FEATURE SELECTION WITH APPLICATIONS TO TEXT CLASSIFICATION

BY DAVID JOSEPH NEU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Operations Research

Written under the direction of

Endre Boros

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE DISSERTATION

FEATURE SELECTION WITH APPLICATIONS TO TEXT CLASSIFICATION

by DAVID JOSEPH NEU

Dissertation Director: Endre Boros

Application of a feature selection algorithm to a textual data set can improve the performance of some classifiers. Due to the characteristics, specifically the size, of textual data sets researchers have traditionally relied on a family of simple heuristics to perform feature selection. These heuristics, which in practice are quite effective, use functions of individual feature statistics, that we refer to as feature ranking functions, to order the feature set.

We are interested in identifying the most effective feature ranking functions. To do this we begin by defining a feature set evaluation methodology. Traditionally the performance of feature selection algorithms has been measured by comparing the performance of classification algorithms before and after feature selection. Instead, we measure various criteria of the selected feature set itself, including measures of separation, noise, size, and robustness. We demonstrate that many of these criteria are competing, and show how the tools of multicriteria optimization can be employed to rank the performance of feature selection algorithms.

Using this methodology we evaluate the performance of a large set of feature ranking functions, including a function that measures the rareness of a feature assuming that relevant and irrelevant documents are generated by two independent stochastic

processes. Motivated by the results, we identify the defining characteristics of the functions that are most successful, noting that many of these can be written as ratios of measures of separation to measures of noise.

Next we introduce a set of axioms which we believe that feature ranking functions should satisfy, and study the set of these functions that can be represented as a linear combination of some finite set of basis functions. We demonstrate that many of the functions or approximations to the functions that we studied are members of this set. Next consider the set of coefficient vectors of this set and show that it is convex, bounded, and not empty. We conclude by investigating the performance of other approaches to feature selection including greedy and ensemble algorithms that use feature ranking functions.

Acknowledgements

I would like to begin by expressing my sincere appreciation to my advisor, Dr. Endre Boros. I feel fortunate to have studied with such gifted mathematician and teacher who was generously willing to share his ideas as well as to patiently work through mine. Throughout this long journey, he maintained an unwaveringly commitment to *my* goal, and for this I thank him.

I would also like to thank Dr. Wanpracha Art Chaovalitwongse, Dr. Vladimir Gurvich, Dr. Myong K. Jeong, and Dr. Paul B. Kantor for contributing to this dissertation by being on my committee. I would specifically like to acknowledge Dr. Paul B. Kantor for introducing me to the field of information retrieval and supporting me on the AntWorld project.

I would like to mention some of the other professors, staff, and students who have made the RUTCOR community so special. First, I would like to mention Dr. Peter Hammer who worked so hard to build RUTCOR. His Boolean functions class was among my favorites, and thoughts of his good humor still bring me a smile. I tremendously enjoyed Dr. András Prékopa's classes. I appreciated the obvious joy he brings to teaching and the concern he has for each of his students. I would like to thank Dr. Stefan Schmieta and Dr. Pierangela Veneziana for their friendship and support. Finally, Dr. Lynn Agre, Mrs. Teresa Hart, and Mrs. Clare Smietana, were always ready to provide words of encouragement and to offer expert guidance on navigating through the university. I hope they all know how much I truly appreciate their help.

I would also like to remember my friend Dr. Divyendu Sinha who passed away. He was a talented researcher, excellent teacher, and a good friend, whom I miss.

I would like to thank my entire family for all of their support. I truly appreciated my brother Ted traveling to spend time with my children, and the constant encouragement

offered by my sister Carolyn. As for my parents, only after becoming a parent myself have I begun to realize how much I owe my them – thank you.

To my children, Amalia, Matthew, and Ryan, I thank for all of your love and support as I worked to achieve this goal and look forward to helping you meet all of yours. Finally, I would like to thank my wife Anne for her love and encouragement. I know that at times my challenges became hers, and I am grateful to her for helping me shoulder them.

Dedication

To my parents without whose support this and so many other things would not have been impossible.

Table of Contents

Abstract ii

Acknowledgements iv

Dedication vi

List of Tables xi

List of Figures xiii

1. Introduction 1

1.1. Textual Feature Selection 1

1.2. Document, Feature, and Collection Models 10

1.3. ROC Curves and the Wilcoxon Rank Sum Test 16

1.4. Multicriteria Optimization 29

1.5. Overview 40

2. Feature Set Evaluation 41

2.1. The Collection 41

2.2. Data Set Reduction 42

2.3. Preparation 43

2.4. Fold Construction for Cross Validation 44

2.5. Topics 45

2.6. Methodology 48

2.7. Separation Measures 49

2.8. Noise Measures 52

2.9. Semantic Measures 54

2.10. Robustness Measures 56

2.11. The Set of Evaluation Criteria 57
2.12. Feature Set Size 59
2.13. Cross Topic / Cross Validation 61
2.14. Result Comparison 62
2.15. Computational Considerations 63

3. Boolean Feature Ranking Functions 66
3.1. Ranking Functions and Algorithms 66
3.2. Probabilistic Model 67
3.3. Term Frequency Model 71
3.4. Separation Model 72
3.5. Information Retrieval Model 74
3.6. Single Feature Classifier Model 75
3.7. Positive and Negative Features 81
3.8. Ranking Functions 86
3.9. Properties of and Relationships Between Ranking Functions 97
3.10. Boolean Feature Ranking Results 103

4. Separation and Noise 106
4.1. Noise Separates 106
4.2. The Impact of Noise 108
4.3. Non-Stopword Noise 111
4.4. Monotone Feature Principle 118
4.5. Characterizing Separation and Noise 127
4.5.1. Class Separation and Noise 127
4.5.2. Collection Noise 131
4.5.3. Strong Collection Noise 135
4.5.4. Collection Noise and Strong Collection Noise 137
4.6. "Noisy" Functions 138
4.7. Separation to Noise Ratios 141

5. The Space of Boolean Feature Ranking Functions 146

5.1. Axioms 146

5.2. Linearization 152

5.3. Power Series Representation 154

5.4. Equality Based Characterizations of $M^*\left[\mathcal{P}_{\eta=1}^{(k)}\right]$ 164

5.5. Experiments 167

5.5.1. $M^*\left[\mathcal{P}_1^{(k)}, Q_q\right]$ 172

5.5.2. $\hat{M}^*\left[\mathcal{P}_1^{(k)}, Q_q\right]$ 181

5.5.3. $\tilde{M}^*\left[\mathcal{P}_1^{(k)}, Q_q\right]$ 190

6. Extensions 196

6.1. Real-Valued Feature Ranking Functions 196

6.1.1. Ranking Functions and Algorithms 196

6.1.2. Real-Valued Feature Ranking Function Properties 199

6.1.3. Real-valued Feature Ranking Results 202

6.2. Boolean Greedy Algorithms 203

6.2.1. μ-GREEDY Results 209

6.3. Lexicographic Greedy Algorithms 215

6.3.1. h_k-GREEDY Results 218

6.4. Ensemble Methods 222

6.4.1. μ-ENSEMBLE Results 227

7. Conclusions 232

References 236

Appendix A. Correlation Coefficient 242

Appendix B. Gini Split Criterion 245

Appendix C. Information Gain 248

Appendix D. Rareness 251

Appendix E. Fisher's Linear Discriminant 254

Appendix F. Proof that Function Sets are Closed 256

Appendix G. SMART Stopword List 258

Appendix H. μ-RANKING Results Stopwords Included Not Discounted 262

Appendix I. μ-RANKING Results Stopwords Included Discounted . . 271

Appendix J. μ-RANKING Results Stopwords Excluded Not Discounted 280

Appendix K. μ-RANKING Results Noise Growth Rate 289

Appendix L. μ-RANKING Results Robustness 292

Appendix M. μ-GREEDY Results Stopwords Included Not Discounted 293

Appendix N. μ-GREEDY Results Stopwords Included Discounted . . . 302

Appendix O. μ-GREEDY Results Noise Growth Rate 311

Appendix P. π-RANKING Results Stopwords Included Not Discounted 314

Appendix Q. π-RANKING Results Stopwords Included Discounted . . 322

Appendix R. h_k-GREEDY Results Stopwords Included Not Discounted 330

Appendix S. h_k-GREEDY Results Stopwords Included Discounted . . 338

Appendix T. Ensemble Results Stopwords Included Not Discounted . 346

Appendix U. Ensemble Results Stopwords Included Discounted 353

Vita 360

List of Tables

1.1. Relationships Between Wilcoxon Rank Sum Statistics and ROCs 25

2.1. Topics 47

2.2. Set $\mathcal{C}$ of Feature Set Evaluation Criteria 58

2.3. Monotonically Non-Decreasing Measures 58

2.4. K Required for Support Set 60

3.1. Separation Model Example 73

3.2. μ-RANKING Not Discounted Stopwords Included 105

4.1. μ-RANKING Discounted Stopwords Included 107

4.2. $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic 109

4.3. μ-RANKING Not Discounted Stopwords Excluded 119

4.4. $\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic 120

4.5. $\mho_K = \omega(V, K, x\text{-RANKING}, \uparrow) \cup \omega(V, K, y\text{-RANKING}, \uparrow)$ 122

4.6. Stopwords vs Non-stopwords for $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ 139

5.1. Axioms Satisfied by the Sample Taylor Polynomials 170

5.2. Taylor Polynomial Coefficients 171

5.3. $M^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$ 173

5.4. $\mathcal{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$ 179

5.5. $\mathcal{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$ 180

5.6. $\hat{M}^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$ 182

5.7. Selected elements of $\hat{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ and functions from $\hat{\mathcal{M}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ 187

5.8. $\widehat{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$ 188

5.9. $\widehat{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$ 189

5.10. $\tilde{M}^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$ 191

5.11. Selected elements of $\tilde{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ and functions in $\tilde{\mathcal{M}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ 193

5.12. $\tilde{\mathcal{R}}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$ 195
5.13. $\tilde{\mathcal{R}}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$ 195
6.1. π-RANKING Not Discounted Stopwords Included 203
6.2. π-RANKING Discounted Stopwords Included 207
6.3. μ-GREEDY Not Discounted Stopwords Included 210
6.4. μ-GREEDY Discounted Stopwords Included 211
6.5. h_k-GREEDY Not Discounted Stopwords Included 220
6.6. h_k-GREEDY Discounted Stopwords Included 221
6.7. Lex-Max-Min Solution 226
6.8. μ-ENSEMBLE Not Discounted Stopwords Included 228
6.9. μ-ENSEMBLE Discounted Stopwords Included 228
H.1. μ-RANKING Stopwords Included Not Discounted 264
I.1. μ-RANKING Stopwords Included Discounted 273
J.1. μ-RANKING Stopwords Excluded Not Discounted 282
K.1. μ-RANKING Noise Growth Rate 291
L.1. μ-RANKING Stopwords Included Robustness 292
M.1. μ-GREEDY Stopwords Included Not Discounted 295
N.1. μ-GREEDY Stopwords Included Discounted 304
O.1. μ-GREEDY Noise Growth Rate 313
P.1. π-RANKING Stopwords Included Not Discounted 315
Q.1. π-RANKING Stopwords Included Discounted 323
R.1. h_k-GREEDY Stopwords Included Not Discounted 331
S.1. h_k-GREEDY Stopwords Included Discounted 339
T.1. μ-ENSEMBLE Stopwords Included Not Discounted 346
U.1. μ-ENSEMBLE Stopwords Included Discounted 353

List of Figures

1.1. ROC Examples 30

1.2. Examples of the random classifier $\bar{g}_{\tau}^{+}$ 31

1.3. Multicriteria Examples 35

3.1. A Boolean ROC Curve for g^{+} 78

3.2. A Boolean ROC Curve for g^{+} 78

3.3. A Boolean ROC Curve for g^{-} 80

3.4. A Boolean ROC Curve for g^{-} 80

3.5. Boolean ROC Curve Positive Feature 83

3.6. Boolean ROC Curve Negative Feature 83

3.7. Boolean ROC Curve and μ_4 101

4.1. $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic 110

4.2. μ-RANKING Stopwords Included Not Discounted: σ_K vs ν_K 112

4.3. μ-RANKING Stopwords Included Not Discounted: θ_K vs ν_K 113

4.4. μ-RANKING Stopwords Included Not Discounted: ξ vs ν_K 114

4.5. μ-RANKING Stopwords Included Not Discounted: $\Delta(\sigma_K)$ vs ν_K 115

4.6. μ-RANKING Stopwords Included Not Discounted: $\Delta(\theta_K)$ vs ν_K 116

4.7. μ-RANKING Stopwords Included Not Discounted: φ_K vs ν_K 117

4.8. $\omega(V, K, x\text{-RANKING}, \uparrow) \cup \omega(V, K, y\text{-RANKING}, \uparrow)$ 126

4.9. Measures of Separation and Class Noise 129

4.10. η_1 with $\varrho = 0.25$ 132

4.11. Distance to Collection Noise Line 134

4.12. Strong Collection Noise 136

4.13. Collection Noise versus Strong Collection Noise 138

4.14. μ_8 140

4.15. Examples of functions in $\mathcal{M}^{\Psi/\aleph}$ 143
4.16. Examples of functions in $\mathcal{M}^{\Psi^{\pm}/\aleph}$ 145
5.1. $\mathcal{V}^*\big[\mathcal{P}_1^{(2)}, Q_*\big]$ 174
5.2. Center of $M^*\big[\mathcal{P}_1^{(3)}, Q_{50}\big]$ 177
5.3. $\mathcal{V}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ for $\lambda_1 \approx \lambda_2 \approx 0$ 178
5.4. $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$ in Reduced Dimension Space 184
5.5. Three selected functions from $\hat{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ 185
5.6. Three additional functions from $\hat{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ 186
5.7. $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$ in Reduced Dimension Space 194
6.1. π-RANKING Stopwords Included Not Discounted: σ_K vs ν_K 204
6.2. π-RANKING Stopwords Included Not Discounted: θ_K vs ν_K 205
6.3. π-RANKING Stopwords Included Not Discounted: φ_K vs ν_K 206
6.4. μ-GREEDY Stopwords Included Not Discounted: σ_K vs ν_K 212
6.5. μ-GREEDY Stopwords Included Not Discounted: θ_K vs ν_K 213
6.6. μ-GREEDY Stopwords Included Not Discounted: φ_K vs ν_K 214
6.7. h_k-GREEDY Stopwords Included Not Discounted: σ_K vs ν_K 223
6.8. h_k-GREEDY Stopwords Included Not Discounted: θ_K vs ν_K 224
6.9. h_k-GREEDY Stopwords Included Not Discounted: φ_K vs ν_K 225
6.10. μ-ENSEMBLE Stopwords Included Not Discounted: σ_K vs ν_K 229
6.11. μ-ENSEMBLE Stopwords Included Not Discounted: θ_K vs ν_K 230
6.12. μ-ENSEMBLE Stopwords Included Not Discounted: φ_K vs ν_K 231
K.1. μ-RANKING Noise Growth Rate Examples 289
O.1. μ-GREEDY Noise Growth Rate Examples 311

Chapter 1

Introduction

We begin this chapter with an overview of textual feature selection. Then, the models that are used to represent documents and features are introduced. Next, selected material related to Receiver Operating Characteristics (ROC) Curves, the Wilcoxon rank sum statistical test and multicriteria optimization, that will be used in the sequel is reviewed. We then provide a survey of other work related to textual feature selection, and conclude the chapter with a overview of the material contained in the remainder of this dissertation.

1.1 Textual Feature Selection

Consider a set of n-dimensional vectors, known as the *collection*, with each vector representing a *document* which is labeled as either *relevant* or *irrelevant* for a given *topic*, and the components of the vectors, known as *features*, corresponding to the terms or some function of the terms in the documents. *Feature selection* is the identification of the subset or subsets of features that best satisfy some specified *criteria*. Feature selection has been studied extensively in the machine learning, pattern recognition, data mining, information retrieval, and text categorization literature. A survey of the field is provided in [43].

While much of the material in this dissertation is independent of the data set under consideration, in this section we will discuss the unique characteristics of textual data, the challenges presented by these characteristics and the feature selection algorithms that they suggest. In addition, we shall introduce several concepts that will be central to the discussion in the sequel.

High Dimensional. Textual data is inherently *high dimensional* with the number of

features being potentially equal to the number of words in the English language. This characteristic, combined with the fact that the number of feature subsets is equal to 2^n, result in textual feature selection being a very large scale problem. Some of the difficulties of working with such data sets include

- the curse of dimensionality – the volume of a space increases exponentially with the number of dimensions and some methods need exponentially more data to obtain comparable results,
- spurious correlations – since the data is not completely random, as the number of dimensions increases, so do the number of correlations among variables with the issue being how to determine which of these correlations are meaningful,
- the collapse of distance metrics – points tend to become equidistant as the number of dimensions increases, thereby impacting the effectiveness of distance based algorithms,

(see e.g. [20, 25, 28, 78]). Ironically, feature selection is one of the techniques used to mitigate these issues.

Noise. Misspellings, the improper, or colloquial use of words, words such as stopwords (e.g. "a", "and", "the") that have low semantic content regardless of the topic, are all examples of *noise* terms. These terms are virtually useless for distinguishing relevant documents from irrelevant ones, but paradoxically, some noise terms such as stopwords occur very frequently. As a result, documents typically have many of these terms in common; a fact which serves to exacerbate the aforementioned tendency of documents to be equidistant because of the high dimensional nature of the data. Due to the prevalence of noise, it would perhaps be more accurate to refer to feature selection in textual data as feature elimination.

Many different descriptions of textual noise features have been offered. In [60] it is stated that a "*noise feature* is one that, when added to the document representation, increases the classification error on new data." In [12], it is suggested that what constitutes noise includes a "perfectly random/irrelevant attribute", an "(almost) constant

attribute", or "dependent attributes". *We shall consider a feature to be noise if neither its presence in a document nor its absence from a document depends on the document's class,* or equivalently, *we shall consider a feature* noise *if its neither its presence in a document nor its absence from a document provides evidence that the document is relevant or irrelevant.*

Distribution. The distribution of textual data tends to follows a power law known as Zipf's law or arguably more accurately, an extension of it known as Mandelbrot's law. Given a list of terms ordered by their frequency of occurrence in a collection, these laws indicate that the frequency of occurrence of any given term is inversely proportional to its rank in the list. The most common terms are, per our discussion above, "noise" features; further, most features, even those not considered "noise" features, are not needed to distinguish relevant documents from irrelevant documents. As a result, textual feature selection can be seen to involve identifying a very small subset of features from a very large one. Quoting from, [61], p.29, "what makes frequency-based approaches to language hard is that almost all words are rare."

Positive and Negative Features. We now introduce the idea of positive and negative features. A feature is said to be *positive* if its presence in a document provides evidence that the document is relevant, and its absence from a document provides evidence that the document is irrelevant. A feature is said to be *negative* if its presence in a document provides evidence that the document is irrelevant, and its absence from a document provides evidence that the document is relevant. We shall revisit positive and negative features during our discussion of the next several topics.

Class Homogeneity. A set of documents is *class homogeneous* if all documents in the set are labeled as relevant for exactly one topic. The set of relevant documents is class homogeneous by definition. If both the sets of relevant and irrelevant documents are class homogeneous, then the set complement operation applied to documents is idempotent, that is, documents which are not relevant are irrelevant, and documents which are not irrelevant are relevant.

In textual data sets, the set of irrelevant documents is typically not class homogeneous. For example, if the topic under consideration is "fuel" and the set of irrelevant

documents is not class homogeneous, then the relevant documents are those that have been labeled to be about "fuel", but the irrelevant documents are not known to be about some other specific topic; the only thing we can say is that they are not about "fuel". This situation results in an asymmetry in that the presence of a positive term in a document provides affirmative evidence that the document is about "fuel", but the presence of a negative term does not provide evidence that the document is about a specific topic. Rather, it only provides evidence that the document is *not* relevant. Such a term can be thought of "blocking" the document from being relevant.

One of the issues with the set of irrelevant documents being class inhomogeneous is related to negative features. Features that appear in many irrelevant documents, and only a few relevant documents seem to meet the criteria for a negative feature. It should be noted, however, that when the set of irrelevant documents is inhomogeneous such features consistently appear in documents regardless of their topic, which means they exactly coincide with the noise features.

Class Skew. Adapting the definition given in [35], the *class skew* is the ratio of the number of relevant documents to the number of irrelevant documents in a collection. In textual data sets, typically there is significant class skew, with the number of irrelevant documents for any given topic greatly exceeding the number of relevant documents. Having a very small number of relevant documents can cause feature selection algorithms to encounter the same sort of difficulties as classification algorithms when presented with such data sets (see e.g. [40, 34, 35, 33]). Examples of these difficulties include overfitting and the inability to reach an adequate level of confidence in results. Further, in data sets with significant class skew it can be difficult for classification algorithms to achieve better performance than the simple strategy of assigning all documents to the majority class, that is, of classifying all documents as irrelevant (see e.g. [40], p.1290). In feature selection the analogous problem is that algorithms may have a tendency to select features that appear in many irrelevant documents, and only a few relevant documents. While such features seem to meet the criteria for a negative feature, since they appear in a very large number of documents, it should be mentioned that they also coincide with our characterization of noise features.

Relevant Centric. As discussed, the fact that in textual data sets, the set of irrelevant documents is not class homogeneous and that there is significant class skew suggests that some features that seem to meet the criteria for a negative feature, specifically, features that appears in many irrelevant documents, and only a few relevant documents, may also be noise features.

However, note that there can in fact be negative features that are not noise features. Suppose that there exists a set of features that appear frequently in relevant documents and also appear frequently in some relatively small subset of irrelevant documents but rarely in irrelevant documents outside of this small subset. Such a feature set can be considered to consist of *positive* features, and can be thought of as defining a cluster of irrelevant documents that we can assume are all about some topic that is different than the one which defines the set of relevant documents. A feature which appears in many of the documents in this cluster of irrelevant documents, but in very few relevant documents can be seen to separate the cluster from the relevant documents and is a *negative* feature that is *not* a noise feature.

For example, suppose that the topic under consideration is "fuel" and that the feature "oil" appears in many relevant documents as well as a small set of irrelevant documents. Suppose that most of members of this small set of irrelevant documents are relevant for the "veg-oil" topic, then if the feature "olive" separates the documents that are relevant for the "fuel" topic from those that are relevant for the "veg-oil" topic, it is a negative feature that is not a noise feature.

It is interesting that this feature being *negative* is dependent on the existence of a set of *positive* features. We now argue that it is difficult to find a negative feature that is independent of a set of positive features and is not a noise feature. There are two situations to consider. First, consider a feature that appears in many irrelevant documents and few relevant documents. Since the set of irrelevant documents contains multiple topics, such features appear in documents regardless of their topic and therefore are noise. Second, consider a feature that appears in few irrelevant documents and few relevant documents. If such a feature is not associated with a set of positive features as described above, then it does not separate the relevant and irrelevant documents well

and is also a noise feature.

Given our assumptions about the characteristics of textual data sets, we claim that all *negative* features that are not noise features are similarly dependent on a set of *positive* features. This claim virtually necessitates adoption of a relevant document centric view of the feature selection problem in which we are more interested in the information contained in relevant than irrelevant documents and are specifically interested in positive features.

Algorithms. Textual feature selection problems can be naturally formulated as combinatorial optimization problems which are typically NP-hard. While it might be possible to find optimal solutions to such problems using tools such as integer programming, owing to the high dimensional nature of the data set, in this dissertation we shall study two classes of heuristic algorithms to solve these problems. The classes of algorithms we shall study are called *ranking* algorithms and *greedy* algorithms.

Ranking algorithms begin by applying some function to each feature that induces a complete ordering on the feature set. The functions we will consider are real-valued and will be referred to as *feature ranking functions*. These functions provide a measure of the degree to which each feature satisfies some criteria. At each iteration, the ranking algorithm iteratively selects the next highest ranked feature until some stopping condition is met. A large proportion of the research on feature selection on textual data sets has involved empirical studies that evaluate the performance of ranking algorithms using different feature ranking functions (see e.g. [32], [40], [54], [3], and [81]).

Greedy algorithms work with a set of selected features and some specified objective function defined on this set. At each iteration they select the feature whose inclusion in the set of selected features most improves the value of the objective function until some stopping condition is met. The use of greedy algorithms for textual feature selection is much less studied than the use of ranking algorithms. There is however, a large body of research on the use of greedy algorithms for feature selection in other domains. In the Logical Analysis of Data (LAD), feature selection was formulated as a set-covering problem (SCP) with the goal of finding minimal subsets of features that preserved the ability to distinguish positive and negative examples (see e.g. [24]). In this context a

greedy heuristic has typically been employed (see e.g. [19] and [11], as well as related research in [58]). In [13], a set of efficient greedy heuristics, some having performance guarantees were proposed. These heuristics were were based on optimization problems defined in terms of measures of separation of the sets of positive and negative examples. Greedy feature selection algorithms were also studied in [1], [2], and [16].

In this dissertation, both ranking and greedy algorithms will stop when a total of K features, for some positive integer K, have been selected. The determination of K will be discussed in §2.12.

We shall let $\omega(V, K, f)$ denote the sequence containing the first K features selected from V by the feature selection algorithm f. Since feature ranking functions assign a real value to each feature, and for some such functions, larger values indicate better features, while for some functions the opposite is true; we extend this notation. We write $\omega(V, K, f, \downarrow)$ to indicate that the features are in descending order and write $\omega(V, K, f, \uparrow)$ to indicate that they are in ascending order.

It should be mentioned that the names of these algorithms are somewhat misleading in that both ranking and greedy algorithms pursue a "greedy" strategy in that they iteratively select the next best feature based their selection criteria, further, both ranking and greedy algorithms return a ranked set of features. We also mention that while anecdotal evidence suggests that the algorithms seem to retrieve features that describe each topic, they do not utilize any semantic, part of speech, structure of language, or similar information.

One important distinguishing characteristic of ranking and greedy algorithms is their run time complexity. Assuming the use of a sorting algorithm such as quicksort, with average runtime complexity of $O(n \log_2 n)$, ranking algorithms are $O(n \log_2 n)$ since they sort the set of n features once. Greedy algorithms on the otherhand are $O(n^2 \log_2 n)$. To see this note that such algorithms sort the set of n features on the first iteration, $n - 1$ features on the second iteration, and so on, until K features have

been selected. Therefore, the runtime complexity is given as

$$\begin{aligned}\sum_{i=n-K}^{n+1} i \log_2 i \leq & \int_{n-K}^{n+1} i \log_2 i \, di \\ = & \frac{i^2 \ln i}{2 \ln 2} - \frac{i^2}{4 \ln 2} \Big|_{i=n-K}^{n+1} \\ = & O(n^2 \log_2 n).\end{aligned}$$

Further, since we also have that

$$\begin{aligned}\sum_{i=n-K}^{n+1} i \log_2 i \geq & \int_{n-K-1}^{n} i \log_2 i \, di \\ = & \frac{i^2 \ln i}{2 \ln 2} - \frac{i^2}{4 \ln 2} \Big|_{i=n-K+1}^{n} \\ = & \Omega(n^2 \log_2 n),\end{aligned}$$

the greedy algorithms are actually $\Theta(n^2 \log_2 n)$.

Classification. Given a set of documents labeled as either relevant or irrelevant for a given topic, *document classification* involves "learning" a function called a *classifier* that assigns documents a label of relevant or irrelevant. Feature selection algorithms are frequently used with classification algorithms. In this context, the literature divides feature selection algorithms into those that utilize the *filter model* and those that utilize the *wrapper model*, (see e.g. [49], [43], and [82]). A feature selection algorithm which, independent of the classification algorithm, selects a feature set which is then presented to the classification algorithm is said to follow the filter model, while a feature selection algorithm which selects its features in conjunction with the classification algorithm is said to follow the wrapper model. We will only consider algorithms that follow the filter model.

As mentioned, one obvious advantage of employing feature selection is that it reduces the dimensionality of the model and since the time complexity of many classification algorithms is highly dependent on the number of features in the data set, this reduction

can result in dramatic decreases in their running time. The combination of overall dimensionality reduction and specifically noise reduction can improve the interpretability of the output of classification algorithms. In addition, there is evidence that utilization of feature selection can improve the accuracy of many classification algorithms (see e.g. [1] and [60], p.251).

While document classification has some similarities to feature selection, it should be mentioned that the nature of these problems is fundamentally different in that feature selection problems have no set of features labeled as desirable or undesirable available as a reference; instead we are left to decide what characteristics are desirable in a feature set. An overview of the set of characteristics we have decided upon is now introduced.

Evaluation. Typically the evaluation of feature selection algorithms involves comparison of the performance of one or more classification algorithms on a data set before and after feature selection has been performed (see e.g. [40], [44][1], [32] and [81]). This approach has the advantage of assessing feature sets upon one of the most common reasons for utilizing feature selection, that is, to improve the accuracy and speed of classification algorithms. The disadvantage of this approach is that the results obviously only apply to the classification algorithms that were included in the evaluation. Further, some classification algorithms can only be run on data with limited size feature sets and therefore cannot be included in such evaluations. Also, this approach does not evaluate any inherent characteristics of the selected feature set that might be useful in assessing its appropriateness for use in other tasks such as document summarization, or which might simply be of interest in their own right.

The criteria that we shall use to evaluate feature sets includes measures of

- the amount of *separation* between the set of relevant and irrelevant document sets,
- the amount of *noise* included in the feature set and,
- the *size* of the selected feature set.

[1]The reported results in this paper did include the "Fraction of features selected", which is a measure of the feature set size.

The *robustness* of feature selection algorithms with respect to the separation and noise measures is also assessed. In contrast, to the aforementioned approach, these criteria are independent of any particular classification algorithm.

1.2 Document, Feature, and Collection Models

In this section we introduce the models that we use to represent documents, features, and collections, as well as the notation associated with these models.

Document Model. As mentioned, we represent documents as vectors of features. It it important to note that modeling documents as vectors results in a significant loss of information. Specifically, it does not preserve information about the order in which terms appear in the document, a fact that leads to it often being referred to as a *bag-of-words* representation. In addition, while we do follow the *vector space model* in that we represent documents as vectors in which each component corresponds to a term and the values of the components are dependent on the number of times the corresponding term appears in a document, we depart from one of the hallmarks of this model in that we do not use the cosine of the angle between two document vectors as a measure of their similarity or the distance between them.

Let $\mathbb{B} = \{0, 1\}$, $\mathbb{R}$ be the set of real numbers, $\mathbb{Z}$ be the set of integers, $\mathbb{Z}^+$ be the set of positive integers and $\mathbb{Z}_+$ be the set of nonnegative integers. We assume that we are given a pair (T, F), with T and F respectively being the set of relevant and irrelevant documents for some topic. We shall associate an index in $W = \{1, 2, \ldots, m\}$ with each of these documents. $W_T = \{1, 2, \ldots, |T|\} \subseteq W$ will denote the set of indices of the relevant documents and $W_F = \{|T| + 1, |T| + 2, \ldots, m\} \subseteq W$ will denote the set of indices of the irrelevant documents. We shall refer to the set $T \cup F$ as the *collection*, will denote the total number of documents in the collection as $m = |T| + |F|$ and will assume that $T \cap F = \varnothing$, that is, there does not exist $u \in T$ and $v \in F$ such that $u = v$.

It should be mentioned that it is possible for two documents to contain very similar text, but because of the sensitivity of the meaning of natural language to minor variations, one document could be considered relevant for a given topic, the other document

could be considered irrelevant, and the vector space representation of the documents could actually be identical. Factors that can affect the occurrence of such anomalous situations are the length of documents, the specific steps used to prepare documents, and the inevitable inconsistencies in the document labeling process given that human judges are utilized. The issue is less likely to occur in longer documents since the terms associated with the given topic are usually repeated, thereby differentiating documents that are relevant for a topic from those that are not. Document preparation involving substantial exclusion or modification of terms, sometimes in an attempt to remove noise, can increase the likelihood of the issue arising. The document preparation we use is intentionally limited and is discussed in §2.3. A now popular example of this problem is discussed in [75] where it is shown that even differences in punctuation can impact meaning. Even in view of the possibility of a relevant and an irrelevant document being identical in the vector space model we shall assume that $T \cap F = \varnothing$.

Feature Models. As mentioned, we represent documents as vectors with each component corresponding to one of the distinct terms in the document collection. A great deal of research in information retrieval, text categorization and related fields has been devoted to studying different *term weighting* functions that are used to specify the values of the components of each vector. We do not repeat the research investigating the characteristics of these functions, but will for reasons mentioned in the sequel, simply select from a well known family of such functions.

Researchers who worked on the SMART system studied a set of term weighting functions that were the product of some function of the *term frequency*, some function of the *document frequency* and some *normalization* factor. The term frequency of a term for a given document is simply the number of times the term appears in the document. The document frequency of a term is a collection based statistic and is the number of documents in the collection in which the term appears. Normalization factors attempt to mitigate the undesirable situation in which two documents that contain a very "similar" set of terms, but have drastically different lengths are considered to have very different content by some metrics (see e.g., [69], [15], [60] and [73]).

The SMART researchers denoted specific term weighting functions as *tdn* where *t*

corresponded to a specific function of the term frequency, d corresponded to a specific function of the document frequency, and n corresponded to a specific normalization factor. We shall utilize two of these functions: the *bnn* which we shall refer to as the *Boolean* model and the *lnc* which we shall refer to as the *real-valued* model.

The rationale for selecting the Boolean and lnc functions is that they are well studied and commonly used, result in two substantially different feature sets to study, and that they only use information local to each document, that is, they do not include a document frequency factor such as the inverse document frequency (i.e. the idf) that implicitly performs feature selection by decreasing the weight of frequently used terms and increasing the weight of rare terms, and they both have the desirable property that their range is the set $[0, 1]$.

Before discussing the details of the Boolean and real-valued models we mention that just as the set W indexes the set of documents, we shall associate an index in $V = \{1, 2, \ldots, n\}$ with each of the features and that we assume that each feature appears in at least one document.

Boolean Feature Model. For $j \in V$, let f_j be the number of times a term j appears in a document vector u, then in the Boolean model, the value of this component in the document vector is

$$u_j = \begin{cases} 1 & \text{if } f_j > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

So, in the Boolean model, the components of each vector take the value 1 if the corresponding term is present in the document, the value 0 if the term is absent from the document and the set of relevant and irrelevant example vectors are $T \subseteq \mathbb{B}^n$ and $F \subseteq \mathbb{B}^n$ respectively. In terms of the tdn notation, bnn uses the Boolean term frequency factor and no, hence the n, document frequency or normalization factors.

The information associated with our Boolean formulation can be represented by associating a 2×2 *contingency table* with each feature. If $Z_T = \{0, 1, \cdots, |T|\}$ and $Z_F = \{0, 1, \cdots, |F|\}$, then a 2×2 contingency table for a feature $j \in V$ for a given topic is a mapping

$$\tau \colon Z_T \times Z_F \mapsto Z_T \times Z_F \times Z_T \times Z_F$$

such that for each $a_j \in Z_T$ and $b_j \in Z_F$

$$\tau(a_j, b_j) = (a_j, b_j, c_j, d_j)$$

where $c_j = |T| - a_j$ and $d_j = |F| - b_j$. We shall denote the set of all 2×2 contingency tables by $\boxplus$ and will let $\boxplus_j \in \boxplus$ represent the specific contingency table associated with a feature $j \in V$.

For each feature $j \in V$, the four quantities a_j, b_j, c_j and d_j are defined as

$$\begin{aligned}
a_j &\triangleq \text{the number of relevant documents containing feature } j \\
b_j &\triangleq \text{the number of irrelevant documents containing feature } j \\
c_j &\triangleq \text{the number of relevant documents which do not contain feature } j \\
d_j &\triangleq \text{the number of irrelevant documents which do not contain feature } j
\end{aligned}$$

and can be viewed as the number of true positives, false positives, false negatives and true negatives, respectively.

For each feature $j \in V$, the relationship between a_j, b_j, c_j and d_j and the document collection is given by the following 2×2 contingency table

	$\mathbf{u} \in \mathbf{T}$	$\mathbf{u} \in \mathbf{F}$	
$\mathbf{u_j = 1}$	a_j	b_j	$a_j + b_j$
$\mathbf{u_j = 0}$	c_j	d_j	$c_j + d_j$
	$a_j + c_j = \|T\|$	$b_j + d_j = \|F\|$	m

where $u \in T \cup F$, marginals $a_j + b_j$ and $c_j + d_j$ represent the number of documents that contain feature j and the number of documents which do not contain feature j respectively. Obviously for a given topic, $|T|$ and $|F|$ are constant for all features while $a_j + b_j$ and $c_j + d_j$ vary for each feature, $j \in V$. As mentioned, the total number of documents in the collection is $m = a_j + b_j + c_j + d_j$ which is obviously also a constant for a given topic.

In computations involving contingency tables, a zero value for a_j, b_j, c_j or d_j can result in modeling or computational problems such as division by zero errors. In order

to avoid such issues, it is common practice to employ one of a set of techniques that are referred to as *smoothing*. As explained in [65] the term "smoothing" follows from the fact that in the context of probability theory, the technique involves "smoothing over the parts of the probability distribution that would have been zero". In all of the experiments that are described in §2 we shall employ a technique commonly known as add one smoothing or Laplace smoothing where rather than using the values a_j, b_j, c_j and d_j, we use $a_j + 1$, $b_j + 1$, $c_j + 1$ and $d_j + 1$. This transformation has an obvious impact on the row marginals which become $a_j + b_j + 2$ and $c_j + d_j + 2$, the column marginals which become $a_j + c_j + 2$ and $b_j + d_j + 2$, and m which becomes $m + 4$.

It is important to emphasize that while it is often convenient to write the contingency table for a Boolean feature $j \in V$ for a given topic in terms of the four quantities a_j, b_j, c_j and d_j, since $c_j = |T| - a_j$ and $b_j = |F| - d_j$, and $|T|$ and $|F|$ are constants, the function τ is clearly only a function of the *two* quantities a_j and b_j. Motivated by this fact, we let

$$x_j = \frac{a_j}{|T|} \quad \text{and} \quad y_j = \frac{b_j}{|F|}$$

and will often find it convenient to apply the transformations

$$a_j = x_j|T| \quad \text{and} \quad b_j = y_j|F|, \tag{1.2}$$

which will sometimes allow us to develop models that are independent of $|T|$ and $|F|$. Also note that we will omit the subscript on these quantities a_j, b_j, c_j and d_j when there is no need to refer to a specific feature.

Real-valued Feature Model. In the real-valued model, the value corresponding to feature $j \in V$ in a document vector u is

$$u_j = \frac{t_j}{\sqrt{t_1^2 + t_2^2 + \cdots t_n^2}} \tag{1.3}$$

where

$$t_j = \begin{cases} \log(f_j) + 1 & \text{if } f_j > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

The components of each vector are weighted by the logarithm and then normalized by the resulting vector length so that $0 \leq u_j \leq 1$ and therefore the set of relevant and irrelevant example vectors are $T \subseteq [0,1]^n$ and $F \subseteq [0,1]^n$ respectively. The logarithmic weighting is based on the theory that the importance of a term does not grow linearly with the number of times it appears in a document, but rather increases more slowly. In terms of the tdn notation, lnc uses the logarithmic term frequency factor, no document frequency factor and the *cosine* normalization factor. The information associated with our real-valued formulation can be represented by associating two vectors with each feature $j \in V$, namely the vector of the feature's lnc values for all relevant documents and the vector of the feature's lnc values for all irrelevant documents.

Projections. Feature selection can be seen as the *projection* of a data set onto a subspace of smaller dimension. Motivated by this fact we now introduce some notation that will facilitate working with projections. Since we indexed the set of features by the elements of V we can adopt a functional notation for representing the elements and subsets of the set of vectors in the collection. In this notation, vectors are viewed as functions. Specifically, $\mathbb{B}^V$ is the set of all functions from V to $\mathbb{B}$, and writing $u \in \mathbb{B}^V$ is equivalent to writing $u \in \mathbb{B}^n$, with similar notation applying to the real-valued model. This functional notation is, particularly convenient when working with projections onto subspaces. For example, if $S \subseteq V$, and $u \in \mathbb{B}^V$, we shall let $u[S] \in \mathbb{B}^S$ indicate the projection of u onto S and for $X \subseteq \mathbb{B}^V$ we shall write $X[S]$ as the projection of X on S, that is, $X[S] = \{u[S] \mid u \in X\}$. For a subset $S \subseteq V$ let us denote by $\chi^S \in \mathbb{B}^n$ its *characteristic vector*, i.e.

$$\chi_j^S = \begin{cases} 1 & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

In the sequel, the set S will usually correspond to a subset of the feature set V after application of a feature selection algorithm.

Collection Models. We will employ a well known matrix representation of the collection. A matrix in which the rows correspond to documents and columns correspond to features will be referred to as a *document-term matrix*. The row indices of such a matrix will be a set $W' \subseteq W$ and the column indices will be a set $V' \subseteq V$. For a

document-term matrix $A = [a_{ij}]$ with $i \in W'$ and $j \in V'$ each entry will take on the corresponding value of either the Boolean or the real-valued model for feature $j \in V'$ as it appears in document $i \in W'$. When $W' = W$ and $V' = V$, the $m \times n$ matrix can be seen to represent the relationship between each document and each feature in the collection. If $V' \subseteq V$ and $W' \subseteq W$, then we will let $A[W', V']$ denote the projection of the matrix A onto the rows in W' and the columns in V'. If A is a document-term matrix we will let $A_{(T)} = A[W_T, V]$ and $A_{(F)} = A[W_F, V]$ denote the matrices which only contain the rows in W_T and W_F respectively.

Space Complexity. Clearly, the information used by the Boolean model can be stored in less space than the information used by the real-valued model. Let us assume for the sake of simplicity, that an integer and a floating point number each require δ units of space. Storing a contingency table for each feature requires $4 \times n \times \delta$ units of space, while storing an $m \times n$ real-valued document-term matrix requires $m \times n \times \delta$ units of space. Therefore, storage of the real-valued document-term matrix takes $m/4$ times more space than the storage of the contingency tables. For a large collection the difference in the amount of required storage can obviously be substantial.

1.3 ROC Curves and the Wilcoxon Rank Sum Test

In this section we discuss Receiver Operating Characteristics (ROC) curves (see e.g. [34] and [33]) and the closely related Wilcoxon rank sum statistics and tests[2] (see e.g. [64] and [72]) which will be used throughout the sequel.

Consider a *real-valued classifier* $f^+ \colon T \cup F \mapsto \mathbb{R}$, that for each document $i \in W$ assigns a *score*

$$f^+(i) = \lambda_i \tag{1.5}$$

which represents its estimate, with larger scores being considered indicative of relevant documents, that the document is relevant. We will let $\lambda \in \mathbb{R}^m$ be the vector of these scores, and $\Lambda \subseteq \mathbb{R}$ be the corresponding set of these scores. The classifier can be a sophisticated classifier that for each document perhaps returns a probability as the

[2]The Wilcoxon rank sum test is equivalent to the Mann-Whitney test.

score, or a simple classifier that for each document just returns a parameter such as the Boolean model value or real-valued model value of some feature.

Given the vector of scores λ, since larger scores are considered to indicate a higher estimate of *relevance*, for a given $\tau \in \Lambda$ we can construct a *binary classifier* $g_\tau^+ : T \cup F \mapsto \mathbb{B}$ as

$$g_\tau^+(i) = \begin{cases} 1 & \text{if } f^+(i) = \lambda_i \geq \tau \\ 0 & \text{otherwise,} \end{cases} \tag{1.6}$$

where $i \in W$. Further, we can define a *family* of binary classifiers

$$\mathcal{G}^+ = \{g_\tau^+ : \tau \in \Lambda\} \tag{1.7}$$

by considering all values of the threshold $\tau \in \Lambda$.

Now consider $\lambda[W_T]$, the vector of scores that f^+ assigns to relevant documents and $\lambda[W_F]$, the vector of scores that f^+ assigns to irrelevant documents. If λ_T and λ_F are randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively, then the probability that λ_T and λ_F are ordered correctly,

$$\mathrm{P}(\lambda_T > \lambda_F)$$

provides a measure of the ability of f^+ to rank relevant documents with respect to irrelevant documents, and a measure of the ability of classifiers in $\mathcal{G}^+$ to correctly assign documents to T and F. A classifier, f^+ with

$$\mathrm{P}(\lambda_T > \lambda_F) = 1$$

has perfect performance in that it ranks each relevant document above each irrelevant document. In this case, there exists exactly one classifier $g_\tau^+ \in \mathcal{G}^+$ which is always correct, i.e. it correctly assigns each document to either T or F. A classifier, f^+ with

$$\mathrm{P}(\lambda_T > \lambda_F) = 0$$

has perfectly imperfect performance in that it ranks each irrelevant document above each relevant document. In this case, there is exactly one classifier $g_\tau^+ \in \mathcal{G}^+$ which is always incorrect, i.e. for each document, its assignment to T or F is always wrong. The performance of a classifier f^+ with

$$\mathrm{P}(\lambda_T > \lambda_F) = \frac{1}{2}$$

is the same as if it had randomly assigned scores to each document.

Since there are $|T||F|$ pairs in $\lambda[W_T] \times \lambda[W_F]$ to consider, if we assume that no $\lambda_T \in \lambda[W_T]$ and $\lambda_F \in \lambda[W_F]$ are equal, clearly

$$\mathrm{P}(\lambda_T > \lambda_F) = \frac{|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_T > \lambda_F\}|}{|T||F|},$$

which to account for equality of $\lambda_T \in \lambda[W_T]$ and $\lambda_F \in \lambda[W_F]$, we can refine to be

$$\mathrm{P}(\lambda_T > \lambda_F) + \frac{1}{2}\mathrm{P}(\lambda_T = \lambda_F) = \frac{U_+}{|T||F|}$$

where

$$\begin{aligned} U_+ = & \Big|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_T > \lambda_F\}\Big| \\ & + \frac{1}{2}\Big|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_T = \lambda_F\}\Big|. \end{aligned} \tag{1.8}$$

While the $\mathrm{P}(\lambda_T > \lambda_F)$ provides a measure of the performance of an arbitrary real-valued classifier f^+, whenever f^+ is such that $|\mathcal{G}^+| = 1$, it also provides a measure the performance of the sole binary classifier contained in $\mathcal{G}^+$. To see this, note that $|\mathcal{G}^+| = 1$ with $\mathcal{G}^+ = \{g_\tau^+\}$ for some $\tau \in \Lambda$ implies that f^+ and g_τ^+ are isomorphic,[3] which in turn shows that $\mathrm{P}(\lambda_T > \lambda_F)$ provides a measure of the performance of the binary classifier g_τ^+. Further, since whenever $|\mathcal{G}^+| = 1$, we have that f^+ and g_τ^+ are isomorphic, we can interpret $\mathrm{P}(\lambda_T > \lambda_F)$ to be the probability that g_τ^+ will correctly classify both a randomly selected relevant document and a randomly selected irrelevant

[3]That is, the score vector λ and the vector $[g_\tau^+(i)]$ for $i \in W$ are isomorphic.

document.

A perhaps more commonly employed measure of the performance of a single binary classifier $g_\tau^+ \in \mathcal{G}^+$ for a given $\tau \in \Lambda$ is a measure called the *accuracy* which is defined as

$$\mathrm{ACC}(g_\tau^+) = \frac{|\{i \in W_T : \ g_\tau^+(i) = 1\}| + |\{i \in W_F : \ g_\tau^+(i) = 0\}|}{|T| + |F|},$$

that is, the ratio of the number of documents correctly classified to the total number of documents. If $i \in W$ is a randomly selected document, the accuracy can be seen to be the

$$\mathrm{P}(\{g_\tau^+(i) = 1, i \in W_T\} \cup \{g_\tau^+(i) = 0, i \in W_F\}).$$

that is, the accuracy of a classifier is the probability that it correctly classifies a randomly selected document. For a given $\tau \in \Lambda$, clearly, if $\mathrm{ACC}(g_\tau^+) = 1$ then g_τ^+ has perfect performance, if $\mathrm{ACC}(g_\tau^+) = 0$ then g_τ^+ has perfectly imperfect performance, and if $\mathrm{ACC}(g_\tau^+) = 1/2$ then the performance is of g_τ^+ is the same as randomly guessing whether each document is relevant or irrelevant. In some contexts rather than seeking a large value of the accuracy, it might be more natural to seek a small value for the *error rate* which is equal to one minus the accuracy.

Now, suppose that $\mathrm{P}(\lambda_T > \lambda_F) < 1/2$, then clearly there are more pairs $(\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F]$ with $\lambda_F > \lambda_T$ than vice versa. Further,

Proposition 1.1. *If λ_T and λ_F are randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively, then $P(\lambda_T > \lambda_F) = 1 - P(\lambda_F > \lambda_T)$ and $P(\lambda_T > \lambda_F) < 1/2$ if and only if the $P(\lambda_F > \lambda_T) > 1/2$.*

Proof. Follows from the fact that the events $\{\lambda_T > \lambda_F\}$ and $\{\lambda_F > \lambda_T\}$ are complementary. ■

This result motivates consideration of the real-valued *complementary classifier* to f^+ which we shall denote as f^-. The classifier f^- outputs the same score vector λ and score set Λ as f^+, but differs in that larger scores are considered indicative of irrelevant documents. Since larger scores indicate a higher estimate of *irrelevance*, for a given $\tau \in \Lambda$, we can construct the binary complementary classifier to $g_\tau^+ \in \mathcal{G}^+$,

$g_\tau^- : T \cup F \times \mapsto \mathbb{B}$ as

$$g_\tau^-(i) = \begin{cases} 1 & \text{if } f^-(i) = f^+(i) = \lambda_i \leq \tau \\ 0 & \text{otherwise,} \end{cases} \tag{1.9}$$

or equivalently, it can be constructed from g_τ^+ as

$$g_\tau^-(i) = \begin{cases} 1 & \text{if } g^+(i) = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1.10}$$

where $i \in W$. Thus, $g_\tau^- \in \mathcal{G}^-$ can be constructed from $g_\tau^+ \in \mathcal{G}^+$ or vice versa by inverting all classifications, that is, by applying the transformation

$$g_\tau^-(i) \leftarrow 1 - g_\tau^+(i)$$

for all $i \in W$ (see e.g. [34] and [39]). Also, while we have not made use of it in this section, there is a 2×2 contingency table associated with every binary classifier $g^+ \in \mathcal{G}^+$ and $g^- \in \mathcal{G}^-$. For example, the following table is the contingency table for the classifier $g^+ \in \mathcal{G}^+$ for an arbitrary $\tau \in \Lambda$.

