, WA. February 2022.

Kayla Moquin (presenter), **Cami Fuglsby**, JenaMarie Baldaino, Danica Ommen (presenter), Christopher Saunders, Jack Hietpas, JoAnn Buscaglia. Automated Particle Micromorphometry and Statistical Scoring for Improved Characterization of Aluminum (Al) Powders in Improvised Explosive Devices (IEDs). American Academy of Forensic Sciences 74^{th} Annual Scientific Meeting. Seattle, WA. February 2022.

Cami Fuglsby (presenter), Christopher P. Saunders, Danica M. Ommen, JoAnn Buscaglia. Black-Box Metamodeling Between Feature Combinations for Paired Comparisons of Documents of Forensic Handwriting Data. American Academy of Forensic Sciences 75^{th} Anniversary Conference. Orlando, FL. February 2023.

Student Mentoring

Callie Duque née Sleep, 2017 on exact upper confidence bounds for random match probabilities, mentored for undergraduate senior capstone experience.

Sierra Lutz, 2018-2019 on false discovery rates and large-scale inference, mentored for undergraduate senior capstone experience and directed her research support for NIJ Grant 2017-DN-BX-0148.

Janean Hanka, 2021-2023 on atypicality-based approaches to classification and machine learning, mentored for undergraduate senior capstone experience and directed her work as part of an NSF-funded REU project, paired with my NRT fellowship.

Cole Patten, August 2022 - May 2023 GRA on NSF funded grant No. 2139349. Supervised to review topological data analysis and manifold learning.

Leadership and Service

National Service

2022-*present* Served as a member on the Forensic Document Examination Consensus Body of the American Academy of Forensic Sciences Standards Board.

2021-2022 Served as a member on the Organization of Scientific Area Committees for Forensic Science Scientific & Technical Review Panels,

OSAC 2022-S-0001, Standard Guide for Image Comparison Conclusions/Opinions

OSAC 2021-S-0036, Standard Guide for Image Authentication

2020-2022 Served as a member on the Forensic Document Examination subcommittee for the Organization of Scientific Area Committees for Forensic Science.

2018, 2019 Served as the Executive Secretary for the Forensic Document Examination subcommittee for the Organization of Scientific Area Committees for Forensic Science.

2017 Participated in the spring meeting for the Forensic Document Examination subcommittee for the Organization of Scientific Area Committees for Forensic Science as a note taker.

2017-2018 Supported the Expert Working Group on Human Factors in Handwriting Examinations Interpretation and Technology subgroup for the National Institute of Standards and Technology.

Refereeing

2017, 2018, 2022 Journal of Forensic Sciences

2021, 2022 Journal of Statistical Computation and Simulation

Departmental Organization

2014-2015 Served as the president of the SDSU Math Club

Volunteering

2014, 2016, 2018 Northeastern SD Chapter of Math Counts

2016-2020 Ready-SET-Go! Camp at South Dakota State University

2019 GEMS Workshop at South Dakota State University

Professional Service

2017 Session Chair, International Conference on Forensic Inference and Statistics.

2018 Session Chair, Joint Statistical Meetings.

2017 International Conference on Forensic Inference and Statistics as a conference volunteer, chaired two sessions.

2018 Invited participant on the CSAFE sponsored Symposium on Error Rates for Evidence Interpretation.

2018 Chair of a SPEED session, Joint Statistical Meetings.

2019 International Workshop on Simulation and Statistics (Organize Student Poster Session.

Memberships

American Academy of Forensic Science - Questioned Documents

American Statistical Association

Institute of Mathematical Statistics

International Society for Bayesian Analysis