	$\boldsymbol{i \in W_T}$	$\boldsymbol{i \in W_F}$	
$\boldsymbol{\lambda_i \geq \tau}$	a_τ	b_τ	$a_\tau + b_\tau$
$\boldsymbol{\lambda_i < \tau}$	c_τ	d_τ	$c_\tau + d_\tau$
	$a_\tau + c_\tau = \lvert T \rvert$	$b_\tau + d_\tau = \lvert F \rvert$	m

Since the first row of this table corresponds to documents that the classifier predicts are relevant and the second row of this table corresponds to documents that the classifier predicts are irrelevant, clearly an alternate way of viewing the construction of $g^- \in \mathcal{G}^-$ from $g^+ \in \mathcal{G}^+$ is that it involves interchanging the rows in this table. Further, we can define a *family* of binary classifiers

$$\mathcal{G}^- = \{g_\tau^- : \tau \in \Lambda\} \tag{1.11}$$

by considering all values of the threshold $\tau \in \Lambda$.

The performance of f^- and the classifiers in $\mathcal{G}^-$ is related to the $\mathrm{P}(\lambda_F > \lambda_T)$, the probability that a randomly selected pair from $\lambda[W_T] \times \lambda[W_F]$ is ordered correctly. If we assume that no $\lambda_T \in \lambda[W_T]$ and $\lambda_F \in \lambda[W_F]$ are equal, clearly

$$\mathrm{P}(\lambda_F > \lambda_T) = \frac{|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_F > \lambda_T\}|}{|T||F|},$$

which to account for equality of $\lambda_T \in \lambda[W_T]$ and $\lambda_F \in \lambda[W_F]$, we can refine to be

$$\mathrm{P}(\lambda_F > \lambda_T) + \frac{1}{2}\mathrm{P}(\lambda_T = \lambda_F) = \frac{U_-}{|T||F|}$$

where

$$\begin{aligned} U_- = & \left|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_F > \lambda_T\}\right| \\ & + \frac{1}{2}\left|\{(\lambda_T, \lambda_F) : (\lambda_T, \lambda_F) \in \lambda[W_T] \times \lambda[W_F] \text{ and } \lambda_T = \lambda_F\}\right|. \end{aligned} \tag{1.12}$$

From Proposition 1.1 and the construction of f^- it can be seen that if the performance of f^+ is worse than random score assignment, that is, if $\mathrm{P}(\lambda_T > \lambda_F) < 1/2$, then the performance of f^- is better than random score assignment, that is $\mathrm{P}(\lambda_F > \lambda_T) = 1 - \mathrm{P}(\lambda_T > \lambda_F) > 1/2$ and vice versa. Similar remarks also clearly apply to the performance of the binary classifiers g_τ^+ and g_τ^- for $\tau \in \Lambda$ as measured by the $\mathrm{P}(\lambda_T > \lambda_F)$ and $\mathrm{P}(\lambda_F > \lambda_T)$ respectively, when $|\mathcal{G}^+| = |\mathcal{G}^-| = 1$, $\mathcal{G}^+ = \{g_\tau^+\}$ and $\mathcal{G}^- = \{g_\tau^-\}$. In addition, if $\mathrm{ACC}(g_\tau^+) < 1/2$, then $\mathrm{ACC}(g_\tau^-) = 1 - \mathrm{ACC}(g_\tau^+) > 1/2$ and vice versa. To see this, notice that

$$\mathrm{ACC}(g_\tau^+) = \frac{a_\tau + d_\tau}{|T| + |F|}$$

and inverting all classifications made by g_τ^+ to get g_τ^- is accomplished by letting $a_\tau \to c_\tau$ and $d_\tau \to b_\tau$ which yields

$$\mathrm{ACC}(g_\tau^-) = \frac{c_\tau + b_\tau}{|T| + |F|} = \frac{|T| - a_\tau + |F| - d_\tau}{|T| + |F|} = 1 - \frac{a_\tau + d_\tau}{|T| + |F|}.$$

Our discussion of the $\mathrm{P}(\lambda_T > \lambda_F)$ and the $\mathrm{P}(\lambda_F > \lambda_T)$ is based on the Wilcoxon

rank sum statistical test (see e.g [64], [72], [46], [5], [79]). While this test can be applied to any two sets of numbers that satisfy some simple conditions (see e.g. [??]), we shall continue our discussion in terms of the sets of scores $\lambda[W_T]$ and $\lambda[W_F]$. The quantity U_- is the statistic used in the one-sided Wilcoxon rank sum test

$$\begin{aligned} H_0 &: \mathrm{P}(\lambda_T > \lambda_F) = \frac{1}{2} \\ H_{a^+} &: \mathrm{P}(\lambda_T > \lambda_F) > \frac{1}{2} \Leftrightarrow \\ H_{a^+} &: \mathrm{P}(\lambda_F > \lambda_T) < \frac{1}{2}. \end{aligned} \tag{1.13}$$

In this test the null hypothesis is that the performance of f^+ is no better than if the the classifier had randomly assigned scores to documents and the one-sided alternative is that the performance of f^+ is better than random score assignment. The quantity U_+ is the statistic used in the one-sided Wilcoxon rank sum test

$$\begin{aligned} H_0 &: \mathrm{P}(\lambda_F > \lambda_T) = \frac{1}{2} \\ H_{a^-} &: \mathrm{P}(\lambda_F > \lambda_T) > \frac{1}{2} \Leftrightarrow \\ H_{a^-} &: \mathrm{P}(\lambda_T > \lambda_F) < \frac{1}{2}. \end{aligned} \tag{1.14}$$

In this test the null hypothesis is that the performance of f^- is no better than if the classifier had randomly assigned scores to documents and the one-sided alternative is that the performance of f^- is worse than random score assignment. The two-sided test

$$\begin{aligned} H_0 &: \mathrm{P}(\lambda_T > \lambda_F) = \frac{1}{2} \\ H_a &: \mathrm{P}(\lambda_T > \lambda_F) \neq \frac{1}{2}. \end{aligned} \tag{1.15}$$

uses $\min(U_-, U_+)$ as its statistic. In this test the null hypothesis is that the performance of either f^+ or f^- is no better than if the classifier had randomly assigned scores to documents and the alternative hypothesis is that the performance of either f^+ or f^- is better than random score assignment. [4]

[4] A statistic related to the Wilcoxon statistics, equal to $\frac{U_+}{|T||F|}$, and denoted as ρ, not to be confused

The somewhat counterintuitive reason that U_- is used in the one-sided test in (1.13) that includes the alternative hypothesis $\mathrm{P}(\lambda_T > \lambda_F)$ is that the test rejects the null hypothesis when U_- is unusually small, and since $U_+ + U_- = |T|F|$ is a constant, when U_- is unusually small, U_+ is unusually large, which means that many scores in $\lambda[W_T]$ exceed the scores in $\lambda[W_F]$ as desired. A similar comment can be made about why U_+ is used in the one side test in (1.14) that includes the alternative hypothesis $\mathrm{P}(\lambda_F > \lambda_T)$. Since the two-sided test uses the $\min(U_-, U_+)$ as its statistic, it is testing whether either U_+ or U_- is unusually small and therefore either U_+ or U_- is unusually large.

ROC curves provide another tool for understanding the performance of classifiers. Consider the family of classifiers $\mathcal{G}^+$ and notice that for each value of $\tau \in \Lambda$ we can calculate the *true positive rate*

$$\mathrm{TPR}_\tau = \frac{|\{i \in W_T \,:\, g_\tau^+(i) = 1\}|}{|T|}$$

and the *false positive rate*

$$\mathrm{FPR}_\tau = \frac{|\{i \in W_F \,:\, g_\tau^+(i) = 1\}|}{|F|}$$

where $g_\tau^+ \in \mathcal{G}^+$. That is, the TPR_τ for classifier $g_\tau^+ \in \mathcal{G}^+$ is the ratio of the number of relevant documents that were classified correctly to the total number of relevant documents and the FPR_τ is the ratio of the number of irrelevant documents that were classified incorrectly to the total number of irrelevant documents. It is natural to envision these quantities being computed by first sorting λ into ascending order and then considering the set of documents with $\lambda_i \geq \tau$ for $i \in W$. The relevant documents in this set define the numerator of the TPR_τ and the irrelevant documents in this set form the numerator of the FPR_τ.

One way to understand the performance of such a family of classifiers is to study how the TPR and FPR vary with τ. An ROC curve plots the TPR on the y-axis and the FPR on the x-axis for various values of τ and thereby provides a framework for

---

with Spearman's rank correlation coefficient, was introduced in [47]. See also [79].

such studies. If Λ is the set of scores from a classifier, then the corresponding ROC curve is constructed from the set of points

$$\left\{ (\mathrm{FPR}_\tau, \mathrm{TPR}_\tau) \colon \tau \in \Lambda \right\}$$

and the resulting graph is a step function.

In terms of the classifiers we have been discussing, each *point* on an ROC curve corresponds to a specific $g_\tau \in \mathcal{G}$ where $\mathcal{G}$ is a family of binary classifiers, Λ is the set of scores of the associated real-valued classifier and $\tau \in \Lambda$. We will use the notation $\mathrm{ROC}(\mathcal{G})$ to refer to the ROC curve for a specific family of binary classifiers $\mathcal{G}$ and when it will not cause any confusion we will simply write ROC. We will often adopt the short hand notation ROC_+ for $\mathrm{ROC}(\mathcal{G}^+)$ and ROC_- for $\mathrm{ROC}(\mathcal{G}^-)$.

Now, if $g_\tau, g_{\tau'} \in \mathcal{G}$ where $\mathcal{G}$ is a family of binary classifiers, Λ is the set of scores of the associated real-valued classifier and $\tau, \tau' \in \Lambda$, then if $\mathrm{TPR}_\tau > \mathrm{TPR}_{\tau'}$ and $\mathrm{FPR}_\tau < \mathrm{FPR}_{\tau'}$, the performance of g_τ, as measured by both the AUC and the accuracy, is clearly better than that of $g_{\tau'}$. In such cases we shall write $g_{\tau'} \preccurlyeq_{\mathrm{ROC}} g_\tau$ and notice that the set of all classifiers given as points in ROC space and the relation $\preccurlyeq$ form a partial order.

The *area under the ROC curve* is the central performance measure in this framework. We will use the notation $\mathrm{AUC}(\mathcal{G})$ to refer to the area under the curve for a specific family of binary classifiers $\mathcal{G}$ and when it will not cause any confusion we will simply write AUC. We will often adopt the short hand notation AUC_+ for $\mathrm{AUC}(\mathcal{G}^+)$ and AUC_- for $\mathrm{AUC}(\mathcal{G}^-)$. The most important properties of the AUC are that for the family of binary classifiers $\mathcal{G}^+$

$$\mathrm{AUC}(\mathcal{G}^+) = \mathrm{P}(\lambda_T > \lambda_F)$$

and for the family of binary classifiers $\mathcal{G}^-$

$$\mathrm{AUC}(\mathcal{G}^-) = \mathrm{P}(\lambda_F > \lambda_T)$$

where λ_T and λ_F are randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively (see

e.g [46] and [5]). In light of these properties, all of our earlier discussions related to the $P(\lambda_T > \lambda_F)$ and the $P(\lambda_F > \lambda_T)$ in the context of the Wilcoxon rank sum tests obviously apply to the AUC, and therefore these properties justify the statement that the ROC curves and the Wilcoxon rank sum tests and statistics are closely related.

Table 1.1 states some of the relationships between the Wilcoxon rank sum statistics and ROCs.

WILCOXON STATISTICS		ROC CURVES
$0 \leq U_+, U_- \leq \lvert T\rvert F\rvert$	$\Leftrightarrow$	$0 \leq \mathrm{AUC}_+, \mathrm{AUC}_- \leq 1$
$U_+ + U_- = \lvert T\rvert\lvert F\rvert$	$\Leftrightarrow$	$\mathrm{AUC}_+ + \mathrm{AUC}_- = 1$
$P(\lambda_T > \lambda_F) = \dfrac{U_+}{\lvert T\rvert\lvert F\rvert}$	$\Leftrightarrow$	$P(\lambda_T > \lambda_F) = \mathrm{AUC}_+$
$P(\lambda_F > \lambda_T) = \dfrac{U_-}{\lvert T\rvert\lvert F\rvert}$	$\Leftrightarrow$	$P(\lambda_F > \lambda_T) = \mathrm{AUC}_-$

Table 1.1: Relationships Between Wilcoxon Rank Sum Statistics and ROCs

The obvious distinguishing property of the ROC curves when compared to the Wilcoxon statistics is the geometric characterization they provide. We now state some observations in terms of the geometry of the ROC.

In view of our comments above, the most obvious of these observations are related to the AUC. For example, consider the family of binary classifiers $\mathcal{G}^+$. The associated curve ROC_+ will have $\mathrm{AUC}_+ > 1/2$ when $P(\lambda_T > \lambda_F) > 1/2$, $\mathrm{AUC}_- > 1/2$ when $P(\lambda_T > \lambda_F) < 1/2$ and, $\mathrm{AUC}_- = 1/2$ when $P(\lambda_T > \lambda_F) = 1/2$. Obviously, similar comments can be made about the family of binary classifiers $\mathcal{G}^-$.

In order to consider these comments in more detail, we now focus on the relationship between the performance of a single binary classifier and the AUC. To do this, we again assume that $|\mathcal{G}^+| = 1$ and $\mathcal{G}^+ = \{g_\tau^+\}$ for $\tau \in \Lambda$ and notice that other than the points $(0,0)$ and $(1,1)$, ROC_+ consists of exactly one point, namely the point $p = (\mathrm{FPR}_\tau, \mathrm{TPR}_\tau)$. If $\mathrm{TPR}_\tau > \mathrm{FPR}_\tau$, then p lies above the line FPR = TPR, the $\mathrm{AUC}_+ > 1/2$, the $P(\lambda_T > \lambda_F) > 1/2$, and the performance of g_τ^+ is better than random guessing. Similarly, if $\mathrm{FPR}_\tau > \mathrm{TPR}_\tau$, then p lies below the line FPR = TPR, the

$\text{AUC}_+ < 1/2$, the $\text{P}(\lambda_T > \lambda_F) < 1/2$, and the performance of g_τ^+ is worse than random guessing. Clearly, if $\text{TPR}_\tau = \text{FPR}_\tau$, then p is on the line FPR = TPR, the $\text{AUC}_+ = 1/2$ and therefore the $\text{P}(\lambda_T > \lambda_F) = 1/2$, so the performance of $g^+ \in \mathcal{G}^+$ is no better than random guessing. Similar comments can be made about the binary classifier $g_\tau^- \in \mathcal{G}^-$ for a given $\tau \in \Lambda$.

Following Proposition 1.1 we saw that for a given $\tau \in \Lambda$, if the performance of $g_\tau^+ \in \mathcal{G}^+$ (or $g_\tau^- \in \mathcal{G}^-$) was worse than random guessing, that the complementary classifier $g_\tau^- \in \mathcal{G}^-$ (or $g_\tau^+ \in \mathcal{G}^+$) would be better than random guessing. We now continue that discussion, presenting it in terms of the geometry of the ROC. To do this, we again assume that $|\mathcal{G}^+| = 1$ and $\mathcal{G}^+ = \{g_\tau^+\}$ for $\tau \in \Lambda$ and consider the point on the ROC_+ that corresponds to g_τ^+, namely the point $p^+ = (\text{FPR}_\tau, \text{TPR}_\tau)$. If $\text{FPR}_\tau > \text{TPR}_\tau$ then as was just stated p^+ lies below the line FPR = TPR, the $\text{AUC}_+ < 1/2$, the $\text{P}(\lambda_T > \lambda_F) < 1/2$, and the performance of g_τ^+ is worse than random guessing. In this case, the point $p^- = (1 - \text{FPR}_\tau, 1 - \text{TPR}_\tau)$ corresponds to the classifier $g_\tau^- \in \mathcal{G}^-$ with p^- lying above the line FPR = TPR. The point p^- along with $(0, 0)$ and $(1, 1)$ form ROC_- and we have that $\text{AUC}_- > 1/2$, the $\text{P}(\lambda_F > \lambda_T) = 1 - \text{P}(\lambda_T > \lambda_F) > 1/2$, and the performance of g_τ^- is better than random guessing.

Notice that this observation can be extended from the case where $|\mathcal{G}^+| = 1$ to that where $\mathcal{G}^+$ is of arbitrary cardinality. Suppose we are given the curve ROC_+, we can construct ROC_- by translating each point $(\text{FPR}_\tau, \text{TPR}_\tau)$ for $\tau \in \Lambda$ on ROC_+ to $(1 - \text{FPR}_\tau, 1 - \text{TPR}_\tau)$ and obviously, given the curve ROC_-, ROC_+ can be constructed similarly. Now if the performance of $g_\tau^+ \in \mathcal{G}^+$ for all $\tau \in \Lambda$ are such that $\text{AUC}_+ < 1/2$, then we can construct $g_\tau^- \in \mathcal{G}^-$ for all $\tau \in \Lambda$ and $\text{AUC}_- > 1/2$.

Proposition 1.2. $\max(AUC_+, 1 - AUC_+) = \max(AUC_+, AUC_-) = \max(AUC_-, 1 - AUC_-) \geq 1/2$ *and equality holds if and only if* $AUC_+ = AUC_- = 1/2$.

Proof. Follows immediately from the fact that $\text{AUC}_+ + \text{AUC}_- = 1$. ■

Therefore, if $\text{AUC}_+ < 1/2$ then $1 - \text{AUC}_+ = \text{AUC}_- > 1/2$. Figure 1.1 provides an example that demonstrates this result, which we will make use of in §3.8.

We will now show how to construct new binary classifiers whose expected performance can be stated in terms of the classifiers in $\mathcal{G}^+$.

Proposition 1.3. *If $g^+_{\tau_1}$,$g^+_{\tau_2} \in \mathcal{G}^+$ are two binary classifiers with $FPR_{\tau_1} < FPR_{\tau_2}$ and $TPR_{\tau_1} < TPR_{\tau_2}$ then for the random classifier*

$$\bar{g}^+_{\bar{\tau}}(i) = \begin{cases} 1 & \textit{with probability } \gamma \textit{ if } \lambda_i \geq \tau_1 \\ 1 & \textit{with probability } 1-\gamma \textit{ if } \lambda_i \geq \tau_2 \\ 0 & \textit{otherwise,} \end{cases} \tag{1.16}$$

$E(\bar{\tau}) = \gamma\tau_1 + (1-\gamma)\tau_2$ and the expected performance is given by the fact that $E(FPR_{\bar{\tau}}) = \gamma FPR_{\tau_1} + (1-\gamma)FPR_{\tau_2}$ and $E(TPR_{\bar{\tau}}) = \gamma\, TPR_{\tau_1} + (1-\gamma)\, TPR_{\tau_2}$.

Proof. For each document $i \in W$, $\bar{g}^+_{\bar{\tau}}$ first randomly selects the threshold to be τ_1 with probability γ and τ_2 with probability $1-\gamma$ and then predicts that i is relevant when λ_i exceeds the selected threshold and otherwise predicts i to be irrelevant. Now, notice that we can write $\bar{g}^+_{\bar{\tau}}(i)$ as

$$\bar{g}^+_{\bar{\tau}}(i) = \begin{cases} 1 & \text{with probability } \gamma \text{ if } g^+_{\tau_1}(i) = 1 \\ 1 & \text{with probability } 1-\gamma \text{ if } g^+_{\tau_2}(i) = 1 \\ 0 & \text{otherwise} \end{cases}$$

which we use define the random variables

$$\begin{aligned} \mathrm{TPR}_{\bar{\tau}} &= \tfrac{1}{|T|} \left|\{i \in W_T \colon g^+_{\bar{\tau}}(i) = 1\}\right| \\ \mathrm{FPR}_{\bar{\tau}} &= \tfrac{1}{|F|} \left|\{i \in W_F \colon g^+_{\bar{\tau}}(i) = 1\}\right|. \end{aligned}$$

Since the $\mathrm{TPR}_{\bar{\tau}}$ and $\mathrm{FPR}_{\bar{\tau}}$ are determined solely by the cases where $\bar{g}^+_{\bar{\tau}}(i) = 1$ for $i \in W$, and that γ is the proportion of these cases that are determined by $g^+_{\tau_1}$ and $1-\gamma$ is the proportion of these cases that are determined by $g^+_{\tau_2}$ we have

$$\begin{aligned} \mathrm{E}(\mathrm{TPR}_{\bar{\tau}}) &= \gamma \left|\{i \in W_T \colon g^+_{\tau_1}(i) = 1\} + (1-\gamma) \#\{i \in W_T \colon g^+_{\tau_2}(i) = 1\}\right| \\ &= \gamma\ \mathrm{TPR}_{\tau_1} + (1-\gamma)\ \mathrm{TPR}_{\tau_2} \end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\mathrm{FPR}_{\bar{\tau}}) &= \gamma \left|\{i \in W_F \colon g^+_{\tau_1}(i) = 1\}\right| + (1-\gamma) \left|\{i \in W_F \colon g^+_{\tau_2}(i) = 1\}\right| \\ &= \gamma\, \mathrm{FPR}_{\tau_1} + (1-\gamma)\, \mathrm{FPR}_{\tau_2}.\end{aligned}$$

■

This result shows that even though $\mathcal{G}^+$ only contains a finite number of classifiers, i.e. one for each element of Λ, that it is possible to construct a new binary classifier $\bar{g}^+_{\bar{\tau}}$ for any $\bar{\tau} \in [\min(\Lambda), \max(\Lambda)]$. If $g^+_{\tau_1}, g^+_{\tau_2} \in \mathcal{G}^+$ are as in Proposition 1.3 then $\tau_1 > \bar{\tau} > \tau_2$ and the point $(\mathrm{E}(\mathrm{TPR}_{\bar{\tau}}), \mathrm{E}(\mathrm{TPR}_{\bar{\tau}}))$ lies on the convex combination of the points $(\mathrm{TPR}_{\tau_1}\ \mathrm{FPR}_{\tau_1})$ and $(\mathrm{TPR}_{\tau_2}\ \mathrm{FPR}_{\tau_2})$ and therefore the expected performance of $\bar{g}^+_{\bar{\tau}}$ will, in terms of the partial order $\preccurlyeq_{\mathrm{ROC}}$, be no worse than that of classifiers $g^+_{\tau_1}$ and $g^+_{\tau_2}$. As an example, again consider the classifier family $\mathcal{G}^+$ in Figure 1.1 and suppose that we would like to create a classifier $\bar{g}^+_{\tau}$ where $\bar{\tau} = 10.25$. To do this we use the classifiers $g^+_{\tau_1}$ with $\tau_1 = 10.5$ and $g^+_{\tau_2}$ with $\tau_2 = 10.1$. Solving the equation

$$\bar{\tau} = \gamma 10.5 + (1-\gamma)10.1 = 10.25$$

yields $\gamma = 0.38$ and therefore the desired convex combination of $(\mathrm{FPR}_{\tau_1}, \mathrm{TPR}_{\tau_1})$ and $(\mathrm{FPR}_{\tau_2}, \mathrm{TPR}_{\tau_2})$ is

$$0.38(0, 0.33) + (1 - 0.38)(0.25, 0.5) = (0.16, 0.44)$$

as shown in Figure 1.2a.

Further, if $g^+_{\tau_1}, g^+_{\tau_2} \in \mathcal{G}^+$ are as in Proposition 1.3, $\tau_1 > \tau > \tau_2$, and $g^+_{\tau} \in \mathcal{G}^+$, then if the performance of g^+_{τ} in terms of the partial order $\preccurlyeq_{\mathrm{ROC}}$ is worse than $g^+_{\tau_1}$ and $g^+_{\tau_2}$, replacing it by $\bar{g}^+_{\bar{\tau}}$ will result in an increase in the AUC_+. As an example, again consider the classifier family $\mathcal{G}^+$ in Figure 1.1 and suppose that we would like to replace the classifier g^+_{τ} with $\tau = 10.1$ and $(\mathrm{TPR}_{\tau}, \mathrm{FPR}_{\tau}) = (0.25, 0.33)$ with a better performing classifier. We use the classifiers $g^+_{\tau_1}$ and $g^+_{\tau_2}$ with $\tau_1 = 10.5$ and $\tau_2 = 9$ which

have $(\mathrm{TPR}_{\tau_1}, \mathrm{FPR}_{\tau_1}) = (0.0, 0.33)$ and $(\mathrm{TPR}_{\tau_2}, \mathrm{FPR}_{\tau_2}) = (0.5, 0.83)$ respectively. Each point on the convex combination of these two points will satisfy

$$\gamma(0.0, 0.33) + (1 - \gamma)(0.5, 0.83) = (0.5 - 0.5\gamma, 0.83 - 0.5\gamma).$$

Solving

$$\mathrm{FPR}_{\tau} = 0.5 - 0.5\gamma = 0.25$$

yields $\gamma = 0.5$ and therefore $\mathrm{TPR}_{\bar{\tau}} = 0.83 - (0.5)(0.5) = 0.58$ as shown in Figure 1.2b. The natural extension of this strategy is to replace ROC_+ with its convex hull (see e.g [34], [35] and [39]) and Proposition 1.3 shows that construction of the convex hull can be accomplished by application of a set of appropriately chosen random classifiers $\bar{g}^+_{\bar{\tau}}$.

We close this section with a comment about the relationship between the AUC and accuracy. It is important to note that the AUC and accuracy (or error rate) are fundamentally different measures in that the former is applied to a *ranking* while the later is applied to a *set*. In fact, consider a binary classifier $g^+_{\tau} \in \mathcal{G}^+$ with $\tau \in \Lambda$ and notice that both sets of documents

$$\{i \in W \colon \lambda_i \geq \tau\} \text{ and } \{i \in W \colon \lambda_i < \tau\}$$

can be independently reordered without changing $\mathrm{ACC}(g^+_{\tau})$. Hence, a single binary classifier which achieves some level of accuracy, actually corresponds to many different real-valued classifiers and therefore many different values of the AUC (see e.g. [23]). We will consider the AUC to be more appropriate than the accuracy when measuring performance in ranking tasks.

1.4 Multicriteria Optimization

Consider a *feasible set* X and an *objective function* $f \colon X \mapsto \mathbb{R}$, then

$$\max_{x \in X} f(x)$$

W	χ^{W_T}	λ
-	-	-
1	1	12.0
2	1	10.5
3	0	10.4
4	1	10.1
5	1	9.0
6	1	9.0
7	0	9.0
8	0	5.1
9	1	4.0
10	0	3.7

(a) f^+ / $\mathcal{G}^+$

FPR	TPR
0.0	0.0
0.0	0.16
0.0	0.33
0.25	0.33
0.25	0.5
-	-
-	-
0.5	0.83
0.75	0.83
0.75	1.0
1.0	1.0

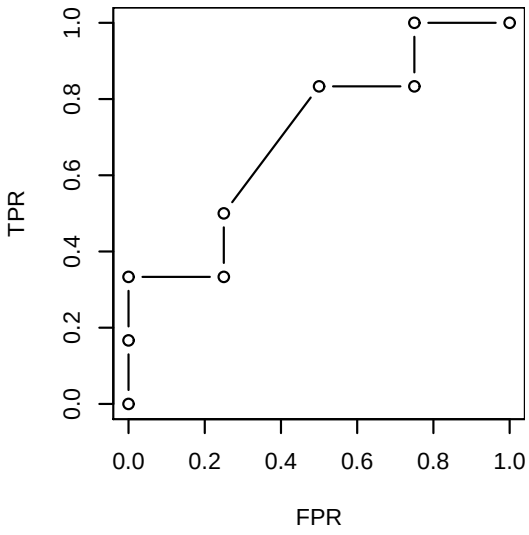
(b) ROC_+

W	χ^{W_T}	λ
-	-	-
10	0	3.7
9	1	4.0
8	0	5.1
7	0	9.0
6	1	9.0
5	1	9.0
4	1	10.1
3	0	10.4
2	1	10.5
1	1	12.0

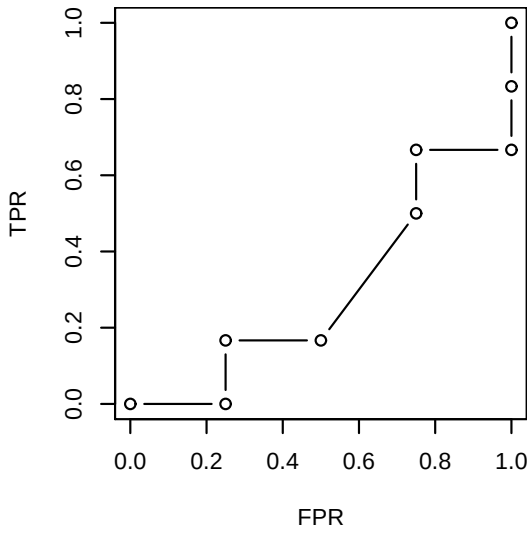
(c) f^- / $\mathcal{G}^-$

FPR	TPR
0.0	0.0
0.25	0.0
0.25	0.16
0.5	0.16
-	-
-	-
0.75	0.5
0.75	0.66
1.0	0.66
1.0	0.83
1.0	1.0

(d) ROC_-

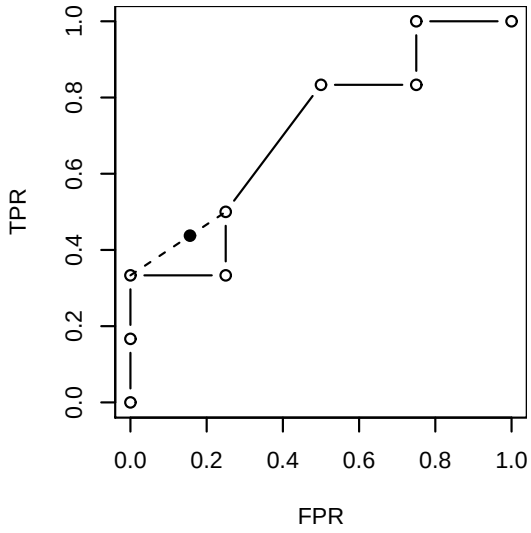


(e) Classifier family $\mathcal{G}^+$ with $AUC_+ = .708$

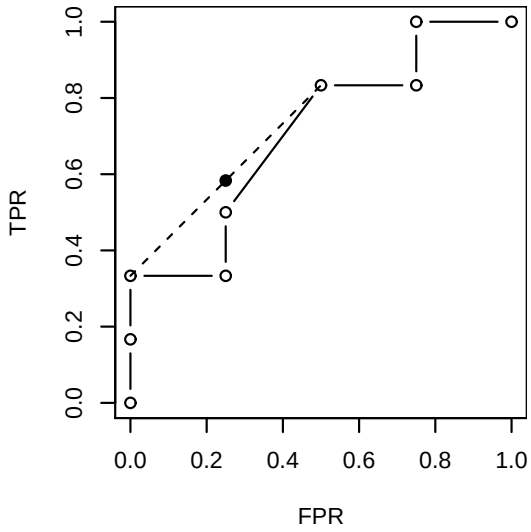


(f) Classifier family $\mathcal{G}^-$ with $AUC_- = .292$

Figure 1.1: ROC Examples



(a) Creating a classifier



(b) Replacing a suboptimal classifier

Figure 1.2: Examples of the random classifier $\bar{g}_{\bar{\tau}}^{+}$

is a *scalar optimization problem.* Since the objective function f is a real-valued function, and the set $\mathbb{R}$ which we refer to as the *objective space* is a total order, an optimal solution to such a problem is a vector $x^* \in X$ such that $f(x^*) \geq f(x)$ for all $x \in X$. Now consider the problem

$$\text{“}\max_{x \in X}\text{”} \ (f_1(x), f_2(x), \cdots, f_p(x)) \tag{1.17}$$

where again X is the *feasible set*, but now the *objective function* f is a vector valued function

$$f \colon X \mapsto \mathbb{R}^p.$$

Such a problem is called a *multicriteria optimization problem.* There is a fair amount of variation in the terminology used in the field of multicriteria optimization. We will closely follow the development used in [27].

One of the defining characteristics of problems such as (1.17) is that the objective functions f_j, for $j = 1, \cdots, p$ are potentially conflicting or competing in that we cannot assume that they are monotonic transformations of each other. If this were not the case, then clearly utilizing only one of them would be sufficient, and we would have a scalar optimization problem. Another defining characteristic is suggested by the fact that we wrote “max” rather than max in (1.17). Since the objective space $\mathbb{R}^p$ is not a total order, we must specify a set and an associated order upon which the problem will be solved before we can describe an optimal solution.

In [27], a framework for classifying multicriteria optimization problems by specifying the

1. *data* – the feasible set, the objective function vector and the objective space,
2. *ordered set* – a set and an ordering relation upon which the optimization problem will be solved and,
3. *model map* – a function that maps each vector in the objective function space to the ordered set

was presented. In this framework, a class of multicriteria optimization problems is denoted as *data/model map/ordered set.* Using this notation we can see that a vector

$x^* \in X$ is an optimal solution to a multicriteria optimization in the class

$$(X, f, \mathbb{R}^p) \,/\, \gamma \,/\, (Y, \preccurlyeq)$$

if there is no $x \in X$ such that $x \neq x^*$ and $\gamma(f(x^*)) \preccurlyeq \gamma(f(x))$ where $(Y, \preccurlyeq)$ is some set and associated order relation and the model map γ is a function $\gamma\colon X \mapsto Y$.

Multicriteria optimization problems can roughly be divided into *scalarization methods* and *non-scalarization methods*. We will discuss two cases of the former, namely the weighted sum method and ϵ-constraint method, and three cases of the later, lexicographic optimality, max-min optimality and lex-max-min optimality. We will see that each class has a different combination of model map and ordered set, thereby introducing a different means of ranking the vectors in the objective space $\mathbb{R}^p$, and as a result, a different notion of optimality. As we describe these classes we do not specify the feasible set X, but will assume that the objective function vector is given by

$$f = (f_1(x), f_2(x), \cdots, f_p(x))$$

where $f_j\colon X \mapsto \mathbb{R}$ for $j = 1, 2, \cdots, p$ and therefore the objective space is $\mathbb{R}^p$.

While the classes we consider do not require it, in our discussion we will assume that the feasible set X consists of finitely many vectors that are explicitly provided as an $m \times p$ matrix in which the rows correspond to the elements of X and the columns correspond to the criteria. Given this assumption, and the fact that each notion of optimality corresponds to a different ordering of the vectors in X, the problems we consider provide us not just with a means of finding the optimal vectors, but of *ranking* the vectors in terms of the associated notion of optimality. We will make use of this fact throughout the sequel. Figure 1.4 contains an example of such a matrix, and includes the results ordered by four of the notions of optimality we consider.

Pareto Optimality. In the following fundamental class of multicriteria optimization problems

$$(X, f, \mathbb{R}^p) \,/\, \mathrm{id} \,/\, (\mathbb{R}^p, \preccurlyeq_E), \tag{1.18}$$

the model map is the the identity function and therefore the ordered set is simply the objective space, $\mathbb{R}^p$. The order relation $\preccurlyeq_E$ is the *componentwise* ordering of the vectors in this space. Specifically, if $x, y \in \mathbb{R}^p$, then $x \preccurlyeq_E y$ when $x_j \leq y_j$ for $j = 1, 2, \cdots, p$ and $x_j < y_j$ for some $j \in \{1, 2, \cdots, p\}$, and similarly, $x \prec_E y$ when $x_j < y_j$ for $j = 1, 2, \cdots, p$.

This ordering relation is the basis of one the most important concepts in multicriteria optimization, namely *Pareto optimality* or what we shall refer to as *efficiency*. A vector $x^* \in X$ is said to be *efficient*, and optimal for (1.18), if there does not exist an $x \in X$ such that $x \neq x^*$ and $f(x^*) \preccurlyeq_E f(x)$. The set of efficient points is therefore the set of maximal points of the componentwise partial order, $\preccurlyeq_E$. If a vector $x^* \in X$ is efficient, then we shall refer to $f(x^*)$ as a *nondominated point*. That is, each nondominated point is a point in the objective space and is the result of applying the objective function to an efficient point in the feasible set. Further, if $x, y \in \mathbb{R}^p$ and $f(x) \preccurlyeq_E f(y)$ then we say that y *dominates* x and $f(y)$ *dominates* $f(x)$. Also, we shall say that a point $x^* \in X$ is *weakly efficient* if there does not exist an $x \in X$ such that $x \neq x^*$ and $f(x^*) \prec_E f(x)$. The corresponding point $f(x^*)$ in the objective space will be called *weakly nondominated*. Figure 1.4 shows examples of efficient and points and dominance relationships.

Even though the relation $\preccurlyeq_E$ is a partial, not a total order, and therefore a problem may have multiple efficient and hence nondominated points, efficiency and dominance do provide a means of ordering vectors in $\mathbb{R}^p$. Intuitively, an efficient point should be preferred to a non-efficient point since the former is better with respect to at least one of the criteria, and no worse with respect to the other criteria than the later. Viewed another way, an efficient point should be preferred to a non-efficient point since while selection of the later may result in an increase with respect to one of the criteria, it will definitely result in a decrease with respect to at least one of the other criteria.

Weighted Sum Method. While some algorithms do seek efficient solutions by operating directly in the objective space (see e.g. Algorithm 2.1 in [27]), a large set of algorithms choose to map each point in the objective space $\mathbb{R}^p$ to the ordered set $\mathbb{R}$ and to pursue solutions there. The most prevalent of such algorithms is referred as the

Data			
i	f_1	f_2	f_3
1	1	1	2
2	7	2	6
3	6	5	7
4	1	1	9

Weighted			
i	f_1	f_2	f_3
3	6	5	7
2	7	2	6
4	1	1	9
1	1	1	2

Lex			
i	f_1	f_2	f_3
2	$\underline{7}$	2	6
3	$\underline{6}$	5	7
4	1	1	$\underline{9}$
1	1	1	$\underline{2}$

Max-Min			
i	f_1	f_2	f_3
3	6	$\underline{5}$	7
2	7	$\underline{2}$	6
1	$\underline{1}$	1	2
4	$\underline{1}$	1	9

Lex-Max-Min			
i	1W	2W	3W
3	5	6	7
2	2	6	7
4	1	1	9
1	1	1	2

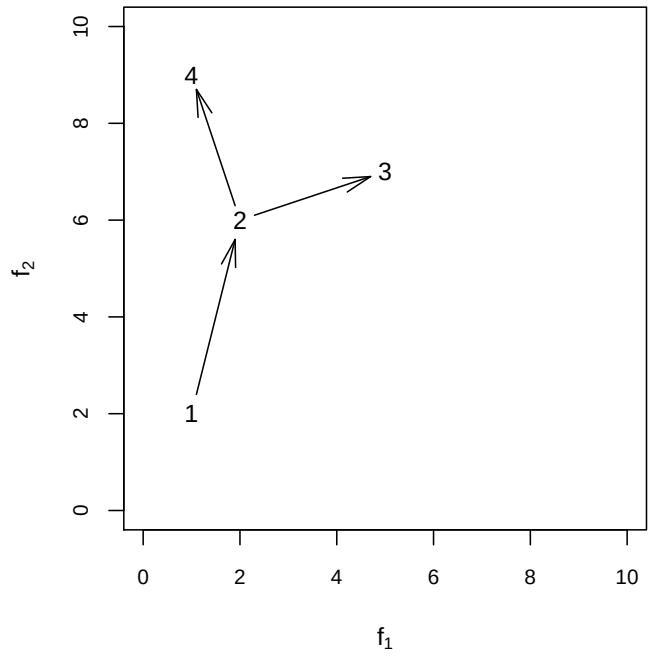


Figure 1.3: Points 3 and 4 are efficient points. The arrows show the dominance partial order, with point 1 being dominated by point 2, which in turn is dominated by points 3 and 4. In the lex-max-min example 1W, 2W, and 3W represent the "1^{rst} worst", "2^{nd} worst", and "3^{rd} worst" criteria.

weighted sum method. It corresponds to the following class of multicriteria optimization problems

$$(X, f, \mathbb{R}^p) \,/\, \langle \lambda, \cdot \rangle \,/\, (\mathbb{R}, \leq). \tag{1.19}$$

The weighted sum method transforms the multicriteria objective function

$$f = (f_1(x), f_2(x), \cdots, f_p(x))$$

into a single criteria objective function

$$g = \sum_{j=1}^{p} \lambda_j f_j(x)$$

and then proceeds to solve the problem

$$\max_{x \in X} g(x).$$

A vector $x^* \in X$ is optimal for (1.19) if there does not exist an $x \in X$ such that $x \neq x^*$ and

$$\sum_{j=1}^{p} \lambda_j f_j(x^*) \leq \sum_{j=1}^{p} \lambda_j f_j(x).$$

It can be shown that if X and f_j for $j = 1, 2, \cdots, p$ are convex, this formulation can be used to find weakly efficient points, another variant of efficiency known as *properly efficient points* and in some cases even efficient points. (See e.g. [27], Theorem 4.1). The example in Figure 1.4 utilizes the average weighting vector $\lambda = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Another method known as the ϵ-constraint method involves optimizing only one of the objective functions $f_{j'}$ at a time, for $j' \in \{1, 2, \cdots, p\}$, while the other objective functions are converted into constraints of the form

$$f_j(x) \geq \epsilon_j$$

for $j = 1, 2, \cdots, p$ with $j \neq j'$ and used along with any constraints used in the definition of X. This method can be used for non-convex problems.

Lexicographic Optimality. In the following class of multicriteria optimization problems

$$(X, f, \mathbb{R}^p) \,/\, \mathrm{id} \,/\, (\mathbb{R}^p, \preccurlyeq_{\mathrm{L}}), \tag{1.20}$$

the model map is the identity function and the ordered set uses the lexicographic ordering to rank vectors in $\mathbb{R}^p$. Recall that if $x, y \in X$ and there exists an index $j' \in \{1, 2, \cdots, p\}$ such that $x_j = y_j$ for $j < j'$ and $x_{j'} < y_{j'}$, then y is said to be lexicographically larger than x which we denote as $x \preccurlyeq_{\mathrm{L}} y$.

A vector $x^* \in X$ is said to be *lexicographically optimal*, and optimal for (1.20), if there does not exist an $x \in X$ such that $x \neq x^*$ and $x^* \preccurlyeq_{\mathrm{L}} x$. Since the lexicographic ordering is a total order, we can provide an affirmative definition of optimality and equivalently say that a vector $x^* \in X$ to be *lexicographically optimal*, and optimal for (1.20), if $x \preccurlyeq_{\mathrm{L}} x^*$ for each $x \in X$. The algorithm to find the optimal solutions to such problems simply places the vectors in X in lexicographic order.

Use of lexicographic optimality is appropriate in applications where the criteria adhere to a known ordering of "importance". It ensures that the optimal solution will have the maximum value for the most important criteria. The example shown in Figure 1.4 assumes that $j = 1$ is the most important criteria, $j = 2$ is the next most important and that $j = 3$ is the least important criteria. Notice that a large value for f_1 means that f_2 and f_3 will not be considered, even when as in the example, $f_2(x_2) > f_2(x_1)$ and $f_3(x_2) > f_3(x_1)$. It can be shown that the optimal solutions to (1.20) are efficient (see e.g. [27], Lemma 5.2).

Max-Min Optimality. In the following class of multicriteria optimization problems

$$(X, f, \mathbb{R}^p) \,/\, \max \,/\, (\mathbb{R}, \leq), \tag{1.21}$$

the model map is the max function, and since it maps the objective space to $\mathbb{R}$, vectors in the ordered set are ranked via the canonical total order on $\mathbb{R}$. If $x, y \in X$ then we say that $x \preccurlyeq_{\mathrm{MM}} y$ if

$$\max\big\{f_1(x), f_2(x), \cdots, f_p(x)\big\} \leq \max\big\{f_1(y), f_2(y), \cdots, f_p(y)\big\}.$$

A vector $x^* \in X$ is said to be *max-min optimal*, and optimal for (1.21), if there does not exist an $x \in X$ such that $x \neq x^*$ and $x^* \preccurlyeq_{\text{MM}} x$.

Problems in this class are frequently formulated as

$$\max_{x \in X} \min_{j=1,2,\cdots,p} f_j(x)$$

and the algorithm to find optimal solutions to this problem places the elements of X in decreasing order of their smallest (i.e. worst) $f_j(x)$ for $j = 1, 2, \cdots, p$. The smallest (i.e worst) $f_j(x)$ in the optimal solution will be the largest (i.e. best) among all the smallest (i.e. worst) $f_j(x)$ for $j = 1, 2, \cdots, p$. If the smallest elements of two vectors are equal, then they are ranked by the vectors' indices. In the example shown in Figure 1.4, this rule is used to used to break the tie between x_1 and x_4. It can be shown that optimal solutions to (1.21) are weakly efficient.

Lex-Max-Min Optimality. If $x \in \mathbb{R}^p$, then the function $\text{sort}_\uparrow(x)$ permutes the indices of x so its components are sorted into ascending order. Now, if $x, y \in \mathbb{R}^p$, then we say that $x \preccurlyeq_{\text{LMM}} y$ if

$$\text{sort}_\uparrow(x) \preccurlyeq_{\text{L}} \text{sort}_\uparrow(y).$$

Now, consider the following class of multicriteria optimization problems

$$(X, f, \mathbb{R}^p) \,/\, \text{sort}_\uparrow \,/\, (\mathbb{R}^p, \preccurlyeq_{\text{L}}). \tag{1.22}$$

We say that a vector $x^* \in X$ is *lex-max-min optimal*, and optimal for (1.22) if there does not exist an $x \in X$ such that $x \neq x^*$ and $x^* \preccurlyeq_{\text{LMM}} x$. As was the case for lexicographical optimality, since the lexicographic order is total, we can also provide an affirmative definition of optimality and equivalently say that a vector $x^* \in X$ is said to be *lex-max-min optimal*, and optimal for (1.22), if $x \preccurlyeq_{\text{LMM}} x^*$ for each $x \in X$. Obviously this hybrid definition of optimality draws from both (1.20) and (1.21).

Following (1.22), algorithms to find the optimal solutions to this problem first, per the model map, sort the components of each vector $x \in X$ into ascending order and then, as dictated by the order set, place the resulting permuted vectors in decreasing

lexicographic order. Arguably the most unique characteristic of this algorithm is the permutation of the components of each vector $x \in X$. As a result of this permutation, the step of lexicographic ordering of vectors $x, y \in X$ may involve comparisons of $f_i(x)$ and $f_j(y)$ for $i, j \in \{1, 2, \cdots, p\}$ and $i \neq j$. Therefore, lex-max-min optimality can be considered an *equitable* notion of optimality since it considers each criterion to be of equal importance (see e.g. [53], [66], and [27], p.147). For this reason it is appropriate for use in applications in which no information about the relative importance of the criteria is available and as such in [8] the lex-max-min criteria is referred to as a means of making "optimal decisions under complete ignorance". It is also interesting to note that it has been shown that lex-max-min optimization problems can also be solved via linear programming (see e.g. [53] and [66]). It can be shown that optimal solutions to (1.22) are efficient.

To clarify some of these points consider an efficient ranking based on lex-max-min that is based on 3 criteria, and that for each of these criteria, larger values are better than smaller ones. Let us assume that this ranking includes two features $i, j \in V$ with i is ranked higher than j. We can be sure that the value of the smallest i.e. worst criteria for i is larger than or equal to the value of the smallest i.e. worst criteria for j, and if they are equal we can be sure that the value of the second smallest i.e. second worst criteria for i is larger than or equal to the value of the second smallest i.e. second worst criteria for j, and if they are equal we can be sure that the value of the third smallest i.e. third worst criteria for i is larger than or equal to the value of the third smallest i.e. third worst criteria for j. Note, however, in this example if the value of the worst criteria of feature i is strictly greater than the value of the worst criteria of feature j that we do not know whether the value of the second worst criteria for feature i is greater than, less than, or equal to the value of the second worst criteria for feature j, nor do we have any such information regarding the third worst criteria. For example, consider the case where feature i is $(10, 11, 14)$ and feature j is $(9, 12, 13)$.

1.5 Overview

In §2 we introduce the data set on which we will run various feature selection experiments. The data set includes fifty different topics, and we present basic collection statistics such as the number of relevant and irrelevant documents as well as the number of features for each topic. In addition, we describe the steps taken to prepare the original data for use in our experiments. We continue by presenting the criteria we use to assess the performance of feature selection algorithms, and explaining the methodology we adopt to rank the performance of different feature selection algorithms. In §3 we present several models that provide the basis for many Boolean feature ranking functions. We also introduce thirty two such functions and we will identify various properties of these functions as well as several relationships between these functions. We conclude the chapter with a brief discussion of the results of experiments which assess the performance of these functions when used in a ranking algorithm. In §4 we take a more detailed look at the result of the experiments that were discussed in §3, as well as additional experiments that were designed to assess the relationship between the separation achieved by a selected feature set and the amount of noise in the set. We also develop formal models of noise, offer a principle that can be used to guide feature selection algorithms, and provide some insight into why some Boolean feature ranking functions work well while others select a substantial amount of noise. In §5 we present a set axioms that we believe all Boolean feature ranking functions should satisfy and use the tools of linear programming and convex analysis to identify characteristics of a special subset of the functions that satisfy these axioms. In §6 present a natural extension of the axioms for Boolean feature ranking functions to real-valued feature ranking functions. In addition, we consider the results of experiments that study the performance of greedy feature ranking algorithms, those based on the real-valued feature model, as well as that utilize tools from multicriteria optimization.

Chapter 2

Feature Set Evaluation

We begin this chapter by describing data set that we will used in all empirical studies of feature selection algorithms. Next we present the methodology used to evaluate feature selection algorithms and the feature sets they select. We then describe the set of criteria upon which these feature sets will be evaluated. We conclude by showing how these criteria can be used to order feature selection algorithms by the degree to which they satisfy these criteria.

2.1 The Collection

The data used in this dissertation is the Reuters-21578 text categorization test collection. The collection and supporting documentation are publicly available in SGML format (see e.g. [55]) and in XML format (see e.g. [59]).

The 21,578 documents in the collection are Reuters newswire items that appeared in 1987. The documents include a variety of tags including tags that identify various parts of the document, tags that indicate whether a document belongs to any of the several subsets of documents that researchers have used in past studies, tags that indicate which topics or other categories a human indexer assigned to a document and tags related to various artifacts of the indexing process. The collection has been indexed on 135 economic topics, 120 of which have at least one relevant document and 57 of which have 20 or more relevant documents.

We shall run experiments on 50 topics, with each topic defining a different feature selection problem. The method of selecting the topics will be described below. For each of these topics the available set of relevant documents will obviously be the documents

the human indexers judged relevant for the topic. To identify a set of irrelevant documents we capitalize on the fact that the README_LEWIS.txt file distributed with the collection states that "a reasonable (though not certain) assumption is that for all TOPICS="YES" stories the indexer at least thought about whether the story belonged to a valid TOPICS category. Thus, the TOPICS="YES" stories with no topics can reasonably be considered negative examples for all 135 valid TOPICS categories." The available set of irrelevant documents for all topics will consist of the 1975 documents in this set, that is, $|F| = 1975$ for all topics.

2.2 Data Set Reduction

The size of the complete data set was too large to complete the empirical studies we wished to perform in a reasonable period of time. For each topic we therefore sought to create a reduced size data set that preserved the salient characteristics of the original data set.

We noticed that the computationally challenging aspect of our studies was the calculation of the measures of separation introduced in §2.7, since they require the calculations of the Hamming distance between each $u \in T$ and $v \in F$, and the product $|T||F|$ in the original data set was quite large for many of the topics.

We therefore decided to construct a subset of the original data set in which we limit the size of the product $|T||F|$ to be less than or equal to some $p > 0$ for each topic, while maintaining the ratio of $|T|$ to $|F|$ for the topic. The new reduced size sets will be denoted as T^* and F^* and the set of features appearing in T^* will be denoted as V^*. In the following, clearly $|T^*|$ $|F^*|$ does not necessarily equal p due to the constraint of maintaining the ratio of $|T|$ to $|F|$ and $|T^*|$ $|F^*|$ being integral.

So we therefore wish to find $|T^*| \in \mathbb{Z}^+$, $|F^*| \in \mathbb{Z}^+$ such that $|T^*||F^*| \leq p$ for fixed $p \in \mathbb{Z}^+$, while preserving the ratio of relevant to irrelevant documents, $\frac{|T^*|}{|F^*|} \approx \frac{|T|}{|F|}$. The constant p was selected so that the resulting computation time for our experiments was manageable, while ensuring that the sets T^* and F^* included a sufficient number of

data points as to be interesting. To achieve this $|T^*|$ and $|F^*|$ were selected using

$$|T^*| \leq \min\left\{|T|, \left\lfloor\sqrt{p\frac{|T|}{|F|}}\right\rfloor\right\} \text{ and } |F^*| = \left\lfloor\frac{|F|}{|T|}|T^*|\right\rfloor$$

with $p = 100,000$.

As will be discussed in §2.3, for each topic we construct T^* and F^* by randomly selecting $|T^*|$ documents from T and $|F^*|$ from F.

2.3 Preparation

All data preparation and experiments are implemented in the R programming language(see e.g. [67]) using its comprehensive collection of packages. We used R's tm (see e.g. [38] and [37]) package to

- Import the Reuters-21578 collection stored as XML into R.
- Perform certain preprocessing of the text including whitespace removal, transformation to lower case, punctuation removal and removal of numbers.
- Identify the set of relevant documents for each topic and the collective set of irrelevant documents.
- Transform the data into a document term matrix representation of the collection.
- Remove rows corresponding to documents, in which there is no text within the `<BODY>` tag.[1]

After these steps were completed, for each topic, $|T^*|$ from $|T|$ and $|F^*|$ rows from $|F|$ are randomly selected from the aforementioned collection document term matrix to create the document term matrices T^* and F^*. At this point any all 0 columns in the matrices T^* and F^* for the given topic, that is, columns corresponding to features that do not appear in any documents in T^* and F^* are eliminated. The document

[1]Text in Reuters-21578 documents is organized with four optional tags `AUTHOR`, `DATELINE`, `TITLE` and `BODY`. Per the README_LEWIS.txt file distributed with the collection, the `BODY` tag contains the "main text of the story". There are 2535 documents in which there is no text within the `BODY` tag and their removal reduces the total number of documents in the collection from 21578 to 19043.

term matrices, T^* and F^* are stored and reused throughout the dissertation either in R's standard matrix data structure or using one of the sparse matrix data structures available in R's Matrix package (see e.g. [6]).

2.4 Fold Construction for Cross Validation

Unless otherwise noted, all of the experiments make use of 10-fold cross validation. For each topic we created sets of relevant and irrelevant document *folds* which we shall denote as T^i, F^i for $1 \leq i \leq 10$ respectively. These folds are used in all experiments that use cross validation. The procedure for constructing these folds is as follows

FOLD CONSTRUCTION

Input: T^* and F^*.

Initialize: Let w_T and w_F respectively be vectors that contain the elements of W_{T^*} and W_{F^*} randomly shuffled.

Step 1: Set $n_T := \lfloor \frac{1}{10}|T^*| \rfloor$, $n_F := \lfloor \frac{1}{10}|F^*| \rfloor$ and $i := 1$.

Step 2: Set $T^i_{test} := \{x^T_j \mid n_T(i-1)+1 \leq j \leq n_T(i+1)\}$ and $T^i_{train} := T^* \setminus T^i_{test}$

Step 3: Set $F^i_{test} := \{x^F_j \mid n_F(i-1)+1 \leq j \leq n_F(i+1)\}$ and $F^i_{train} := F^* \setminus F^i_{test}$

Step 4: Any columns which are all 0 in T^i_{train} and F^i_{train} are deleted from T^i_{train}, F^i_{train}, T^i_{test} and F^i_{test}. That is, columns corresponding to features that do not appear in any training documents for the i^{th} fold are eliminated from all documents in this fold. Additionally, in specified experiments terms that appear in Cornell's SMART list of 571 stopwords [57] are removed.

Step 5: Set $U^i := \left(T^i_{train}, F^i_{train}, T^i_{test}, F^i_{test}\right)$

Step 6: If $i < 10$ then set $i := i + 1$ and goto **Step 2**, otherwise goto **Output**.

Output: U^i for $1 \leq i \leq 10$.

The fact that the ratio of relevant to irrelevant documents is approximately the same in each $T^i_{train} \cup F^i_{train}$, and in each $T^i_{test} \cup F^i_{test}$, as in the original data set, is referred to as *stratified* 10-fold validation (see e.g. [22]).

2.5 Topics

The first step in selecting the topics to be used in this dissertation was to rank them in decreasing order of $|T^*|$, that is in decreasing order of the number of relevant documents in the reduced data set. Next, the two most highly ranked topics were eliminated since after the data reduction process they had more relevant documents than irrelevant documents which is not characteristic of text classification problems. Since we use 10-fold cross-validation and we were are interested in calculating our separation measures between T and F as well as between T and T, we then decided to only utilize topics that had at least 25 relevant documents after the data reduction process so that each test fold would have at least 5 relevant test documents. Inspection of the remaining topics showed that only the top 52 topics had 25 or more relevant documents and so as to have a round number, we decided to select the top 50 topics.

These topics as well as the number of relevant and irrelevant documents and the number of features after the processing described in §2.3 are shown in Table 2.1. Note that the actual number of features in each fold after Step 4 in the fold construction process described in §2.4 will obviously be less than the total number in T^*.

Topic	$\lvert T^*\rvert$	$\lvert F^*\rvert$	$\lvert V^*\rvert$
money-fx	186	536	7983
grain	170	584	8371
crude	169	589	8793
trade	161	617	8827
interest	146	680	8245
ship	122	816	9570
wheat	120	825	9031
corn	106	938	9779

Topic	$\lvert T^* \rvert$	$\lvert F^* \rvert$	$\lvert V^* \rvert$
oilseed	95	1030	9897
sugar	94	1060	9952
dlr	92	1081	9775
gnp	88	1135	10226
coffee	85	1173	10238
gold	82	1217	10403
veg-oil	82	1190	10216
money-supply	79	1238	9958
nat-gas	79	1238	10592
livestock	75	1322	11075
soybean	74	1316	10958
bop	71	1388	10930
cpi	71	1388	10812
copper	62	1590	11542
carcass	61	1606	11922
reserves	60	1623	11404
cocoa	58	1684	11906
jobs	58	1684	11753
rice	58	1709	12026
iron-steel	57	1731	12049
cotton	56	1783	12282
alum	54	1838	12167
yen	54	1838	12206
ipi	53	1836	12038
gas	52	1867	12411
meal-feed	50	1975	12672
rubber	49	1975	12663
barley	48	1975	12584

Topic	$\lvert T^* \rvert$	$\lvert F^* \rvert$	$\lvert V^* \rvert$
zinc	43	1975	12538
palm-oil	42	1975	12616
pet-chem	41	1975	12688
silver	36	1975	12540
lead	35	1975	12514
rapeseed	35	1975	12485
sorghum	34	1975	12628
tin	33	1975	12597
strategic-metal	32	1975	12506
wpi	29	1974	12391
fuel	28	1975	12501
hog	26	1975	12469
soy-meal	26	1975	12548
heat	25	1975	12454

Table 2.1: Topics

After experimentation had begun it was noticed that in topic "money-fx", following Booleanization, the relevant document with NEWID= "13530" and the irrelevant document with NEWID="13416" were were identical. Similarly, it was noticed that in topic "trade", following Booleanization, the relevant document with NEWID="5973" and the irrelevant document with NEWID="06099" were identical. So that the assumption that $T \cap F = \varnothing$ would be satisfied we removed the irrelevant document with NEWID="13416" from the "money-fx" topic resulting in a reduction in the number of irrelevant documents from 537 to 536. Similarly, the irrelevant document with NEWID="06099" from the "trade" topic was removed resulting in a reduction in the number of irrelevant documents from 618 to 617.

2.6 Methodology

Evaluation of a feature selection algorithm and the set of features S it selects begins by using the algorithm to select a set of K features. The set of criteria, $\mathcal{C}$ used for evaluation includes criteria that are direct measures of the

- the *separation* between the projected sets $T[S]$ and $F[S]$,
- the amount of *noise* in S, and
- the *robustness* of the algorithm when presented with small variations in T and F,

as well as criteria that are indirect measures of

- the *quality* of the features in S, and
- the *size* of S.

In the sections that follow, we will discuss these criteria and the measures on which they are based; for now we simply mention that some of the criteria are based on measures that are known to be monotonically non-decreasing with $k = 1, \ldots, K$, while other criteria are based on measures for which no such statement can be made.

Since each of the algorithms we study return an ordered set of features, we can view the selected features either as the set S or as the sequence (s_k). For monotonically non-decreasing measures we will adopt the former view and will compute the value of the measure for each subset $S_k = \bigcup_{i=1}^{k}(s_i)$ with $s_i \in S$ and will use these values to compute the *area under the curve* for the graph of each measure as a function of k and will use this area as the criteria. There are several advantages to utilizing the area under the curve as the criteria with such measures. Since these criteria reward algorithms that select superior features early in the sequence (s_k) and penalize algorithms that instead select inferior features early in the sequence (s_k), they allow us to study how a measure varies as k increases. In addition, since they provide a single number that captures the values for a measure over the range of all values of k, the comparison of algorithms is greatly simplified. For measures that are not monotonically non-decreasing with k we will view the selected features as a set, will compute the value of the measure at each

element of S, and will then use the average of these values as the criteria. We will see in §2.11 that in both cases our approach can be naturally extended to assess the impact of noise on a feature set.

2.7 Separation Measures

If $u, v \in \mathbb{B}^n$ then we shall let

$$u \triangle v = \{j \in S \colon u_j \neq v_j\}$$

denote the set containing the components where the vectors u and v differ. The number of components in which two vectors differ is known as the *Hamming distance* between them. If $y = \chi^S$, then

$$d_y(u, v) = d(u[S], v[S]) = \sum_{j \in u \triangle v} y_j$$

can be seen to be the *Hamming distance* between the vectors $u, v \in \mathbb{B}^V$ projected onto the set S. Clearly, for $u[S], v[S] \in \mathbb{B}^S$ we have $0 \leq d_y(u, v) \leq |S|$. We say that a pair of vectors $u[S], v[S] \in \mathbb{B}^S$ is *separated* if $u[S] \neq v[S]$, that is, if there exists a $j \in S$ such that $j \in u \triangle v$, or equivalently if $d_y(u, v) > 0$ and say they are well separated when $d_y(u, v) > 0$ is in some sense large.

A set $S \subseteq V$ is said to be a *support set* for (T, F) if it has the property that $T[S] \cap F[S] = \varnothing$. That is, S is a support set for (T, F), if the projection of each $u \in T$ onto S is separated from the projection of each $v \in F$ onto S. Clearly, selecting a feature set which is a support set is desirable since by definition the original assumption that $T \cap F = \varnothing$ is preserved in the projected space, however, finding minimal size support sets is computational challenging. In [24] it was shown that the problem of finding minimal size support sets can be formulated as the NP-complete set-covering problem

(SCP)

$$\begin{aligned} \text{minimize} \quad & \sum_{j \in V} y_j \\ \text{subject to} \quad & \sum_{j \in u \triangle v} y_j \geq 1 \text{ , for all } u \in T, v \in F, \\ & y_j \in \mathbb{B} \text{ , for all } j \in V, \end{aligned}$$

and therefore the problem of finding minimal size support sets is NP-hard. Each constraint in this problem corresponds to a pair of $u \in T$ and $v \in F$ and requires that the pair be separated in the resulting projection. Clearly, $S \subseteq V$ is a support set for (T, F) if and only if $d(u[S], v[S]) > 0$ for each $u \in T$ and $v \in F$.

Letting $S \subseteq V$, we now consider four measures of the separation between the projected *sets* $T[S]$ and $F[S]$. The first three of these measures are based directly on the Hamming distance and are given as

$$\rho(y) = \min_{u \in T, v \in F} d_y(u, v) \tag{2.1}$$

$$\sigma(y) = \frac{1}{|T||F|} \sum_{u \in T, v \in F} d_y(u, v) \tag{2.2}$$

$$\theta(y) = \sum_{u \in T, v \in F} \min\{d_y(u, v), 1\} \tag{2.3}$$

where $y = \chi^S$ for some $S \subseteq V$. These measures respectively represent the *minimum* Hamming distance between the elements in $T[S]$ and those in $F[S]$, the mean or *average* Hamming distance between the elements in $T[S]$ and those in $F[S]$, and the *number of pairs* in $T[S]$ and $F[S]$ that are *separated*. Clearly, S is a support set for (T, F) if and only if $\rho(y) > 0$. We now mention that while we would like to measure the number of features each algorithm needs to achieve a support set, the fact that an algorithm may not achieve a support set in K or less features, prevents us from doing so. However, it can be seen that S is a support set for (T, F) if and only if $\theta(y) = |T||F|$, and therefore θ provides an indication of how close S is to being a support set, which we will consider to be a good alternative measure.

In order to define the fifth measure we now introduce the vector

$$\mathbf{h}(y) = [h_o(y), h_1(y), \cdots, h_n(y)]$$

where $y = \chi^S$ for some $S \subseteq V$ and

$$h_k(y) = \Big| \{(u,v) \in T \times F \colon d_y(u,v) = k\} \Big|$$

for $k = 0, 1, \ldots, n$. The components $h_k(y)$ are the number of pairs in $T[S] \times F[S]$ that are exactly at Hamming distance k with $h_k(y) = 0$ for any $k > |S|$. Note that it can be shown that $h_0(y) + h_1(y) + \cdots + h_n(y) = |T||F|$.

Let $S, S' \subseteq V$ and $y = \chi^S$, $y' = \chi^{S'}$, then we write

$$\mathbf{h}(y) \preccurlyeq_{\mathrm{L}} \mathbf{h}(y')$$

and say that $\mathbf{h}(y)$ is *lexicographically smaller* than $\mathbf{h}(y')$ if there exists an integer i with $0 \leq i \leq n$ such that $h_k(y) = h_k(y')$ for $k < i$ and $h_i(y) < h_i(y')$. If $\mathbf{h}(y) \preccurlyeq_{\mathrm{L}} \mathbf{h}(y')$, then the set S will be said to *lexicographically separate* T and F better than the set S'. The motivation for this definition is that the vector $\mathbf{h}(y)$ has more of the pairs in $T \times F$ distributed in the components $h_k(y)$ for $k > i$ which correspond to larger Hamming distances, than the vector $\mathbf{h}(y')$. The condition $\mathbf{h}(y) \preccurlyeq_{\mathrm{L}} \mathbf{h}(y')$ is equivalent to

$$\sum_{k=0}^{n} h_k(y)\, \alpha^k < \sum_{k=0}^{n} h_k(y')\, \alpha^k$$

or

$$\sum_{k=0}^{n} h_k(y)\, (1 - \alpha^k) > \sum_{k=0}^{n} h_k(y')\, (1 - \alpha^k)$$

if $\alpha \geq 0$ is small enough. A result similar to the following was given in [13].

Proposition 2.1. *Suppose* $\alpha \in \big[0, 1/|T||F|\big]$ *then,*

$$\mathbf{h}(y) \preccurlyeq_{\mathrm{L}} \mathbf{h}(y') \text{ if and only if } \sum_{k=0}^{n} h_k(y)\, \alpha^k < \sum_{k=0}^{n} h_k(y')\, \alpha^k.$$

Therefore, the fifth measure of separation between projections of the set T and F onto some set of features $S \subseteq V$ is the *lexicographic Hamming distance*

$$\phi_\alpha(y) = \sum_{k=0}^{n} h_k(y)(1-\alpha^k) = \sum_{u \in T, v \in F} \left(1 - \alpha^{d_y(u,v)}\right) \tag{2.4}$$

where $y = \chi^S$ and $\alpha \geq 0$. The measure ϕ_α is a weighted sum in which larger weights $(1-\alpha^k)$ correspond to larger Hamming distances $k = d_y(u,v)$ between elements of $u \in T$ and $v \in F$ in the projected space. Note that using $\alpha < 1/|T|F|$ when $|T||F|$ is large can cause the computation of ϕ_α to encounter numerical difficulties. Specifically, as k increases, α^k quickly approaches 0 and as a result, most terms in (2.4) receive a weight of 1. However, since using larger values of α avoids this issue while preserving ϕ_α's valuable property of assigning larger weights to larger Hamming distances, we use $\alpha = 0.99$ when using ϕ_α in the evaluation of feature sets and try several different values of α in experiments described in §6.3 where ϕ_α is used as an objective function. The functions in (2.1), (2.2), (2.3) and (2.4), as well as weighted variants of (2.3) and (2.4) were studied in [13], where they were used as objective functions in the problem of finding minimal size support sets that are well separated.

2.8 Noise Measures

We are interested in the number of features selected by a feature selection algorithm that are noise features. Unfortunately, there is no agreed upon definition of a noise feature or set of features that are considered to be noise. In order to count the number of noise features selected in the empirical studies that we conduct, we therefore consider any feature that is in the set of stopwords, as specified in Cornell's SMART list of 571 stopwords (see e.g. [57]) to be a noise feature, and all other features to be non-noise features. The complete list of these stopwords is provided in Appendix G. We will denote the set of stopwords as $J \subset V$, and the number of stopwords in the set of selected features S will be denoted as ν. The obvious shortcomings of this approach are that the set of selected features may contain features that might reasonably be considered noise features, that and for some topics, some stopwords may actually not

be noise features.

In an effort to understand the how noise contributes to σ for a feature selection algorithm f, we use the algorithm NOISE GROWTH RATE to compute the growth rate of σ and σ_K, which we denote as $\dot{\sigma}[J]$ and $\sigma_K[J]$, when f is presented with only noise features. At each of 50 iterations we run f on a different randomly generated variant of (T, F) which has been projected onto the set of stopwords J, and for $k = 1, \ldots, K$ compute the value of σ_k for the first k selected features. For $k = 1, \ldots, K$, we then compute the median of the 50 values of σ_k, and fit a regression line to the resulting K values. Examples of the noise growth data and the associated regression line for a single topic are shown in Figure K.1 and Figure O.1.

NOISE GROWTH RATE

Input: (T, F) and feature selection algorithm f.

Initialize: Set $i := 1$ and let G be a $50 \times K$ matrix.

Step 1: Set $\gamma :=$ a random number in $[0.25, 1]$.

Step 2: Set $n_T := \max\{\lfloor \gamma |T| \rfloor, 2\}$ and $n_F := \max\{\lfloor \gamma |F| \rfloor, 2\}$.

Step 3: Set $W'_T :=$ n_T randomly selected documents from W_T and $W'_F :=$ n_F randomly selected documents from W_F.

Step 5: Set $(s_k) := \omega(J, K, f, \cdot)$ using $(T[W'_T], F[W'_F])$.

Step 6: Set $g_{i,k} := \sigma(s_k)$ for $k = 1, \ldots, K$, where $s_k \in V$.

Step 7: If $i := 50$ then goto **Output**, otherwise set $i := i + 1$ and goto **Step 1**.

Output: Compute the median of each column in G, fit a regression line $\sigma_k[J] = \sigma_0[J] + \dot{\sigma}_k[J]\, k$ to the K median values, use the regression line to compute $\sigma_K[J]$, and output the intercept $\sigma_0[J]$, the slope $\dot{\sigma}_k[J]$, and $\sigma_K[J]$ for this topic.

2.9 Semantic Measures

Inherent in our discussion so far is the tenet that a set of selected features S is in some sense "good", if it does not include any noise features and if $T[S]$ and $F[S]$ are well separated. While we certainly do subscribe to this tenet, one might argue that it is possible for S to satisfy these criteria, but for the individual features it contains to have little or no semantic relationship to the topic under consideration. While there is no agreed upon method of assessing the semantic relationship that a set of features has to a topic, we will assume that the performance of a simple classifier using only the features in S provides one reasonable measure. Specifically, we shall consider the F measure, which will be discussed in §3.5, of a Bernoulli Naive Bayes classifier using on the K features in S as a measure of the semantic quality of S (see e.g. [60], [62]), [56] and [31]. The Bernoulli Naive Bayes classifier, as opposed to to other classifiers such as the Multinomial Naive Bayes classifier, can reasonably be considered an appropriate choice for this task since it known to be very sensitive to the set features presented to it (see e.g. p.243, [60]). We shall denote the F measure as φ.

A naive Bayes classifier $g^{\beta}\colon T \cup F \mapsto \{0, 1\}$ predicts the class of a document $u \in T \cup F$ based on the following rule

$$g^{\beta}(u) = \begin{cases} 1 & \text{if } \mathrm{P}(u \in T \mid u) > \mathrm{P}(u \in F \mid u), \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Thus, it predicts the document to be relevant when, given that it contains a certain set of features, the probability that it is relevant is larger than the probability that it is irrelevant. Estimates for $\mathrm{P}(u \in T \mid u)$ and $\mathrm{P}(u \in F \mid u)$ are based on Bayes rule and are given as

$$\mathrm{P}(u \in T \mid u) = \log(\mathrm{P}(u \in T)) + \sum_{j \in S} \log(\mathrm{P}(u_j = 1 \mid u \in T))u_j + (1 - \log(\mathrm{P}(u_j = 1 \mid u \in T)))(1 - u_j)$$

and

$$\mathrm{P}(u \in F \mid u) = \log(\mathrm{P}(u \in F)) + \sum_{j \in S} \log(\mathrm{P}(u_j = 1 \mid u \in F))u_j + \\ (1 - \log(\mathrm{P}(u_j = 1 \mid u \in F)))(1 - u_j)$$

which can be rewritten as

$$\begin{aligned} \mathrm{P}(u \in T \mid u) &= \log(\mathrm{P}(u \in T)) + \sum_{j \in S} \log(\mathrm{P}(u_j = 1 \mid u \in T))u_j + \\ &\qquad \log(\mathrm{P}(u_j = 0 \mid u \in T))(1 - u_j) \\ \mathrm{P}(u \in F \mid u) &= \log(\mathrm{P}(u \in F)) + \sum_{j \in S} \log(\mathrm{P}(u_j = 1 \mid u \in F))u_j + \\ &\qquad \log(\mathrm{P}(u_j = 0 \mid u \in F))(1 - u_j). \end{aligned} \tag{2.6}$$

Now, for $j \in V$, we will see in (3.1) and (3.2) that

$$\mathrm{P}(u \in T) = \frac{|T|}{|T| + |F|} \text{ and } \mathrm{P}(u \in F) = \frac{|F|}{|T| + |F|},$$

and in (3.5) we will see that

$$\mathrm{P}(u_j = 1 \mid u \in T) = \frac{a_j}{|T|},$$

and therefore

$$\mathrm{P}(u_j = 0 \mid u \in T) = 1 - \mathrm{P}(u_j = 1 \mid u \in T) = 1 - \frac{a_j}{|T|} = \frac{c_j}{|T|},$$

and similarly, in (3.6) we will see that

$$\mathrm{P}(u_j = 1 \mid u \in F) = \frac{b_j}{|F|},$$

and therefore

$$\mathrm{P}(u_j = 0 \mid u \in F) = 1 - \mathrm{P}(u_j = 1 \mid u \in F) = 1 - \frac{b_j}{|F|} = \frac{d_j}{|F|}.$$

Substituting back into (2.6) yields

$$\begin{aligned} \mathrm{P}(u \in T \mid u) &= \log\left[\frac{|T|}{|T|+|F|}\right] + \sum_{j \in S} \log\left[\frac{a_j}{|T|}\right] u_j \ + \log\left[\frac{c_j}{|T|}\right](1-u_j) \\ \mathrm{P}(u \in F \mid u) &= \log\left[\frac{|F|}{|T|+|F|}\right] + \sum_{j \in S} \log\left[\frac{b_j}{|F|}\right] u_j \ + \log\left[\frac{d_j}{|F|}\right](1-u_j). \end{aligned} \tag{2.7}$$

As we discussed in §1.2, we will perform Laplace smoothing, since doing so avoids estimates being void when a feature does not appear in any documents. Therefore (2.7) is actually equivalent to

$$\begin{aligned} \mathrm{P}(u \in T \mid u) &= \log\left[\frac{|T|}{|T|+|F|}\right] + \sum_{j \in S} \log\left[\frac{a'_j+1}{|T|+2}\right] u_j \ + \log\left[\frac{c'_j+1}{|T|+2}\right](1-u_j) \\ \mathrm{P}(u \in F \mid u) &= \log\left[\frac{|F|}{|T|+|F|}\right] + \sum_{j \in S} \log\left[\frac{b'_j+1}{|F|+2}\right] u_j \ + \log\left[\frac{d'_j+1}{|F|+2}\right](1-u_j). \end{aligned} \tag{2.8}$$

where a', b', c', and d' are the values prior to performing Laplace smoothing.

2.10 Robustness Measures

A *robust* feature selection algorithm is one in which the set of selected features S as well as the order in which the features appear in the the sequence (s_i) is relatively unaffected by small changes in T and F. To measure the robustness of a feature selection algorithm we apply the ROBUSTNESS algorithm.

ROBUSTNESS

Input: (T, F) and feature selection algorithm f.

Initialize: Set $i := 1$ and $r_j = 0$ for $j = 1, \ldots, |V|$.

Step 1: Set $\gamma :=$ a random number in $[0.25, 1]$.

Step 2: Set $n_T := \max\{\lfloor \gamma |T| \rfloor, 2\}$ and $n_F := \max\{\lfloor \gamma |F| \rfloor, 2\}$.

Step 3: Set $W'_T :=$ n_T randomly selected documents from W_T and
$W'_F :=$ n_F randomly selected documents from W_F.

Step 5: Set $(s_k) := \omega(V, K, f, \cdot)$ using $(T[W_T'], F[W_F'])$.

Step 6: Set $r_{s_k} := r_{s_k} + 1$ for $k = 1, \ldots, K$ where $s_k \in V$.

Step 7: If $i := 50$ then goto **Output**, otherwise set $i := i + 1$ and goto **Step 1**.

Output: Set $R := \{r_j \colon r_j > 0 \text{ and } j \in V\}$ and $\xi :=$ average of R and output ξ for this topic.

We use this algorithm to compute the *robustness*, denoted as ξ, of a feature selection algorithm which is the average number of times that features appear in the top K features averaged over 50 runs of the algorithm on randomly constructed subsets of $|T|$ and $|F|$, averaged over all topics. Notice that in each iteration we construct and run f on a new randomly generated variant of (T, F). These variants will likely have different values of parameters, such as a,b,c and d, that are used in feature selection algorithms. We consider feature selections algorithms with larger values of ξ to be more robust than those with smaller values since larger values indicate that the same set of features appear frequently in (s_k) even though the algorithm is run on variants of (T, F). Note that this measure obviously does not use cross validation as is described in §2.13 but does average its results over all topics.

2.11 The Set of Evaluation Criteria

As mentioned in §2.6, for monotonically non-decreasing measures we will compute the value of the measure for each subset $S_k = \bigcup_{i=1}^{k}(s_i)$ and will use these values to compute the *area under the curve* for the graph of each measure as a function of $k = 1, \ldots, K$. We therefore begin this section by stating the following simple result.

Proposition 2.2. *The measure functions ρ, σ, ν, θ and ϕ^{α} are monotonically non-decreasing in the size of the projection space, $k = 1, \ldots, K$.*

Rather than using the actual values of the area under the curve for the measures listed in Proposition 2.2, we choose to normalize them so their values are in $[0, 1]$.

We will denote these normalized area under the curve values as $\widehat{\rho}$, $\widehat{\sigma}$, $\widehat{\nu}$, $\widehat{\theta}$ and $\widehat{\phi}^{\alpha}$ respectively. The measure φ is not monotonically non-decreasing in the size of the projection space and therefore we will use the mean of its values over all $k = 1, \ldots, K$, which we will denote as $\bar{\varphi}$.[2] The measure ξ is not a function of $k = 1, \ldots, K$ and we do not normalize its values. Table 2.11 summarizes the criteria that comprise the set $\mathcal{C}$.

$\widehat{\rho}$	AUC of the Minimum Hamming Distance
$\widehat{\sigma}$	AUC of the Average Hamming Distance
$\bar{\vartheta}$	Mean of the Coefficient of Variation of the Hamming Distance
$\widehat{\nu}$	AUC of the Number of Stopwords
$\widehat{\theta}$	AUC of the Number Pairs Separated
$\widehat{\phi}_{\alpha}$	AUC of the Lexicographic Hamming Distance
$\bar{\varphi}$	Mean of the F Measure
ξ	The Robustness

Table 2.2: Set $\mathcal{C}$ of Feature Set Evaluation Criteria

Normalization of measures requires that we know the maximum possible value of the area under the curve for each measure, which is calculated by computing the sum over $k = 1, \ldots, K$ of the maximum possible value of each measure at k. Table 2.3 provides the information required to perform the desired normalization.

Measure	Minimum	Maximum	Maximum at k	Maximum AUC
ρ	0	K	k	$K(K+1)/2$
σ	0	K	k	$K(K+1)/2$
ν	0	K	k	$K(K+1)/2$
θ	0	$\lvert T\rvert\lvert F\rvert$	$\lvert T\rvert\lvert F\rvert$	$K\lvert T\rvert\lvert F\rvert$
ϕ_{α} for $0 \leq \alpha < 1$	0	$\lvert T\rvert\lvert F\rvert$	$\lvert T\rvert\lvert F\rvert$	$K\lvert T\rvert\lvert F\rvert$

Table 2.3: Monotonically Non-Decreasing Measures

Since selection of noise features is undesirable, we are interested in learning what impact such features have on the criteria in $\mathcal{C}$. To this end we define a set of *discounted criteria* $\mathcal{C}' = \{\widehat{\rho}', \widehat{\sigma}', \widehat{\theta}', \widehat{\phi_{\alpha}'}, \bar{\varphi}, \xi\}$. Before discussing these criteria we mention that

[2]The measure φ does not need to be normalized since it already is in $[0, 1]$.

clearly there is no discounted version of $\widehat{\nu}$. Similarly, there is no discounted version of $\bar{\varphi}$; that is, we simply use $\bar{\varphi}$ because inclusion of noise in a feature set already reduces its semantic content and the removal of noise features can reasonably be expected to improve the F measure. We also do not use a discounted variant of ξ.

The definition of the discounted criteria is motivated by the idea that while the value of the criteria that are based on monotonically non-decreasing measures clearly increase with each selected feature, we would like to discount the amount of this increase when a noise feature is selected. While there are certainly many ways to define such discounted criteria, our approach is illustrated by the definition of $\widehat{\sigma}'(y)$ which is defined to be the area under the curve for the graph of $\sigma'(y)$ as a function of $k = 1, \ldots, K$ where

$$\sigma'_k(y) = \begin{cases} \sigma_{k-1}(y) & \text{if } s_k \in J, \\ \sigma_k(y) & \text{otherwise,} \end{cases} \tag{2.9}$$

and $\sigma_k(y)$ is the value of $\sigma(y)$ at $k \in \{1, \ldots, K\}$. Clearly this definition penalizes algorithms that select noise features, with the severity of the penalty being larger when such features are selected early in (s_k). Discounted variants of the other criteria that are based on monotonically non-decreasing measures are defined similarly as are measures that are not monotonically non-decreasing measures. For example, $\bar{\varphi}'(y)$ is the mean of the set

$$\varphi'_k(y) = \begin{cases} \varphi_{k-1}(y) & \text{if } s_k \in J, \\ \varphi_k(y) & \text{otherwise,} \end{cases} \tag{2.10}$$

for all $k = 1, \ldots, K$. Figure 4.1 provides an example of the non-discounted and discounted variants of σ.

2.12 Feature Set Size

Since $\mathcal{C}$ includes criteria that are the area under the curve of monotonically non-decreasing measures, we we must specify the number of features K that are to be retrieved in our experiments by all feature selection algorithms. That is, while ρ, σ, ν, θ, and ϕ_α are all bounded above, Proposition 2.2 states that in the absence of any

Statistic	95%	98%	100%
Minimum of K	1.00	1.00	1.00
First Quartile of K	2.00	4.00	19.00
Median of K	4.00	7.00	33.00
Mean of K	9.41	16.57	117.10
Third Quartile of K	7.00	12.00	66.00
Maximum of K	310.00	370.00	9023.00

Table 2.4: K Required for Support Set in 95%, 98% and 100% of All (μ, topic) Combinations

constraints, the obviously undesirable strategy of simply selecting of all n features in V will guarantee that each of these measures attains its maximum value.

Assuming that feature selection algorithms try to order the features in (s_k) by the degree to which they satisfy the criteria in $\mathcal{C}$, we would like to specify a K so that the feature set selected by "most" algorithms exceeds some threshold for each of these criteria and for all topics. Admitting that any such decision will be somewhat arbitrary we decided adopt a somewhat simplified version of this strategy and to base the value of K on the number of features that are required to obtain a support set.

Since this value varies greatly depending on the topic and the feature selection algorithm, we ran a simple empirical study in order to determine reasonable values. Specifically, we ran the ranking algorithm without cross validation for a subset of the feature ranking functions in §3.8 and for each of the topics listed in §2.5. A summary of results averaged across all topics and all feature ranking functions is listed in Table 2.4.

In addition, we found that for $K \leq 25$

- 93.38% of the (μ, topic) combinations separated 95% of the (T, F) pairs,
- 86.38% of the (μ, topic) combinations separated 98% of the (T, F) pairs, and
- 37.19% of the (μ, topic) combinations separated 100% of the (T, F) pairs.

and based on these results we decided to utilize $K = 25$ in our experiments.

2.13 Cross Topic / Cross Validation

Unless stated otherwise all experiments employ 10-fold cross validation using the sets U^i for $i = 1, \ldots, 10$ discussed in §2.4 as input and essentially proceed as follows

CROSS TOPIC / CROSS VALIDATION

Input: U^i for $i = 1, \ldots, 10$ and all topics $q \in Q$ where the set $Q = \{1, \ldots, 50\}$ indexes set of topics.

Initialize: Set $q := 1$ and $i := 1$.

Step 1: Set $\left(T^i_{train}, F^i_{train}, T^i_{test}, F^i_{test}\right) := U^i$ for topic $q \in Q$.

Step 2: Set $(s_k) := \omega(V, K, f, \cdot)$ using $(T^i_{train}, F^i_{train})$.

Step 3: Calculate the values of $\mathcal{C}$ using the projections $(T^i_{train}[(s_k)], F^i_{train}[(s_k)])$, $(T^i_{test}[(s_k)], F^i_{test}[(s_k)])$, and $(T^i_{\Delta}[(s_k)], F^i_{\Delta}[(s_k)])$ where the last pair contains the percent change between training and testing, and store the results in $\mathcal{C}^i_{train}$, $\mathcal{C}^i_{test}$ and $\mathcal{C}^i_{\Delta}$ respectively.

Step 4: If $i := 10$ then set $\mathcal{C}^q_{train}$ and $\mathcal{C}^q_{test}$ respectively to be the average of criteria $\mathcal{C}^i_{train}$ and $\mathcal{C}^i_{test}$ for $i = 1, \ldots, 10$, and set $q := q + 1$, $i := i + 1$ and goto **Step 1**, otherwise set $i := i + 1$ and goto **Step 1**.

Output Set $\mathcal{C}_{train} := \mathcal{C}^q_{train}$ and $\mathcal{C}_{test} := \mathcal{C}^q_{test}$ the mean of the criteria for $q \in Q$ and output $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$.

The reason for normalizing the elements of $\mathcal{C}$ should now be clear. There may be substantial variation in the value of measures for the same feature selection algorithm between different topics. For example, larger or smaller values of $|T|$ and $|F|$ could clearly result in larger or smaller values of θ. Normalizing the elements of $\mathcal{C}$, for example we divide the area under the curve of θ by $|T||F|$, is a means of mitigating the influence of individual topics.

2.14 Result Comparison

For a given experiment, the output of the cross topic/cross validation method described in §2.13, can be viewed as a two dimensional table in which each row corresponds to a feature selection algorithm and each column corresponds to one of the criteria in $\mathcal{C}$. Presented with these results we would like to order the feature selection algorithms on the basis of some combination of the criteria. In the absence of any information indicating that any of these potentially competing criteria is more important than another, we can create an efficient ordering of feature ranking algorithms by treating the problem as an multicriteria optimization problem and solving it using the lex-max-min algorithm introduced in §1.4. Unless otherwise stated we shall use the set

$$\mathcal{C}^* = \{\widehat{\nu}, \widehat{\theta}, \widehat{\sigma}, \bar{\varphi}\} \tag{2.11}$$

as well as the discounted variant

$$\mathcal{C}^{*\prime} = \{\widehat{\nu}, \widehat{\theta}', \widehat{\sigma}', \bar{\varphi}\} \tag{2.12}$$

of criteria to this end. The reason that $\widehat{\rho}$ and ϕ_α are not included in $\mathcal{C}^*$ is that they did not differ significantly in preliminary runs of the experiments in §3.10. We also order the feature ranking algorithms by the lex algorithm in which order of importance of the criteria is $\langle \widehat{\nu}, \widehat{\theta}, \widehat{\sigma}, \bar{\varphi} \rangle$, and by the result of simply averaging the four criteria. Since smaller values of $\hat{\nu}$ are better, we use $1 - \hat{\nu}$ rather than $\hat{\nu}$. Also, we round the values of the criteria to two significant digits.

One refinement of the lex-max-min algorithm that we utilize in this application is that, before rounding, we scale each column of criteria in the aforementioned table by adding the difference between 1 and the maximum value of the column, to each entry in the column, resulting in the maximum value being 1 and others being less than 1. The reason we do this is to prevent a criteria whose values are relatively less than the other criteria from having undo influence in the ordering of the feature selection algorithms.

While we will not use this information in the lex-max-min, lex, and average ranking

of algorithms, we also compute the values at K, for each measure in $\mathcal{C}^*$, i.e.

$$\mathcal{C}^*_K = \{\nu_K, \theta_K, \sigma_K, \varphi_K\}$$

and

$$\mathcal{C}^{*\prime}{}_K = \{\nu'_K, \theta'_K, \sigma'_K, \varphi'_K\}.$$

Also, for each measure $c \in \mathcal{C}^*$, we will consider the *stability* of an algorithm with respect to c, to be the difference between the value of c on the training and test sets, and will denote this value as $\Delta(c)$. The stability of the measures in $\mathcal{C}^{*\prime}$, $\mathcal{C}^*_K$, and $\mathcal{C}^{*\prime}{}_K$ will also be computed.

In closing this section we remark that while at times it may be difficult to resist the temptation to do so, our purpose in creating these orderings of feature selection algorithms is *not* to identify the single "best" algorithm, but to identify the characteristics of the family of algorithms that exhibit the best (and the worst) performance in terms of the criteria in $\mathcal{C}^*$.

2.15 Computational Considerations

Calculation of the measures discussed in §2.7 require that the Hamming distance between all pairs of projected vectors in $(T[S], F[S])$ be calculated. In this section we shall discuss some simple observations that can be used to perform these computations efficiently. We begin by stating two following properties of the Hamming distance without proof.

Proposition 2.3. *If* $S, S' \subseteq V$, $S \subseteq S'$ *and* $u, v \in \mathbb{B}^n$ *then* $d(u[S], v[S]) \leq d(u[S'], v[S'])$.

Proposition 2.3 states that the Hamming distance is a non-decreasing function of the size of the projection space.

Proposition 2.4. *If* $1 < k \leq n$, $S = \{1, 2, \ldots, k\}$, $S' = \{k+1, k+2, \ldots, n\}$ *and* $u, v \in \mathbb{B}^n$ *then* $d(u, v) = d(u[S], v[S]) + d(u[S'], v[S'])$.

Proposition 2.4 shows that the calculation of the Hamming distance between two vectors can be decomposed into calculations of the Hamming distance of projections of

the vectors onto a subspace. The following algorithm, which is based on Proposition 2.4, efficiently computes the Hamming distance between two vectors at each of p projections onto a nested subspace when projected onto each element of a collection of subsets that partition V.

HAMMING DISTANCE

Input: $u, v \in \mathbb{B}^n$, $1 \leq p \leq n$ and $k_0 = 0 < k_1 < k_2 \cdots < k_p = n$.

Initialize: $i \leftarrow 0$ and $w \leftarrow [0, 0, \cdots, 0] \in \mathbb{Z}_+^p$.

Step 1: If $i = p$ goto **Output**, otherwise

set $S_i \leftarrow V \setminus \{k_i, k_i + 1, \cdots, k_{i+1}\}$, $y_i \leftarrow \chi^{S_i}$, $w_{i+1} \leftarrow d_{y_i}(u, v) + w_i$ and goto **Step 1**.

Output: Output w and stop.

Note that

$$w_i = d(u[S_1 \cup S_2 \cup \cdots \cup S_i], v[S_1 \cup S_2 \cup \cdots \cup S_i])$$

for each $1 \leq i \leq p$. A naive algorithm would recompute $d(u[S_i], v[S_i])$ when computing $d(u[S_{i+1}], v[S_{i+1}])$ resulting in $O(n^2)$ summations over the components of u and v. Proposition 2.4 implies that using the fact that by

$$d\Big(u\Big[\bigcup_{i=1}^{n} S_i\Big], v\Big[\bigcup_{i=1}^{n} S_i\Big]\Big) = \sum_{i=1}^{n} d(u[S_i], v[S_i])$$

the algorithm can be improved to make only $O(n)$ summations.

Finally, the following result shows that it is possible to efficiently compute the average Hamming distance between all pairs of projected vectors in $(T[S], F[S])$.

Proposition 2.5.

$$\sigma(y) = \frac{1}{|T||F|} \sum_{j \in S} a_j d_j + b_j c_j$$

Proof. If $y = \chi^S$, then we recall that average Hamming distance was given in (2.2) as

$$\sigma(y) = \frac{1}{|T||F|} \sum_{u \in T, v \in F} d_y(u, v),$$

which can be written as

$$\sigma(y) = \frac{1}{|T||F|} \sum_{u \in T, v \in F} \sum_{j \in S} |u_j - v_j|$$

which after interchanging the summation is equal to

$$\sigma(y) = \frac{1}{|T||F|} \sum_{j \in S} \sum_{u \in T, v \in F} |u_j - v_j|.$$

For a given $j \in S$, the inner summation can be written as

$$\sum_{u \in T, v \in F} |u_j - v_j| = a_j d_j + b_j c_j$$

which is the number of differences between pairs for feature j, and allows us to write

$$\sigma(y) = \frac{1}{|T||F|} \sum_{j \in S} |u_j - v_j|.$$

■

Therefore, using the precomputed information contained in $\boxplus_j \in \boxplus$ for each $j \in S$ we can efficiently compute $\sigma(y)$.

Chapter 3

Boolean Feature Ranking Functions

We begin this chapter by formally defining Boolean feature ranking functions and ranking algorithms. We then introduce some common feature ranking functions from the literature, and identify some of their properties and some relationships between them. We conclude by reviewing the results of an empirical study of the relative performance of ranking algorithms based on each of these feature ranking functions.

3.1 Ranking Functions and Algorithms

A *Boolean feature ranking function* is a function

$$\mu : \boxplus \mapsto \mathbb{R}$$

that for each feature $j \in V$ maps the values a_j, b_j, c_j and d_j to the set of real numbers. The notation μ will be used for a generic Boolean feature ranking function and specific Boolean feature ranking functions will be denoted as μ_i for some integer i. We shall denote the set of Boolean feature ranking functions as $\mathcal{M}$. Since the elements of $\mathcal{M}$ can be written as a function of x_j and y_j, of a_j and b_j, or of a_j, b_j, c_j and d_j, we will write $\mu(x_j, y_j)$, $\mu(a_j, b_j)$ or $\mu(a_j, b_j, c_j, d_j)$ when we wish to highlight the fact that a particular μ is written as a function of two or four variables respectively, and will write $\mu(\boxplus_j)$ and sometimes $\mu(j)$ when this distinction is not important.

Ranking algorithms are based on feature ranking functions and in fact each $\mu \in \mathcal{M}$ defines a different ranking algorithm, and therefore the set $\mathcal{M}$ defines an entire class of ranking algorithms. We denote the ranking algorithm for a particular $\mu \in \mathcal{M}$ as μ-RANKING and provide the definition of this algorithm below.

μ-RANKING

Input: The set V, a function $\mu \in \mathcal{M}$, a sort order $\updownarrow \in \{\downarrow, \uparrow\}$ and an integer K.

Step 1: Set sequence (s_k) to the set V, sorted by function μ, in $\updownarrow$ order.

Step 2: Set $(s_k) \leftarrow s_1, s_2, \cdots, s_K$.

Output: Output (s_k)

While there are obviously myriad feature ranking functions, since the performance of μ-RANKING is related to how well the given μ measures the degree to which each feature satisfies the criteria in $\mathcal{C}$, commonly used functions tend to be based on a few key ideas for achieving this goal, and these ideas are implemented using a small set of core models of $\boxplus$. Five of the core models of $\boxplus$ used to implement feature ranking functions are discussed in §3.2, §3.3, §3.4, §3.5 and §3.6.

3.2 Probabilistic Model

In this section we assume that the probability of certain events related to the occurrence of feature $j \in V$ in the collection can be approximated using the frequencies a_j, b_j, c_j and d_j, and list the probabilities and odds of these events.

1. Consider a document $u \in T \cup F$, then the probability that the document is relevant is

$$\mathrm{P}(u \in T) = \frac{|T|}{|T| + |F|} \tag{3.1}$$

   and the probability that the document is irrelevant is

$$\mathrm{P}(u \in F) = \frac{|F|}{|T| + |F|}. \tag{3.2}$$

2. Consider a document $u \in T \cup F$ and a feature $j \in V$, then the probability that

the feature appears in the document and the document is relevant is

$$\mathrm{P}(u_j = 1, u \in T) = \frac{a_j}{|T| + |F|} \tag{3.3}$$

and the probability that the feature appears in the document and the document is irrelevant is

$$\mathrm{P}(u_j = 1, u \in F) = \frac{b_j}{|T| + |F|}. \tag{3.4}$$

3. Consider a document $u \in T \cup F$ and a feature $j \in V$, then the probability that the feature appears in the document given that document is relevant is

$$\mathrm{P}(u_j = 1 \mid u \in T) = \frac{\mathrm{P}(u_j = 1, u \in T)}{\mathrm{P}(u \in T)} = \frac{\frac{a_j}{|T|+|F|}}{\frac{|T|}{|T|+|F|}} = \frac{a_j}{|T|} \tag{3.5}$$

and the probability that the feature appears in the document given that document is irrelevant is

$$\mathrm{P}(u_j = 1 \mid u \in F) = \frac{\mathrm{P}(u_j = 1, u \in F)}{\mathrm{P}(u \in F)} = \frac{\frac{b_j}{|T|+|F|}}{\frac{|F|}{|T|+|F|}} = \frac{b_j}{|F|}. \tag{3.6}$$

As discussed in §1.3, (3.5) and (3.6) are the TPR and FPR respectively.

4. Consider a document $u \in T \cup F$ and a feature $j \in V$, then

$$\frac{\mathrm{P}(u_j = 1 \mid u \in T)}{\mathrm{P}(u_j = 1 \mid u \in F)} = \frac{\frac{a_j}{|T|}}{\frac{b_j}{|F|}} \tag{3.7}$$

which is the ratio of the TPR to the FPR is the *positive likelihood ratio*. It can also be seen to be the slope of an ROC curve as can be seen in Figure 3.29 (see e.g [17]).

5. Consider a document $u \in T \cup F$ and a feature $j \in V$, then the probability that the feature appears in the document, regardless of whether the document is relevant

or irrelevant, is

$$\begin{aligned} \mathrm{P}(u_j = 1) &= \mathrm{P}(\{u_j = 1 \mid u \in T\} \cup \{ u_j = 1 \mid u \in F\}) \\ &= \mathrm{P}(u \in T)\,\mathrm{P}(u_j = 1 \mid u \in T) + \mathrm{P}(u \in F)\,\mathrm{P}(u_j = 1 \mid u \in F) \\ &= \frac{|T|}{|T|+|F|}\,\frac{a_j}{|T|} + \frac{|F|}{|T|+|F|}\,\frac{b_j}{|F|} = \frac{a_j + b_j}{|T|+|F|}. \end{aligned} \tag{3.8}$$

6. Consider documents $u \in T$ and $v \in F$ and a feature $j \in V$, then the probability that the feature appears in both u *and* v is

$$\begin{aligned} &\mathrm{P}(u_j = 1 \mid u \in T \ \cap \ v_j = 1 \mid v \in F) \\ &\quad = \mathrm{P}(u_j = 1 \mid u \in T)\ \mathrm{P}(v_j = 1 \mid v \in F) \\ &\quad = \frac{a_j b_j}{|T||F|}. \end{aligned} \tag{3.9}$$

7. Consider a document $u \in T \cup F$ and a feature $j \in V$, then the probability that the document is relevant given that the feature appears in the document is

$$\mathrm{P}(u \in T \mid u_j = 1) = \frac{\mathrm{P}(u_j = 1, u \in T)}{\mathrm{P}(u_j = 1)} = \frac{\frac{a_j}{|T|+|F|}}{\frac{a_j+b_j}{|T|+|F|}} = \frac{a_j}{a_j + b_j} \tag{3.10}$$

and the probability that the document is irrelevant given that the feature appears in the document is

$$\mathrm{P}(u \in F \mid u_j = 1) = \frac{\mathrm{P}(u_j = 1, u \in F)}{\mathrm{P}(u_j = 1)} = \frac{\frac{b_j}{|T|+|F|}}{\frac{a_j+b_j}{|T|+|F|}} = \frac{b_j}{a_j + b_j}. \tag{3.11}$$

8. We will also make use of the *odds* that a document is relevant which is the probability that the document is relevant divided by the probability that the document is irrelevant. Specifically, we are interested in the *odds* that a document that contains a feature is relevant. It is the probability that a document that contains the feature is relevant divided by the probability that a document contains a feature is irrelevant. Formally, consider a document $u \in T \cup F$ and a feature $j \in V$, then

given that $u_j = 1$, the *odds* that $u \in T$ are

$$\frac{\mathrm{P}(u \in T \mid u_j = 1)}{\mathrm{P}(u \in F \mid u_j = 1)} = \frac{\frac{a_j}{a_j+b_j}}{\frac{b_j}{a_j+b_j}} = \frac{a_j}{b_j}. \tag{3.12}$$

9. Similarly, the *odds* that a document that does not contain a feature is relevant is the probability that a document that does not contain the feature is relevant divided by the probability that a document does not contain the feature is irrelevant. Formally, consider a document $v \in T \cup F$ and a feature $j \in V$, then given that $v_j = 0$, the *odds* that $v \in T$ are

$$\frac{\mathrm{P}(v \in T \mid v_j = 0)}{\mathrm{P}(v \in F \mid v_j = 0)} = \frac{\frac{c_j}{c_j+d_j}}{\frac{d_j}{c_j+d_j}} = \frac{c_j}{d_j}, \tag{3.13}$$

10. The odds in (3.12) and (3.13) allow us to compute the *odds ratio* which compares the odds of a document that contains a feature being relevant with the odds of a document that does not contain the feature being relevant. It is defined as

$$\frac{\frac{\mathrm{P}(u \in T \mid u_j=1)}{\mathrm{P}(u \in F \mid u_j=1)}}{\frac{\mathrm{P}(v \in T \mid v_j=0)}{\mathrm{P}(v \in F \mid v_j=0)}} = \frac{\frac{a_j}{b_j}}{\frac{c_j}{d_j}} = \frac{a_j d_j}{b_j c_j}. \tag{3.14}$$

    If the odds ratio is 1, then odds of relevance are the same for both documents that contain the feature and those that do not. If the odds ratio is greater than 1, then the odds of relevance is greater for documents that contain the feature. If the odds ratio is less than 1, then the odds of relevance is greater for documents that do not contain the feature.

The probabilities and odds discussed in this section form the basis of many of the probabilistic models that are used in information retrieval and are used to construct many of the Boolean feature ranking functions that we will consider.

3.3 Term Frequency Model

The calculations in §3.2 employed the traditional interpretation of the quantities a_j, b_j, c_j and d_j in which the occurrence of a feature $j \in V$ in a document is modeled as a Boolean variable, and a_j and b_j are viewed as integer frequencies or *counts* of the number of times the feature appeared in relevant documents and irrelevant documents respectively, with c_j and d_j being viewed as counts of the number of times the feature did not appear in relevant documents and irrelevant documents respectively.

It is possible to develop an alternate interpretation of these quantities in which the occurrence of the feature j in a document is modeled as a real variable and a_j, b_j, c_j and d_j are viewed as *values* that can be interpreted to mean that feature j occurred[1] 1 time in a_j relevant documents, 1 time in b_j irrelevant documents, 0 times in c_j relevant documents and 0 times in d_j irrelevant documents. We could then envision the existence of hypothetical vectors

$$\mathbf{x}_j \;=\; (1, 1, 1, 1, \ldots, 0, 0)$$

and

$$\mathbf{y}_j \;=\; (1, 1, 1, \ldots, 0, 0, 0, 0, 0),$$

with $\mathbf{x}_j$ having a_j 1s and c_j 0s and $\mathbf{y}_j$ having b_j 1s and d_j 0s. Following this interpretation we could calculate that the expected value of the term frequency for feature j in relevant documents as

$$1\;\frac{a_j}{|T|} + 0\;\frac{c_j}{|T|} = \frac{a_j}{|T|}$$

and the expected value of the term frequency for feature j in irrelevant documents as

$$1\;\frac{b_j}{|F|} + 0\;\frac{d_j}{|F|} = \frac{b_j}{|F|}.$$

We will not make use of this interpretation in this chapter but will refer to it in §6.1

[1]The specific meaning of "occurred", that is, whether it means a simple term frequency or some function of the term frequency, is not important to this discussion – we will simply refer to the value as the term frequency.

when it will allow us to draw some parallels between Boolean and real-valued feature ranking functions.

3.4 Separation Model

In this section we discuss how functions of the quantities a_j, b_j, c_j and d_j relate to the amount of separation provided by a feature $j \in V$.

Consider a feature $j \in V$, then

$$a_j d_j = |\{(u,v) \in T \times F : u_j = 1 \text{ and } v_j = 0\}|$$

is the number of pairs (u,v) in which $u_j = 1$ and $v_j = 0$. For a feature $j \in V$, $a_j d_j$ is the number of relevant–irrelevant document pairs which are separated by the presence of the feature in relevant documents and the absence of the feature from irrelevant documents.

Similarly, a feature $j \in V$, then

$$b_j c_j = |\{(u,v) \in T \times F : u_j = 0 \text{ and } v_j = 1\}|$$

is the number of pairs (u,v) in which $u_j = 0$ and $v_j = 1$. For a feature $j \in V$, $b_j c_j$ is the number of relevant–irrelevant document pairs which are separated by the absence of the feature from relevant documents and the presence of the feature in irrelevant documents.

Therefore, for $j \in V$, the quantity $a_j b_j + c_j d_j$ is the total number of relevant–irrelevant document pairs separated by feature j. For instance, consider the example shown in Table 3.1.

The sets

$$\{(4,6),(4,7),(5,6),(5,7)\} \tag{3.15}$$

and

$$\{(1,8),(2,8),(3,8)\} \tag{3.16}$$

	Document	Feature j
T	1	0
	2	0
	3	0
	4	1
	5	1
F	6	0
	7	0
	8	1

Table 3.1: Separation Model Example

are the sets of relevant–irrelevant documents pairs which are separated by feature j and have cardinality $a_j d_j$ and $b_j c_j$ respectively. The sets

$$\{(4,8),(5,8)\} \tag{3.17}$$

and

$$\{(1,6),(1,7),(2,6),(2,7),(3,6),(3,7)\} \tag{3.18}$$

are the sets of relevant–irrelevant documents pairs which are not separated by feature j. In practice, for a feature $j \in V$, we would like the quantities $a_j d_j$, $b_j c_j$ and $a_j d_j + b_j c_j$ being normalized by the total number of relevant–irrelevant document pairs and will utilize

$$\sigma_j^+ = \frac{a_j d_j}{|T||F|} \tag{3.19}$$

$$\sigma_j^- = \frac{b_j c_j}{|T||F|} \tag{3.20}$$

$$\sigma_j = \frac{a_j d_j + b_j c_j}{|T||F|} \tag{3.21}$$

which we shall refer to as the *positive incremental average Hamming distance*, *negative incremental average Hamming distance* and *incremental average Hamming distance*, respectively.

3.5 Information Retrieval Model

In this section we present a model of the quantities a_j, b_j, c_j and d_j for $j \in V$ that is motivated by the field information retrieval.

Assume that for each feature $j \in V$, there exists an information retrieval algorithm IR(j) which simply retrieves all documents from the collection $T \cup F$ that contain this feature. In this section we adopt the following interpretations

$$\begin{aligned} a_j &\triangleq \text{the number of relevant documents documents retrieved by IR}(j) \\ b_j &\triangleq \text{the number of irrelevant documents retrieved by IR}(j) \\ c_j &\triangleq \text{the number of relevant documents not retrieved by IR}(j) \\ d_j &\triangleq \text{the number of irrelevant documents not retrieved by IR}(j) \end{aligned}$$

for feature $j \in V$.

Two fundamental performance metrics in information retrieval that we now present using the notation above are

$$P = \frac{a_j}{a_j + b_j} \tag{3.22}$$

which is referred to as the *precision* and

$$R = \frac{a_j}{a_j + c_j} \tag{3.23}$$

is referred to as the *recall*. As seen in (3.10), the precision can be seen to correspond to the $\mathrm{P}(u \in T \mid u_j = 1)$, and as seen in (3.5), the recall can be seen to correspond to the $\mathrm{P}(u_j = 1 \mid u \in T)$.

There are many information retrieval metrics that are based on the precision and recall. One commonly used family of such metrics are the F_α metrics which are defined as

$$F_\alpha = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}},$$

which are the harmonic means of the precision and the recall weighted by parameter α (see e.g. [60], p.144), with $F_{1/2}$ being commonly referred to simply as the F measure.

Use of the harmonic mean of the precision and recall as opposed to their arithmetic mean has the advantage of not rewarding the strategy of simply retrieving all documents in order to ensure a perfect recall score (see e.g. p.144, [60]).

3.6 Single Feature Classifier Model

In this section we show that each feature can be used to create two simple classifiers and present a model of the quantities a_j, b_j, c_j and d_j for $j \in V$ that is based on these classifiers.

Consider the classifier $f_j \colon T \cup F \mapsto \mathbb{R}$ defined as

$$f_j(u) = \begin{cases} 1 & \text{if } u_j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

where $u \in T \cup F$ and $j \in V$. For each document $u \in T \cup F$ it assigns a score of 1 if u contains feature $j \in V$ and a score of 0 if it does not. This classifier is a special case of the classifier given in (1.5) and as in that section we shall let $\lambda \in \mathbb{R}^m$ be the vector of the scores generated by f_j and will assume that each score represents the classifier's estimate that the document is relevant.

In §1.3 we saw that the score vector from a classifier such as f_j can be used to define two complementary Boolean classifiers. Following this approach we define the classifier $g_j^+ : T \cup F \mapsto \mathbb{B}$ as

$$g_j^+(i) = \begin{cases} 0 & \text{if } \lambda_i \geq 1 \\ 1 & \text{otherwise,} \end{cases} \tag{3.25}$$

and define the classifier $g_j^- : T \cup F \mapsto \mathbb{B}$ as

$$g_j^-(i) = \begin{cases} 0 & \text{if } \lambda_i \leq 1 \\ 1 & \text{otherwise.} \end{cases} \tag{3.26}$$

where $j \in V$ and $i \in W$. The classifier g_j^+ assumes that larger scores represent a higher estimate of relevance and classify each document that contains the feature $j \in V$ as relevant, while the classifier g_j^- assumes that larger scores represent a higher estimate

of irrelevance and classify each document that contains the feature $j \in V$ as irrelevant.

Now for $j \in V$ we can see that

$$\begin{aligned} a_j &= |\{i \in W_T : g^+(i) = 1\}| \\ b_j &= |\{i \in W_F : g^+(i) = 1\}| \\ c_j &= |\{i \in W_T : g^+(i) = 0\}| \\ d_j &= |\{i \in W_F : g^+(i) = 0\}| \end{aligned}$$

and that

$$\begin{aligned} a_j &= |\{i \in W_T : g^-(i) = 0\}| \\ b_j &= |\{i \in W_F : g^-(i) = 0\}| \\ c_j &= |\{i \in W_T : g^-(i) = 1\}| \\ d_j &= |\{i \in W_F : g^-(i) = 1\}|. \end{aligned}$$

So, for example we have that a_j is the number of relevant documents that g_j^+ classified as relevant *and* it is the number of relevant documents that g_j^- classified as irrelevant. Similarly, b_j is the number of irrelevant documents that g_j^+ classified as relevant *and* it is the number of irrelevant documents that g_j^- classified as irrelevant.

Notice that for both classifiers g_j^+ and g_j^-, the values a_j, b_j, c_j, and d_j correspond to a contingency table $\boxplus_j \in \boxplus$ as was discussed in §1.2. There are a few interesting cases that will play an important role in future discussions. When $b_j = c_j = 0$, we say that the corresponding contingency table $\boxplus_j \in \boxplus$ contains *perfect information* and the classifier g_j^+ classifies all documents correctly and therefore has *perfect performance*. When $a_j = d_j = 0$, we also say that the corresponding contingency table $\boxplus_j \in \boxplus$ contains *perfect information* and the classifier g_j^- classifies all documents correctly and therefore has *perfect performance*. When $c_j = d_j = 0$, the feature is present in every document regardless of whether its relevant or irrelevant and we say that the corresponding contingency table $\boxplus_j \in \boxplus$ contains *no information*. When $a_j = b_j = 0$, the feature is absent from every document regardless of whether its relevant or irrelevant and we say again

that the corresponding contingency table $\boxplus_j \in \boxplus$ contains *no information*. Although not a degenerate case like the two no information cases just mentioned, in the case when $a_j/(a_j + c_j) = b_j/(b_j + d_j)$, i.e. when the feature j is present in the same proportion of relevant and irrelevant documents, the corresponding contingency table $\boxplus_j \in \boxplus$ can also be considered to contain *no information*.

While the classifiers given in (1.11) and (1.10) each define a different Boolean classifier for each unique threshold τ and therefore actually correspond to a *family* of Boolean classifiers for a given $j \in V$, the classifiers g_j^+ and g_j^- defined in (3.25) and (3.26) effectively have the threshold τ fixed at 1 and hence each only defines a single classifier. As a result, the ROC curves for g_j^+ and g_j^- only consist of only one point other than $(0, 0)$ and $(1, 1)$.

Considering the classifier g_j^+, we see that other than the points $(0, 0)$ and $(1, 1)$, the ROC curve consists of the single point with TPR given as

$$\frac{|\{i \in W_T : g^+(i) = 1\}|}{|T|} = \frac{a_j}{a_j + c_j}$$

and the FPR given as

$$\frac{|\{i \in W_F : g^+(i) = 1\}|}{|T|} = \frac{b_j}{b_j + d_j}.$$

Therefore, the ROC curve for the classifier g_j^+ for $j \in V$, is a piecewise linear curve constructed from the points

$$\left\{(0, 0), \quad \left(\frac{b}{b+d}, \frac{a}{a+c}\right), \quad (1, 1)\right\} \tag{3.27}$$

as shown in Figure 3.1 and Figure 3.2.

The AUC for these curves simply consists of the sum areas of the regions labeled, 1, 2 and 3, which is given as

$$\text{AUC}_+ = \frac{1}{2}\left(\frac{b}{b+d}\right)\left(\frac{a}{a+c}\right) + \left(1 - \frac{b}{b+d}\right)\left(\frac{a}{a+c}\right) + \frac{1}{2}\left(1 - \frac{b}{b+d}\right)\left(1 - \frac{a}{a+c}\right),$$

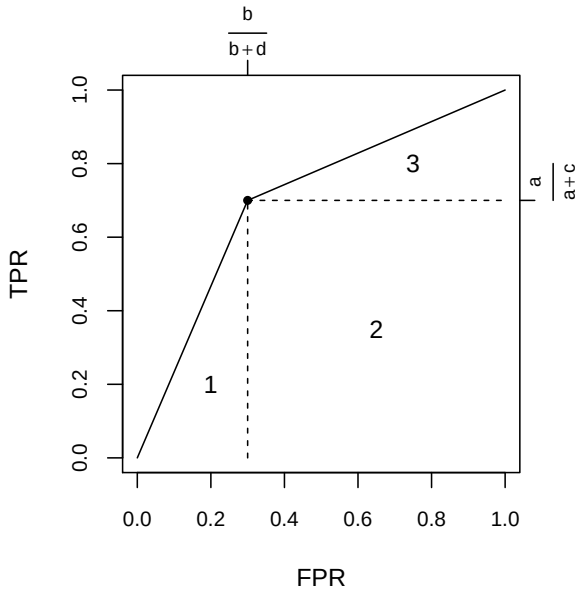


Figure 3.1: A Boolean ROC Curve for g^+

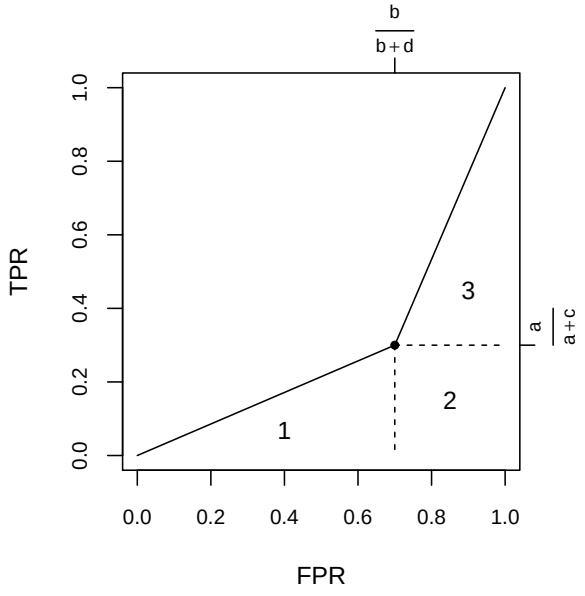


Figure 3.2: A Boolean ROC Curve for g^+

which simplifies to

$$\mathrm{AUC}_+ = \frac{ab + 2ad + cd}{2(a+c)(b+d)}. \tag{3.28}$$

Proceeding similarly for the classifier g_j^- we see that other than the points $(0,0)$ and $(1,1)$, the ROC consists of the single point with TPR given as

$$\frac{|\{i \in W_T : g^-(i) = 1\}|}{|T|} = \frac{c_j}{a_j + c_j}$$

and FPR given as

$$\frac{|\{i \in W_F : g^-(i) = 1\}|}{|T|} = \frac{d_j}{b_j + d_j}.$$

Therefore, in the Boolean case, the ROC curve for the classifier g_j^- for $j \in V$, is a piecewise linear curve constructed from the points

$$\left\{(0,0), \quad \left(\frac{a}{a+c}, \frac{b}{b+d}\right), \quad (1,1)\right\} \tag{3.29}$$

as shown in Figure 3.3 and Figure 3.4.

Notice that the point $(\frac{a}{a+c}, \frac{b}{b+d})$ in (3.4) is the reflection over the line TPR = FPR of the point $(\frac{b}{b+d}, \frac{a}{a+c})$ in (3.1) and that the point $(\frac{a}{a+c}, \frac{b}{b+d})$ in (3.3) is the reflection over the line TPR = FPR of the point $(\frac{b}{b+d}, \frac{a}{a+c})$ in (3.2). The AUC for these curves simply consists of the sum areas of the regions labeled, 1, 2 and 3, which is given as

$$\mathrm{AUC}_- = \frac{1}{2}\left(\frac{a}{a+c}\right)\left(\frac{b}{b+d}\right) + \left(1 - \frac{a}{a+c}\right)\left(\frac{b}{b+d}\right) + \frac{1}{2}\left(1 - \frac{a}{a+c}\right)\left(1 - \frac{b}{b+d}\right),$$

which simplifies to

$$\mathrm{AUC}_- = \frac{ab + 2bc + cd}{2(a+c)(b+d)}. \tag{3.30}$$

In view of (3.28) and (3.30) we now state the following result.

Proposition 3.1. *If $j \in V$, λ is the vector of scores generated by classifier f defined in (3.24) and λ_T and λ_F are randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively,*

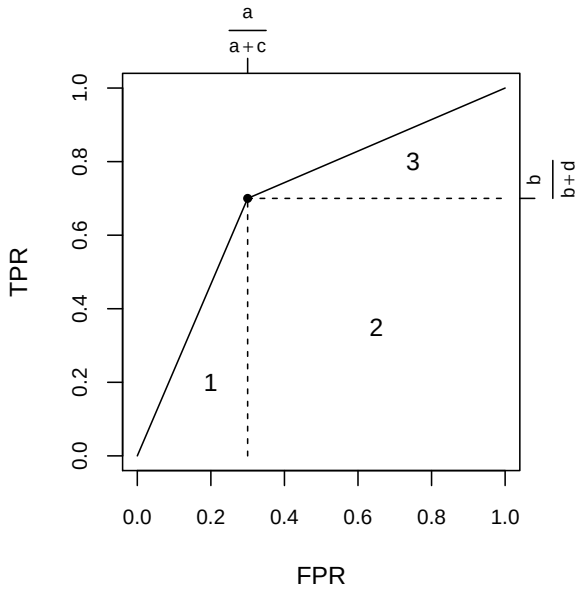


Figure 3.3: A Boolean ROC Curve for g^-

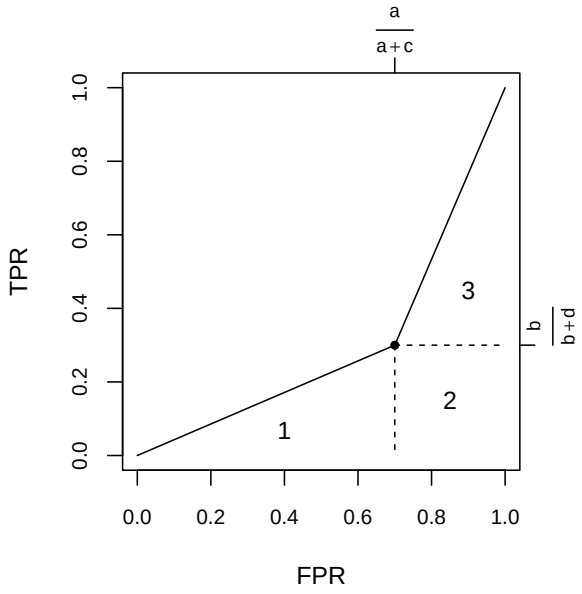


Figure 3.4: A Boolean ROC Curve for g^-

then corresponding to the Boolean classifier g_j^+ *defined in (3.25) we have*

$$P(\lambda_T > \lambda_F) = \frac{U_+}{|T||F|} = AUC_+ = \frac{ab + 2ad + cd}{2|T||F|}$$

and corresponding to the Boolean classifier g_j^- *defined in (3.26) we have*

$$P(\lambda_F > \lambda_T) = \frac{U_-}{|T||F|} = AUC_- = \frac{ab + 2bc + cd}{2|T||F|}.$$

Proof. Follows from (3.28) and (3.30) and the relationships listed in Table 1.1. Rather than using (3.28) and (3.30), this can also be shown by noting that the set $\lambda[W_T]$ contains a 1s and c 0s, and the set $\lambda[W_F]$ contains b 1s and d 0s. Therefore, ad is the number of pairs in $T \times F$ with the score for the relevant document exceeding that of the irrelevant document and $ab + cd$ is the number of pairs in $T \times F$ with equal scores which yields

$$U_+ = ad + \frac{1}{2}(ab + cd).$$

Following a similar approach we get

$$U_- = bc + \frac{1}{2}(ab + cd)$$

and the result follows by using the relationships listed in Table 1.1. ■

3.7 Positive and Negative Features

In §1.1 we informally defined a *positive* feature as a feature whose presence in a document is evidence that the document is relevant, and defined a *negative* feature as a feature whose presence in a document is evidence that the document is irrelevant.

In terms of the probabilistic model discussed in §3.2, we shall call a Boolean feature $j \in V$ *positive separating* or simply *positive* if

$$\frac{a_j}{|T|} > \frac{b_j}{|F|} \tag{3.31}$$

and will call a feature *negative separating* or simply *negative* if

$$\frac{a_j}{|T|} < \frac{b_j}{|F|}. \tag{3.32}$$

In terms of the separation model discussed in §3.4 we shall call a Boolean feature $j \in V$ positive if

$$\frac{a_j d_j}{|T|} > \frac{b_j c_j}{|F|} \tag{3.33}$$

and will call the feature negative if

$$\frac{a_j d_j}{|T|} < \frac{b_j c_j}{|F|}. \tag{3.34}$$

In terms of the single feature classifier model discussed in §3.6, we shall call a Boolean feature $j \in V$ positive if, as shown in Figure 3.5,

$$\mathrm{AUC}_+ > \mathrm{AUC}_- \tag{3.35}$$

or equivalently if

$$\mathrm{AUC}_+ > \frac{1}{2}$$

and will call the feature negative if, as shown in Figure 3.6,

$$\mathrm{AUC}_- > \mathrm{AUC}_+ \tag{3.36}$$

or equivalently if

$$\mathrm{AUC}_- > \frac{1}{2}$$

where AUC_+ is the AUC for the classifier g^+ given in (3.25) and AUC_- is the AUC for the classifier g^- given in (3.26).

The following result shows that these definitions coincide.

Proposition 3.2. *The characterizations of positive and negative Boolean features in terms of the probabilistic, separation and single feature classifier models are all equivalent.*

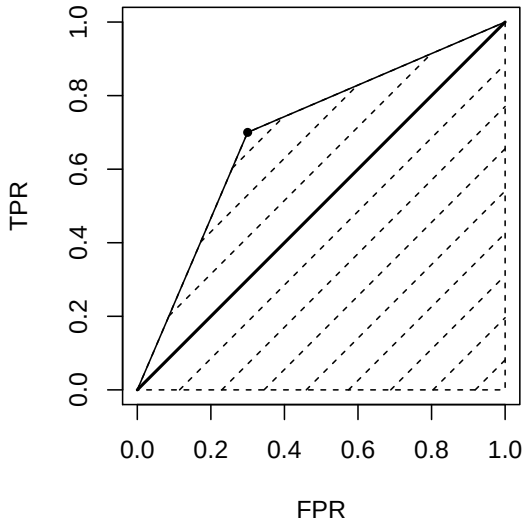


Figure 3.5: Boolean ROC Curve Positive Feature

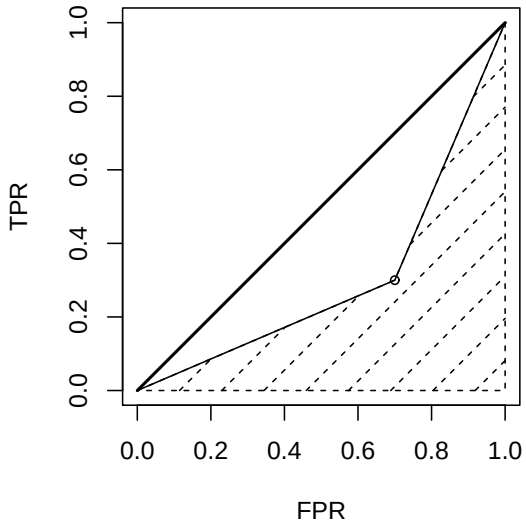


Figure 3.6: Boolean ROC Curve Negative Feature

Proof. Noting that for $j \in V$, (3.31) can be written as

$$\frac{a_j}{|T|} > \frac{b_j}{|F|} \Leftrightarrow \frac{a_j}{|T|} - \frac{b_j}{|F|} > 0,$$

(3.33) can be written as

$$\frac{a_j d_j}{|T||F|} > \frac{b_j c_j}{|T||F|} \Leftrightarrow \frac{a_j d_j - b_j c_j}{|T||F|} > 0 \Leftrightarrow \frac{a_j}{|T|} - \frac{b_j}{|F|} > 0,$$

(3.35) can be written as

$$\begin{aligned} \mathrm{AUC}_+ - \mathrm{AUC}_- > 0 \quad &\Leftrightarrow \quad \frac{ab + 2ad + cd}{2|T||F|} - \frac{ab + 2bc + cd}{2|T||F|} > 0 \\ &\Leftrightarrow \quad \frac{2ad - 2bd}{2|T|F|} > 0 \\ &\Leftrightarrow \quad \frac{a_j}{|T|} - \frac{b_j}{|F|} > 0 \end{aligned}$$

and that similar statements hold for (3.32), (3.34) and (3.36) shows the claim. ■

Corollary 3.1. *Consider a feature $j \in V$. If*

$$\frac{a_j}{|T|} - \frac{b_j}{|F|} > 0$$

then j is a positive feature and if

$$\frac{a_j}{|T|} - \frac{b_j}{|F|} < 0$$

then j is a negative feature.

Given features $i, j \in V$, Corollary 3.1 motivates us to consider feature i to be "more positive" than feature j when

$$\frac{a_i}{|T|} - \frac{b_i}{|F|} > \frac{a_j}{|T|} - \frac{b_j}{|F|}$$

and "more negative" than feature j when

$$\frac{a_i}{|T|} - \frac{b_i}{|F|} < \frac{a_j}{|T|} - \frac{b_j}{|F|}.$$

Clearly, when considering only positive features, the most positive features will be the best at separating T and F. Likewise, when considering only negative features, the most negative features will be the best at separating T and F.

In view of our objective of identifying features that separate T and F, and the relevant document centric strategy discussed in §1.1, we will be interested in feature ranking functions that give a high rank to features that are very positive and give a low rank to features that are very negative. Such feature ranking functions can informally be thought of as acting like one-sided statistical tests and we shall refer to them as *one-sided* feature ranking functions.

However, since the amount of separation provided by a feature is independent of whether the feature is positive or negative, we shall also be interested in feature ranking functions that give a high rank to features that are either very positive *or* very negative. Such feature ranking functions can informally be thought of as acting like two-sided statistical tests and we shall refer to them as *two-sided* feature ranking functions.

While some feature ranking functions are two-sided by construction, Corollary 3.1 suggests that some one-sided feature ranking functions that are dependent on

$$\frac{a}{|T|} - \frac{b}{|F|}$$

can be transformed to be two-sided by application of an appropriate function, such as the absolute value or the square, that maps the range of the function from $\mathbb{R}$ to $\mathbb{R}_+$. In our studies of feature ranking functions we extend this idea to a larger set of functions. If $\mu \in \mathcal{M}$ such that

$$\frac{a}{|T|} > \frac{b}{|F|} \;\Rightarrow\; \mu(a,b) > 0 \;\text{ and }\; \frac{a}{|T|} < \frac{b}{|F|} \;\Rightarrow\; \mu(a,b) < 0, \tag{3.37}$$

that is, if a feature ranking function is positive when $\frac{a}{|T|} > \frac{b}{|F|}$ and negative when

$\frac{a}{|T|} < \frac{b}{|F|}$, then we shall use both the feature ranking function *and* its absolute value.

3.8 Ranking Functions

In this section we shall consider *thirty two* Boolean feature ranking functions. Since the set $\mathcal{C}$ contains multiple, potentially conflicting criteria, it is not *a priori* obvious how to identify a single $\mu \in \mathcal{M}$ that will perform well. While the collection of functions we study is not intended to be exhaustive, in order to hopefully increase the chance of finding a function that does actually perform well, we have chosen functions from a variety of fields including information retrieval, information theory, logical analysis of data, machine learning, probability theory, ROC curves, statistics and text categorization. One restriction we did impose on the functions we study is, that with the exception of one function that uses of the factorial and one that uses the logarithm, we only consider algebraic functions of a, b, c and d.

If $\mu_i, \mu_j \in \mathcal{M}$ and $\mu_i = f(\mu_j)$ for some monotonic function f, then clearly

$$\omega(V, K, \mu_i\text{-RANKING}, \cdot) = \omega(V, K, \mu_j\text{-RANKING}, \cdot).$$

That is, if two feature ranking functions only differ in a monotonic transformation, then the sequence returned by μ_i-RANKING and μ_j-RANKING are identical. When we identify such relationships between feature ranking functions we will consider one of the functions to be superfluous and will not consider it in the sequel. We will see that *eight* of the original thirty two functions are superfluous.

Some feature ranking functions we study only take on non-negative values, while others take on both positive and negative values. In the later case we will study the function and its absolute value. We will study the absolute value of *eight* of the original functions, which restores the total number of functions that will be considered in the sequel to *thirty two*.

To simplify the presentation, we drop the subscripts on a_j, b_j, c_j, d_j, in the remainder of this section, however, the resulting a, b, c, d will still be considered to be associated with a single feature $j \in V$.

μ_0. The function, μ_0 simply returns a random number sampled from the continuous uniform distribution U(0,1). It is included as a baseline for comparison with other functions.

μ_1. The functions, $\widehat{\mu}_1$, μ_1, μ_2 and μ_3, either directly correspond to, or are motivated by the term weighting functions F1[2], F2, F3 and F4 that were introduced in [68]. The function F1 is defined as

$$F1(a,b,c,d) = \frac{\frac{a}{a+c}}{\frac{a+b}{a+b+c+d}} = \frac{\frac{a}{|T|}}{\frac{a+b}{|T|+|F|}} = \frac{|T|+|F|}{|T|}\frac{a}{a+b} \tag{3.38}$$

and quoting from [68] "represents the ratio of the proportion of relevant documents in which t occurs to the proportion of the entire collection in which it occurs". Since for a given topic, $\big(|T|+|F|\big)/|T|$ is constant and therefore monotonically transforms $a/a+b$, rather than using F1 directly we shall use

$$\mu_1(a,b,c,d) = \mu_1(a,b) = \frac{a}{a+b} \tag{3.39}$$

$$\mu_1(x,y) = \frac{|T|x}{|T|x+|F|y} \tag{3.40}$$

since per (3.10) it has the theoretically desirable interpretation of being equal to the $\mathrm{P}(u \in T \mid u_j = 1)$. It is also the *precision* as discussed in (3.22).

$\widehat{\mu}_1$. The function $\widehat{\mu}_1$ is the term weighting function F2. It is defined as

$$\widehat{\mu}_1(a,b,c,d) = \widehat{\mu}_1(a,b) = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{a}{|T|}}{\frac{b}{|F|}} = \frac{|F|}{|T|}\frac{a}{b} \tag{3.41}$$

$$\widehat{\mu}_1(x,y) = \frac{x}{y}. \tag{3.42}$$

Quoting from [68], it "represents the ratio of the proportion of relevant documents to that of non-relevant documents". As discussed in §3.2, it can also be seen to be the ratio of the TPR to the FPR, as given in (3.5) and (3.6) which is the *positive likelihood ratio* as given in (3.7).

---

[2]It should be noted that μ_1 is not the F1 function that is the harmonic mean of the precision and the recall, from information retrieval. This function, appears later in this chapter as μ_{16}.

μ_2. The function μ_2 is the term weighting function F3. It is defined as

$$\begin{aligned}
\mu_2(a,b,c,d) &= \frac{\frac{a}{c}}{\frac{a+b}{c+d}} = \frac{a\ (c+d)}{c\ (a+b)} && (3.43)\\
\mu_2(a,b) &= \frac{a(|T|+|F|-(a+b))}{(|T|-a)(a+b)} && (3.44)\\
\mu_2(x,y) &= \frac{x\big(|T|(1-x)+|F|(1-y)\big)}{(1-x)(|T|x+|F|y)}. && (3.45)
\end{aligned}$$

Quoting from [68], it "represents the ratio between the 'relevance odds' for the feature (i.e. the ratio between the number of relevant documents in which it does occur and the number in which it does not occur) and the 'collection odds' for t".

μ_3. The function μ_3 is the term weighting function F4. It is defined as

$$\begin{aligned}
\mu_3(a,b,c,d) &= \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc} && (3.46)\\
\mu_3(a,b) &= \frac{a(|F|-b)}{b(|T|-a)} && (3.47)\\
\mu_3(a,b) &= \frac{x(1-y)}{y(1-x)}. && (3.48)
\end{aligned}$$

Quoting from [68], it "represents the ratio between the feature's relevance odds and its 'non-relevance' odds". As discussed in §3.2, it can also be seen to be the *odds ratio*, as given in (3.14).

$\widehat{\mu}_4$. The functions $\widehat{\mu}_4$, μ_4, μ_5 and μ_6 are defined as the difference between the numerator and denominator of the functions F1, F2, F3 and F4. The function $\widehat{\mu}_4$, is the difference between F1's numerator and denominator. It is defined as

$$\begin{aligned}
\widehat{\mu}_4(a,b,c,d) &= \frac{a}{|T|} - \frac{a+b}{|T|+|F|} = \frac{ad-bc}{|T|\,(|T|+|F|)} && (3.49)\\
\widehat{\mu}_4(a,b) &= \frac{a|F|-b|T|}{|T|\,(|T|+|F|)} && (3.50)\\
\widehat{\mu}_4(x,y) &= \frac{|F|\,(x-y)}{|T|+|F|} && (3.51)
\end{aligned}$$

Since $\widehat{\mu}_4$ satisfies (3.37), we shall study both $\widehat{\mu}_4$ and $|\widehat{\mu}_4|$.

μ_4. The function μ_4 is the difference between F2's numerator and denominator. It is

defined as

$$\begin{aligned}\mu_4(a,b,c,d) &= \frac{a}{|T|} - \frac{b}{|F|} = \frac{ad-bc}{|T||F|} && (3.52)\\ \mu_4(a,b) &= \frac{a|F|-b|T|}{|T||F|} && (3.53)\\ \mu_4(x,y) &= x-y && (3.54)\end{aligned}$$

It can be written as the difference between the TPR and the FPR or as the difference between positive incremental average Hamming distance and negative incremental average Hamming distance, and motivated by the later expression, we shall refer to it as the *Hamming difference*[3]. Since μ_4 satisfies (3.37), we shall study both μ_4 and $|\mu_4|$.

$\widetilde{\mu}_4$. The function is defined as

$$\begin{aligned}\widetilde{\mu}_4(a,b,c,d) &= \frac{ab+2ad+cd}{2|T||F|} && (3.55)\\ \widetilde{\mu}_4(a,b) &= \frac{a|F|-b|T|+|T||F|}{2|T||F|} && (3.56)\\ \widetilde{\mu}_4(x,y) &= \frac{x-y+1}{2}. && (3.57)\end{aligned}$$

It is the AUC_+ for the Boolean classifier g_j^+ defined in (3.25). The reader is referred to §3.6 for the definition of this classifier and the derivation of this function.

μ_4'. The function μ_4' is defined as

$$\begin{aligned}\mu_4'(a,b,c,d) &= \max\{\mathrm{AUC}_+, \mathrm{AUC}_-\}\\ &= \max\left\{\frac{ab+2ad+cd}{2|T||F|}, \frac{ab+2bc+cd}{2|T||F|}\right\} && (3.58)\\ \mu_4'(a,b) &= \max\left\{\frac{a|F|-b|T|+|T||F|}{2|T||F|}, \frac{-a|F|+b|T|+|T||F|}{2|T||F|}\right\} && (3.59)\\ \mu_4'(x,y) &= \max\left\{\frac{x-y+1}{2}, \frac{-x+y+1|}{2}\right\} && (3.60)\end{aligned}$$

where AUC_+ and AUC_- are the AUC for the Boolean classifiers g_j^+ and g_j^- defined in

[3]Note that there does not seem to be a commonly agreed upon name for this function, e.g. in [40], after taking the absolute value it is referred to as the *balanced accuracy*, while in [42], it is referred to as *percent difference*.

(3.25) and (3.26) respectively. It is the two-sided variant of $\widehat{\mu}_4$ since by the definition of the max it clearly gives high ranks to features that are very positive *or* very negative.

$\boldsymbol{\mu_5}$. Function μ_5 is the difference between F3's numerator and denominator. It is defined as

$$\begin{aligned}
\mu_5(a,b,c,d) &= \frac{a}{c} - \frac{a+b}{c+d} = \frac{ad-bc}{c(c+d)} && (3.61)\\
\mu_5(a,b) &= \frac{a|F| - b|T|}{(|T|-a)(|T|+|F|-(a+b))} && (3.62)\\
\mu_5(x,y) &= \frac{|F|(x-y)}{(1-x)\big(|T|(1-x)+|F|(1-y)\big)}. && (3.63)
\end{aligned}$$

Since μ_5 satisfies (3.37), we shall study both μ_5 and $|\mu_5|$.

$\boldsymbol{\mu_6}$. Function μ_6 is the difference between F4's numerator and denominator. It is defined as

$$\begin{aligned}
\mu_6(a,b,c,d) &= \frac{a}{c} - \frac{b}{d} = \frac{ad-bc}{cd} && (3.64)\\
\mu_6(a,b) &= \frac{a|F| - b|T|}{(|T|-a)(|F|-b))} && (3.65)\\
\mu_6(x,y) &= \frac{x-y}{(1-x)(1-y)} && (3.66)
\end{aligned}$$

Since μ_6 satisfies (3.37), we shall study both μ_6 and $|\mu_6|$.

$\boldsymbol{\mu_7}$. Function μ_7 is defined as

$$\begin{aligned}
\mu_7(a,b,c,d) &= \frac{a+d}{|T|+|F|} && (3.67)\\
\mu_7(a,b) &= \frac{a-b}{|T|+|F|} + \frac{|F|}{|T|+|F|} && (3.68)\\
\mu_7(x,y) &= \frac{|T|x+|F|(1-y)}{|T|+|F|} = \varrho x + (1-\varrho)(1-y) && (3.69)
\end{aligned}$$

it is the *accuracy* of the classifier g^+ defined in (3.25). Since $\widehat{\mu}_7$ satisfies (3.37), we shall study both $\widehat{\mu}_7$ and $|\widehat{\mu}_7|$.

$\boldsymbol{\widehat{\mu}_7}$. Function $\widehat{\mu}_7$ is defined as

$$\widehat{\mu}_7(a,b,c,d) = \frac{(a+d)-(b+c)}{|T|+|F|} \qquad (3.70)$$

$$\widehat{\mu}_7(a,b) = \frac{2a-2b}{|T|+|F|} - \frac{|T|-|F|}{|T|+|F|} \tag{3.71}$$

$$\widehat{\mu}_7(x,y) = \frac{|T|(2x-1)-|F|(2y-1)}{|T|+|F|}. \tag{3.72}$$

The function $\widehat{\mu}_7$ is the difference between the number of documents that are correctly classified by the classifier g^+ defined in (3.25) and the number of documents which were correctly classified by the classifier g^- defined in (3.26), normalized by the total number of documents. Since $\widehat{\mu}_7$ satisfies (3.37), we shall study both $\widehat{\mu}_7$ and $|\widehat{\mu}_7|$.

μ_8. The function μ_8 is defined as

$$\mu_8(a,b,c,d) = \frac{ad+bc}{(a+c)(b+d)} = \frac{ad+bc}{|T||F|} \tag{3.73}$$

$$\mu_8(a,b) = \frac{a|F|+b|T|-2ab}{|T|F|} \tag{3.74}$$

$$\mu_8(x,y) = x+y-2xy \tag{3.75}$$

and is the *incremental average Hamming distance.*

μ_9. For a document $u \in V$, let X to be the indicator random variable defined by

$$X = \begin{cases} 1 & \text{if } u_j = 1, \\ 0 & \text{otherwise} \end{cases}$$

and Y be the indicator random variable defined by

$$Y = \begin{cases} 1 & \text{if } u \in T, \\ 0 & \text{otherwise.} \end{cases}$$

The function μ_9 is the *Pearson Product Moment Correlation* coefficient or simply the *correlation coefficient* and is defined as

$$\mu_9(a,b,c,d) = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{3.76}$$

$$\mu_9(a,b) = \frac{a|F|-b|T|}{\sqrt{(a+b)(|T|+|F|-(a+b))|T||F|}} \tag{3.77}$$

$$\mu_9(x,y) = \frac{\sqrt{(|T||F|)}(x-y)}{\sqrt{(x|T|+y|F|)(|T|+|F|-(x|T|+y|F|))}}. \tag{3.78}$$

It measures the degree to which variables X and Y have a linear relationship. The reader is referred to the Appendix A for its derivation. Since μ_9 satisfies (3.37), we shall study both μ_9 and $|\mu_9|$.

$\widehat{\boldsymbol{\mu}}_{\mathbf{9}}$. The function $\widehat{\mu}_9$ is defined as

$$\widehat{\mu}_9(a,b,c,d) = \frac{(|T|+|F|)\ (ad-bc)^2}{|T||F|(a+b)(c+d)} \tag{3.79}$$

$$\widehat{\mu}_9(a,b) = \frac{(|T|+|F|)(a|F|-b|T|)^2}{|T||F|(a+b)(|T|+|F|-(a+b))}. \tag{3.80}$$

It is the χ^2 *statistic* for the variables X and Y as defined in the description of μ_9. It is a statistical measure of the association of X and Y.

$\boldsymbol{\mu}_{\mathbf{10}}$. The function μ_{10}, is defined as

$$\mu_{10}(a,b,c,d) = \frac{(|T|+|F|)(|ad-bc|-\frac{1}{2}(|T|+|F|))^2}{|T||F|(a+b)(c+d)} \tag{3.81}$$

$$\mu_{10}(a,b) = \frac{(|T|+|F|)(|T|+|F|-2|a|F|-b|T||)^2}{4(a+b)(|T|+|F|-(a+b))|T||F|} \tag{3.82}$$

It is *Yate's continuity correction of the χ^2 statistic* for the variables X and Y as defined in the description of μ_9 and the χ^2 statistic (see e.g. [29]).

$\boldsymbol{\mu}_{\mathbf{11}}$. The function μ_{11} is the *Gini split criterion.* It is defined as

$$\mu_{11}(a,b,c,d) = \frac{2}{m}\left[\frac{ab}{a+b}+\frac{cd}{c+d}\right] \tag{3.83}$$

$$\mu_{11}(a,b) = \frac{2}{(|T|+|F|)}\ \frac{(|T||F|a+|T||F|b-b^2|T|-a^2|F|)}{(a+b)(|T|+|F|-(a+b))} \tag{3.84}$$

$$\mu_{11}(x,y) = \frac{2|T||F|}{(|T|+|F|)}\ \frac{|T|x(1-x)+|F|y(1-y)}{(x|T|+y|F|)(|T|+|F|-(x|T|+y|F|))} \tag{3.85}$$

It is based on the Gini impurity and has been used as a split criterion in tree-based classification methods (see e.g. [14]). It is one of only two feature ranking functions (the other being μ_{21}) that we study for which smaller values are better. The reader is referred to Appendix B for its derivation.

μ_{12}. The function μ_{12} is the *information gain* and it is defined as follows

$$\mu_{12}(a,b,c,d) = H(a,b,c,d) - \left(\frac{a+b}{|T|+|F|}H_\theta(a,b,c,d) + \frac{c+d}{|T|+|F|}H_{\bar{\theta}}(a,b,c,d)\right) \quad (3.86)$$

where

$$H(a,b,c,d) = -\left(\frac{|T|}{|T|+|F|}\right)\log_2\left(\frac{|T|}{|T|+|F|}\right) - \left(\frac{|F|}{|T|+|F|}\right)\log_2\left(\frac{|F|}{|T|+|F|}\right),$$

$$H_\theta(a,b,c,d) = -\left(\frac{a}{a+b}\right)\log_2\left(\frac{a}{a+b}\right) - \left(\frac{b}{a+b}\right)\log_2\left(\frac{b}{a+b}\right),$$

and

$$H_{\bar{\theta}}(a,b,c,d) = -\left(\frac{c}{c+d}\right)\log_2\left(\frac{c}{c+d}\right) - \left(\frac{d}{c+d}\right)\log_2\left(\frac{d}{c+d}\right).$$

It is an entropy-based measure and has been used as a split criterion in tree-based classification methods (see e.g. [14]). The reader is referred to Appendix C for its derivation.

μ_{13}. The function μ_{13} is defined as

$$\mu_{13}(a,b,c,d) = \frac{ad}{|T||F|} \quad (3.87)$$

$$\mu_{13}(a,b) = \frac{a(|F|-b)}{|T||F|} \quad (3.88)$$

$$\mu_{13}(x,y) = x(1-y) \quad (3.89)$$

and is the *positive incremental average Hamming distance*.

μ_{14}. The function μ_{14} is defined as

$$\mu_{14}(a,b,c,d) = \frac{bc}{|T||F|} \quad (3.90)$$

$$\mu_{14}(a,b) = \frac{b(|T|-a)}{|T||F|} \quad (3.91)$$

$$\mu_{14}(x,y) = y(1-x) \quad (3.92)$$

and is the *negative incremental Hamming distance*.

μ_{15}. The function μ_{15} is defined as

$$\mu_{15}(a,b,c,d) = \frac{4a}{4a+b+3c} \tag{3.93}$$

$$\mu_{15}(a,b) = \frac{4a}{a+b+3|T|} \tag{3.94}$$

$$\mu_{15}(x,y) = \frac{4|T|x}{x|T|+y|F|+3|T|}. \tag{3.95}$$

As discussed in §3.5, it is the information retrieval metric F_α with parameter $\alpha = \frac{1}{4}$.

μ_{16}. The function μ_{16} is defined as

$$\mu_{16}(a,b,c,d) = \frac{2a}{2a+b+c} \tag{3.96}$$

$$\mu_{16}(a,b) = \frac{2a}{a+b+|T|} \tag{3.97}$$

$$\mu_{16}(x,y) = \frac{2|T|x}{|T|x+|F|y+|T|} \tag{3.98}$$

As discussed in §3.5, it is the information retrieval metric F_α with parameter $\alpha = \frac{1}{2}$.

μ_{17}. The function μ_{17} is defined as

$$\mu_{17}(a,b,c,d) = \frac{4a}{4a+3b+c} \tag{3.99}$$

$$\mu_{17}(a,b) = \frac{4a}{3a+3b+|T|} \tag{3.100}$$

$$\mu_{17}(x,y) = \frac{4|T|x}{3|T|x+3|F|y+|T|} \tag{3.101}$$

As discussed in §3.5, it is the information retrieval metric F_α with parameter $\alpha = \frac{3}{4}$.

μ_{18}. The function μ_{18} is defined as

$$\mu_{18}(a,b,c,d) = \mu_{18}(a,b) = \frac{a}{|T|} \tag{3.102}$$

$$\mu_{18}(x,y) = x \tag{3.103}$$

and is the *true positive rate* as discussed in §1.3, the $\mathrm{P}(u_j = 1 \mid u \in T)$ as given in (3.5), and the *recall* as given in (3.23).

$\widehat{\boldsymbol{\mu}}_{\mathbf{18}}$**.** The function $\widehat{\mu}_{18}$ is defined as

$$\widehat{\mu}_{18}(a,b,c,d) = \widehat{\mu}_{18}(a,b) = a \tag{3.104}$$

$$\widehat{\mu}_{18}(x,y) = |T|x \tag{3.105}$$

and is simply the number of relevant documents containing the given feature.

$\boldsymbol{\mu}_{\mathbf{19}}$**.** The function μ_{19} is defined as

$$\mu_{19}(a,b,c,d) = \mu_{19}(a,b) = \frac{b}{|F|} \tag{3.106}$$

$$\mu_{19}(x,y) = y \tag{3.107}$$

and is the *false positive rate* as discussed in §1.3 and the $\mathrm{P}(u_j = 1 \mid u \in F)$ as given in (3.6).

$\widehat{\boldsymbol{\mu}}_{\mathbf{19}}$**.** The function $\widehat{\mu}_{19}$ is defined as

$$\widehat{\mu}_{19}(a,b,c,d) = \widehat{\mu}_{19}(a,b) = b \tag{3.108}$$

$$\widehat{\mu}_{19}(a,b,c,d) = |F|y \tag{3.109}$$

and is simply the number of irrelevant documents containing the given feature.

$\boldsymbol{\mu}_{\mathbf{20}}$**.** The function, μ_{20} is another measure of association and is referred to as *Yule's Q* and is defined as

$$\mu_{20}(a,b,c,d) = \frac{ad-bc}{ad+bc} \tag{3.110}$$

$$\mu_{20}(a,b) = \frac{a|F|-b|T|}{a|F|+b|T|-2ab} \tag{3.111}$$

$$\mu_{20}(x,y) = \frac{x-y}{x+y-2xy}. \tag{3.112}$$

Note that it is the ratio of μ_4 to μ_8. Since μ_{20} satisfies (3.37), we shall study both μ_{20} and $|\mu_{20}|$.

$\boldsymbol{\mu}_{\mathbf{21}}$**.** The function, μ_{21} is a new measure that we introduce and refer to as the *rareness*.

It is defined as

$$\mu_{21}(a,b,c,d) = \frac{|T|!\ |F|!}{a!\ b!\ c!\ d!}\left(\frac{a+b}{|T|+|F|}\right)^{a+b}\left(\frac{c+d}{|T|+|F|}\right)^{c+d} \tag{3.113}$$

$$\mu_{21}(a,b) = \frac{|T|!\ |F|!}{a!\ b!\ (|T|-a)!\ (|F|-b)!}\left(\frac{a+b}{|T|+|F|}\right)^{a+b}\left(\frac{|T|+|F|-(a+b)}{|T|+|F|}\right)^{m-(a+b)}. \tag{3.114}$$

Based on the assumption that relevant and irrelevant documents are "generated" by two independent stochastic processes, it provides a measure of how unlikely a feature is to occur. The reader is referred to Appendix D for its derivation. It is one of only two feature ranking functions (the other being μ_{11}) that we study for which smaller values are better.

μ_{22}. The function, μ_{22} is defined as

$$\mu_{22}(a,b,c,d) = \frac{\frac{a}{|T|}-\frac{b}{|F|}}{\frac{a}{|T|}+\frac{b}{|F|}} = \frac{ad-bc}{ad+bc+2ab} \tag{3.115}$$

$$\mu_{22}(a,b) = \frac{a|F|-b|T|}{a|F|+b|T|} \tag{3.116}$$

$$\mu_{22}(x,y) = \frac{x-y}{x+y}. \tag{3.117}$$

Since μ_{22} satisfies (3.37), we shall study both μ_{22} and $|\mu_{22}|$.

μ_{23}. The function, μ_{23} is *Fisher's linear discriminant*. It is defined as

$$\mu_{23}(a,b,c,d) = \frac{\left(\frac{a}{|T|}-\frac{b}{|F|}\right)^2}{\frac{a}{|T|}\left(1-\frac{a}{|T|}\right)+\frac{b}{|F|}\left(1-\frac{b}{|F|}\right)} \tag{3.118}$$

$$\mu_{23}(x,y) = \frac{(x-y)^2}{x(1-x)+y(1-y)}. \tag{3.119}$$

The reader is referred to Appendix E for its derivation.

μ_{24}. The function, is similar to Fisher's linear discriminant. It is defined as

$$\mu_{24}(a,b,c,d) = \frac{\frac{a}{|T|}-\frac{b}{|F|}}{\sqrt{\frac{a}{|T|}\left(1-\frac{a}{|T|}\right)}+\sqrt{\frac{b}{|F|}\left(1-\frac{b}{|F|}\right)}} \tag{3.120}$$

$$\mu_{24}(x,y) = \frac{x-y}{\sqrt{x(1-x)}+\sqrt{y(1-y)}}. \tag{3.121}$$

The reader is referred to Appendix E for its derivation. Since μ_{24} satisfies (3.37), we shall study both μ_{24} and $|\mu_{24}|$.

3.9 Properties of and Relationships Between Ranking Functions

In this section we state some observations about the feature ranking functions introduced in §3.8.

The following lemma shows that the functions μ_1, $\widehat{\mu}_1$, μ_2 and μ_3 all involve a ratio of ad to bc or similarly a ratio of $a|F|$ to $b|T|$.

Proposition 3.3. *Each of the functions* $\frac{|F|}{|T|}\mu_1$, $\widehat{\mu}_1$, μ_2 *and* μ_3 *can be written in the form*

$$\frac{ad+\gamma}{bc+\gamma}$$

where γ *is a function of a,b,c and d and in the form*

$$\frac{a|F|-ab+\delta}{b|T|-ab+\delta}$$

where δ *is a function of a and b.*

Proof. Let γ_1 be the function for $\frac{|F|}{|T|}\mu_1$, $\widehat{\gamma}_1$ be the function for $\widehat{\mu}_1$, and γ_2 and γ_3 be the functions for μ_2 and μ_3 respectively. By elementary calculations it can be shown that

$$\begin{aligned}
\gamma_1 &= a^2+ab+ac \\
\widehat{\gamma}_1 &= ab \\
\gamma_2 &= ac \\
\gamma_3 &= 0.
\end{aligned}$$

Let δ_1 be the function for $\frac{|F|}{|T|}\mu_1$, $\widehat{\delta}_1$ be the function for $\widehat{\mu}_1$, and δ_2 and δ_3 be the functions

for μ_2 and μ_3 respectively. By elementary calculations it can be shown that

$$\begin{aligned}
\delta_1 &= a|T| + ab \\
\widehat{\delta}_1 &= ab \\
\delta_2 &= a|T| - a^2 \\
\delta_3 &= 0.
\end{aligned}$$

■

The following lemma shows that the functions μ_4, $\widehat{\mu}_4$, $\widetilde{\mu}_4$, μ_4', μ_5 and μ_6 all include the difference of ad and bc or similarly the difference of $a|F|$ and $b|T|$.

Proposition 3.4. *Each of the functions μ_4, $\widehat{\mu}_4$, $\widetilde{\mu}_4$, μ_4', μ_5 and μ_6 can be written in the form*

$$\frac{ad - bc}{\gamma}$$

where γ is a function of a,b,c and d and in the form

$$\frac{a|F| - b|T|}{\delta}$$

where δ is a function of a and b.

Proof. Let $\widehat{\gamma}_4$ be the function γ for $\widehat{\mu}_4$, $\widetilde{\gamma}_4$ be the function γ for $\widetilde{\mu}_4$, and γ_i be the function γ for μ_i for $i \in \{4, 5, 6\}$. By elementary calculations it can be shown that

$$\begin{aligned}
\widehat{\gamma}_4 &= (a+c)(a+b+c+d) = m|T| \\
\gamma_4 &= (a+c)(b+d) = |T||F| \\
\widetilde{\gamma}_4 &= 2(a+c)(b+d) = 2|T||F| \\
\gamma_5 &= c(c+d) \\
\gamma_6 &= cd
\end{aligned}$$

Let $\widehat{\delta}_4$ be the function δ for $\widehat{\mu}_4$, $\widetilde{\delta}_4$ be the function δ for $\widetilde{\mu}_4$, and δ_i be the function δ for

μ_i for $i \in \{4, 5, 6\}$. By elementary calculations it can be shown that

$$\begin{aligned}
\widehat{\delta}_4 &= m|T| \\
\delta_4 &= |T||F| \\
\widetilde{\delta}_4 &= 2|T||F| \\
\delta_5 &= (|T| - a)^2 + (|T| - a)(|F| - b) \\
\delta_6 &= (|T| - a)(|F| - b)
\end{aligned}$$

■

The following result shows that all but two of the feature ranking functions introduced in §3.8 can be written in a common form.

Proposition 3.5. *Each of the feature ranking functions introduced in §3.8, with the exception of the information gain, μ_{12} and the rareness, μ_{21}, can be written as, or only differs in a monotonic transformation from, a quadratic rational function of a and b.*

The fact that feature ranking functions can be written as the ratio of two quadratic degree two polynomials in a and b will be used extensively in the sequel.

We now state several lemmas that show which feature ranking functions are related by a monotonic transformation.

Proposition 3.6. *The functions μ_1 and $\widehat{\mu}_1$ differ only in a monotonic transformation.*

Proof. The factor $\frac{|F|}{|T|}$ in $\widehat{\mu}_1$ is a constant and can therefore be eliminated, leaving

$$\mu_1(a, b) = \frac{a}{a + b} \quad \text{and} \quad \widehat{\mu}_1(a, b) \sim \frac{a}{b}. \tag{3.122}$$

Next we recall (see [7]) that for any $w, x, y, z \in \mathbb{R} \geq 0$, with not both w and x being zero, and not both y and z being zero

$$\frac{w}{x} \leq \frac{y}{z} \Leftrightarrow \frac{w}{w + x} \leq \frac{y}{y + z}. \tag{3.123}$$

That μ_1 and $\widehat{\mu}_1$ only differ in a monotonic transformation now follows from (3.123) and the fact that it is not possible to have both $a = 0$ and $b = 0$ since we assume that each feature appears in at least one document. ■

Proposition 3.7. *The functions $\widehat{\mu}_4$, μ_4 and $\widetilde{\mu}_4$ differ only in a monotonic transformation.*

Proof. Since m and $|F|$ are constants we have

$$\widehat{\mu}_4(a, b) = \frac{|F|}{m}\, \mu_4(a, b)$$

and by writing $\widetilde{\mu}_4$ as

$$\widetilde{\mu}_4(a, b) = \frac{a|F| - b|T|}{2|T||F|} + \frac{|T|F|}{2|T||F|}$$

we see that

$$\widetilde{\mu}_4(a, b) = \frac{1}{2}\mu_4(a, b) + \frac{1}{2}. \tag{3.124}$$

■

Proposition 3.8. *The functions $|\mu_4|$ and μ_4' only differ in a monotonic transformation.*

Proof. Recall the following relationship between the maximum and the absolute value (see [26]). For any $x, y \in \mathbb{R}$

$$\max(x, y) = \frac{x + y + |x - y|}{2}. \tag{3.125}$$

Using (3.125), we can write μ_4' as

$$\begin{aligned}
\max(\text{AUC}_+, \text{AUC}_-) &= \max\left(\frac{ab + 2ad + cd}{2|T||F|}, \frac{ab + 2bc + cd}{2|T||F|}\right) \\
&= \frac{1}{2}\left(\frac{ab + 2ad + cd}{2|T||F|} + \frac{ab + 2bc + cd}{2|T||F|}\right. \\
&\quad \left. + \left|\frac{ab + 2ad + cd}{2|T||F|} - \frac{ab + 2bc + cd}{2|T||F|}\right|\right) \\
&= \frac{1}{2}\left(\frac{2(a + c)(b + d)}{2|T||F|} + \left|\frac{2(ad - bc)}{2|T||F|}\right|\right) \\
&= \frac{1}{2} + \frac{1}{2}\left|\mu_4\right|. \qquad (3.126)
\end{aligned}$$

■

In view of (3.124) and (3.126), we can interpret $\frac{1}{2}|\mu_4|$, as shown in Figure 3.7, to be the probability above or below random guessing (i.e. above or below $\frac{1}{2}$), that the score the classifier g_j^+ defined in (3.25) or the classifier g_j^- defined (3.26) for $j \in V$, assigns to a randomly selected relevant document will be larger than the score it assigns to a randomly selected irrelevant document.

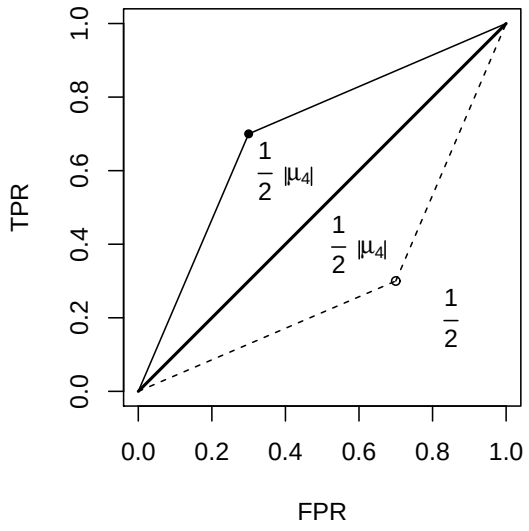


Figure 3.7: Boolean ROC Curve and μ_4

Proposition 3.9. *The functions μ_7 and $\widehat{\mu}_7$ differ only in a monotonic transformation.*

Proof. Follows from the fact that after eliminating the last terms in (3.68) and (3.71), which are constants, μ_7 and $\widehat{\mu}_7$ only differ by a factor of 2. ■

Proposition 3.10. *The functions μ_9 and $\widehat{\mu}_9$ differ only in a monotonic transformation.*

Proof. Follows from the fact that $\widehat{\mu}_9 = m\mu_9^2$. ■

Proposition 3.11. *The functions μ_{18} and $\widehat{\mu}_{18}$ differ only in a monotonic transformation.*

Proof. Follows from the fact that $\widehat{\mu}_{18} = \mu_{18}|T|$. ■

Proposition 3.12. *The functions μ_{19} and $\widehat{\mu}_{19}$ differ only in a monotonic transformation.*

Proof. Follows from the fact that $\widehat{\mu}_{19} = \mu_{19}|F|$. ■

As a result of Proposition 3.6, Proposition 3.7, Proposition 3.8, Proposition 3.9, Proposition 3.10, Proposition 3.11, and Proposition 3.12 we will not refer to $\widehat{\mu}_1$, $\widehat{\mu}_4$, $\widetilde{\mu}_4$, μ'_4, $\widehat{\mu}_7$, $\widehat{\mu}_9$, $\widehat{\mu}_{18}$, or $\widehat{\mu}_{19}$ in the sequel.

We now state several results related to the feature ranking function μ_4.

Proposition 3.13. $AUC_+ - AUC_- = \frac{U_+}{|T||F|} - \frac{U_-}{|T||F|} = \mu_4$.

Proof. Follows from elementary calculations used in the proof of Proposition 3.2. ■

Proposition 3.14. $\max(AUC_+, AUC_-) = AUC_+$ *if and only if* $\frac{a}{|T|} - \frac{b}{|F|} = \mu_4 > 0$.

Proof. Assume that $\max(\mathrm{AUC}_+, \mathrm{AUC}_-) = \mathrm{AUC}_+$. This is equivalent to

$$\mathrm{AUC}_+ - \mathrm{AUC}_- > 0,$$

from which we have

$$\frac{ab + 2ad + cd}{2|T||F|} - \frac{ab + 2bc + cd}{2|T||F|} > 0 \Leftrightarrow \frac{2(ad - bc)}{2|T||F} > 0 \Leftrightarrow \mu_4 > 0.$$

Now assume that $\frac{a}{|T|} - \frac{b}{|F|} > 0$. Then we have

$$\frac{a}{|T|} - \frac{b}{|F|} = \left| \frac{a}{|T|} - \frac{b}{|F|} \right| = |\mu_4| > 0.$$

Therefore, it is also the case that $\frac{1}{2} + \frac{1}{2}|\mu_4| > 0$ and the result follows from Proposition 3.8. ■

Corollary 3.2. *Consider a feature $j \in V$. If*

$$\max(AUC_+, AUC_-) = AUC_+$$

then j is a positive feature and if

$$\max(AUC_+, AUC_-) = AUC_-$$

then j is a negative feature.

This result shows the equivalence of two characterizations of negative and positive features, namely that based on μ_4 given by Corollary 3.1 and that given by μ_4' .

3.10 Boolean Feature Ranking Results

In this section, we use the methodology presented in §2 to evaluate the relative performance of μ-RANKING for each of the feature ranking functions discussed in §3.8. The detailed results are provided in Appendix H. Table 3.2 summarizes the results by listing seven orderings of the feature ranking functions based on the non-discounted measures in $\mathcal{C}^*$. Based on the data in this table we make the following observations

- The top of the lex-max-min and avg multicriteria rankings included many of the classic feature ranking functions such as the correlation coefficient, the entropy, F_α, Fisher's linear discriminant and its variants, and the Gini criteria.
- The new rareness measure μ_{21}, also appeared near the top of the lex-max-min and avg multicriteria rankings.
- The feature ranking functions appearing at the very top of the lex multicriteria ranking, in which $\widehat{\nu}$ was the most important criteria, included functions that generally did not perform well with respect to $\widehat{\sigma}$ and $\widehat{\theta}$.
- The performance of feature ranking functions and their absolute values was not substantially different.
- The feature ranking functions ranked highest in terms of $\widehat{\sigma}$, notably μ_8, the incremental average Hamming distance, were among the worst performing features ranking functions as measured by $\widehat{\nu}$.

We reiterate that our goal is *not* to determine which feature ranking functions exhibit the best or worst performance with μ-RANKING, but rather to identify the characteristics of feature ranking functions that exhibit good as well as bad performance when used in this algorithm and we will pursue this goal further in §4.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
μ_{23}	μ_2	μ_{23}	μ_1	0.00	μ_8	0.99	μ_8	0.54	μ_2	0.90
$\lvert\mu_4\rvert$	μ_3	μ_{12}	μ_2	0.00	$\lvert\mu_4\rvert$	0.98	μ_{13}	0.49	μ_7	0.90
μ_{15}	μ_{20}	$\lvert\mu_{24}\rvert$	μ_3	0.00	μ_{13}	0.98	$\lvert\mu_4\rvert$	0.46	$\lvert\mu_7\rvert$	0.90
$\lvert\mu_{24}\rvert$	$\lvert\mu_{20}\rvert$	μ_{21}	μ_{20}	0.00	μ_{23}	0.98	μ_5	0.45	μ_3	0.89
μ_4	μ_1	μ_{24}	$\lvert\mu_{20}\rvert$	0.00	μ_4	0.97	μ_6	0.45	μ_9	0.89
μ_{21}	μ_{22}	μ_{10}	μ_{22}	0.00	μ_{12}	0.97	μ_4	0.43	$\lvert\mu_9\rvert$	0.89
μ_{24}	$\lvert\mu_{22}\rvert$	$\lvert\mu_4\rvert$	$\lvert\mu_{22}\rvert$	0.00	μ_{15}	0.97	μ_{23}	0.43	μ_{10}	0.89
μ_{12}	μ_{12}	μ_9	μ_7	0.01	μ_{16}	0.97	$\lvert\mu_5\rvert$	0.41	μ_{11}	0.89
μ_{16}	μ_{10}	$\lvert\mu_9\rvert$	$\lvert\mu_7\rvert$	0.01	μ_{21}	0.97	$\lvert\mu_6\rvert$	0.41	μ_{12}	0.89
μ_{10}	μ_9	μ_{11}	μ_9	0.01	μ_{24}	0.97	μ_{15}	0.41	μ_{17}	0.89
μ_9	$\lvert\mu_9\rvert$	μ_{16}	$\lvert\mu_9\rvert$	0.01	$\lvert\mu_{24}\rvert$	0.97	μ_{18}	0.41	μ_{20}	0.89
$\lvert\mu_9\rvert$	μ_{11}	μ_{17}	μ_{10}	0.01	μ_5	0.96	μ_{14}	0.40	$\lvert\mu_{20}\rvert$	0.89
μ_{11}	μ_{17}	μ_2	μ_{11}	0.01	μ_6	0.96	$\lvert\mu_{24}\rvert$	0.40	μ_{16}	0.88
μ_{17}	μ_7	μ_7	μ_{12}	0.01	μ_9	0.96	μ_{21}	0.38	μ_{21}	0.88
μ_2	$\lvert\mu_7\rvert$	$\lvert\mu_7\rvert$	μ_{17}	0.01	$\lvert\mu_9\rvert$	0.96	μ_{24}	0.38	μ_{23}	0.88
μ_7	μ_{21}	μ_{15}	μ_{21}	0.02	μ_{10}	0.96	μ_{12}	0.37	$\lvert\mu_{24}\rvert$	0.88
$\lvert\mu_7\rvert$	μ_{24}	μ_4	μ_{24}	0.02	μ_{11}	0.96	μ_{16}	0.37	μ_1	0.87
μ_3	$\lvert\mu_{24}\rvert$	μ_3	$\lvert\mu_{24}\rvert$	0.03	μ_{14}	0.96	μ_{19}	0.36	μ_{22}	0.87
μ_{20}	μ_0	μ_{20}	μ_0	0.04	μ_{17}	0.96	μ_{10}	0.35	$\lvert\mu_{22}\rvert$	0.87
$\lvert\mu_{20}\rvert$	μ_{23}	$\lvert\mu_{20}\rvert$	μ_{16}	0.05	μ_2	0.95	μ_9	0.34	μ_{24}	0.87
μ_1	μ_{16}	μ_1	μ_{23}	0.05	μ_7	0.95	$\lvert\mu_9\rvert$	0.34	$\lvert\mu_4\rvert$	0.86
μ_{22}	$\lvert\mu_4\rvert$	μ_{22}	$\lvert\mu_4\rvert$	0.12	$\lvert\mu_7\rvert$	0.95	μ_{11}	0.34	μ_{15}	0.86
$\lvert\mu_{22}\rvert$	μ_{15}	$\lvert\mu_{22}\rvert$	μ_{15}	0.13	μ_3	0.92	μ_{17}	0.33	μ_4	0.84
μ_{13}	μ_4	μ_{13}	μ_4	0.14	μ_{20}	0.92	μ_2	0.29	μ_{13}	0.79
μ_5	μ_{14}	μ_8	μ_{14}	0.32	$\lvert\mu_{20}\rvert$	0.92	μ_7	0.29	μ_8	0.78
μ_6	μ_{13}	μ_5	μ_{13}	0.37	$\lvert\mu_5\rvert$	0.89	$\lvert\mu_7\rvert$	0.29	μ_5	0.77
$\lvert\mu_5\rvert$	μ_5	μ_6	μ_5	0.38	$\lvert\mu_6\rvert$	0.89	μ_3	0.24	μ_6	0.77
$\lvert\mu_6\rvert$	μ_6	$\lvert\mu_5\rvert$	μ_6	0.38	μ_{18}	0.88	μ_{20}	0.24	$\lvert\mu_5\rvert$	0.74
μ_8	$\lvert\mu_5\rvert$	$\lvert\mu_6\rvert$	$\lvert\mu_5\rvert$	0.41	μ_1	0.84	$\lvert\mu_{20}\rvert$	0.24	$\lvert\mu_6\rvert$	0.73
μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	$\lvert\mu_6\rvert$	0.41	μ_{19}	0.84	μ_1	0.20	μ_{18}	0.68
μ_{18}	μ_8	μ_{18}	μ_8	0.50	μ_{22}	0.84	μ_{22}	0.20	μ_{14}	0.35
μ_{19}	μ_{18}	μ_{19}	μ_{18}	0.59	$\lvert\mu_{22}\rvert$	0.84	$\lvert\mu_{22}\rvert$	0.20	μ_{19}	0.27
μ_0	μ_{19}	μ_0	μ_{19}	0.59	μ_0	0.15	μ_0	0.01	μ_0	0.06

Table 3.2: μ-RANKING Not Discounted Stopwords Included

Chapter 4

Separation and Noise

Notable in our discussion of the μ-RANKING experimental results in §3.10 was the fact that the feature ranking function, namely μ_8, whose selected feature sets achieved the best separation, was also among those that contained the most noise. In this chapter we will use this characteristic of μ_8 to study the relationship between separation and noise and the impact it has on feature selection algorithms. These studies will involve consideration of several assertions which we support by reviewing the results of experiments that employ the feature set evaluation methodology from §2. In addition, so that we can consider specific feature sets, we will also offer anecdotal evidence associated with the *fuel* topic to support various assertions. When considering such evidence, it should be mentioned that the associated results will be computed using all documents associated with this topic and will not employ cross fold validation.

4.1 Noise Separates

In this section we assert that the reason that the feature sets selected by μ_8 achieve the best separation is because they include stopwords. To support this assertion, we will use the discounted variants of the measures in $\mathcal{C}^*$ to assess the performance of μ-RANKING for each feature ranking function. These results are listed in detail in Appendix I, and are summarized in Table 4.1.

For the training folds, we note that $\widehat{\sigma} = 0.54$ and $\sigma_K = 12.43$ drop to $\widehat{\sigma}\,' = 0.29$ and $\sigma_K' = 5.71$ which represent declines of 46% and 54% respectively. Further, the value of σ_K' is actually below the corresponding mean and median of this measure which are 6.25 and 6.62 respectively. These observations support the assertion that the reason that the feature sets selected by μ_8 achieve the best separation is because they include

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}'$		$\widehat{\sigma}'$		$\bar{\varphi}$	
μ_{21}	μ_2	μ_{12}	μ_1	0.00	$\lvert\mu_4\rvert$	0.98	$\lvert\mu_4\rvert$	0.41	μ_2	0.90
μ_{24}	μ_3	μ_{23}	μ_2	0.00	μ_{23}	0.98	μ_{23}	0.40	μ_7	0.90
$\lvert\mu_{24}\rvert$	μ_{20}	$\lvert\mu_{24}\rvert$	μ_3	0.00	μ_4	0.97	$\lvert\mu_{24}\rvert$	0.39	$\lvert\mu_7\rvert$	0.90
μ_{12}	$\lvert\mu_{20}\rvert$	μ_{21}	μ_{20}	0.00	μ_8	0.97	μ_4	0.37	μ_3	0.89
μ_{23}	μ_1	μ_{24}	$\lvert\mu_{20}\rvert$	0.00	μ_{12}	0.97	μ_{12}	0.37	μ_9	0.89
μ_{16}	μ_{22}	μ_9	μ_{22}	0.00	μ_{21}	0.97	μ_{21}	0.37	$\lvert\mu_9\rvert$	0.89
μ_{10}	$\lvert\mu_{22}\rvert$	$\lvert\mu_9\rvert$	$\lvert\mu_{22}\rvert$	0.00	μ_{24}	0.97	μ_{24}	0.37	μ_{10}	0.89
μ_9	μ_{12}	μ_{10}	μ_7	0.01	$\lvert\mu_{24}\rvert$	0.97	μ_{15}	0.35	μ_{11}	0.89
μ_{11}	μ_9	μ_{11}	$\lvert\mu_7\rvert$	0.01	μ_9	0.96	μ_{16}	0.35	μ_{12}	0.89
$\lvert\mu_9\rvert$	$\lvert\mu_9\rvert$	μ_{17}	μ_9	0.01	$\lvert\mu_9\rvert$	0.96	μ_9	0.34	μ_{17}	0.89
μ_{17}	μ_{10}	μ_2	$\lvert\mu_9\rvert$	0.01	μ_{10}	0.96	$\lvert\mu_9\rvert$	0.34	μ_{20}	0.89
μ_2	μ_{11}	μ_{16}	μ_{10}	0.01	μ_{11}	0.96	μ_{10}	0.34	$\lvert\mu_{20}\rvert$	0.89
μ_7	μ_{17}	$\lvert\mu_4\rvert$	μ_{11}	0.01	μ_{13}	0.96	μ_{11}	0.34	μ_{16}	0.88
$\lvert\mu_7\rvert$	μ_7	μ_7	μ_{12}	0.01	μ_{15}	0.96	μ_{17}	0.33	μ_{21}	0.88
$\lvert\mu_4\rvert$	$\lvert\mu_7\rvert$	$\lvert\mu_7\rvert$	μ_{17}	0.01	μ_{16}	0.96	μ_{13}	0.31	μ_{23}	0.88
μ_{15}	μ_{21}	μ_3	μ_{21}	0.02	μ_{17}	0.96	μ_5	0.30	μ_{24}	0.88
μ_4	μ_{24}	μ_{20}	μ_{24}	0.02	μ_2	0.95	μ_6	0.30	$\lvert\mu_{24}\rvert$	0.88
μ_3	$\lvert\mu_{24}\rvert$	$\lvert\mu_{20}\rvert$	$\lvert\mu_{24}\rvert$	0.03	μ_7	0.95	μ_2	0.29	μ_1	0.87
μ_{20}	μ_0	μ_4	μ_0	0.04	$\lvert\mu_7\rvert$	0.95	μ_7	0.29	$\lvert\mu_4\rvert$	0.87
$\lvert\mu_{20}\rvert$	μ_{23}	μ_{15}	μ_{16}	0.05	μ_5	0.93	$\lvert\mu_7\rvert$	0.29	μ_{15}	0.87
$\lvert\mu_{22}\rvert$	μ_{16}	μ_1	μ_{23}	0.05	μ_6	0.93	μ_8	0.29	μ_{22}	0.87
μ_1	$\lvert\mu_4\rvert$	μ_{22}	$\lvert\mu_4\rvert$	0.12	μ_{14}	0.93	$\lvert\mu_5\rvert$	0.28	$\lvert\mu_{22}\rvert$	0.87
μ_{22}	μ_{15}	$\lvert\mu_{22}\rvert$	μ_{15}	0.13	μ_3	0.92	$\lvert\mu_6\rvert$	0.28	μ_4	0.86
μ_{13}	μ_4	μ_{13}	μ_4	0.14	μ_{20}	0.92	μ_{14}	0.26	μ_{13}	0.85
μ_5	μ_{14}	μ_5	μ_{14}	0.32	$\lvert\mu_{20}\rvert$	0.92	μ_3	0.24	μ_5	0.82
μ_6	μ_{13}	μ_6	μ_{13}	0.37	$\lvert\mu_5\rvert$	0.87	μ_{20}	0.24	μ_6	0.82
$\lvert\mu_5\rvert$	μ_5	μ_8	μ_5	0.38	$\lvert\mu_6\rvert$	0.87	$\lvert\mu_{20}\rvert$	0.24	μ_8	0.81
$\lvert\mu_6\rvert$	μ_6	$\lvert\mu_5\rvert$	μ_6	0.38	μ_1	0.84	μ_1	0.20	$\lvert\mu_5\rvert$	0.75
μ_8	$\lvert\mu_5\rvert$	$\lvert\mu_6\rvert$	$\lvert\mu_5\rvert$	0.41	μ_{22}	0.84	μ_{18}	0.20	$\lvert\mu_6\rvert$	0.74
μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	$\lvert\mu_6\rvert$	0.41	$\lvert\mu_{22}\rvert$	0.84	μ_{22}	0.20	μ_{18}	0.71
μ_{18}	μ_8	μ_{18}	μ_8	0.50	μ_{18}	0.82	$\lvert\mu_{22}\rvert$	0.20	μ_{14}	0.29
μ_{19}	μ_{18}	μ_{19}	μ_{18}	0.59	μ_{19}	0.75	μ_{19}	0.15	μ_{19}	0.17
μ_0	μ_{19}	μ_0	μ_{19}	0.59	μ_0	0.12	μ_0	0.01	μ_0	0.05

Table 4.1: μ-RANKING Discounted Stopwords Included

stopwords.

In Proposition 2.5 we saw that the average Hamming distance of the set of selected features, S can be computed as the sum of the incremental average Hamming distance, μ_8, of each feature in S. Therefore, we can compute, the proportion of the average Hamming distance of S attributed to the set of selected non-stopwords, $V \setminus J \cap S$, and the proportion of average Hamming distance of S attributed to the set of selected stopwords, $J \cap S$. We can also compute the maximum possible average Hamming distance that can be achieved by any set of K features simply by ranking V by the incremental average Hamming distance and summing its value for the K top ranked features. Similarly, we can compute the maximum possible average Hamming distance that can be achieved by any set of K features that does not include any stopwords.

As an example of the impact of stopwords on separation we now consider the *fuel* topic. The set of features corresponding to $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ is listed in Table 4.2 and Figure 4.1 shows a plot of σ and σ' as a function of k. Using this information we see that the maximum possible value of σ_K when stopwords are included is 12.24, 5.85 or 47.76% of which was contributed by non-stopwords, and 6.40 or 52.24% of which was contributed by stopwords. These observations also support the assertion that the reason that the feature sets selected by μ_8 achieve the best separation is because they include stopwords. It also indicates that separation and noise are *competing criteria*.

4.2 The Impact of Noise

Having seen the impact of the selection of noise on the separation achieved by $\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)$, we are now interested in understanding how noise affects the separation achieved by the other Boolean feature ranking functions we studied. To do this we again consider the summary of the results of the experiments from §3.10 which can be found in Table 3.2. Using these results we created Figure 4.2[1] which shows that σ_K increases with ν_K for all of the Boolean feature ranking functions that we studied.

[1]Each integer i in the plots of evaluation measures versus ν_K corresponds to the feature ranking function μ_i. Also, the position of some labels in this these plots and the plot shown in Figure 4.8 have been slightly adjusted to avoid overlap of labels for feature ranking functions with similar performance.

Rank	Feature	Stopword	μ_8
1	fuel	No	0.83
2	oil	No	0.65
3	prices	No	0.59
4	mln	No	0.56
5	petroleum	No	0.53
6	*from*	Yes	0.51
7	*will*	Yes	0.51
8	*with*	Yes	0.50
9	*its*	Yes	0.49
10	dlrs	No	0.49
11	pct	No	0.49
12	*one*	Yes	0.48
13	*was*	Yes	0.47
14	corp	No	0.46
15	*were*	Yes	0.45
16	*two*	Yes	0.44
17	*are*	Yes	0.44
18	*has*	Yes	0.43
19	*new*	Yes	0.43
20	*that*	Yes	0.42
21	company	No	0.42
22	*for*	Yes	0.42
23	year	No	0.42
24	*would*	Yes	0.41
25	barrel	No	0.40
$\sigma\big(\omega(V, K, \mu_8\text{-RANKING}, \uparrow)\big)$			12.24

Table 4.2: $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic

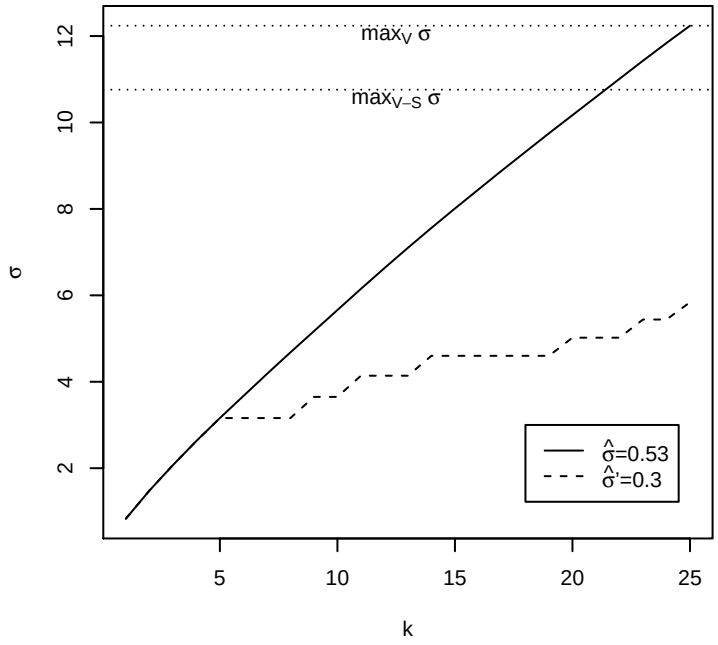


Figure 4.1: $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic

When only using σ_K and ν_K as evaluation criteria, the functions with relatively low values of ν_K and relatively large values of σ_K are clearly the most desirable and the plot depicts what might informally be considered an *efficient frontier* which identifies the best function for a given value of ν_K. While Figure 4.2 plotted the mean of σ_K for all topics, Figure H.3 shows the values of σ_K and ν_K for all topics and therefore depicts the differences in variation between the feature ranking functions. As would be expected, the results showed a similar relationship between $\widehat{\sigma}$ and $\widehat{\nu}$.

Continuing our discussion of the results shown in Table 3.2, we see that σ_K is not the only one of our measures from $\mathcal{C}$ that increases with ν_K. In fact, as seen in Figure 4.3 and Figure 4.4 that similar statements can be made about θ_K and ξ.[2] In addition, as can be seen in Figure 4.5 and Figure 4.6, $\Delta(\sigma_K)$ and $\Delta(\theta_K)$ also increase with ν_K. As one might expect, similar statements can be made about $\widehat{\theta}$, $\Delta(\widehat{\sigma})$, and $\Delta(\widehat{\theta})$. Similar

[2]As an interesting aside, note that the robustness of μ_9 and μ_{24} was significantly better than that of their absolute value variants $|\mu_9|$ and $|\mu_{24}|$. The detailed robustness data for μ-RANKING is in Table L.1.

to the conclusion we reached about separation and noise at the end of §4.1, we now conclude that noise and size (as measured by θ), and noise and robustness are also *competing criteria*.

Since the selection of noise features, and specifically stopwords, does in fact result in an increase in $\widehat{\sigma}$, $\Delta(\widehat{\sigma})$, σ_K, and $\Delta(\sigma_K)$, as well as in $\widehat{\theta}$, $\Delta(\widehat{\theta})$, θ_K, and $\Delta(\theta_K)$, as well as in ξ, it seems reasonable to ask why we we would like to identify the feature selection algorithms that yield feature sets that do not include them. The answer to this question is two fold. First, the semantic content of a feature set, as measured by φ_K, decreases with the number of stopwords it includes, as can be seen in Figure 4.7, with a similar statement holding for $\bar{\varphi}$. Second, if stopwords were the only noise features present in textual data sets, we could simply delete them prior to feature selection and they would not be a concern. As we shall see in the next section, however, the presence of stopwords in a selected feature set indicates that the feature selection algorithm has likely selected other non-stopword noise terms with relatively low semantic content.

4.3 Non-Stopword Noise

A natural question to ask at this point is whether there is a flaw in our use of stopwords as a measure of the noise in a feature set, and whether or not we could eliminate the problem of noise selection by simply a priori deleting all stopwords from the feature set. We claim that simply deleting all stopwords from the feature set will not adequately address the issue of noise selection by feature selection algorithms. To support this claim we removed the stopwords from V and then reran the μ-RANKING experiments that were discussed in §3.10. The detailed results of these experiments are listed in Appendix J and they are summarized in Table 4.3.

Comparison of these results with those in Table H.1 shows that for μ_8, not unexpectedly, there was a decrease in both $\widehat{\sigma}$ and σ_K. However, given the substantial 46% difference between $\widehat{\sigma}$ and $\widehat{\sigma}'$, and the 54% difference between σ_K and σ_K' shown in Table H.1 and Table I.1, what was arguably more interesting was the fact that the 11% decrease in $\widehat{\sigma}$ from 0.54 to 0.48 and the 15.37% decrease in σ_K from 12.43 to 10.52,

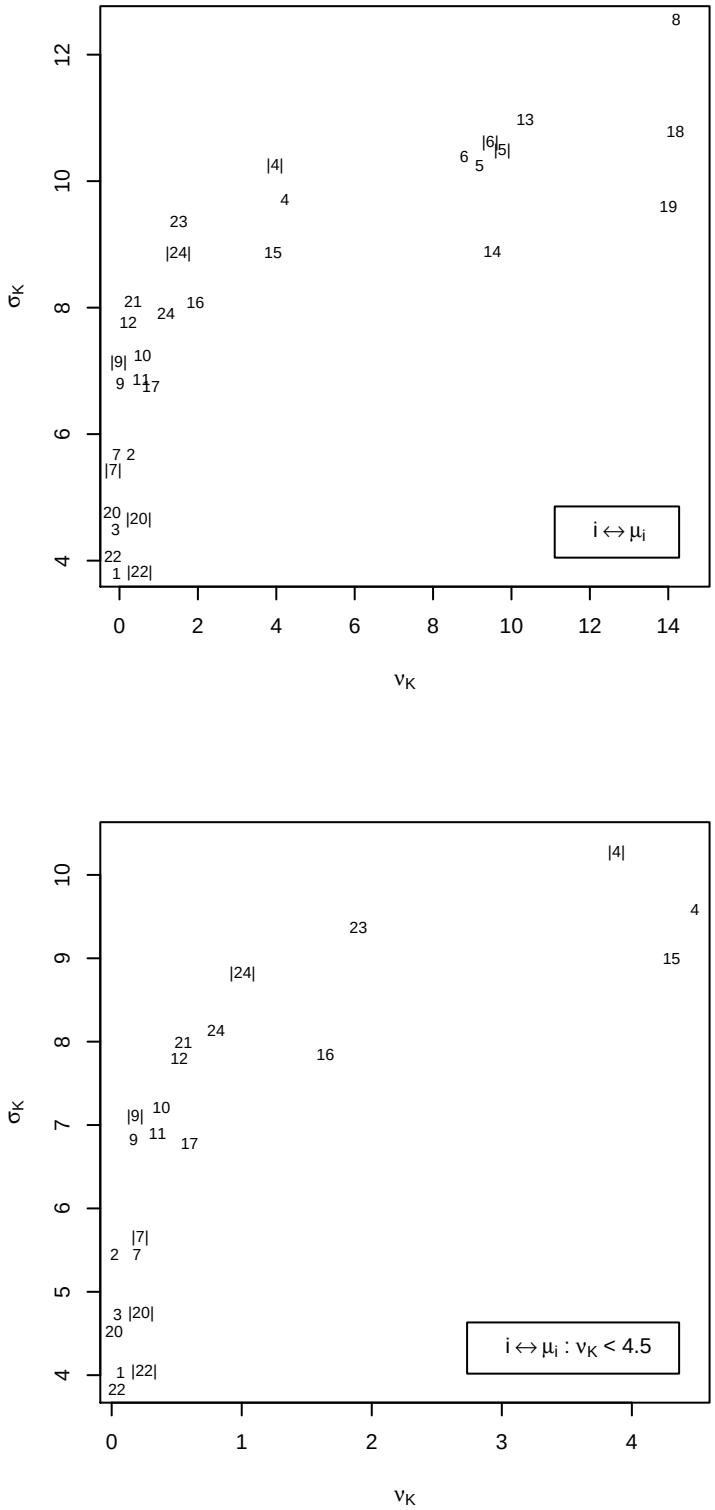


Figure 4.2: μ-RANKING Stopwords Included Not Discounted: σ_K vs ν_K

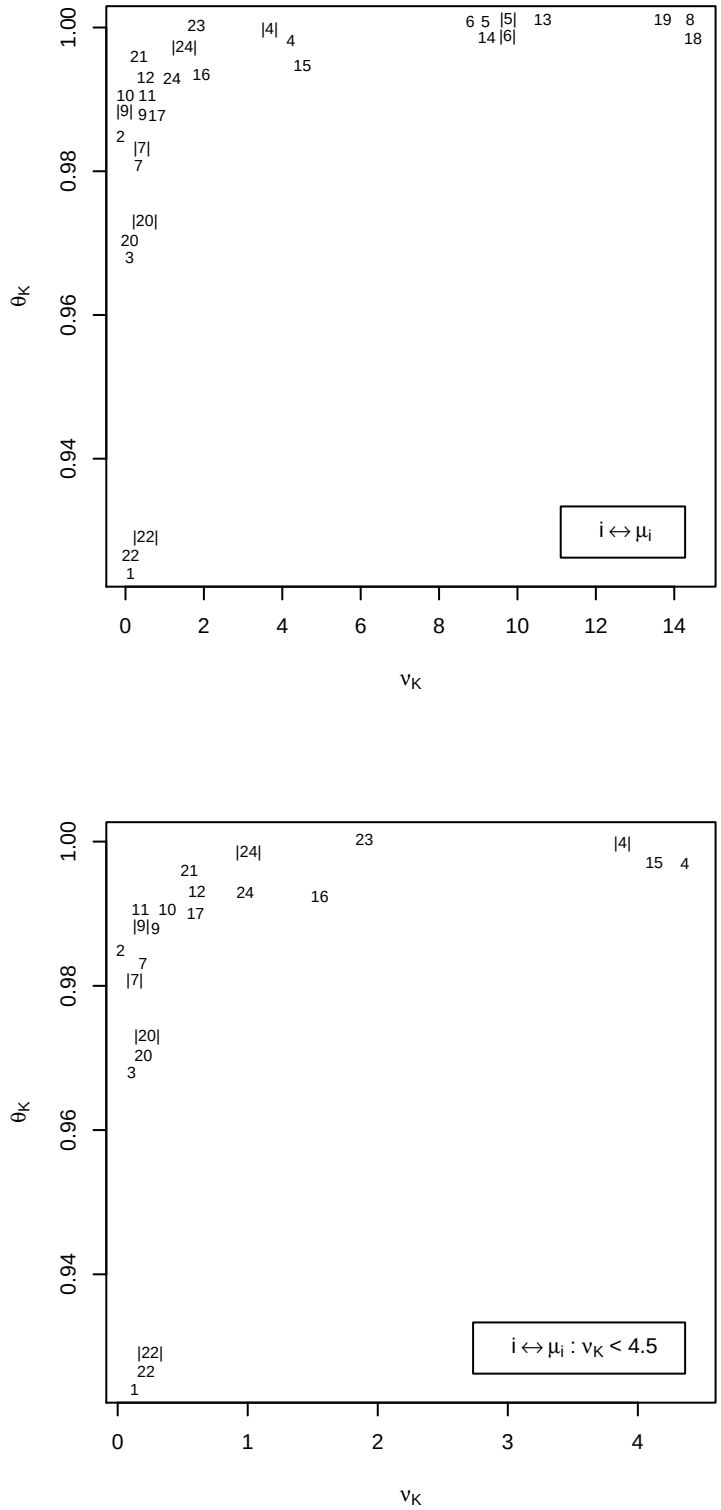


Figure 4.3: μ-RANKING Stopwords Included Not Discounted: θ_K vs ν_K

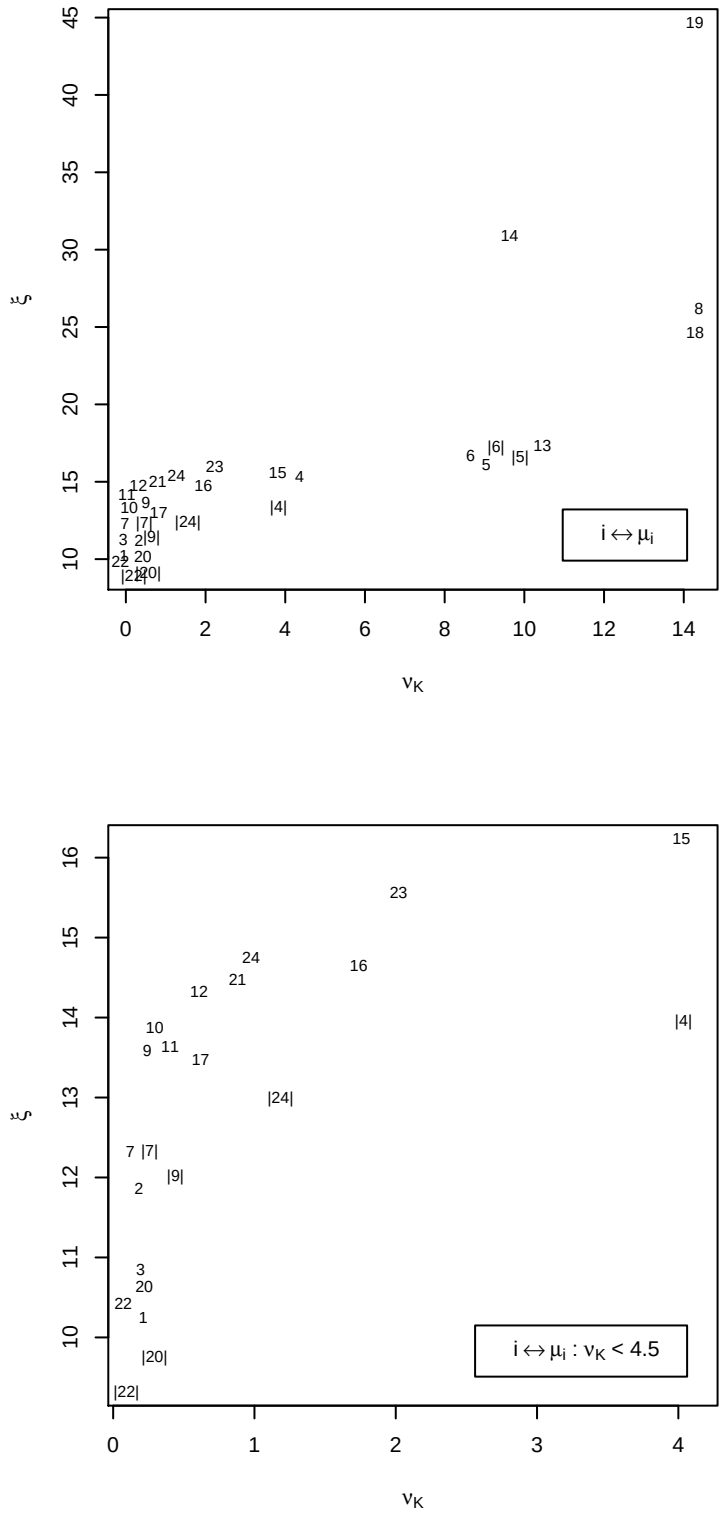


Figure 4.4: μ-RANKING Stopwords Included Not Discounted: ξ vs ν_K

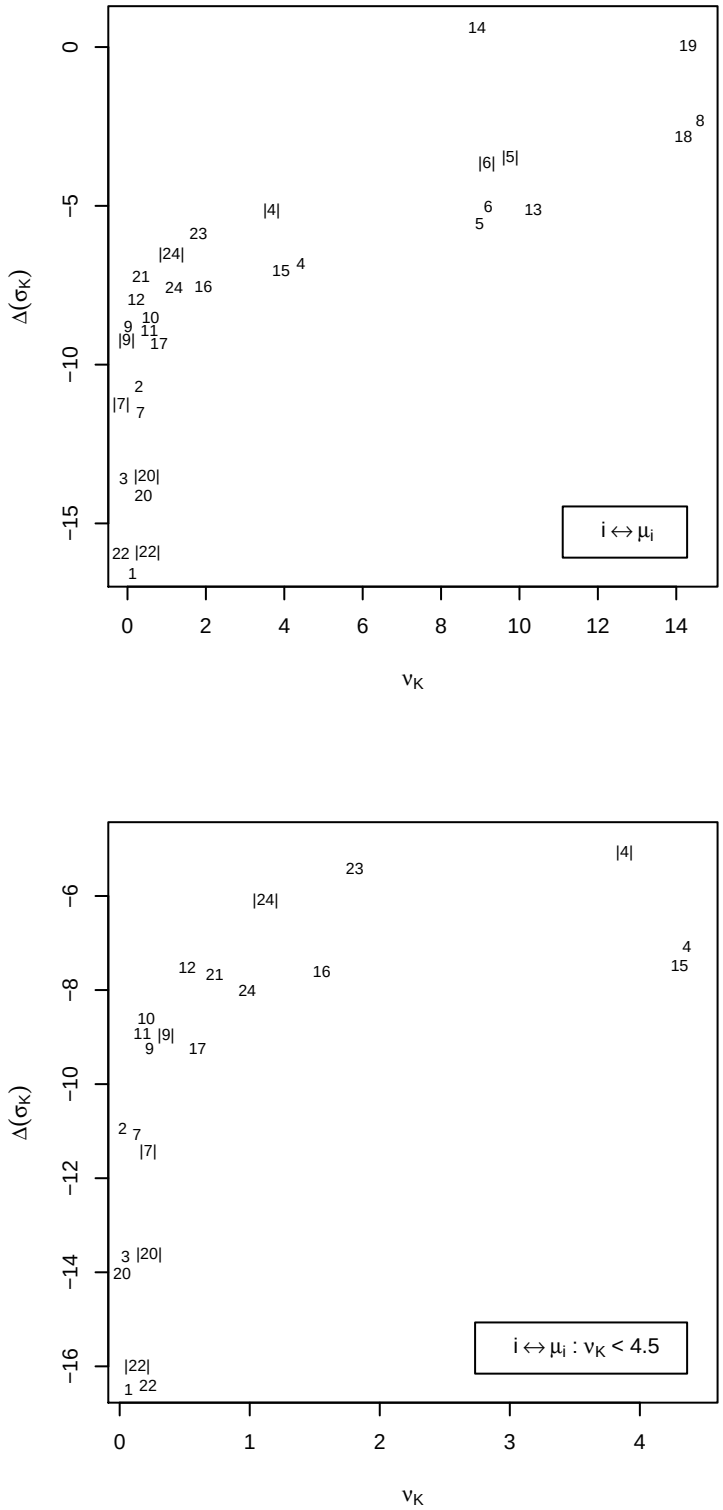


Figure 4.5: μ-RANKING Stopwords Included Not Discounted: $\Delta(\sigma_K)$ vs ν_K

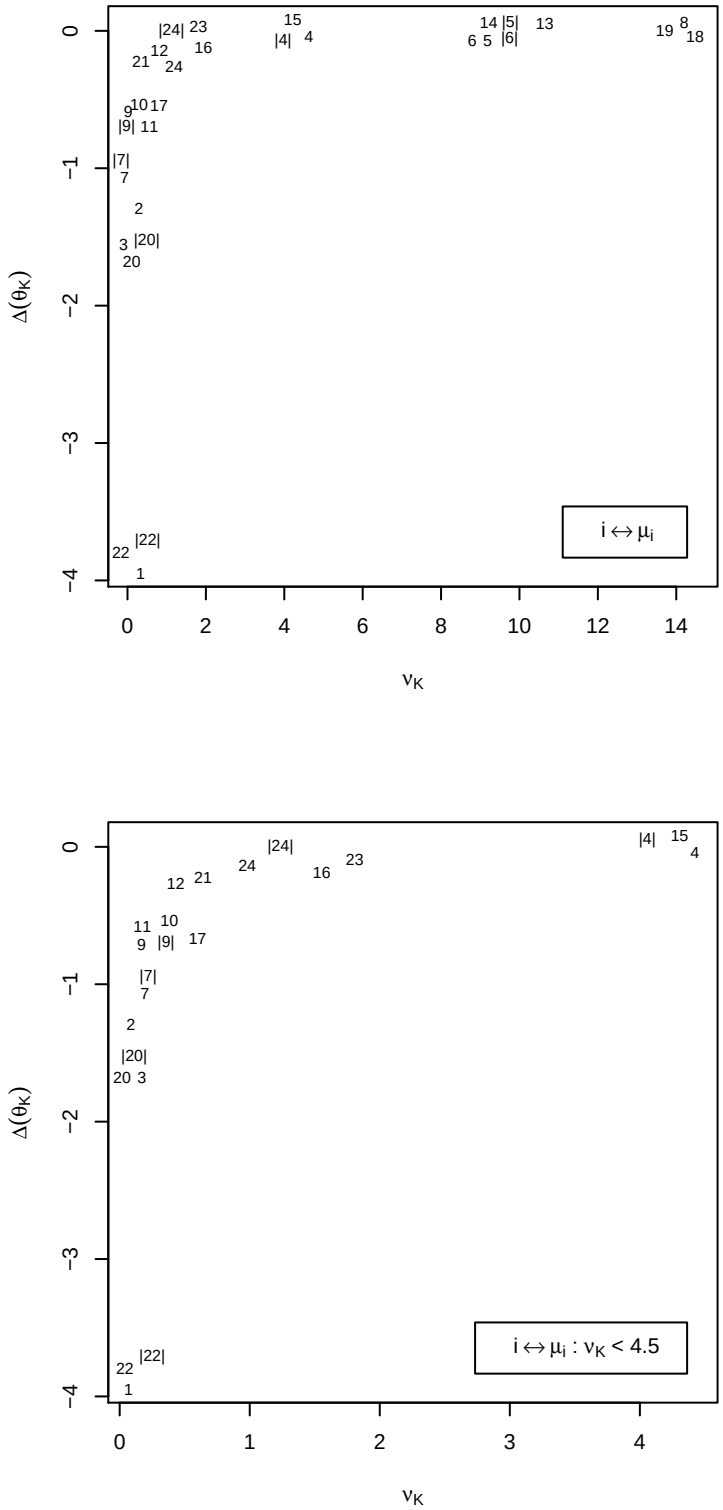


Figure 4.6: μ-RANKING Stopwords Included Not Discounted: $\Delta(\theta_K)$ vs ν_K

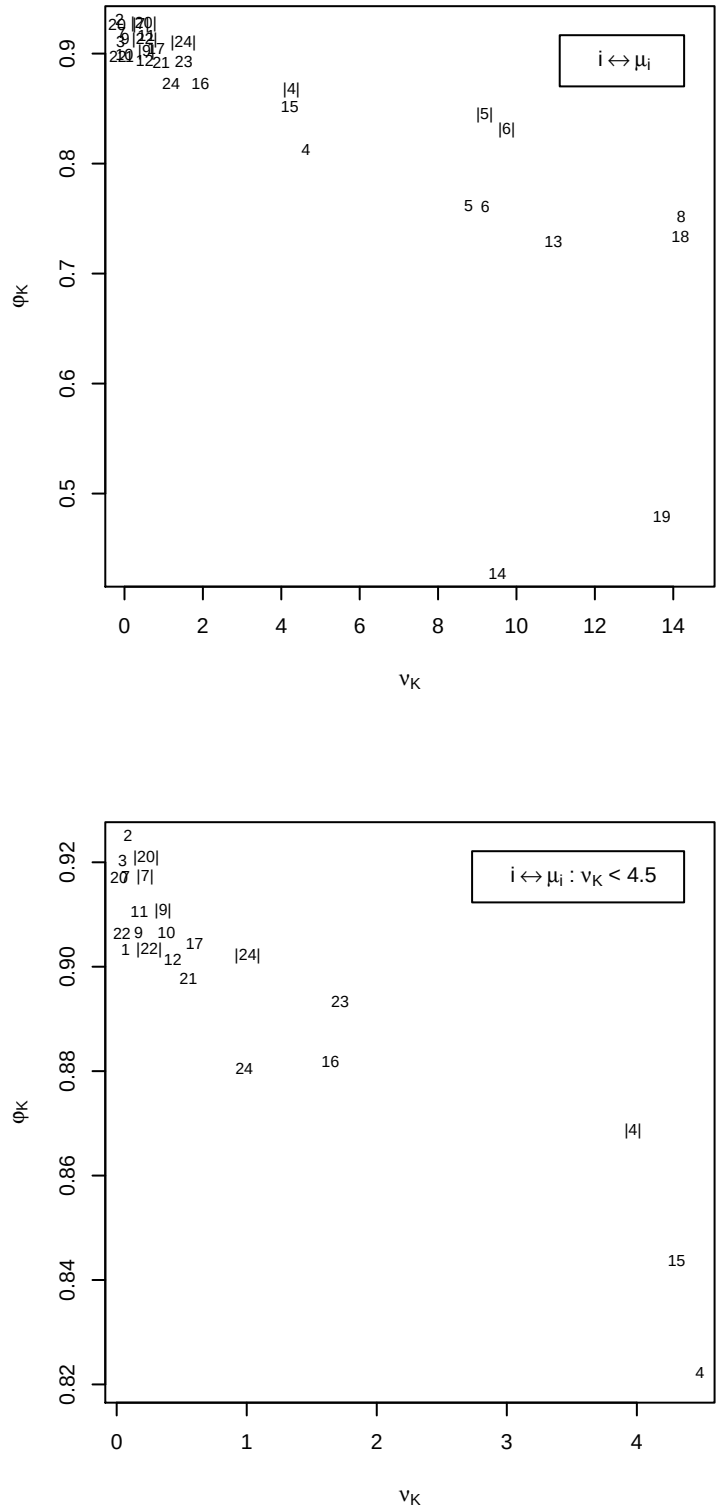


Figure 4.7: μ-RANKING Stopwords Included Not Discounted: φ_K vs ν_K

resulting from the removal of stopwords, was so small by comparison.

We now offer another assertion as a possible explanation for these results. We claim that stopwords are not the only noise features, and when μ_8 is applied to V it selects these other noise features, and when it is applied to $V \setminus J$ it selects even more of them. Further, the relatively small decreases in $\widehat{\sigma}$ and σ_K can be explained by the fact that these other noise features have relatively large average incremental Hamming distance values that are comparable to the values associated with the stopwords in $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$. To support this claim, we now consider the set of features corresponding to $\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)$ which are listed in Table 4.4. Using this information we see that the maximum possible value of σ_K when stopwords are excluded is 10.78. However, nine of the features (listed in italic font) in $\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)$, namely *mln*, *dlrs*, *pct*, *corp*, *bank*, *dlr*, *billion*, *due*, and *debt*, are also in $\omega(V \setminus J, K, a+b\text{-RANKING}, \uparrow)$. While it seems clear that the frequent occurrence of these features is due to the fact that the Reuters-21578 collection contains articles from a financial publication, we submit that per our discussion in §1.1 they are noise features, and the fact that they account for 4.33 or 40% of 10.78, serves to support our assertion that, the strategy of deleting stopwords from the feature set prior to performing feature selection will not eliminate the issue of noise selection.

4.4 Monotone Feature Principle

In an attempt to identify the characteristics that distinguish noise features from non-noise features we now turn our attention to $\omega(V, K, x\text{-RANKING}, \uparrow)$, the set of most frequently occurring features in relevant documents, and $\omega(V, K, y\text{-RANKING}, \uparrow)$, the set of most frequently occurring features in irrelevant documents. We shall let

$$\mho_K = \omega(V, K, x\text{-RANKING}, \uparrow) \cup \omega(V, K, y\text{-RANKING}, \uparrow)$$

and

$$\Omega_K = \omega(V, K, x\text{-RANKING}, \uparrow) \cap \omega(V, K, y\text{-RANKING}, \uparrow).$$

[3]The notation r(·) will be used in this chapter to indicate the rank of a feature in a sequence.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
$\lvert\mu_4\rvert$	μ_8	μ_8	μ_0	0.00	$\lvert\mu_4\rvert$	0.98	μ_8	0.48	μ_2	0.90
μ_8	$\lvert\mu_4\rvert$	$\lvert\mu_4\rvert$	μ_1	0.00	μ_8	0.98	$\lvert\mu_4\rvert$	0.44	μ_7	0.90
μ_{23}	μ_{23}	μ_{23}	μ_2	0.00	μ_{13}	0.98	$\lvert\mu_5\rvert$	0.43	$\lvert\mu_7\rvert$	0.90
μ_{13}	μ_{13}	$\lvert\mu_{24}\rvert$	μ_3	0.00	μ_{23}	0.98	$\lvert\mu_6\rvert$	0.43	μ_3	0.89
μ_5	μ_5	μ_{13}	μ_4	0.00	μ_4	0.97	μ_{13}	0.42	μ_9	0.89
μ_6	μ_6	μ_4	$\lvert\mu_4\rvert$	0.00	μ_5	0.97	μ_{18}	0.42	$\lvert\mu_9\rvert$	0.89
$\lvert\mu_{24}\rvert$	$\lvert\mu_{24}\rvert$	μ_{12}	μ_5	0.00	μ_6	0.97	μ_{23}	0.42	μ_{10}	0.89
μ_4	μ_4	μ_{15}	$\lvert\mu_5\rvert$	0.00	μ_{12}	0.97	μ_5	0.41	μ_{11}	0.89
$\lvert\mu_5\rvert$	μ_{15}	μ_{21}	μ_6	0.00	μ_{15}	0.97	μ_6	0.41	μ_{12}	0.89
$\lvert\mu_6\rvert$	μ_{21}	μ_{24}	$\lvert\mu_6\rvert$	0.00	μ_{21}	0.97	μ_4	0.40	μ_{17}	0.89
μ_{15}	μ_{24}	μ_5	μ_7	0.00	μ_{24}	0.97	$\lvert\mu_{24}\rvert$	0.40	μ_{20}	0.89
μ_{18}	μ_{12}	μ_6	$\lvert\mu_7\rvert$	0.00	$\lvert\mu_{24}\rvert$	0.97	μ_{15}	0.39	$\lvert\mu_{20}\rvert$	0.89
μ_{21}	μ_{16}	μ_{16}	μ_8	0.00	μ_9	0.96	μ_{21}	0.38	$\lvert\mu_{24}\rvert$	0.89
μ_{12}	μ_{10}	μ_{10}	μ_9	0.00	$\lvert\mu_9\rvert$	0.96	μ_{24}	0.38	μ_{16}	0.88
μ_{24}	μ_9	$\lvert\mu_5\rvert$	$\lvert\mu_9\rvert$	0.00	μ_{10}	0.96	μ_{12}	0.37	μ_{21}	0.88
μ_{16}	$\lvert\mu_9\rvert$	$\lvert\mu_6\rvert$	μ_{10}	0.00	μ_{11}	0.96	μ_{16}	0.37	μ_{23}	0.88
μ_{10}	μ_{11}	μ_9	μ_{11}	0.00	μ_{16}	0.96	μ_{10}	0.35	μ_{24}	0.88
μ_9	μ_{17}	$\lvert\mu_9\rvert$	μ_{12}	0.00	μ_{17}	0.96	μ_9	0.34	μ_1	0.87
$\lvert\mu_9\rvert$	μ_{14}	μ_{11}	μ_{13}	0.00	μ_2	0.95	$\lvert\mu_9\rvert$	0.34	$\lvert\mu_4\rvert$	0.87
μ_{11}	μ_2	μ_{17}	μ_{14}	0.00	μ_7	0.95	μ_{11}	0.34	μ_{15}	0.87
μ_{17}	μ_7	μ_{18}	μ_{15}	0.00	$\lvert\mu_7\rvert$	0.95	μ_{14}	0.34	μ_{22}	0.87
μ_2	$\lvert\mu_7\rvert$	μ_2	μ_{16}	0.00	μ_{14}	0.95	μ_{19}	0.34	$\lvert\mu_{22}\rvert$	0.87
μ_7	$\lvert\mu_5\rvert$	μ_7	μ_{17}	0.00	$\lvert\mu_5\rvert$	0.94	μ_{17}	0.33	μ_4	0.86
$\lvert\mu_7\rvert$	$\lvert\mu_6\rvert$	$\lvert\mu_7\rvert$	μ_{18}	0.00	$\lvert\mu_6\rvert$	0.94	μ_2	0.29	μ_{13}	0.85
μ_3	μ_{18}	μ_3	μ_{19}	0.00	μ_{18}	0.94	μ_7	0.29	μ_5	0.84
μ_{20}	μ_{19}	μ_{20}	μ_{20}	0.00	μ_3	0.92	$\lvert\mu_7\rvert$	0.29	μ_6	0.84
$\lvert\mu_{20}\rvert$	μ_3	$\lvert\mu_{20}\rvert$	$\lvert\mu_{20}\rvert$	0.00	μ_{19}	0.92	μ_3	0.24	μ_8	0.84
μ_1	μ_{20}	μ_1	μ_{21}	0.00	μ_{20}	0.92	μ_{20}	0.24	$\lvert\mu_5\rvert$	0.82
μ_{22}	$\lvert\mu_{20}\rvert$	μ_{22}	μ_{22}	0.00	$\lvert\mu_{20}\rvert$	0.92	$\lvert\mu_{20}\rvert$	0.24	$\lvert\mu_6\rvert$	0.82
$\lvert\mu_{22}\rvert$	μ_1	$\lvert\mu_{22}\rvert$	$\lvert\mu_{22}\rvert$	0.00	μ_1	0.84	μ_1	0.20	μ_{18}	0.80
μ_{19}	μ_{22}	μ_{14}	μ_{23}	0.00	μ_{22}	0.84	μ_{22}	0.20	μ_{19}	0.33
μ_{14}	$\lvert\mu_{22}\rvert$	μ_{19}	μ_{24}	0.00	$\lvert\mu_{22}\rvert$	0.84	$\lvert\mu_{22}\rvert$	0.20	μ_{14}	0.32
μ_0	μ_0	μ_0	$\lvert\mu_{24}\rvert$	0.00	μ_0	0.13	μ_0	0.01	μ_0	0.06

Table 4.3: μ-RANKING Not Discounted Stopwords Excluded

Rank	Feature	μ_8	$\mathrm{r}(a+b)^3$
1	fuel	0.83	892
2	oil	0.65	267
3	prices	0.59	232
4	*mln*	0.56	5
5	petroleum	0.53	766
6	*dlrs*	0.51	7
7	*pct*	0.51	9
8	*corp*	0.5	16
9	company	0.49	37
10	year	0.49	26
11	barrel	0.49	1407
12	*bank*	0.48	13
13	*dlr*	0.47	25
14	effective	0.46	516
15	*billion*	0.45	21
16	crude	0.44	1156
17	gasoline	0.44	1988
18	today	0.43	128
19	*due*	0.43	15
20	products	0.42	583
21	cts	0.42	890
22	price	0.42	69
23	*debt*	0.42	19
24	april	0.41	40
25	international	0.4	43
$\sigma\big(\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)\big)$			10.78

Table 4.4: $\omega(V \setminus J, K, \mu_8\text{-RANKING}, \uparrow)$ for the Fuel Topic

The features in $\mho_K$ are shown in Table 4.5 and we begin by noticing that

$$\mho_K \cap \omega(V, K, a+b\text{-RANKING}, \uparrow) =$$
$$\{reuter, said, the, and, for, pct, dlrs, from, its, will, was, corp,$$
$$mln, with, bank, due, has, that, are, debt, issue, billion, inc, which, dlr\}$$

and per §1.1, these 25 features are noise features. However, what seems potentially more interesting is that 12 of these features (those shown in italic font in Table 4.5) are also in

$$\Omega_K = \{reuter, said, the, and, for, pct, dlrs, from, its, will, was, corp\}.$$

If we continue this investigation, we see that a total of 30 (the additional 18 features are shown in normal font in Table 4.5) of the 38 features in $\mho_K$ appear in $\omega(V, 2K, a+b\text{-RANKING}, \uparrow)$, and so are noise features, and 23 of these features also appear in Ω_{2K}. These observations motivate the following characterization of noise features.

Observation 4.1. *If $j \in V$ and $j \in \Omega_K$ for some small $K \in \mathbb{Z}^+$, then j is a noise feature.*

Returning to Table 4.5 we also notice that each of the 12 features in Ω_K and each of the 23 features in Ω_{2K} have relatively small values of $|x - y|$. These observations motivate the following characterization of noise features.

Observation 4.2. *If $j \in V$, $j \in \mho_K$, and $|x_j - y_j| < \epsilon$ for some small $\epsilon > 0$, then j is a noise feature.*

In §1 we mentioned that feature frequency in textual data tends to follow a power law. Let P be a set of real numbers, and (p_k) be the corresponding sequence in decreasing order. If $(p_k) = \alpha(\gamma + k)^{-\beta}$, then we say that P follows the Zipf-Mandelbrot law with parameters α, β and γ, which we will denote as $P \sim ZM(\alpha, \beta, \gamma)$. When $\gamma = 0$, we have $(p_k) = \alpha k^{-\beta}$, and we say that P follows Zipf's law, which we will denote as $P \sim Z(\alpha, \beta)$. Classically, it has been stated with $\beta = 1$. If, for a set of real numbers

Feature	r(x)	x	r(y)	y	r($x-y$)	$x-y$	r(μ_8)	μ_8
reuter	1	0.97	1	0.99	12396	-0.03	2132	0.04
said	2	0.97	3	0.97	12291	0.00	1203	0.06
the	3	0.93	2	0.98	12424	-0.04	539	0.09
fuel	4	0.83	3132	0.00	1	0.83	1	0.83
and	5	0.77	4	0.88	12482	-0.11	60	0.30
for	6	0.77	6	0.65	89	0.12	22	0.42
oil	7	0.67	349	0.04	2	0.63	2	0.65
pct	8	0.67	9	0.53	59	0.14	11	0.49
prices	9	0.60	271	0.05	3	0.55	3	0.59
dlrs	10	0.53	7	0.59	12447	-0.06	10	0.49
from	11	0.53	12	0.40	63	0.14	6	0.51
petroleum	12	0.53	1155	0.01	4	0.52	5	0.53
its	13	0.47	11	0.41	673	0.06	9	0.49
one	14	0.47	27	0.23	20	0.23	12	0.48
will	15	0.47	8	0.59	12486	-0.12	7	0.51
was	16	0.43	21	0.29	55	0.15	13	0.47
barrel	17	0.40	2993	0.00	5	0.40	25	0.40
corp	18	0.40	16	0.32	208	0.08	14	0.46
two	19	0.40	47	0.19	23	0.21	16	0.44
were	20	0.40	28	0.23	35	0.17	15	0.45
company	21	0.37	37	0.21	50	0.16	21	0.42
crude	22	0.37	1806	0.01	7	0.36	32	0.37
effective	23	0.37	621	0.02	8	0.34	30	0.37
gasoline	24	0.37	12398	0.00	6	0.37	33	0.37
new	25	0.37	31	0.23	60	0.14	19	0.43
mln	26	0.33	5	0.67	12501	-0.33	4	0.56
are	27	0.33	18	0.31	11163	0.03	17	0.44
with	28	0.30	10	0.51	12496	-0.21	8	0.50
has	29	0.30	15	0.33	12414	-0.03	18	0.43
that	30	0.30	17	0.31	12355	-0.01	20	0.42
which	31	0.27	24	0.28	12359	-0.01	27	0.40
dlr	32	0.23	25	0.26	12402	-0.03	29	0.37
billion	33	0.20	22	0.29	12466	-0.09	31	0.37
inc	34	0.17	23	0.28	12485	-0.12	37	0.35
bank	35	0.07	13	0.36	12499	-0.30	28	0.38
debt	36	0.07	19	0.30	12497	-0.24	46	0.33
due	37	0.03	14	0.34	12500	-0.30	38	0.35
issue	38	0.03	20	0.29	12498	-0.25	55	0.30

Table 4.5: $\mho_K = \omega(V, K, x\text{-RANKING}, \uparrow) \cup \omega(V, K, y\text{-RANKING}, \uparrow)$

P, it is assumed that $P \sim Z(\alpha, \beta)$ $(p_k) = \alpha k^{-\beta}$, the values of the parameters α and β have historically been identified by fitting a regression line to

$$\log(p_k) = \log(\alpha) - \beta k.$$

Alternatively, fitting to "binned" data or to the CDF has been shown to provide better results. More sophisticated methods for fitting data to a power law are studied in [21] and fitting data to the Zipf-Mandelbrot law is considered in [30]. Whether or not feature frequency in textual data tends to follow Zipf's law or the Zipf-Mandelbrot law has been discussed extensively (see e.g. [61]). In the following result we assume Zipf's law holds.

Proposition 4.1. *Let $j \in V$, with $j \in \mho_K$ for some small $K \in \mathbb{Z}^+$, and assume that $\omega(V, K, x\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ and $\omega(V, K, y\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 1$. Then, $j \in \Omega_K$ if and only if*

$$|x_j - y_j| \leq \alpha \left(\frac{K^\beta - 1}{K^\beta} \right).$$

Proof. We show the "if" part first. Let $k_{x,j}$ and $k_{y,j}$ be the rank of feature j in $\omega(V, |V|, x\text{-RANKING}, \uparrow)$ and $\omega(V, |V|, y\text{-RANKING}, \uparrow)$ respectively. Since $\omega(V, K,$ $x\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ and $\omega(V, K, y\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ we have

$$|x_j - y_j| = \left| \frac{\alpha}{k_{x,j}^\beta} - \frac{\alpha}{k_{y,j}^\beta} \right|.$$

The assumption that $j \in \mho_K$ only tells us that $1 \leq k_{x,j} \leq K$, or $1 \leq k_{y,j} \leq K$, or both. However, because we also assumed that $j \in \Omega_K$, we know that both $1 \leq k_{x,j} \leq K$ *and* $1 \leq k_{y,j} \leq K$. Therefore,

$$0 \leq \left| \frac{\alpha}{k_{x,j}^\beta} - \frac{\alpha}{k_{y,j}^\beta} \right| \leq \alpha \left(\frac{K^\beta - 1}{K^\beta} \right)$$

with the lower bound occurring when $k_{x,j} = k_{y,j}$, and the upper bound occurring when

$k_{x,j} = 1$ (or $k_{y,j} = 1$) and $k_{y,j} = K$ (or $k_{x,j} = K$). For the "only if" part we are given

$$|x_j - y_j| \leq \alpha \left(\frac{K^\beta - 1}{K^\beta} \right)$$

and because $\omega(V, K, x\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ and $\omega(V, K, y\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$, we have

$$\left| \frac{\alpha}{k_{x,j}^\beta} - \frac{\alpha}{k_{y,j}^\beta} \right| \leq \alpha \left(\frac{K^\beta - 1}{K^\beta} \right).$$

As before, the assumption that $j \in \mho_K$ tells us that $1 \leq k_{x,j} \leq K$, or $1 \leq k_{y,j} \leq K$, or both. To prove that $j \in \Omega$, we must show that both $1 \leq k_{x,j} \leq K$ *and* $1 \leq k_{y,j} \leq K$. Assuming to the contrary, gives us that $1 \leq k_{x,j} \leq K$ (or $1 \leq k_{y,j} \leq K$) and $k_{y,j} > K$ (or $k_{x,j} > K$). WLOG, consider $k_{x,j} = 1$ and $k_{y,j} = K + 1$, so

$$\left| \frac{\alpha}{k_{x,j}^\beta} - \frac{\alpha}{k_{y,j}^\beta} \right| = \alpha \left(\frac{(K+1)^\beta - 1}{(K+1)^\beta} \right) \not\leq \alpha \left(\frac{K^\beta - 1}{K^\beta} \right).$$

■

Corollary 4.1. *Let $j \in V$, assume that $\omega(V, |V|, x\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ and $\omega(V, |V|, y\text{-RANKING}, \uparrow) \sim Z(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 1$, and let $|x_j - y_j| \leq \epsilon$ for some given $0 < \epsilon \leq \alpha$. Then, $j \in \Omega_K$ if and only if $K \geq \left\lceil \left(\frac{\alpha}{\alpha - \epsilon} \right)^{\frac{1}{\beta}} \right\rceil$.*

The result shows that the two characterizations of noise given by Observation 4.1 and Observation 4.2 are equivalent. It should be mentioned that finding such a connection is only possible because of the relationship that Zipf's laws provides between feature frequencies and the rank of these frequencies when placed in decreasing order.

Notice that if we follow the original formulation of Zipf's law in which it was assumed that $\beta = 1$, our bound becomes

$$|x_j - y_j| \leq \alpha \left(1 - \frac{1}{K} \right).$$

Also, note that following the notation used in the proof, that if for a given $j \in V$, we

know the values of $k_{x,j}$ and $k_{y,j}$, we can improve the bound to be

$$|x_j - y_j| \leq \alpha \left(\frac{\hat{K}^\beta - 1}{\hat{K}^\beta} \right)$$

where $\hat{K} = \max\{k_{x,j}, k_{y,j}\} \leq K$.

One might reasonably question the assumption that both $\omega(V, K, x\text{-RANKING}, \uparrow)$ and $\omega(V, K, y\text{-RANKING}, \uparrow)$ follow Zipf's law with the same parameters (α, β), and in fact when all documents in $|T|$ and $|F|$ are used for the fuel topic, this does not seem to be the case. However, when all of the documents in T are used, and when $|T|$ documents are randomly sampled from F, the assumption does seem to hold.

Now, if features that occur frequently in both relevant documents and irrelevant documents are noise features, then we must conclude that of the frequently occurring features, those that are non-noise features must either occur frequently in relevant documents, *or* in irrelevant documents, but *not* both. However, since we consider features that appear frequently in irrelevant documents to be noise features, we must instead conclude that those that are non-noise features must only occur frequently in relevant documents. In fact, we can see that the 8 features (those shown in bold font), in Table 4.5 which we have not yet concluded are noise features, do in fact meet this criteria for non-noise features. These features, *fuel*, *oil*, *prices*, *petroleum*, *barrel*, *crude*, *effective*,[4] and *gasoline*, are in $\omega(V, K, x\text{-RANKING}, \uparrow)$ and are not in $\omega(V, K, y\text{-RANKING}, \uparrow)$. In addition, these features appear in $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ with ranks 1, 2, 3, 5, 25, 32, 30, and 33, respectively, and therefore each makes a relatively large contribution to the separation. Recalling that we refer to those features with $x - y > 0$ as positive features and those with $x - y < 0$ as negative features provides us with an alternative way to state our observations about non-noise features.

Observation 4.3. ***(Monotone Feature Principle)*** *Non-noise features in textual data sets, that provide substantial separation, are highly ranked positive features.*

[4]While the feature "effective" does not have the intuitively obvious association with the "fuel" topic that the other seven features do, a search of the data set shows that, phrases such as "effective crude prices" are common.

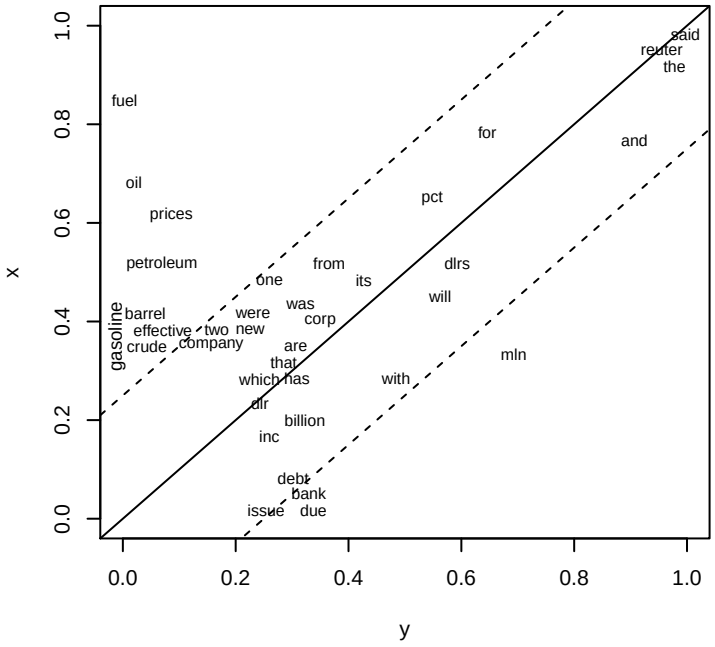


Figure 4.8: $\omega(V, K, x\text{-RANKING}, \uparrow) \cup \omega(V, K, y\text{-RANKING}, \uparrow)$

As expected, we see that the features in Table 4.5 that we identified as non-noise features have relatively large positive values of $x-y$, while those features with relatively small values of $|x-y|$ have been previously identified as noise features. Figure 4.8 depicts this situation with features with relatively large values of x and small values of y being non-noise features, and those features falling in the region where $|x-y|$ is relatively small being noise features.

Observation 4.3 states that non-noise features in textual data sets, are features with a high frequency in relevant documents and a low frequency in irrelevant documents. It should be mentioned that for a textual data set in which F is class homogeneous and $|T| \approx |F|$, it is revised to read, *"Non-noise features in textual data sets, that provide substantial separation, are either highly ranked positive features, or highly ranked negative features, but not both."*

While the advantages of identifying variables in two class classification problems on binary data sets as being monotone have been studied, (see e.g. page 295, [11]), it is

interesting to note that Observation 4.3 lies in contrast to the situation for many non-textual data sets (see e.g. [45]) in which machine learning algorithms generate rules that contain negated variables.

4.5 Characterizing Separation and Noise

In this section, motivated by the experimental results and discussions earlier in this chapter, we identify the common characteristics of the $\mu \in \mathcal{M}$ which select relatively few stopwords while achieving relatively high separation. We will use this information to provide characterizations of separation and noise features which will motivate definitions of what we shall refer to as *separation functions* and *noise functions.* We will make use of Proposition 3.5, noting that after applying the transformations in (1.2), with the exception of the non-algebraic μ_{12} and μ_{21}, all of the $\mu \in \mathcal{M}$ presented in 3.8 can be written as

$$\mu(x, y) = \frac{\psi(x, y)}{\eta(x, y)},$$

where ψ and η are quadratic functions of x and y. In addition, for $j \in V$, we will make use of the fact that $x_j = \mathrm{P}(u_j = 1 \mid u \in T)$ and $y_j = \mathrm{P}(u_j = 1 \mid u \in F)$ and, since x_j is the TPR and y_j is the FPR, we will present some of this discussion in terms of the ROC framework.

4.5.1 Class Separation and Noise

We begin by considering the elementary function $x - y$ which we denoted as μ_4. This function is distinguished by the fact that in the experiments in §3.10 it selected relatively few stopwords and achieved relatively high separation.

As discussed in §3.7, $x - y$ provides a measure of the separation associated with a given feature. Following (3.37), it takes on positive values for positive features and negative values for negative features. For a positive feature, the larger the value of $x-y$, the more separation the feature provides and the best separating positive features correspond to the point $\operatorname{argmax} x - y = (1, 0)$ where $x - y = 1$. Similarly, for a negative feature, the smaller the value of $x-y$, the more separation the feature provides and the

best separating negative features correspond to the point $\operatorname{argmin} x - y = (0, 1)$ where $x - y = -1$. If $j \in V$, notice that $\operatorname{argmax} x - y = (1, 0)$ has the nice interpretation that $\mathrm{P}(u_j = 1 \mid u \in T) = 1$ and $\mathrm{P}(u_j = 1 \mid u \in F) = 0$ and that $\operatorname{argmin} x - y = (0, 1)$ has the nice interpretation that $\mathrm{P}(u_j = 1 \mid u \in T) = 0$ and $\mathrm{P}(u_j = 1 \mid u \in F) = 1$. Further, in the ROC framework these points correspond to the case where the Boolean classifiers g_j^+ and g_j^- defined in (3.25) and (3.26) respectively have perfect performance.

While large values of $x - y$ are indicative of separating features, as the following result shows, small values are indicative of a certain type of *noise* feature.

Proposition 4.2. *Let $j \in V$, then if $x_j - y_j = 0$, j is a noise feature.*

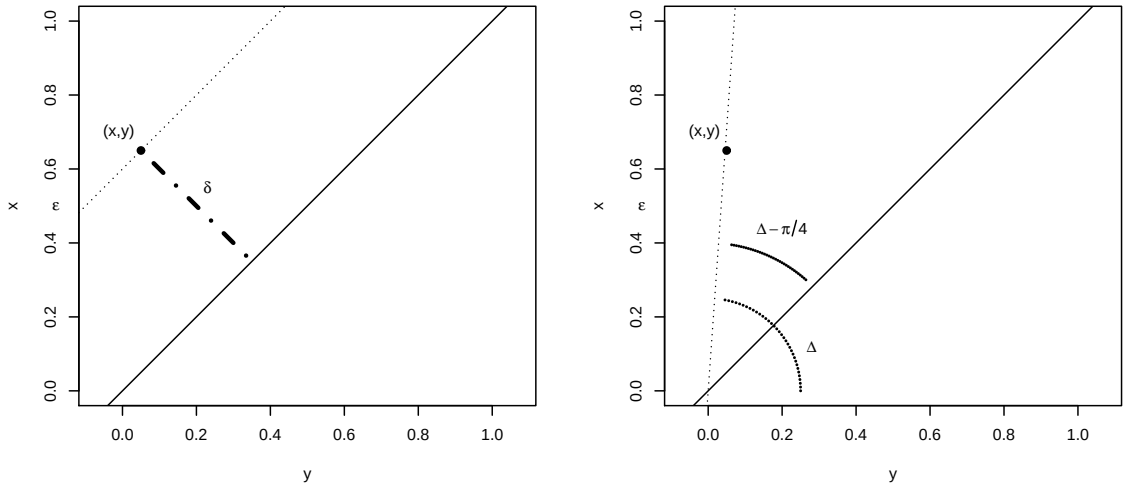
Proof. Since $x_j = \mathrm{P}(u_j = 1 \mid u \in T)$ and $y_j = \mathrm{P}(u_j = 1 \mid u \in F)$, we have that $\mathrm{P}(u_j = 1 \mid u \in F) = \mathrm{P}(u_j = 1 \mid u \in T)$. Since,

$$\begin{aligned}
\mathrm{P}(u_j = 1) &= \mathrm{P}(u \in T)\mathrm{P}(u_j = 1 \mid u \in T) + \mathrm{P}(u \in F)\mathrm{P}(u_j = 1 \mid u \in F) \\
&= \mathrm{P}(u \in T)\mathrm{P}(u_j = 1 \mid u \in T) + \mathrm{P}(u \in F)\mathrm{P}(u_j = 1 \mid u \in T) \\
&= (\mathrm{P}(u \in T) + \mathrm{P}(u \in F))(\mathrm{P}(u_j = 1 \mid u \in T))
\end{aligned}$$

and since $\mathrm{P}(u \in T) + \mathrm{P}(u \in F) = 1$, we have $\mathrm{P}(u_j = 1) = \mathrm{P}(u_j = 1 \mid u \in T)$. Since $\mathrm{P}(u_j = 1 \mid u \in T) = \mathrm{P}(u_j = 1 \mid u \in F)$, the probability of the feature j appearing in a document, i.e. $\mathrm{P}(u_j = 1)$, is independent of the document's class; which exactly coincides with j being a noise feature. ■

Let $j \in V$ be a positive feature, when $x_j - y_j = 0$ we shall say that j is a *class noise feature* and when $x_j - y_j = \epsilon$ for small $\epsilon > 0$ we shall say that j is an *ϵ-class noise feature*. These definitions have a geometric interpretation in ROC space. The set of all ϵ-class noise features corresponds to the line $x - y = \epsilon$ and the set of class noise features corresponds to the line $x - y = 0$. The L_2 distance from a point (x, y) to the line of class noise features is $\delta = (x - y)/\sqrt{2} = \epsilon/\sqrt{2}$, (see Figure 4.9a), and because $x - y$ is proportional to δ, it provides a measure of how close a feature is to being a class noise feature.

Note that $x - y$ provides a dual measure of class noise and separation. In the



(a) Measure of Separation and Noise

(b) Another Measure of Separation and Noise

Figure 4.9: Measures of Separation and Class Noise

context of separation we refer to a feature $j \in V$ with $x_j - y_j = \tau$ for large $\tau > 0$ as a τ-separating feature and in the specific case when $x_j - y_j = 0$ we will refer to it as a *zero-separating* feature. Similar comments can be made when $j \in V$ is a negative feature. In summary, *a feature that does not separate the document classes is a class noise feature.*

It is interesting to note that x/y which we denoted as $\widehat{\mu}_1$ similarly provides a measure of the separation associated with a given feature. Let (x, y) be a point in ROC space and the let the angle Δ be as shown in Figure 4.9b. Positive features are those with $1 < x/y < \infty$, $\pi/4 < \Delta \leq \pi/2$, and $0 < \Delta - \pi/4 \leq \pi/4$, and negative features are those with $0 \leq x/y < 1$, $0 \leq \Delta < \pi/4$ and $-\pi/4 \leq \Delta - \pi/4 < 0$.

For positive features, the larger the value of x/y, Δ, and $\Delta - \pi/4$, the more separation the feature provides and the best separating features correspond to the point $\operatorname{argmax} x/y = (1, 0)$ where $x/y = \infty$, i.e. the point $\operatorname{argmax} \tan(\Delta - \pi/4) = \pi/2$ where $\tan(\Delta - \pi/4) = 1$. Similarly, for negative features, the smaller the value of x/y, Δ, $\Delta - \pi/4$, the more separation the feature provides and the best separating negative features correspond to the point $\operatorname{argmax} x/y = (0, 1)$ where $x/y = 0$, i.e. the point $\operatorname{argmax} \tan(\Delta - \pi/4) = 0$ where $\tan(\Delta - \pi/4) = -1$. If $j \in V$, notice that

$\operatorname{argmax} x_j/y_j = (1,0)$ and $\operatorname{argmin} x_j/y_j = (0,1)$ correspond to the points where the likelihood ratio takes its maximum and minimum values respectively.

The following result provides a characterization of noise features in terms of x/y and Δ.

Proposition 4.3. *If $j \in V$, then if $x_j/y_j = 1$ or equivalently if $\Delta = \pi/4$, j is a noise feature.*

Proof. Follows immediately from the fact that $x_j/y_j = 1$ and $\Delta = \pi/4$ if and only if $x_j - y_j = 0$, which by Proposition 4.2 implies that j is a noise feature. ■

Our discussion of x/y and Δ suggests consideration of the function

$$\tan(\Delta - \frac{\pi}{4}) = \frac{\tan(\Delta) - 1}{\tan(\Delta) + 1}, \tag{4.1}$$

which since $\tan(\Delta) = x/y$ is a function of x/y. Since (4.1) is increasing in Δ and x/y, and $|\tan(0 - \frac{\pi}{4})| = 1$, $|\tan(\frac{\pi}{4} - \frac{\pi}{4})| = 0$, and $|\tan(\frac{\pi}{2} - \frac{\pi}{4})| = 1$, the function

$$\left|\tan(\Delta - \frac{\pi}{4})\right| = \left|\frac{\tan(\Delta) - 1}{\tan(\Delta) + 1}\right| = \left|\frac{x/y - 1}{x/y + 1}\right|$$

provides a measure of the distance from a point (x, y) to the line of class noise $x - y = 0$. It is interesting to note that (4.1) can be written as a

$$\left|\frac{x/y - 1}{x/y + 1}\right| = \left|\frac{x - y}{x + y}\right|,$$

which is $|\mu_{22}|$ and offers another example of a Boolean feature ranking function which selected relatively few stopwords, achieved relatively high separation, and is an increasing function of $x - y$. Other examples of such functions include μ_9, μ_{23}, and μ_{24}.

Whether viewed as a measure of separation or as a measure of noise, the value of $x - y$ is obviously the *same* for all points on the $x - y = \tau$. Therefore, each such line can be viewed as an *isocurve* that defines a *family of features*, each having the same value of the measure $x - y$. In §4.5.2 and §4.5.3 we consider $\text{den}(\mu_9)$ and $\text{den}(\mu_{23})$ which provide two additional definitions of noise which suggest that for any such feature family, some

of the features should be considered more desirable than others.

4.5.2 Collection Noise

The feature ranking functions μ_9 and μ_{11} selected relatively few stopwords and achieved relatively high separation in experiments in §3.10. Both of these functions have denominators that are proportional to

$$\big(|T|x + |F|y\big)\big(|T| + |F| - |T|x - |F|y\big). \tag{4.2}$$

Letting

$$\varrho = \frac{|T|}{|T| + |F|},$$

we have

$$(1 - \varrho) = \frac{|F|}{|T| + |F|},$$

and we see that (4.2) is proportional to

$$\eta_1(x, y) = \big(\varrho x + (1 - \varrho)y\big)\big(\varrho(1 - x) + (1 - \varrho)(1 - y)\big). \tag{4.3}$$

We begin our study of η_1 by noting that

$$\nabla\eta_1(x, y) = \begin{bmatrix} \varrho(1 - 2\varrho x - 2y + 2\varrho y) \\ (1 - \varrho)(1 - 2\varrho x - 2y + 2\varrho y) \end{bmatrix}^T$$

and that the system $\nabla\eta_1(x, y) = 0$ has only one linearly independent equation, i.e.

$$1 - 2\varrho x - 2y + 2\varrho y = 0. \tag{4.4}$$

The solutions of (4.4) correspond to the points on the line

$$\varrho x + (1 - \varrho)y = \frac{1}{2} \tag{4.5}$$

which is where η_1 takes its maximum value. The first factor in (4.3) is 0 at $(0,0)$ and the second factor in (4.3) is 0 at $(1,1)$ since $0 \leq x \leq 1$, $0 \leq y \leq 1$, $0 \leq \varrho \leq 1$. Since $\eta_1(x,y)$ is clearly non-negative over the range of x and y, we therefore have that

$$\operatorname{argmin} \eta_1(x,y) = \big\{(0,0),(1,1)\big\}$$

for all $0 \leq x \leq 1$, $0 \leq y \leq 1$, (see Figure 4.10).

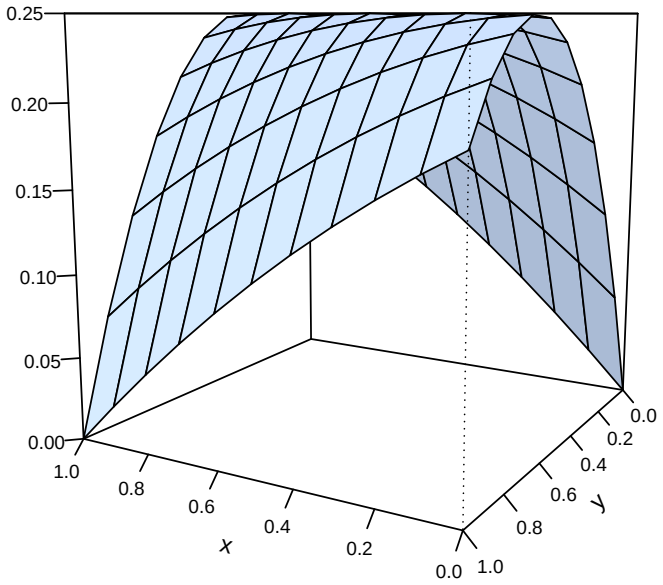


Figure 4.10: η_1 with $\varrho = 0.25$

The function $\eta_1(x,y)$ decreases monotonically as the L_2 distance from the line in (4.5) increases. This fact suggests the following result.

Proposition 4.4. *If $\mu \in \mathcal{M}$, $\mu(x,y) = \psi(x,y)/\eta_1(x,y)$, and $\psi(x,y) = \psi(x',y')$, then*

$$|\varrho x + (1-\varrho)y - \frac{1}{2}| \geq |\varrho x' + (1-\varrho)y' - \frac{1}{2}| \;\Rightarrow\; \mu(x,y) \geq \mu(x',y'). \tag{4.6}$$

Proof. Since the function η_1 decreases monotonically as the L_2 distance from the line in (4.5) increases, when ψ is constant, μ increases monotonically as the L_2 distance

from the line in (4.5) increases, i.e.

$$\frac{|\varrho x + (1-\varrho)y - \frac{1}{2}|}{\sqrt{\varrho^2 + (1-\varrho)^2}} \geq \frac{|\varrho x' + (1-\varrho)y' - \frac{1}{2}|}{\sqrt{\varrho^2 + (1-\varrho)^2}} \;\Rightarrow\; \mu(x,y) \geq \mu(x',y'),$$

and the result follows immediately. ■

If $\mu(x,y) = (x-y)/\eta_1(x,y)$, and points (x,y) and (x',y') appear on the line $x-y = \tau$, the following two examples show how the presence of η_1 affects the relative ranking of these points. First, consider the case of a balanced collection when $|T| = |F|$. In this case, $\varrho = 1/2$, $(1-\varrho) = 1/2$, $\eta_1(x,y)$ is proportional to

$$\widehat{\eta}_1(x,y) = (x+y)(2-x-y), \tag{4.7}$$

and the line in (4.5) becomes $x + y = 1$, the distance from a point (x,y) to the line is $\delta = |x+y-1|/\sqrt{2}$, and

$$|x+y-1| \geq |x'+y'-1| \Rightarrow \mu(x,y) \geq \mu(x',y').$$

The "pure" solution $(\tau, 0)$ and the point $(1, 1-\tau)$ will be ranked the highest, and the point, $((1+\tau)/2, (1-\tau)/2)$ which is on the line, will be ranked the lowest (see Figure 4.11a).

Next, consider the case of a very skewed collection where, $\lim_{\varrho\to 0} \eta_1(x,y)$ is

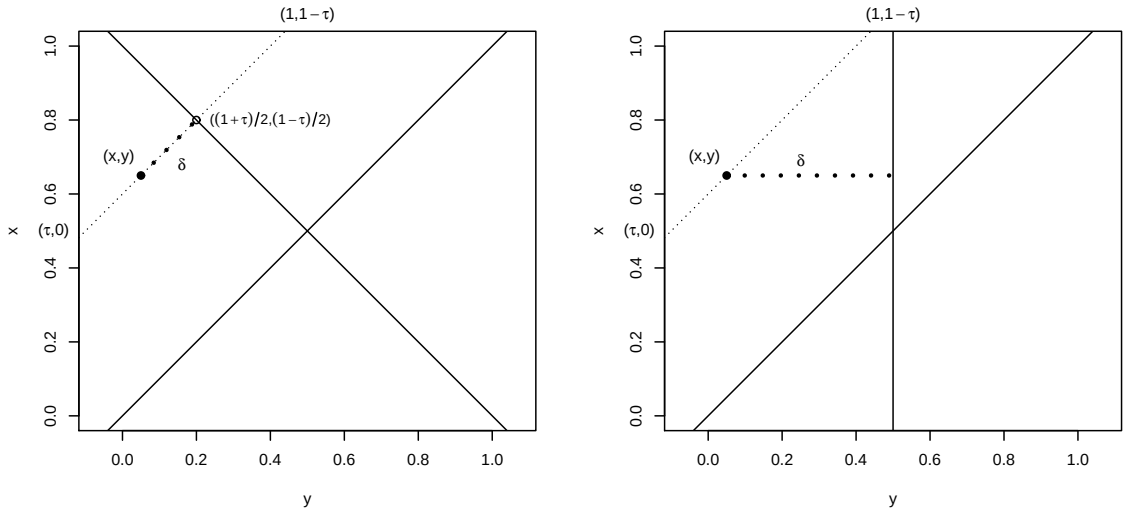
$$\widetilde{\eta}_1(x,y) = y(1-y), \tag{4.8}$$

$$\lim_{\varrho\to 0} \varrho x + (1-\varrho)y - \frac{1}{2} = y - \frac{1}{2}$$

so we have the line $y = \frac{1}{2}$, the distance from a point (x,y) to the line is $|y - 1/2|$, and

$$\left|y - \frac{1}{2}\right| \geq \left|y' - \frac{1}{2}\right| \Rightarrow \mu(x,y) \geq \mu(x',y').$$

The pure solution $(\tau, 0)$ will be the highest ranked point and the lowest ranked point will be $(1, 1-\tau)$ (see Figure 4.11b).



(a) Balanced Collection

(b) Skewed Collection

Figure 4.11: Distance to Collection Noise Line

The following result justifies the behavior of $\mu \in \mathcal{M}$ as in Proposition 4.4.

Proposition 4.5. *Let $j \in V$, then if (x_j, y_j) is on the line $\varrho x_j + (1-\varrho)y_j = \frac{1}{2}$, j is a noise feature.*

Proof. Since

$$\varrho x_j + (1-\varrho)y_j \;=\; \mathrm{P}(u \in T)\,\mathrm{P}(u_j = 1 \mid u \in T) + \mathrm{P}(u \in F)\,\mathrm{P}(u_j = 1 \mid u \in F) \;=\; \mathrm{P}(u_j = 1),$$

$\mathrm{P}(u_j = 1) = \frac{1}{2}$ and so $\mathrm{P}(u_j = 0) = \frac{1}{2}$. Therefore, the feature is equally likely to be present as it is to be absent from any document, regardless of whether it is relevant or irrelevant, which exactly coincides with j being a noise feature. ■

We will refer to the line in (4.5) as the *line of collection noise* and will refer to features that appear on this line as *collection noise features*. It is interesting to note that, as was alluded to in §1.1, for a highly skewed collection, collection noise is only measured using the proportion of *irrelevant* documents in which a feature appears.

4.5.3 Strong Collection Noise

The feature ranking functions μ_{23} and μ_{24} selected relatively few stopwords and achieved relatively high separation in experiments in §3.10. The denominators of these functions are

$$\text{den}(\mu_{23}(x,y)) = x(1-x) + y(1-y) \tag{4.9}$$

and

$$\text{den}(\mu_{24}(x,y)) = \sqrt{x(1-x)} + \sqrt{y(1-y)}.$$

In ROC space, these functions only differ in a monotonic transformation and therefore in this section, we will only consider $\text{den}(\mu_{23}(x,y))$ which we will denote as $\eta_2(x,y)$.

We begin by noting that $\nabla\eta_2(x,y) = [1-2x, 1-2y]$ and that $\nabla\eta_2(x,y) = 0$ has unique solution $(\frac{1}{2},\frac{1}{2})$. Next, since

$$\nabla^2\eta_2(x,y) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

and

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -2x^2 - 2y^2 < 0$$

for all x,y, $\nabla^2\eta_2(x,y)$ is negative definite. Therefore, $\eta_2(x,y)$ is strictly concave and $(\frac{1}{2},\frac{1}{2})$ is its global maximum. It can also be seen that

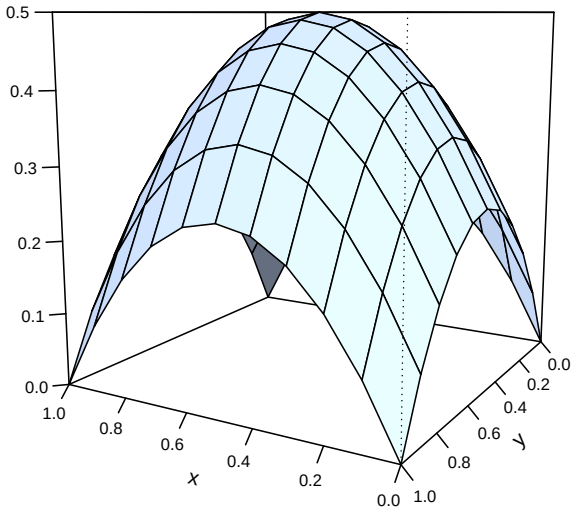
$$\operatorname{argmin}\eta_2(x,y) = \big\{(0,0),(0,1),(1,0),(1,1)\big\}$$

for all $0 \le x \le 1$, $0 \le y \le 1$, (see Figure 4.12a).

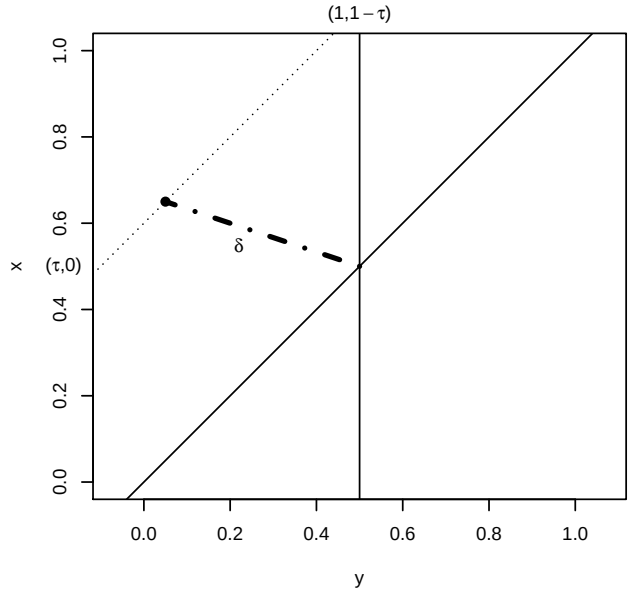
The function $\eta_2(x,y)$ decreases monotonically as the L_2-distance from point $(\frac{1}{2},\frac{1}{2})$ increases. This fact suggests the following result.

Proposition 4.6. *If $\mu \in \mathcal{M}$, $\mu(x,y) = \psi(x,y)/\eta_2(x,y)$, and $\psi(x,y) = \psi(x',y')$, then*

$$\sqrt{\big(x-\frac{1}{2}\big)^2 + \big(y-\frac{1}{2}\big)^2} \ge \sqrt{\big(x'-\frac{1}{2}\big)^2 + \big(y'-\frac{1}{2}\big)^2} \;\Rightarrow\; \mu(x,y) \ge \mu(x',y'). \tag{4.10}$$



(a) η_2



(b) Distance to Strong Collection Noise

Figure 4.12: Strong Collection Noise

Proof. Since the function η_2 decreases monotonically as the L_2 distance from the point $(\frac{1}{2}, \frac{1}{2})$ increases, when ψ is constant, μ increases monotonically as the L_2 distance from the point $(\frac{1}{2}, \frac{1}{2})$ increases. ■

Suppose, $\mu(x, y) = (x - y)/\eta_2(x, y)$, the points (x, y), (x', y') appear on the line $x - y = \tau$, and (x, y) is further from the point $(\frac{1}{2}, \frac{1}{2})$, than (x', y'). In this case, η_2 will result in (x, y) being ranked higher than (x', y'). Of the points on the line $x - y = \tau$ the "pure" solution $(\tau, 0)$ and the point $(1, 1 - \tau)$ will be ranked the highest and the point closest to the point $(\frac{1}{2}, \frac{1}{2})$ the will be ranked the lowest. (see Figure 4.12b). Note that Proposition 4.6 could have been stated in terms of the L_1 distance, i.e.

$$\left|x - \frac{1}{2}\right| + \left|y - \frac{1}{2}\right| \geq \left|x' - \frac{1}{2}\right| + \left|y' - \frac{1}{2}\right| \;\Rightarrow\; \mu(x, y) \geq \mu(x', y').$$

The following result justifies the behavior of $\mu \in \mathcal{M}$ as in Proposition 4.6.

Proposition 4.7. *Let $j \in V$, then if $x_j = \frac{1}{2}$ and $y_j = \frac{1}{2}$, j is a noise feature.*

Proof. Since $x_j = \mathrm{P}(u_j = 1 \mid u \in T)$ and $y_j = \mathrm{P}(u_j = 1 \mid u \in F)$, we have

$$\mathrm{P}(u_j = 1 \mid u \in T) = \mathrm{P}(u_j = 0 \mid u \in T) = \mathrm{P}(u_j = 1 \mid u \in F) = \mathrm{P}(u_j = 0 \mid u \in F) = \frac{1}{2},$$

so that there is equal probability that the feature will appear in *any* document in the collection, which exactly coincides with j being a noise feature. ■

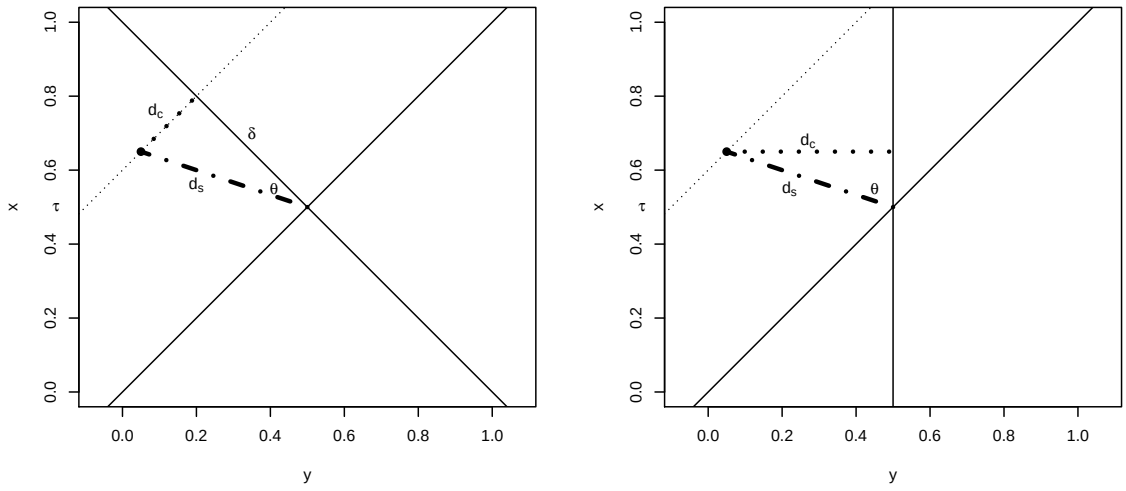
We shall call a feature as in the result a *strong collection noise feature.*

4.5.4 Collection Noise and Strong Collection Noise

In this short section we discuss the relationship between collection noise and strong collection noise. Let us consider a feature family defined by the line $x - y = \tau$ and a point $(\hat{x}, \hat{y})$ that corresponds to a specific feature and is on that line. Let δ denote the distance from the line, $x - y = \tau$ to the line of class noise, $x - y$. The strong collection noise is proportional to the distance from the point $(1/2, 1/2)$ to the point $(\hat{x}, \hat{y})$ which we denote as d_s.

We begin by considering a balanced collection with $|T| = |F|$ and $\varrho = \frac{1}{2}$. In this case η_1 becomes $\widehat{\eta}_1$ (see (4.7)) which is proportional to the distance from the line $x + y = 1$ to the point $(\hat{x}, \hat{y})$. The relationship between the collection noise and strong collection noise is depicted in Figure 4.13a. For a fixed δ, since $d_s = \sqrt{d_c^2 + \delta^2}$, d_s clearly varies monotonically with d_c. Since $\sin(x)$ is close to linear on $[0, \pi/2]$, this can also be seen by noticing that $d_c = d_s \sin(\theta)$.

Next we consider the case of a skewed collection when $\varrho \to 0$. In this case η_1 becomes $\widetilde{\eta}_1$ (see (4.8)) which is proportional to the distance from the line $y = 1/2$ to the point $(\hat{x}, \hat{y})$. The relationship between the collection noise and strong collection noise is depicted in Figure 4.13b. Again, since $\sin(x)$ is close to linear on $[0, \pi/2]$, and $d_c = d_s \sin(\theta)$, d_s clearly varies monotonically with d_c.



(a) Balanced Collection

(b) Skewed Collection

Figure 4.13: Collection Noise versus Strong Collection Noise

4.6 "Noisy" Functions

Using the information developed in the earlier sections, we now take one last look at μ_8. We begin by considering some simple statistics regarding the performance of μ_8 on the *fuel* topic. This topic has 28 relevant documents, 1975 irrelevant documents, a total of

12501 features, of which 383[5] are stopwords and 12118 are non-stopwords. As shown in §4.1, $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$ contained 14 stopwords. Even without considering the fact that an additional four of the selected features were in $\omega(V, K, a+b\text{-RANKING}, \uparrow)$ and therefore could also reasonably be assumed to be noise features, this performance, which is summarized in Table 4.6 strongly suggests that the amount of selected noise

	Stopword	Non-stopword	
$j \in S$	14	11	$K = 25$
$j \notin S$	369	12107	12476
	383	12118	12501

Table 4.6: Stopwords vs Non-stopwords for $\omega(V, K, \mu_8\text{-RANKING}, \uparrow)$

is not a random phenomenon, but due to some inherent characteristic of μ_8. Further support for this suggestion is offered by the data in Appendix K where in Table K.1 and Figure K.2 we see that μ_8 had the largest value of $\sigma_K[J]$, and along with μ_{19} had the second largest value of $\dot{\sigma}[J]$.

One explanation for this situation is offered by the Feature Monotonicity Principle, which says, "Non-noise features in textual data sets, that provide substantial separation, are highly ranked positive features.". However, notice that

$$\begin{aligned}\mu_8(x,y) = &\ x(1-y) + y(1-x)\\ = &\ \mathrm{P}((\{u_j = 1 \mid u \in T\} \cap \{u_j = 0 \mid u \in F\}) \cup\\ &\ (\{u_j = 0 \mid u \in T\} \cap \{u_j = 1 \mid u \in F\})).\end{aligned}$$

which corresponds to the XOR (i.e. the exclusive OR) function, which will include highly ranked positive feature as well as highly ranked *negative* features in the set of features it ranks highly, and it is these negative features that are ostensibly noise features. It should, however, be mentioned that the use of μ_8 *is* perfectly in concert with alternate form of the Feature Monotonicity Principle which applies when F is class homogeneous and $|T| \approx |F|$.

[5]Not all 571 stopwords on the SMART list appear in for the fuel topic.

Now, notice that $\mu_8(1/2, 1/2) = 1/2$, so given the Zipfian distribution of textual data, it assigns a relatively large value to the point of strong collection noise. Further,

$$\nabla\mu_8(x, y) = [1 - 2y, 1 - 2x]^T,$$

$$\nabla^2\mu_8(x, y) = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix},$$

$\det(\nabla^2) = -4$, and therefore as can be seen in Figure 4.14, $(1/2, 1/2)$ is a saddle point. Therefore, not only does μ_8 specifically assign a relatively large value to the point of

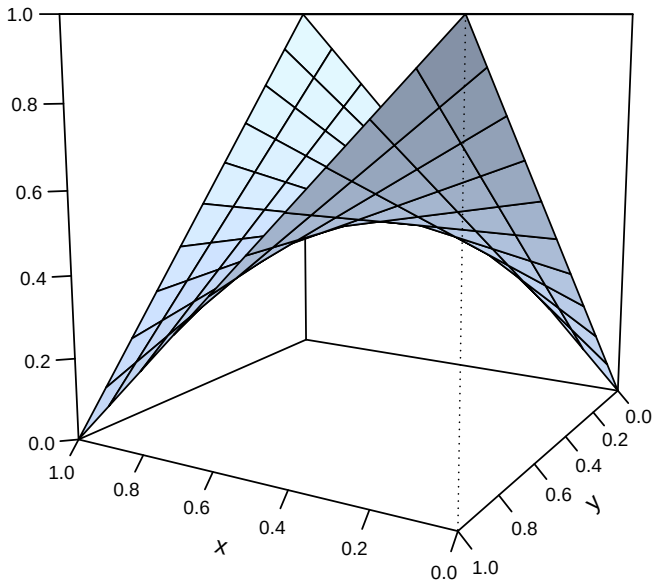


Figure 4.14: μ_8

strong collection noise, it also assigns relatively large values to all points in a non-trivial neighborhood of the point of strong collection noise.

4.7 Separation to Noise Ratios

In this section, motivated by our discussion in §4.5.1, §4.5.2 and §4.5.3, we characterize a subset of the feature ranking functions that selected relatively few stopwords and achieved relatively high separation in experiments in §3.10. We will refer to any finite, positive function $f : [0,1]^2 \rightarrow \mathbb{R}$ whose value increases with the distance from a noise feature as a *separating function*. Similarly, we will refer to any finite, positive function $f : [0,1]^2 \rightarrow \mathbb{R}$ whose value decreases with the distance from a noise feature or set of noise features as a *noise function*. Clearly, if f is a separating function then $-f$ is a noise function and vice versa.

We shall specifically refer to a finite, positive function as a *class separating function* (e.g. $|x - y|$) or *class noise function* if it increases or decreases respectively with the distance from the line of class noise. We shall refer to a finite, positive function as a *collection separating function* or *collection noise function* (e.g. η_1, or in the case of a balanced collection $\widehat{\eta}_1$, or in the case of a skewed collection $\widetilde{\eta}_1$) if it increases or decreases respectively with the distance from the line of collection noise. Finally, we shall refer to a finite, positive function as a *strong collection separating function* or *strong collection noise function* (e.g. η_2) if it increases or decreases respectively with the distance from the point of strong collection noise. We shall denote the set of all separating functions as Ψ and the set of all noise functions as $\aleph$.

Using this terminology, several of the feature ranking functions, e.g. $|\mu_9|$, μ_{23} and $|\mu_{24}|$, that selected relatively few stopwords and achieved relatively high separation in experiments in §3.10, can be seen to be as the *ratio of two distance functions*, and more specifically as the *ratios of separating functions to noise functions*. We shall denote the set of all such functions as $\mathcal{M}^{\Psi/\aleph}$.

It is interesting to note that the situation is similar with μ_{11}. Consider the case of a balanced collection with $T = F$, in which μ_{11} becomes

$$\widehat{\mu}_{11}(x,y) = \frac{x(1-x) + y(1-y)}{(x+y)(2-x-y)} = \frac{\eta_2(x,y)}{\widehat{\eta}_1(x,y)}.$$

That is, $\text{num}(\widehat{\mu}_{11}) = \eta_2$ and so $\text{num}(\widehat{\mu}_{11}) \in \aleph$, and $\text{den}(\widehat{\mu}_{11}) = \widehat{\eta}_1$ and so $\text{den}(\widehat{\mu}_{11}) \in \aleph$.

Therefore, $\widehat{\mu}_{11}$ is the ratio of a strong collection noise function to a collection noise function. However, we claim that this is a bit of a ruse and that $\widehat{\mu}_{11}$ is actually in $\mathcal{M}^{\Psi/\aleph}$. To see this, first note that since features with smaller values of $\widehat{\mu}_{11}$ are considered better than those with larger values, we could equivalently consider

$$\widetilde{\mu}_{11}(x,y) = \frac{-\eta_2(x,y)}{\widehat{\eta}_1(x,y)} = \frac{x(x-1)+y(y-1)}{(x+y)(2-x-y)}$$

for which features with larger values would be considered better than those with smaller values. Next, we note that $-\eta_2$ is a separating function which shows the claim.

As can be seen in Figure 4.15a, Figure 4.15b, Figure 4.15c, the functions in $\mathcal{M}^{\Psi/\aleph}$ are characterized by the fact that

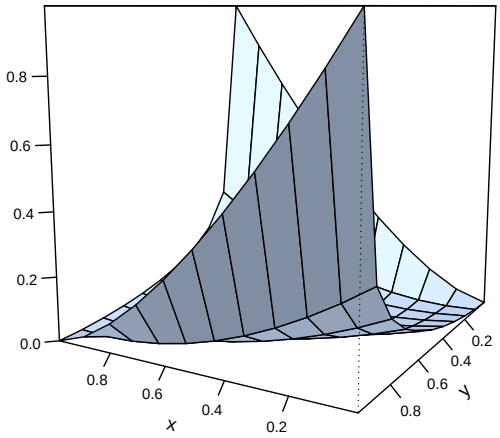
1. They take their minimum value along the line $x = y$.
2. They take their maximum value at $(x, y) = (1, 0)$ and $(x, y) = (0, 1)$.
3. They are monotonically non-decreasing from the line $x = y$ to the points $(x, y) = (1, 0)$ and $(x, y) = (0, 1)$.

While μ_{12} and μ_{21} are not in $\mathcal{M}^{\Psi/\aleph}$ they do share the characteristics of these functions.

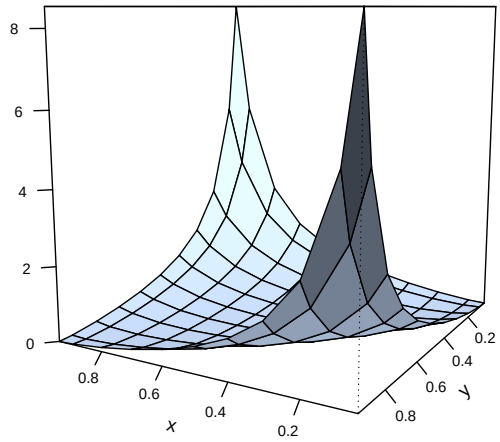
We now address the fact that the terminology we have just introduced does not accommodate functions such as μ_9 and μ_{24} whose numerator $x - y$ is 0 at the line of class noise, and when $x > y$, i.e when the feature is positive, it takes progressively larger positive values as the distance from the line of class noise increases, and when $y > x$, i.e. when the feature is negative, it takes progressively larger positive values as the distance from the line of class noise increases. If ψ is a separating function then, we say that ψ' is a *signed class separating function* if

$$\psi'(x,y) = \begin{cases} \psi(x,y) & \text{if } x \geq y, \\ -\psi(x,y) & \text{otherwise.} \end{cases}$$

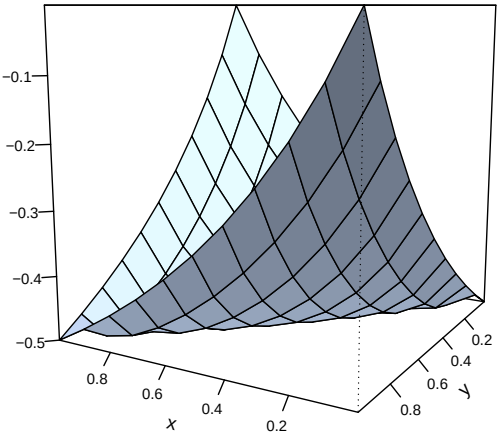
So for example, if $\psi(x, y) = |x-y|$, then $\psi'(x, y) = x-y$. When k is even, then $(x-y)^k$ is a class separating function and when k is odd it is a signed class separating function. Note, that if ψ' is a signed class separating function, then $\psi'(x, y) = -\psi'(y, x)$, that is



(a) $|\mu_9|$ with $|T| = 28$ and $|F| = 1975$



(b) μ_{23}



(c) $\tilde{\mu}_{11}$

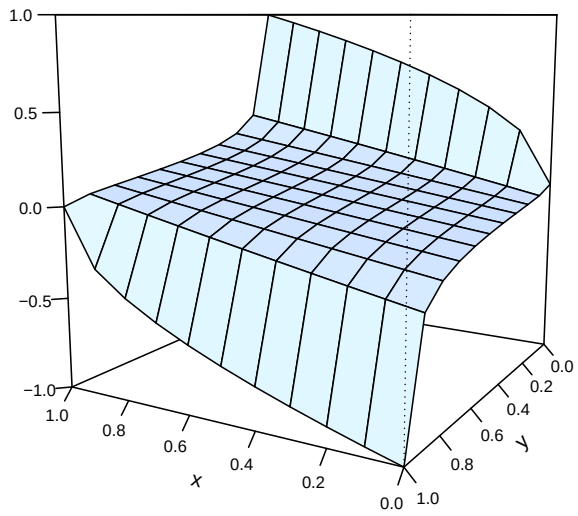
Figure 4.15: Examples of functions in $\mathcal{M}^{\Psi/\aleph}$

the value of a feature is negated when its corresponding point is reflected over the line $x = y$. We will let $\Psi^{\pm}$ represent the set of all signed separating functions. Using this terminology, we note that μ_9 and μ_{24} are *ratios of signed separating functions to noise functions*. We will denote the set of all such functions as $\mathcal{M}^{\Psi^{\pm}/\aleph}$.

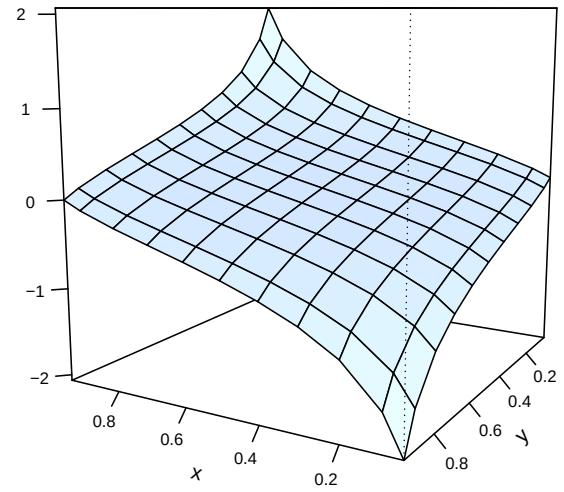
As can be seen in Figure 4.16a and Figure 4.16b, the functions in $\mathcal{M}^{\Psi^{\pm}/\aleph}$ are characterized by the fact that

1. They take their maximum value at $(x, y) = (1, 0)$.
2. They take their minimum value at $(x, y) = (0, 1)$.
3. They take the value of 0 along the line $x = y$.
4. They are monotonically non-decreasing from the point $(x, y) = (0, 1)$ to the point $(x, y) = (0, 1)$.

We close this section by remarking that it is interesting to note the even though the foundations of these functions are quite varied, that most of the Boolean feature ranking functions that selected relatively few stopwords and achieved relatively high separation in experiments in §3.10 can be written as the ratios of separating functions to noise functions. Since separation and noise are competing criteria, we cannot reasonably expect to find a single feature ranking function whose value increases as the separation increases and the noise decreases for all $(x, y) \in [0, 1]^2$. For example, let $(x_i, y_i) = (0.25, 0)$ and $(x_j, y_j) = (0.5, 0)$. Clearly (x_j, y_j) provides an improvement in separation over (x_i, y_i), but it is also closer to the point of strong collection noise. The functions in $\mathcal{M}^{\Psi/\aleph}$ and $\mathcal{M}^{\Psi^{\pm}/\aleph}$ represent an approach for addressing this situation which we will discuss further in §6.4.



(a) μ_9 with $|T| = 28$ and $|F| = 1975$



(b) μ_{24}

Figure 4.16: Examples of functions in $\mathcal{M}^{\Psi^{\pm}/\aleph}$

Chapter 5

The Space of Boolean Feature Ranking Functions

In this chapter we will characterize the space of *desirable* Boolean feature ranking functions. Motivated by our discussions in previous chapters we identify a set of axioms which we believe that all feature ranking functions should satisfy. We then consider a very general subset of the set of functions which satisfy these axioms, namely the set of all such functions that can be represented as a linear combination of the elements of some finite basis set of real-valued functions on $[0, 1]$. Next we consider a specific basis set and state some results which facilitate the study, through the use of linear programming theory and tools, of the space of all desirable functions it generates.

5.1 Axioms

In this section we present a set of axioms that we believe feature ranking functions should satisfy. Our goal in selecting these axioms was to a set of conditions that is

- *minimal* – removal of any axiom will yield a class of feature ranking functions which are in some sense *not* desirable,
- *incontrovertible* – each axiom should clearly represent a desirable property of a feature ranking function,
- *relatively weak* – the set of axioms should not be so specific as to define an overly restricted class of feature ranking functions, and
- *linear* – the axioms should be linear in x and y.

Our intent is that these axioms will not to only allow us to generate the feature ranking functions that performed well in the experiments in §3.10, but that they will identify

a larger family of functions whose characteristics would result in similar or better performance. We shall assume that all *desirable* feature ranking functions, $\mu\colon [0,1]^2 \to \mathbb{R}$, satisfy the following axioms

$$\mu(1,0) = 1 \tag{A1}$$

$$\mu(0,1) = -1 \tag{A2}$$

$$x = y \;\Rightarrow \mu(x,y) = 0 \tag{A3}$$

$$x \geq x' \text{ and } y \leq y' \;\Rightarrow \mu(x,y) \geq \mu(x',y') \tag{A4}$$

where $(x,y) \in [0,1]^2$ and will denote the set of all such functions as $\mathcal{M}^*$. Additionally, we will consider it desirable for feature ranking functions to satisfy

$$\mu(x,y) = -\mu(y,x) \tag{A5}$$

$$\mu(x,y) = \mu(1-y, 1-x) \tag{A6}$$

and will denote functions that satisfy (A1)–(A5) as $\hat{\mathcal{M}}^*$ and those that satisfy (A1)–(A6) as $\tilde{\mathcal{M}}^*$.

The points $(1,0)$ and $(0,1)$ correspond to the most positive and most negative features as well as those that achieve the largest separation of T and F and (A1) and (A2) simply require that these features be assigned positive and negative values respectively. It should be noted that the specific choices of 1 and -1 are not important and that any positive and negative values would suffice. The following result shows that a feature ranking function which satisfies these axioms is scaled between -1 and 1 and that its maximum and minimum values occur at $(1,0)$ and $(0,1)$ respectively.

Proposition 5.1. *If* $\mu\colon [0,1]^2 \to \mathbb{R}$ *satisfies (A1), (A2) and (A4) then*

$$\mu(x,y) \leq 1 \tag{5.1}$$

$$\operatorname{argmax} \mu(x,y) = (1,0) \tag{5.2}$$

$$\mu(x,y) \geq -1 \tag{5.3}$$

$$\operatorname{argmin} \mu(x,y) = (0,1) \tag{5.4}$$

for all $(x, y) \in [0, 1]^2$.

Proof. (5.1) and (5.2) follow immediately from the fact that from (A1) and (A4) we have that

$$\mu(1 - \Delta x, 0 + \Delta y) \leq \mu(1, 0) = 1$$

for $\Delta x, \Delta y \geq 0$, and (5.3) and (5.4) follow similarly. ■

(A3) requires that features which appear with the same relative frequency in T and F be assigned a value of 0. This coincides with the results in §4.5.1 that showed that such features are class noise features that do not separate T and F. The following result shows that a feature which either appears in *no* documents or in *all* documents should be assigned a value of 0.

Proposition 5.2. *If* $\mu\colon [0, 1]^2 \to \mathbb{R}$ *satisfies (A3) then*

$$\mu(0, 0) = 0 \tag{5.5}$$

$$\mu(1, 1) = 0. \tag{5.6}$$

Proof. Follows immediately from the fact that $(0, 0)$ and $(1, 1)$ satisfy $x = y$. ■

(A4) can be written as

$$\mu(x, y) \leq \mu(x + \Delta x, y - \Delta y)$$

where $\Delta x, \Delta y \geq 0$ and therefore it requires that a feature which appears in relatively more documents in T and relatively fewer documents in F than another feature, should be assigned values that are at least as large as those assigned to the other feature. (A4) can also be written as

$$\mu(x, y) \geq \mu(x - \Delta x, y + \Delta y)$$

where $\Delta x, \Delta y \geq 0$ and therefore it also requires that features which appear in relatively fewer documents in T and relatively more documents in F than other features, should be assigned values that are no larger than those assigned to the other feature.

If (α, β) is some direction with $\alpha, \beta \geq 0$ and if $\epsilon > 0$, then (A4) can be written as

$$\mu(x, y) \leq \mu(x + \alpha\epsilon, y - \beta\epsilon),$$

or equivalently as

$$\mu(x, y) \geq \mu(x - \alpha\epsilon, y + \beta\epsilon).$$

By writing

$$\lim_{\epsilon \to 0} \frac{\mu(x + \alpha\epsilon, y - \beta\epsilon) - \mu(x, y)}{\epsilon} \geq 0$$

or equivalently

$$\lim_{\epsilon \to 0} \frac{\mu(x - \alpha\epsilon, y + \beta\epsilon) - \mu(x, y)}{\epsilon} \leq 0,$$

we see that (A4) says that the directional derivative in the direction $(\alpha, -\beta)$ is non-negative and is non-positive in the direction $(-\alpha, \beta)$.

Geometrically, (A4) requires that a feature corresponding to a point (x, y) be assigned a value at least as large a feature corresponding to the point (x', y') when (x, y) lies to the *northwest* of (x', y'). Similarly, it requires that a feature corresponding to a point (x', y') be assigned a value at no larger than a feature corresponding to the point (x, y) when (x', y') lies to the *southwest* of (x, y). It is also interesting to note that if (A4) were strengthened to be

$$x \geq x' \text{ and } y \leq y' \text{ and } (x, y) \neq (x', y') \Rightarrow \mu(x, y) \geq \mu(x', y')$$

or equivalently,

$$x > x' \text{ and } y \leq y' \text{ or } x \geq x' \text{ and } y < y' \Rightarrow \mu(x, y) \geq \mu(x', y')$$

that, following §1.4, the antecedent would be equivalent to requiring that (x, y) *dominated* (x', y').

While restricting ourselves to linear axioms obviously limits the characteristics that we can require of feature ranking functions, the functions that satisfy (A4), which we will denote as $\mathcal{M}^{(A4)}$, can be seen to share properties of some of the classes of functions

we discussed in §4. For example, the following result shows that $\Psi^{\pm} \subseteq \mathcal{M}^{(A4)}$.

Proposition 5.3. *If $\mu\colon [0,1]^2 \to \mathbb{R}$ satisfies (A4), then μ is a signed separating function.*

Proof. It is enough to show that if $x \geq x'$ and $y \leq y'$ then $x - y \geq x' - y'$. Writing the antecedent of (A4) as $x - x' \geq 0$ and $y' - y \geq 0$ and adding the inequalities yields $x - x' + y' - y \geq 0$, which can be rewritten as $x - y \geq x' - y'$. ■

In order to provide another example of a class of functions that $\mathcal{M}^{(A4)}$ includes we now remark that

$$\psi(x,y) \geq \psi(x',y') \text{ and } \eta(x,y) \leq \eta(x',y') \Rightarrow \mu(x,y) \geq \mu(x',y') \tag{5.7}$$

and

$$\psi(x,y) \leq \psi(x',y') \text{ and } \eta(x,y) \leq \eta(x',y') \Rightarrow \mu(x,y) \leq \mu(x',y') \tag{5.8}$$

are necessary conditions for $\mu \in \mathcal{M}^{\Psi^{\pm}/\aleph}$, and therefore the sets of functions that satisfy these conditions are important subsets of $\mathcal{M}^{\Psi^{\pm}/\aleph}$. On certain subsets of $[0,1]^2$, the set of functions which satisfy (A4) and (5.7) are identical. For example, suppose we are given a skewed collection, the feature ranking function

$$\mu(x,y) = \frac{x-y}{\tilde{\eta_1}(x,y)} = \frac{x-y}{\sqrt{y(1-y)}}$$

and let

$$X_1 = \{(x,y) \,:\, x \geq y \text{ and } y \leq 1/2\},$$

and

$$X_2 = \{(x,y) \,:\, x \leq y \text{ and } y \geq 1/2\},$$

then clearly the set of functions which satisfy (A4) and (5.7) on X_1 are identical as are the set of functions that satisfy (A4) and (5.8) on X_2. As another example, suppose

we are given the feature ranking function

$$\mu(x,y) = \frac{x-y}{\tilde{\eta}_2(x,y)} = \frac{x-y}{\sqrt{x(1-x)}+\sqrt{y(1-y)}}$$

and let

$$Y_1 = \{(x,y) : x \geq 1/2 \text{ and } y \leq 1/2\},$$

and

$$Y_2 = \{(x,y) : x \leq 1/2 \text{ and } y \geq 1/2\},$$

then clearly the set of functions which satisfy (A4) and (5.7) on Y_1 are identical as are the set of functions which satisfy (A4) and (5.8) on Y_2. So, on certain subsets of $[0,1]^2$, $(A4)$ captures an important characteristic of functions in $\mathcal{M}^{\Psi^{\pm}/\aleph}$; the value assigned to features should increase as the separation increases and the noise decreases.

The following shows that if $\mu \in \mathcal{M}^*$, when (x,y) corresponds to a negative feature, then $\mu(x,y) < 0$ and when (x,y) corresponds to a positive feature, then $\mu(x,y) > 0$.

Proposition 5.4. *If $\mu\colon [0,1]^2 \to \mathbb{R}$ satisfies A3 and A4 then $\mu(x,y) \leq 0$ when $x < y$ and $\mu(x,y) \geq 0$ when $x > y$ with equality in both cases occurring when $x = y$.*

Proof. Let (x',y') be any point in $[0,1]$ such that $x' > y'$. Clearly there exist $x, \Delta x, \Delta y \in [0,1]$ so that any such (x',y') can be written as $(x+\Delta x, x-\Delta y)$. Now from A4 and A3 we have $\mu(x',y') = \mu(x+\Delta x, x-\Delta y) \geq \mu(x,x) = 0$. The case where $x < y$ follows similarly. ■

(A5) says that μ must be an *alternating polynomial*. It strengthens Proposition 5.4 and requires that a positive feature reflected over the line of class noise, i.e. a negative feature, have the same value but the opposite sign and vice versa. It can also be considered to generalize (A1) and (A2). Additional examples that motivate this axiom include $\mu(1,x) = -\mu(x,1)$ and $\mu(0,x) = -\mu(x,0)$ for $x \in [0,1]$. The first example corresponds to a feature which appears in all relevant documents and some proportion x of the irrelevant documents, and a feature which appears in all of the irrelevant documents and some proportion x of the relevant documents. The second example corresponds to a

feature which appears in none of the relevant documents and some proportion x of the irrelevant documents, and a feature which appears in some proportion x of the relevant documents and none of the irrelevant documents. Note that to support "discounting" negative features, this axiom could be generalized to

$$\mu(x, y) = -\lambda\mu(y, x)$$

where $y > x$ and $0 \leq \lambda \leq 1$.

(A6) requires that features whose corresponding points are reflections over the line of collection noise have same value. As discussed in 4.5.2, when considering a family of features, $x - y = \tau$, the value of some of the best performing feature ranking functions, increased with the distance from the line of collection noise. Such functions (e.g. μ_{23} and μ_9 when $|T| = |F|$) are independent of the class skew, that is, they are not dependent on $|T|$ and $|F|$. Motivated by this fact, this axiom requires that this increase be symmetric about the line of collection noise. In contrast, (A6) does not hold for μ_9 with arbitrary $|T|$ and $|F|$.

5.2 Linearization

Having defined the set of functions $\mathcal{M}^*$ in §5.1, in this section we will construct and begin to study a rich subset of $\mathcal{M}^*$. Consider a finite set

$$\mathcal{F} = \{f \ : \ f \colon [0, 1]^2 \to \mathbb{R}\}$$

and let

$$\mathcal{M}[\mathcal{F}] = \{\mu_{\mathbf{c}} \ : \ \mu_{\mathbf{c}}(x, y) = \sum_{f \in \mathcal{F}} c_f f(x, y) \text{ and } \mathbf{c} \in \mathbb{R}^{|\mathcal{F}|}\}.$$

That is, $\mathcal{F}$ is some finite set of real-valued functions on $[0, 1]^2$, $\mathcal{M}[\mathcal{F}] \subseteq \mathcal{M}$ is the set all functions in $\mathcal{M}$ which can be written as a linear combination of the elements of $\mathcal{F}$, and the set $\mathcal{F}$ forms a basis of $\mathcal{M}[\mathcal{F}]$. In this section we will consider $\mathcal{M}^*[\mathcal{F}] = \mathcal{M}^* \cap \mathcal{M}[\mathcal{F}]$, the set of all functions which can be represented as a linear combination of the elements of $\mathcal{F}$ and which that satisfy the axioms presented in §5.1. We will also study the set

$\hat{\mathcal{M}}^*[\mathcal{F}] = \mathcal{M}^* \cap \hat{\mathcal{M}}^*[\mathcal{F}]$ and the set $\tilde{\mathcal{M}}^*[\mathcal{F}] = \mathcal{M}^* \cap \tilde{\mathcal{M}}^*[\mathcal{F}]$.

We now notice that the axioms presented in §5.1 can be written as a system of linear equations and inequalities in $\mathbf{c} \in \mathbb{R}^{|\mathcal{F}|}$. Specifically, (A1) and (A2) can both be written as a linear equation, (A3), (A5), and (A6), can be written as an infinite set of inequalities, one for each $x \in [0,1]$, and (A4) can be written as a infinite set of inequalities, one for each combination of $(x,y) \in [0,1]^2$ and $(x',y') \in [0,1]^2$ that satisfies its antecedent.

Clearly there is a one-to-one correspondence between the functions in $\mathcal{M}[\mathcal{F}]$ and the set of vectors

$$M[\mathcal{F}] = \{\mathbf{c} : \mathbf{c} \in \mathbb{R}^{|\mathcal{F}|}\},$$

so in addition to being interested in the algebraic properties of the sets of functions $\mathcal{M}^*[\mathcal{F}]$, $\hat{\mathcal{M}}^*[\mathcal{F}]$, and $\tilde{\mathcal{M}}^*[\mathcal{F}]$, we are also interested in the geometric properties of the corresponding sets of vectors, i.e. we are also interested in the sets

$$M^*[\mathcal{F}] = \{\mathbf{c} \in \mathbb{R}^{|\mathcal{F}|} : \mu_{\mathbf{c}} \in \mathcal{M}^*\},$$

$$\hat{M}^*[\mathcal{F}] = \{\mathbf{c} \in \mathbb{R}^{|\mathcal{F}|} : \mu_{\mathbf{c}} \in \hat{\mathcal{M}}^*\},$$

and

$$\tilde{M}^*[\mathcal{F}] = \{\mathbf{c} \in \mathbb{R}^{|\mathcal{F}|} : \mu_{\mathbf{c}} \in \tilde{\mathcal{M}}^*\}$$

respectively, which are defined as the intersection of the linear inequalities corresponding to the associated axioms. We now state a basic result that demonstrates the relationship between a geometric property of the sets of vectors and an algebraic property of the sets of functions that we have just introduced.

Proposition 5.5. *The sets of vectors $M^*[\mathcal{F}]$, $\hat{M}^*[\mathcal{F}]$, and $\tilde{M}^*[\mathcal{F}]$ are convex.*

Proof. Follows immediately from the fact that these sets are defined as the intersection of a set of linear inequalities. ■

Corollary 5.1. *The sets of functions $\mathcal{M}^*[\mathcal{F}]$, $\hat{\mathcal{M}}^*[\mathcal{F}]$, and $\tilde{\mathcal{M}}^*[\mathcal{F}]$ are closed under convex combination.*

While Corollary 5.1 follows immediately from Proposition 5.5 and the fact that there is a one-to-one correspondence between the functions in $\mathcal{M}^*[\mathcal{F}]$, $\hat{\mathcal{M}}^*[\mathcal{F}]$, and $\tilde{\mathcal{M}}^*[\mathcal{F}]$ and the vectors in $M^*[\mathcal{F}]$, $\hat{M}^*[\mathcal{F}]$, and $\tilde{M}^*[\mathcal{F}]$, the result can be shown directly, and a proof is included in Appendix F. These results provide some justification for our constructive approach. Upon deciding to study the sets of functions $\mathcal{M}^*[\mathcal{F}]$, $\hat{\mathcal{M}}^*[\mathcal{F}]$, and $\tilde{\mathcal{M}}^*[\mathcal{F}]$, we did not know (and still do not know) how many functions in $\mathcal{M}^*$ would be in $\mathcal{M}^*[\mathcal{F}]$. However, closure under convex combination indicates that the space is in some sense of a non-trivial size, since once a set of functions in one of these spaces has been identified, we can generate an infinite number of other functions that are also in the space.

As will be seen, we are interested in identifying the extremal points in the sets $M^*[\mathcal{F}]$, $\hat{M}^*[\mathcal{F}]$, and $\tilde{M}^*[\mathcal{F}]$, however, the fact that these sets are defined in terms of *infinite* sets of equations and inequalities precludes using linear programming theory and tools to explore their structure. For a given $q \in \mathbb{Z}_+$, we will therefore approximate the interval $[0, 1]$ by the set

$$Q_q = \Big\{ \frac{p}{q} \ : \ 0 \leq p \leq q \Big\},$$

and will approximate the set $[0, 1]^2$ by the the rational grid formed by $Q_q \times Q_q$. We will let $M^*\big[\mathcal{F}, Q_q\big]$, $\hat{M}^*\big[\mathcal{F}, Q_q\big]$ and $\tilde{M}^*\big[\mathcal{F}, Q_q\big]$ denote approximations of the sets $M^*[\mathcal{F}]$, $\hat{M}^*[\mathcal{F}]$, and $\tilde{M}^*[\mathcal{F}]$ in which (A3), (A5), and (A6) are written as *finite* sets of inequalities, one for each $x \in Q_q$, and (A4) is written as a *finite* set of inequalities, one for each combination of $(x, y) \in Q_q \times Q_q$ that satisfies its antecedent.

5.3 Power Series Representation

The definitions of the various sets of functions and vectors we discussed in §5.2 were dependent on the basis set $\mathcal{F}$ which was given to be an arbitrary finite set of real-valued

functions on $[0,1]^2$. In this section we consider the case in which the basis set is

$$\mathcal{P}_\eta^{(k)} = \left\{ \frac{\left(x-\frac{1}{2}\right)^i \left(y-\frac{1}{2}\right)^j}{\eta(x,y)} \ : \ 0 \le i,j \le k,\ i+j \le k \right\},$$

where following §4.7, $\eta : [0,1]^2 \to \mathbb{R}$ is a given (i.e. fixed) polynomial with $0 < \eta(x,y) < \infty$ for all $(x,y) \in [0,1]^2$, and the generated set of functions is

$$\mathcal{M}[\mathcal{P}_\eta^{(k)}] = \left\{ \mu_{\mathbf{c}} \ : \ \mu_{\mathbf{c}}(x,y) \ = \sum_{\substack{0 \le i,j \le k \\ i+j \le k}} \frac{(x-\frac{1}{2})^i \, (y-\frac{1}{2})^j}{\eta(x,y)} \ c_{ij} \text{ and } \mathbf{c} \in \mathbb{R}^{|\mathcal{P}_\eta^{(k)}|} \right\}.$$

The set $\mathcal{M}[\mathcal{P}_\eta^{(k)}] \subseteq \mathcal{M}$, which contains all linear combinations of the functions in the basis set $\mathcal{P}_\eta^{(k)}$, is the set of two-dimensional rational functions where the numerator is a two dimensional power series of degree k about the point $(1/2, 1/2)$, and the denominator is some fixed, finite, positive polynomial.

Justifying our choice of basis set is the fact that clearly, Proposition 5.5 holds for $M^*[\mathcal{P}_\eta^{(k)}]$, $\hat{M}^*[\mathcal{P}_\eta^{(k)}]$, and $\tilde{M}^*[\mathcal{P}_\eta^{(k)}]$, and Corollary 5.1 holds for $\mathcal{M}^*[\mathcal{P}_\eta^{(k)}]$, $\hat{\mathcal{M}}^*[\mathcal{P}_\eta^{(k)}]$, and $\tilde{\mathcal{M}}^*[\mathcal{P}_\eta^{(k)}]$. Further justification is provided by the fact that per Proposition 3.5, all but two of the feature ranking functions we have studied can be exactly written as a linear combination of the elements of the basis set. Finally, motivation for this choice is given by the fact that good approximations of many real-valued functions can be written as a linear combination of the elements of the basis set. For example, any function which is k differentiable at the point $(1/2, 1/2)$ can be approximated at this point by its degree-k Taylor polynomial. Therefore, using $\eta(x,y) = 1$, allows us to view the functions that we identify in $\mathcal{M}^*[\mathcal{P}_1^{(k)}]$, $\hat{\mathcal{M}}^*[\mathcal{P}_1^{(k)}]$, and $\tilde{\mathcal{M}}^*[\mathcal{P}_1^{(k)}]$ as Taylor polynomials, and to compare these functions with the actual Taylor polynomials of classic functions. This approach may assist in our understanding of the set of functions that satisfy the axioms listed in §5.1.

The following result shows a fundamental property of the set $M^*[\mathcal{P}_\eta^{(k)}]$.

Proposition 5.6. *The set $M^*[\mathcal{P}_\eta^{(k)}]$ is bounded and not empty.*

Proof. Assume to the contrary that $M^*[\mathcal{P}_\eta^{(k)}]$ is unbounded and therefore contains a line, i.e. assume there is a vector $\mathbf{d}$ and a direction $\mathbf{v} \neq \mathbf{0}$ such that $\mathbf{d} + \mathbf{v}t \in M^*[\mathcal{P}_\eta^{(k)}]$ for all $-\infty < t < \infty$. Then, because of the one-to-one relationship between the vectors in $M^*[\mathcal{P}_\eta^{(k)}]$ and the functions in $\mathcal{M}^*[\mathcal{P}_\eta^{(k)}]$, we equivalently have that

$$\mu_{\mathbf{d}+\mathbf{v}t}(x,y) \;=\; \sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^i\,(y-\frac{1}{2})^j}{\eta(x,y)}\;\left[d_{i,j}+v_{i,j}t\right] \in \mathcal{M}^*[\mathcal{P}_\eta^{(k)}].$$

From Proposition 5.1 we know that $\mu_{\mathbf{d}+\mathbf{v}\mathbf{t}}$ is bounded with $-1 \le \mu_{\mathbf{d}+\mathbf{v}t}(x,y) \le 1$, i.e.,

$$-\eta(x,y) \;\le \sum_{\substack{0\le i,j\le k\\ i+j\le k}} (x-\frac{1}{2})^i\,(y-\frac{1}{2})^j d_{i,j} + \sum_{\substack{0\le i,j\le k\\ i+j\le k}} (x-\frac{1}{2})^i\,(y-\frac{1}{2})^j v_{i,j}t \;\le\; \eta(x,y) \quad (5.9)$$

for each $(x,y) \in [0,1]^2$ and each $-\infty < t < \infty$. Therefore, we must have

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} (x-\frac{1}{2})^i\,(y-\frac{1}{2})^j v_{i,j} = 0 \quad (5.10)$$

for all $(x,y) \in [0,1]^2$, since otherwise the second sum in (5.9) would be unbounded.

Now, suppose that in (5.10) we fix x to be some arbitrary value in $[0,1]$. Then we have a polynomial in y, which from the Fundamental Theorem of Algebra, either is the zero polynomial, or has a finite number of zeroes, and since (5.10) must hold for all $y \in [0,1]$, it must be the zero polynomial. If we write (5.10) as

$$\begin{aligned}
&(y-\tfrac{1}{2})^k \left[\,(x-\tfrac{1}{2})^0\, v_{0,k}\,\right] +\\
&(y-\tfrac{1}{2})^{k-1} \left[\,(x-\tfrac{1}{2})^0\, v_{0,k} + (x-\tfrac{1}{2})^1\, v_{1,k-1}\,\right] +\\
&\qquad\qquad\vdots\\
&(y-\tfrac{1}{2})^0 \left[\,(x-\tfrac{1}{2})^0\, v_{0,k} + (x-\tfrac{1}{2})^1\, v_{1,k-1} + \cdots + (x-\tfrac{1}{2})^k\, v_{k,0}\,\right] \;=\; 0,
\end{aligned}$$

then since each factor $(y-\frac{1}{2})^k$ is itself a polynomial in y and as such has a finite number

of zeroes, we must have

$$\begin{aligned}
&\big(x-\frac{1}{2}\big)^0 v_{0,k} \;= 0\\
&\big(x-\frac{1}{2}\big)^0 v_{0,k} + \big(x-\frac{1}{2}\big)^1 v_{1,k-1} \;= 0\\
&\big(x-\frac{1}{2}\big)^0 v_{0,k} + \big(x-\frac{1}{2}\big)^1 v_{1,k-1} + \big(x-\frac{1}{2}\big)^2 v_{2,k-2} \;= 0\\
&\qquad\qquad\qquad\vdots\\
&\big(x-\frac{1}{2}\big)^0 v_{0,k} + \big(x-\frac{1}{2}\big)^1 v_{1,k-1} + \big(x-\frac{1}{2}\big)^2 v_{2,k-2} + \cdots + \big(x-\frac{1}{2}\big)^k v_{k,0} \;= 0.
\end{aligned} \tag{5.11}$$

Now, (5.11) can be written as the system $A\mathbf{v} = \mathbf{0}$ where the rows of the matrix A are the vectors

$$\Big[\big(x-\tfrac{1}{2}\big)^0, \big(x-\tfrac{1}{2}\big)^1, \big(x-\tfrac{1}{2}\big)^2, \cdots, \big(x-\tfrac{1}{2}\big)^j, 0, 0, \cdots, 0\Big], \tag{5.12}$$

for $j = 0, \ldots, k$. Since the set of vectors in (5.12) is linearly independent, the only solution to $A\mathbf{v} = \mathbf{0}$ is $\mathbf{v} = 0$, i.e. there does not exist a direction $\mathbf{v} \neq 0$ which satisfies (5.11) and this yields the desired contradiction showing that $M^*[\mathcal{P}_\eta^{(k)}]$ is bounded. Finally, to show that it is not empty, we refer the reader to Proposition 5.12.

■

Obviously, the conclusions of this result also apply to $\hat{M}^*[\mathcal{P}_\eta^{(k)}]$ and $\tilde{M}^*[\mathcal{P}_\eta^{(k)}]$. In addition, is important to note that this result relies on the fact that the sets $M^*[\mathcal{P}_\eta^{(k)}]$, $\hat{M}^*[\mathcal{P}_\eta^{(k)}]$, and $\tilde{M}^*[\mathcal{P}_\eta^{(k)}]$ are constructed using infinite sets of equations and inequalities that correspond to the axioms specified in §5.1 and therefore the conclusions may not hold for the approximations $M^*\big[\mathcal{P}_\eta^{(k)}, Q_q\big]$, $\hat{M}^*\big[\mathcal{P}_\eta^{(k)}, Q_q\big]$, and $\tilde{M}^*\big[\mathcal{P}_\eta^{(k)}, Q_q\big]$ with finite $q \in \mathbb{Z}_+$.

Given basis $\mathcal{P}_\eta^{(k)}$, the axioms from §5.1 become

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{1}{\eta(1,0)}\,\frac{-1^j}{2^{i+j}}\,c_{i,j} = 1 \tag{A1-$\mathcal{P}_\eta^{(k)}$}$$

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{1}{\eta(0,1)}\,\frac{-1^i}{2^{i+j}}\,c_{i,j} = -1 \tag{A2-$\mathcal{P}_\eta^{(k)}$}$$

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^i(x-\frac{1}{2})^j}{\eta(x,x)}\, c_{ij} = 0 \text{ for each } x\in[0,1] \tag{A3-$\mathcal{P}_\eta^{(k)}$}$$

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} \ge \sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2}-\Delta x)^i(y-\frac{1}{2}+\Delta y)^j}{\eta(x-\Delta x,y+\Delta y)}\, c_{ij}$$

$$\text{for each } (x,y)\in[0,1]^2 \text{ and } \Delta x,\Delta y\ge 0 \tag{A4-$\mathcal{P}_\eta^{(k)}$}$$

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} = -\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(y-\frac{1}{2})^i(x-\frac{1}{2})^j}{\eta(y,x)}\, c_{ij}$$

$$\text{for each } (x,y)\in[0,1]^2 \tag{A5-$\mathcal{P}_\eta^{(k)}$}$$

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} = -\sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(1-y-\frac{1}{2})^i(1-x-\frac{1}{2})^j}{\eta(1-y,1-x)}\, c_{ij}$$

$$\text{for each } (x,y)\in[0,1]^2 \tag{A6-$\mathcal{P}_\eta^{(k)}$}$$

Notice that, while (A3-$\mathcal{P}_\eta^{(k)}$) is written using *infinitely* many inequalities, the following result shows that only a *finite* number of them are actually required and therefore in the approximations based on Q_q that were discussed in §5.2, this axiom can be written in a manner that is independent of the choice of $q\in\mathbb{Z}_+$.

Proposition 5.7. *$\mu_{\mathbf{c}}(x,x)=0$ for each $x\in[0,1]$ if and only if*

$$\sum_{\substack{0\le i,j\\ i+j=l}} c_{ij} = 0.$$

Proof. First note that (A3-$\mathcal{P}_\eta^{(k)}$) can be written as

$$\begin{aligned}\mu_{\mathbf{c}}(x,x) &= \sum_{\substack{0\le i,j\le k\\ i+j\le k}} \frac{(x-\frac{1}{2})^{i+j}}{\eta(x,x)}\, c_{ij}\\ &= \sum_{l=0}^{k} \frac{(x-\frac{1}{2})^l}{\eta(x,x)} \sum_{\substack{0\le i,j\\ i+j=l}} c_{ij}\\ &= \frac{1}{\eta(x,x)} \sum_{l=0}^{k} \Big(x-\frac{1}{2}\Big)^l \sum_{\substack{0\le i,j\\ i+j=l}} c_{ij} = 0. \end{aligned} \tag{5.13}$$

Clearly, (5.13) shows that

$$\sum_{\substack{0\le i,j\\ i+j=l}} c_{ij} = 0 \Rightarrow \mu_{\mathbf{c}}(x,x) = 0 \text{ for each } x \in [0,1].$$

To show that

$$\mu_{\mathbf{c}}(x,x) = 0 \text{ for each } x \in [0,1] \Rightarrow \sum_{\substack{0\le i,j\\ i+j=l}} c_{ij} = 0$$

notice that since we assumed that $\eta(x,x) < \infty$, the RHS of (5.13) can only be 0 when at least one of the other two factors is 0. Since each

$$\sum_{\substack{0\le i,j\\ i+j=l}} \left(x - \frac{1}{2}\right)^l$$

for $l = 0, 1, \ldots, k$ is a polynomial in x of degree at most k, and can therefore have at most k roots, (5.13) can only be 0 at all $x \in [0,1]$ when

$$\sum_{\substack{0\le i,j\\ i+j=l}} c_{ij} = 0.$$

■

Proposition 5.7 shows that (A3-$\mathcal{P}_\eta^{(k)}$) can be written as

$$\begin{aligned}
&c_{00} = 0\\
&c_{10} + c_{01} = 0\\
&c_{20} + c_{11} + c_{02} = 0\\
&c_{30} + c_{21} + c_{12} + c_{03} = 0\\
&\cdots\\
&c_{k,0} + c_{k-1,1} + \cdots + c_{1,k-1} + c_{0,k} = 0
\end{aligned}$$

which is a system of $k+1$ equations which when combined with (A1-$\mathcal{P}_\eta^{(k)}$) and (A2-$\mathcal{P}_\eta^{(k)}$) yields a set of $k+3$ linear equations, whose solution set correspond to the $\mu_{\mathbf{c}} \in \mathcal{M}[\mathcal{P}_\eta^{(k)}]$

which satisfy these three axioms and do not involve an approximation based on Q_q and a choice of $q \in \mathbb{Z}_+$.

The following result shows that for an important choice of $\eta(x,y)$ that (A4-$\mathcal{P}_\eta^{(k)}$) can be simplified.

Proposition 5.8. *If $\eta(x,y) = 1$, then for each $(x,y) \in [0,1]^2$*

$$\mu_{\mathbf{c}}(x,y) \geq \mu_{\mathbf{c}}(x - \Delta x, y + \Delta y)$$

where $\Delta x, \Delta y \geq 0$, if and only if

$$\frac{\partial \mu_{\mathbf{c}}}{\partial x} \geq 0 \text{ and } \frac{\partial \mu_{\mathbf{c}}}{\partial y} \leq 0. \tag{5.14}$$

Proof. To show that (A4-$\mathcal{P}_\eta^{(k)}$) implies (5.14) we note (A4-$\mathcal{P}_\eta^{(k)}$) can be written as

$$\sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} \frac{(x - \frac{1}{2})^i (y - \frac{1}{2})^j}{\eta(x,y)} \; c_{ij} \geq \sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} \frac{(x - \frac{1}{2} - \alpha\epsilon)^i (y - \frac{1}{2} + \beta\epsilon)^j}{\eta(x - \alpha\epsilon, y + \beta\epsilon)} \; c_{ij}$$

for each $(x,y) \in [0,1]^2$, where (α, β) is some direction with $\alpha, \beta \; \geq 0$, and $\epsilon > 0$. Since we assumed $\eta(x,y) = 1$, this can be simplified to be

$$\sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} (x - \frac{1}{2})^i (y - \frac{1}{2})^j \; c_{ij} \geq \sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} (x - \frac{1}{2} - \alpha\epsilon)^i (y - \frac{1}{2} + \beta\epsilon)^j \; c_{ij}$$

and rearranging yields

$$\sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} (x - \frac{1}{2} - \alpha\epsilon)^i (y - \frac{1}{2} + \beta\epsilon)^j \; c_{ij} - \sum_{\substack{0 \leq i,j \leq k \\ i+j \leq k}} (x - \frac{1}{2})^i (y - \frac{1}{2})^j \; c_{ij} \; \leq \; 0.$$

By use of the Binomial Theorem we have

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}\left[\sum_{q=0}^{i}\binom{i}{q}\Big(x-\frac{1}{2}\Big)^{i-q}(-\alpha\epsilon)^q\right]\left[\sum_{r=0}^{j}\binom{j}{r}\Big(x-\frac{1}{2}\Big)^{j-r}(\beta\epsilon)^r\right] - \sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}(x-\frac{1}{2})^i(y-\frac{1}{2})^j \;\le\; 0,$$

and expanding yields

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}\left[\Big(x-\frac{1}{2}\Big)^{i}+i\Big(x-\frac{1}{2}\Big)^{i-1}(-\alpha\epsilon)+\binom{i}{2}\Big(x-\frac{1}{2}\Big)^{i-2}(-\alpha\epsilon)^2+\cdots\right] \times\left[\Big(y-\frac{1}{2}\Big)^{j}+j\Big(y-\frac{1}{2}\Big)^{j-1}(\beta\epsilon)+\binom{j}{2}\Big(y-\frac{1}{2}\Big)^{j-2}(\beta\epsilon)^2+\cdots\right] - \sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}(x-\frac{1}{2})^i(y-\frac{1}{2})^j \;\le\; 0. \tag{5.15}$$

By dividing (5.15) by ϵ and computing the limit as $\epsilon \to 0$ we have

$$\sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}\left[i\Big(x-\frac{1}{2}\Big)^{i-1}\Big(y-\frac{1}{2}\Big)^{j}(-\alpha)+j\Big(y-\frac{1}{2}\Big)^{j-1}\Big(x-\frac{1}{2}\Big)^{i}(\beta)\right]\le 0, \tag{5.16}$$

and if we also assume that (α,β) are such that $\sqrt{\alpha^2+\beta^2}=1$, then (5.16) is the directional derivative of $\mu_{\mathbf{c}}$ in direction (α,β). Considering $(\alpha=1,\beta=0)$ yields the partial derivative of $\mu_{\mathbf{c}}$ in the x direction

$$\frac{\partial\mu_{\mathbf{c}}}{\partial x}=\sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}\left[i\Big(x-\frac{1}{2}\Big)^{i-1}\Big(y-\frac{1}{2}\Big)^{j}\right]\ge 0, \tag{5.17}$$

and considering $(\alpha=0,\beta=1)$ yields the partial derivative of $\mu_{\mathbf{c}}$ in the y direction

$$\frac{\partial\mu_{\mathbf{c}}}{\partial y}=\sum_{\substack{0\le i,j\le k\\ i+j\le k}} c_{ij}\left[j\Big(y-\frac{1}{2}\Big)^{j-1}\Big(x-\frac{1}{2}\Big)^{i}\right]\le 0. \tag{5.18}$$

To show the converse, we assume that (5.17) and (5.18) hold and note that a linear

combination of these two inequalities weighted by α and β respectively, with $\alpha, \beta \geq 0$ and $\sqrt{\alpha^2 + \beta^2} = 1$, is equivalent to the directional derivative (5.16) which was shown to be equivalent to (A4-$\mathcal{P}_\eta^{(k)}$). ■

As a follow up to Proposition 5.8 it is interesting to note that when $\eta(x, y) = 1$, that (A4-$\mathcal{P}_\eta^{(k)}$) which involves changes in both x and y, can be written as two conditions in which the changes in x and y are separated. An important consequence of this fact is that in approximations based on Q_q, (A4-$\mathcal{P}_\eta^{(k)}$) can be written in terms of $2|Q_q|^2$ inequalities by writing (5.17) and (5.18) for each $(x, y) \in Q_q \times Q_q$. By contrast, directly writing (A4-$\mathcal{P}_\eta^{(k)}$) requires $O(|Q_q|^4)$ inequalities. We now recall that a subset of $\mathbb{R}^d$ for some $d > 0$ is said to be *polyhedral* if it is the intersection of a finite number of halfspaces. Such sets have a finite number of vertices and a finite number of facets. The following related result follows from Proposition 5.8.

Proposition 5.9. *If $k > 2$ then $M^*[\mathcal{P}_\eta^{(k)}]$ is not polyhedral.*

As a consequence of Proposition 5.9, it follows that neither $\hat{\mathcal{M}}^*[\mathcal{P}_\eta^{(k)}]$, nor $\tilde{\mathcal{M}}^*[\mathcal{P}_\eta^{(k)}]$ are polyhedral.

Similar to Proposition 5.7, the following result shows that while (A5-$\mathcal{P}_\eta^{(k)}$) is written using *infinitely* many inequalities, only a *finite* number of them are actually required.

Proposition 5.10. *$\mu_{\mathbf{c}}(x, y) = -\mu_{\mathbf{c}}(y, x)$ for each $(x, y) \in [0, 1]^2$ if and only if $\eta(x, y) = 1$ and $c_{ij} = -c_{ji}$ for each $0 \leq i, j \leq k, i + j \leq k$.*

Proof. If $\mu_{\mathbf{c}}(x, y) = -\mu_{\mathbf{c}}(y, x)$ then we have

$$\begin{aligned}
&\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} = -\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \frac{(y-\frac{1}{2})^i(x-\frac{1}{2})^j}{\eta(y,x)}\, c_{ij}\\
&\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} = -\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(y,x)}\, c_{ji}\\
&\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(x,y)}\, c_{ij} + \frac{(x-\frac{1}{2})^i(y-\frac{1}{2})^j}{\eta(y,x)}\, c_{ji} = 0\\
&\sum_{\substack{0\leq i,j\leq k\\ i+j\leq k}} \Big(x-\frac{1}{2}\Big)^i\Big(y-\frac{1}{2}\Big)^j \left[\frac{c_{ij}}{\eta(x,y)} + \frac{c_{ji}}{\eta(y,x)}\right] = 0 \qquad (5.19)
\end{aligned}$$

for each $(x, y) \in [0, 1]^2$. In the proof of Proposition 5.6 we showed that (5.10) cannot hold for all $(x, y) \in [0, 1]^2$. Therefore, (5.19) can only hold when $c_{ij}\eta(y, x) = -c_{ji}\eta(x, y)$ for each $0 \leq i, j \leq k, i + j \leq k$ and for all $(x, y) \in [0, 1]^2$. Since we assumed that $\eta(x, y) = 1$ this implies that $c_{ij} = -c_{ij}$ for each $0 \leq i, j \leq k, i + j \leq k$. The other direction follows immediately after noticing that when $\eta(x, y) = 1$ and $c_{ij} = -c_{ji}$ for each $0 \leq i, j \leq k, i + j \leq k$, that for each $0 \leq i, j \leq k, i + j \leq k$

$$\big(x - \frac{1}{2}\big)^i \big(y - \frac{1}{2}\big)^j \; c_{ij} = -\big(x - \frac{1}{2}\big)^i \big(y - \frac{1}{2}\big)^j \; c_{ji}.$$

■

Corollary 5.2. *$c_{ii} = 0$ for each $2i \leq k$.*

Proposition 5.11. *(A5-$\mathcal{P}_\eta^{(k)}$) can be written as a system of $O(k^2)$ equations.*

Proof. Proposition 5.10 and Corollary 5.2 show that (A5-$\mathcal{P}_\eta^{(k)}$) can be written as

$$\begin{aligned}
&2c_{00} = 0 \\
&c_{10} + c_{01} = 0 \\
&c_{20} + c_{02} = 0 \\
&2c_{11} = 0 \\
&c_{30} + c_{03} = 0 \\
&c_{21} + c_{12} = 0 \\
&\dots
\end{aligned}$$

which is a system of

$$f(k) = \sum_{i=0}^{k} \left\lceil \frac{i+1}{2} \right\rceil \tag{5.20}$$

equations. Note that (5.20) is the sum of the partition numbers of size 2 for the integers $i = 0, 1, \dots, k$. For example, it is (1+1) + (2+2) + 3, ..., for even k, and it is (1+1) + (2+2) + (3+3), ..., for odd k. To find a closed form solution to (5.20) we note that for both even and odd k, the recurrence relation, $f(k+2) = f(k) + k + 3$, holds. When

k is even, its solution can be seen to be

$$f(k) = \Big(\frac{k}{2} + 1\Big)^2.$$

Next we notice that when k is odd that $f(k) = f(k-1) + \sqrt{f(k-1)}$ and therefore

$$f(k) = \Big(\frac{k-1}{2} + 1\Big)^2 + \Big(\frac{k-1}{2} + 1\Big).$$

when k is odd. ■

The system of equations corresponding to (A1-$\mathcal{P}_\eta^{(k)}$), (A2-$\mathcal{P}_\eta^{(k)}$), (A3-$\mathcal{P}_\eta^{(k)}$), and (A5-$\mathcal{P}_\eta^{(k)}$) contains

$$\Big(\frac{k}{2} + 1\Big)^2 + k + 3 \;=\; \frac{1}{4}\big(k^2 + 8k + 16\big)$$

equations when k is even, and

$$\Big(\frac{k-1}{2} + 1\Big)^2 + \Big(\frac{k-1}{2} + 1\Big) + k + 3 \;=\; \frac{1}{4}\big(k^2 + 8k + 15\big)$$

equations when k is odd.

5.4 Equality Based Characterizations of $M^*\big[\mathcal{P}_{\eta=1}^{(k)}\big]$

If we let $k = 1$, $\eta(x, y) = 1$, and only assume (A1-$\mathcal{P}_\eta^{(k)}$), (A2-$\mathcal{P}_\eta^{(k)}$), and (A3-$\mathcal{P}_\eta^{(k)}$), the equations corresponding to these axioms yield the following 4×3 system

$$\begin{array}{rrrcr} c_{0,0} & c_{1,0} & c_{0,1} & & \\ 1 & \frac{1}{2} & -\frac{1}{2} & = & 1 \\ 1 & -\frac{1}{2} & \frac{1}{2} & = & -1 \\ 1 & 0 & 0 & = & 0 \\ 0 & 1 & 1 & = & 0 \end{array}$$

which has a rank of 3. The unique solution of this system is $\mathbf{c} = [0, 1, -1]$ which corresponds to $\mu_{\mathbf{c}}(x, y) = x - y$ and gives us the following result.

Proposition 5.12. *$M^*\big[\mathcal{P}^{(1)}_{\eta=1}\big] = \{[0,1,-1]\}$ and $\mathcal{M}^*\big[\mathcal{P}^{(1)}_{\eta=1}\big] = \{x-y\}$.*

If we let $k=2$, $\eta(x,y)=1$, and only assume (A1-$\mathcal{P}^{(k)}_\eta$), (A2-$\mathcal{P}^{(k)}_\eta$), and (A3-$\mathcal{P}^{(k)}_\eta$), the equations corresponding to these axioms yields the following 5×6 system

$$\begin{array}{cccccccc} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & & \\ 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & = & 1 \\ 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & = & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & = & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & = & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & = & 0 \end{array}$$

which has a rank of 5 and the solution set

$$\begin{aligned} c_{0,0} &= 0 \\ c_{1,0} &= 1 \\ c_{0,1} &= -1 \\ c_{2,0} &= c_{2,0} \\ c_{1,1} &= 0 \\ c_{0,2} &= -c_{2,0} \end{aligned}$$

which corresponds to

$$\mu_{\mathbf{c}}(x,y) = \Big(x-\frac{1}{2}\Big) - \Big(y-\frac{1}{2}\Big) + \lambda\Big(x-\frac{1}{2}\Big)^2 - \lambda\Big(y-\frac{1}{2}\Big)^2$$

with the free parameter $\lambda \in \mathbb{R}$. Note that additionally assuming (A5-$\mathcal{P}^{(k)}_\eta$) does not change the solution.

It should be mentioned that for arbitrary values of λ that (A4-$\mathcal{P}^{(k)}_\eta$) and the conclusions of Proposition 5.1 may not be satisfied, however, by Proposition 5.8 we have that

$$\frac{\partial \mu_{\mathbf{c}}(x,y)}{\partial x} = 1 + 2\lambda\big(x-\frac{1}{2}\big) \geq 0$$

from which we can conclude that when $-1 \leq \lambda \leq 1$, the axiom and therefore the proposition are satisfied. This allows to write the following result.

Proposition 5.13. *$M^*\big[\mathcal{P}^{(2)}_{\eta=1}\big] = \big\{\,[0, 1-\lambda, \lambda-1, \lambda, 0, -\lambda] \;:-1 \leq \lambda \leq 1\,\big\}$ and $\mathcal{M}^*\big[\mathcal{P}^{(2)}_{\eta=1}\big] = \big\{\,(1-\lambda)(x-y) + \lambda(x^2-y^2) \;:-1 \leq \lambda \leq 1\,\big\}$.*

If we let $k = 3$, $\eta(x, y) = 1$, and only assume (A1-$\mathcal{P}^{(k)}_\eta$), (A2-$\mathcal{P}^{(k)}_\eta$), and (A3-$\mathcal{P}^{(k)}_\eta$), the equations corresponding to these axioms yields the following 6×10 system

$c_{0,0}$	$c_{1,0}$	$c_{0,1}$	$c_{2,0}$	$c_{1,1}$	$c_{0,2}$	$c_{3,0}$	$c_{2,1}$	$c_{1,2}$	$c_{0,3}$		
1	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{1}{8}$	$-\frac{1}{8}$	=	1
1	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{8}$	$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{1}{8}$	=	-1
1	0	0	0	0	0	0	0	0	0	=	0
0	1	1	0	0	0	0	0	0	0	=	0
0	0	0	1	1	1	0	0	0	0	=	0
0	0	0	0	0	0	1	1	1	1	=	0

which has a rank of 6 and therefore 4 free parameters. Rather than pursuing this case we also assume (A5-$\mathcal{P}^{(k)}_\eta$). Adding the corresponding equations yields

$c_{0,0}$	$c_{1,0}$	$c_{0,1}$	$c_{2,0}$	$c_{1,1}$	$c_{0,2}$	$c_{3,0}$	$c_{2,1}$	$c_{1,2}$	$c_{0,3}$		
1	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{1}{8}$	$-\frac{1}{8}$	=	1
1	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{8}$	$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{1}{8}$	=	-1
1	0	0	0	0	0	0	0	0	0	=	0
0	1	1	0	0	0	0	0	0	0	=	0
0	0	0	1	1	1	0	0	0	0	=	0
0	0	0	0	0	0	1	1	1	1	=	0
2	0	0	0	0	0	0	0	0	0	=	0
0	1	1	0	0	0	0	0	0	0	=	0
0	0	0	1	0	1	0	0	0	0	=	0
0	0	0	0	2	0	0	0	0	0	=	0
0	0	0	0	0	0	1	0	0	1	=	0
0	0	0	0	0	0	0	1	1	0	=	0

which is a 12×10 system with a rank of 7 and therefore has 3 free parameters. It has the solution set

$$\begin{aligned}
c_{0,0} &= 0 \\
c_{1,0} &= c_{1,0} \\
c_{0,1} &= -c_{1,0} \\
c_{2,0} &= c_{2,0} \\
c_{1,1} &= 0 \\
c_{0,2} &= -c_{2,0} \\
c_{3,0} &= c_{3,0} \\
c_{2,1} &= 4 + 4c_{1,0} + c_{3,0} \\
c_{1,2} &= -c_{2,1} \\
c_{0,3} &= -c_{3,0}
\end{aligned}$$

which corresponds to

$$\begin{aligned}
\mu(x,y) = {} & \lambda_1\Big((x-\frac{1}{2}) - (y-\frac{1}{2})\Big) + \lambda_2\Big((x-\frac{1}{2})^2 - (y-\frac{1}{2})^2\Big) + \lambda_3\Big((x-\frac{1}{2})^3 - (y-\frac{1}{2})^3\Big) \\
& + (4 + 4\lambda_1 + \lambda_3)\Big((x-\frac{1}{2})^2(y-\frac{1}{2}) - (x-\frac{1}{2})(y-\frac{1}{2})^2\Big)
\end{aligned}$$

with the free parameters $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. As for $k = 2$, for arbitrary values of $\lambda_1, \lambda_2, \lambda_3$ the conclusions of Proposition 5.1 may not be satisfied.

5.5 Experiments

To explore the sets $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ we will use the *lrs* software tools [4]. Given a set of equations and inequalities representing a polyhedron, *lrs* uses a revised version of the reverse search vertex enumeration algorithm to find its vertices. It uses exact rational arithmetic for all computations. We will let $\mathcal{V}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ denote the matrix of vertices identified by *lrs* that correspond to the sets $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$. We

will let $\mathcal{R}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, $\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ denote the *reduced row echelon* form of these matrices.

We now offer a few comments about the tables of results presented in this section. Using the notation presented earlier, the *Class* column indicates the approximation under consideration. As we mentioned, Proposition 5.6 only addresses the question of boundedness for $M^*\big[\mathcal{P}_\eta^{(k)}\big]$, $\hat{M}^*\big[\mathcal{P}_\eta^{(k)}\big]$, and $\tilde{M}^*\big[\mathcal{P}_\eta^{(k)}\big]$ and this question remains open for the approximations $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$. For each of these approximations we will only list results for the smallest $q \in \mathbb{Z}_+$ that is bounded, so it can be assumed that any approximation that utilizes a smaller q than is listed is unbounded. In §5.4 we showed that when $k = 1$, $\mu_{\mathbf{c}}(x) = x - y$ was the unique member of each of the three approximation sets and therefore we will not discuss these sets further in this section. The value of the *Variables* is $(k+1)(k+2)/2$. The *All Equations & Inequalities* column contains the total number of equations and inequalities that were presented to the *redund* program which is part of *lrs*. This program identifies and removes redundant equations and inequalities thereby reducing the size of the problem to that listed in the *Non-redundant Equations* and the *Non-redundant Inequalities* columns. We also include the ranks of vertex matrices associated with the corresponding approximation sets and have made the following related observation.

Observation 5.1. *For a given approximation set $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ and a given k the corresponding reduced row echelon matrix $\mathcal{R}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ is identical for all $q \in \mathbb{Z}_+$ that were considered, and therefore the rank of the vertex matrix $\mathcal{V}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ is identical for all $q \in \mathbb{Z}_+$ that were considered. Further, this rank is less than the number of variables.*

Corresponding observations were made for the approximation sets $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ and $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$. These observations will allow us to map the original problems to a lower dimensional subspace thereby reducing size of the problems that are presented to *lrs*. We will see that there are two reasons for the reduction in the number of dimensions. First, some of the basis variables in the original space are identically zero. Second, the basis variables in the reduced subspace are not simply a subset of the non-zero basis variables in the original space, but rather are linear combinations of those in the

original space; a fact that will aid us in our study of the structure of $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$, and $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$.

When possible, we will try to compare the approximations we identify to selected feature ranking functions that we studied in previous chapters. To accomplish this, for several such functions μ, we will compute the degree k Taylor polynomial in x and y about the point $(1/2, 1/2)$ as given by

$$\begin{aligned}
\tau_k(\mu(x,y)) \;=\; & \mu\big(\tfrac{1}{2},\tfrac{1}{2}\big) \\
& + \;\big(x-\tfrac{1}{2}\big)\mu^x\big(\tfrac{1}{2},\tfrac{1}{2}\big) + \big(y-\tfrac{1}{2}\big)\mu^y\big(\tfrac{1}{2},\tfrac{1}{2}\big) \\
& + \;\frac{1}{2!}\left[\big(x-\tfrac{1}{2}\big)^2\mu^{xx}\big(\tfrac{1}{2},\tfrac{1}{2}\big) + 2\big(x-\tfrac{1}{2}\big)\big(y-\tfrac{1}{2}\big)\mu^{xy}\big(\tfrac{1}{2},\tfrac{1}{2}\big) + \big(y-\tfrac{1}{2}\big)^2\mu^{yy}\big(\tfrac{1}{2},\tfrac{1}{2}\big)\right] \\
& + \;\frac{1}{3!}\left[\big(x-\tfrac{1}{2}\big)^3\mu^{xxx}\big(\tfrac{1}{2},\tfrac{1}{2}\big) + 3\big(x-\tfrac{1}{2}\big)^2\big(y-\tfrac{1}{2}\big)\mu^{xxy}\big(\tfrac{1}{2},\tfrac{1}{2}\big)\right. \\
& \left. + \;3\big(x-\tfrac{1}{2}\big)\big(y-\tfrac{1}{2}\big)^2\mu^{yyx}\big(\tfrac{1}{2},\tfrac{1}{2}\big) + \big(y-\tfrac{1}{2}\big)^3\mu^{yyy}\big(\tfrac{1}{2},\tfrac{1}{2}\big)\right]\cdots
\end{aligned}$$

where

$$\mu^x = \frac{\partial\mu}{\partial x},\ \mu^y = \frac{\partial\mu}{\partial y},\ \mu^{xx} = \frac{\partial\mu}{\partial x\partial x},\ \mu^{xy} = \frac{\partial\mu}{\partial x\partial y},\ \mu^{yy} = \frac{\partial\mu}{\partial y\partial y},\ \cdots.$$

We will let $v_k(\mu)$ denote the vector in $M\big[\mathcal{P}_1^{(k)}, Q_q\big]$ that corresponds to $\tau_k(\mu(x,y))$.

The feature ranking functions we shall use for comparison are

$$\begin{aligned}
\mu_{9,\varrho=1/2}(x,y) &= \frac{x-y}{\sqrt{(x+y)(2-x-y)}} \\
\mu_{9,\lim\varrho\to 0}(x,y) &= \frac{x-y}{\sqrt{y(1-y)}} \\
\mu_{22}(x,y) &= \frac{x-y}{x+y} \\
\mu_{24}(x,y) &= \frac{x-y}{\sqrt{x(1-x)}+\sqrt{y(1-y)}}
\end{aligned}$$

As justification for using the Taylor polynomials for comparison we note that they provide a reasonable approximation of these four feature ranking functions. To see this we compute

$$
\begin{aligned}
|| \ \mu_{9,\varrho=1/2}(x,y), \tau_5(\mu_{9,\varrho=1/2}(x,y)) \ ||_2 &= 0.005 \\
|| \ \mu_{9,\lim \varrho\to 0}(x,y), \tau_5(\mu_{9,\lim \varrho\to 0}(x,y)) \ ||_2 &= 0.002 \\
|| \ \mu_{22}(x,y), \tau_5(\mu_{22}(x,y)) \ ||_2 &= 0.015 \\
|| \ \mu_{24}(x,y), \tau_5(\mu_{24}(x,y)) \ ||_2 &= 0.035
\end{aligned}
$$

where

$$
|| \ \mu(x,y), \tau_k(\mu(x,y)) \ ||_2 = \int_{0+\epsilon}^{1-\epsilon} \int_{0+\epsilon}^{1-\epsilon} (\mu(x,y) - \tau_k(\mu(x,y)))^2 dxdy
$$

and $\epsilon = 0.05$. As expected the quality of the approximation decreases as the distance from the point $(1/2, 1/2)$ increases.

The corresponding degree 3 Taylor polynomials after simplification are

$$
\begin{aligned}
\tau_3(\mu_{9,\varrho=1/2}(x,y)) &= \frac{3}{2}(x-y) - (x^2-y^2) + \frac{1}{2}(x^3-y^3) + \frac{1}{2}(x^2y - xy^2) \\
(1/3)\tau_3(\mu_{9,\lim \varrho\to 0}(x,y)) &= (x-y) - \frac{4}{3}xy + \frac{4}{3}y^2 + \frac{4}{3}xy^2 - \frac{4}{3}y^3 \\
\tau_3(\mu_{22}(x,y)) &= 3(x-y) - 3(x^2-y^2) + (x^3-y^3) + (x^2y - xy^2) \\
(2/3)\tau_3(\mu_{24}(x,y)) &= (x-y) - \frac{2}{3}(x^2-y^2) + \frac{2}{3}(x^3-y^3) - \frac{2}{3}(x^2y - xy^2)
\end{aligned}
$$

and Table 5.2 contains the coefficients for the corresponding Taylor polynomials.

Noting that when necessary we multiply by an appropriate constant to scale the Taylor polynomials to $[0, 1]$, it can be seen that all of these functions satisfy (A1-$\mathcal{P}_\eta^{(k)}$), (A2-$\mathcal{P}_\eta^{(k)}$), (A3-$\mathcal{P}_\eta^{(k)}$), and (A4-$\mathcal{P}_\eta^{(k)}$). Table 5.5 shows which functions satisfy the

$\tau_k(\mu)$	(A5-$\mathcal{P}_\eta^{(k)}$)	(A6-$\mathcal{P}_\eta^{(k)}$)
$\tau_3(\mu_{9,\varrho=1/2})$	✓	✓
$(1/3)\tau_3(\mu_{9,\lim \varrho\to 0})$		
$\tau_3(\mu_{22})$	✓	
$(2/3)\tau_3(\mu_{24})$	✓	✓

Table 5.1: Axioms Satisfied by the Sample Taylor Polynomials

remaining two axioms.

$v_k(\mu)$	$c_{0,0}$	$c_{1,0}$	$c_{0,1}$	$c_{2,0}$	$c_{1,1}$	$c_{0,2}$	$c_{3,0}$	$c_{2,1}$	$c_{1,2}$	$c_{0,3}$	$c_{4,0}$	$c_{3,1}$	$c_{2,2}$	$c_{1,3}$	$c_{0,4}$	$c_{5,0}$	$c_{4,1}$	$c_{3,2}$	$c_{2,3}$	$c_{1,4}$	$c_{0,5}$
$v_5(\mu_{9,\varrho=1/2})$	0	1	-1	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	0	0	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{6}{8}$	$-\frac{6}{8}$	$-\frac{9}{8}$	$-\frac{3}{8}$
$(1/3)v_3(\mu_{9,\lim\varrho\to0})$	0	$\frac{2}{3}$	$-\frac{2}{3}$	0	0	0	0	0	$\frac{4}{3}$	$-\frac{4}{3}$	–	–	–	–	–	–	–	–	–	–	–
$(4/15)v_5(\mu_{9,\lim\varrho\to0})$	0	$\frac{8}{15}$	$-\frac{8}{15}$	0	0	0	0	0	$\frac{16}{15}$	$-\frac{16}{15}$	0	0	0	0	0	0	0	0	0	$\frac{16}{5}$	$-\frac{16}{5}$
$v_5(\mu_{22})$	0	1	-1	-1	0	1	1	1	-1	-1	-1	-2	0	2	1	1	3	2	-2	-3	-1
$(2/3)v_3(\mu_{24})$	0	$\frac{2}{3}$	$-\frac{2}{3}$	0	0	0	$\frac{2}{3}$	$-\frac{2}{3}$	$\frac{2}{3}$	$-\frac{2}{3}$	–	–	–	–	–	–	–	–	–	–	–
$(8/15)v_5(\mu_{24})$	0	$\frac{8}{15}$	$-\frac{8}{15}$	0	0	0	$\frac{8}{15}$	$-\frac{8}{15}$	$\frac{8}{15}$	$-\frac{8}{15}$	0	0	0	0	0	$\frac{16}{15}$	$-\frac{16}{15}$	$\frac{16}{15}$	$-\frac{16}{15}$	$\frac{16}{15}$	$-\frac{16}{15}$

Table 5.2: Taylor Polynomial Coefficients

5.5.1 $M^*\left[\mathcal{P}_1^{(k)}, Q_q\right]$

A summary of the results for $M^*\left[\mathcal{P}_1^{(k)}, Q_q\right]$ are shown in Table 5.3. In experiments related to $M^*\left[\mathcal{P}_1^{(2)}, Q_q\right]$ the following two matrices were identified

$$\mathcal{V}^*\left[\mathcal{P}_1^{(2)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} \\ 0 & 1 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & -1 & 0 & 1 \end{bmatrix}$$

and

$$\mathcal{R}^*\left[\mathcal{P}_1^{(2)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

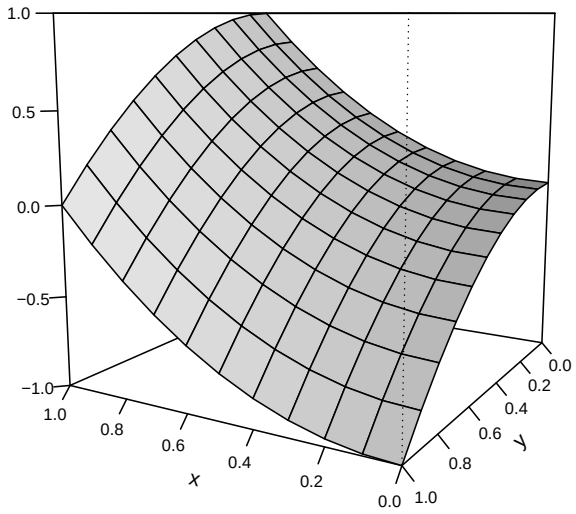
for $q \in \{1, 10, 20, 30, 40, 50\}$. These experimental results match the analytical results in §5.4. Specifically, in Proposition 5.13 it was shown that $\mathcal{M}^*\left[\mathcal{P}_{\eta=1}^{(2)}\right]$ is a particular set of linear combinations of the functions corresponding to the vectors in $\mathcal{R}^*\left[\mathcal{P}_1^{(2)}, Q_*\right]$, and it can be shown that the set of convex combinations of the functions corresponding to the vectors in $\mathcal{V}^*\left[\mathcal{P}_1^{(2)}, Q_*\right]$ is equivalent to the set identified in Proposition 5.13. It is interesting to note that the function corresponding to the "center" of the set defined by the convex combination of the two vertices in $\mathcal{V}^*\left[\mathcal{P}_1^{(2)}, Q_*\right]$ is $x - y$. Plots of the feature ranking functions corresponding to these vertices are shown in Figure 5.1.

As shown in Table 5.3 in experiments related to $M^*\left[\mathcal{P}_1^{(3)}, Q_q\right]$, the number of vertices in $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}, Q_q\right]$, as was expected, increased with q and therefore did not suggest a vertex set that defines $M^*\left[\mathcal{P}_1^{(3)}\right]$. However, in these experiments the following matrix was identified

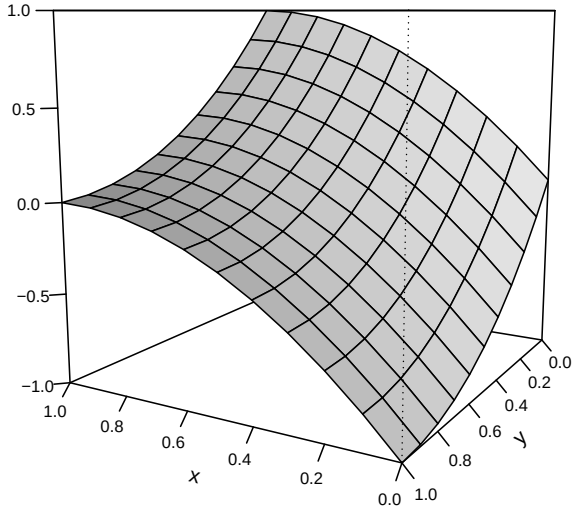
$$\mathcal{R}^*\left[\mathcal{P}_1^{(3)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Class	Variables	Non-redundant Equations	Non-redundant Inequalities	All Equations & Inequalities	Vertices	Rank of Vertex Matrix
$M^*\left[\mathcal{P}_1^{(1)}, Q_1\right]$	3	3	0	12	1	1
$M^*\left[\mathcal{P}_1^{(2)}, Q_1\right]$	6	5	2	13	2	2
$M^*\left[\mathcal{P}_1^{(3)}, Q_2\right]$	10	6	13	24	10	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_3\right]$	10	6	22	38	15	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_4\right]$	10	6	31	56	35	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_5\right]$	10	6	54	78	62	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_6\right]$	10	6	63	104	92	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_7\right]$	10	6	88	134	119	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_8\right]$	10	6	103	168	173	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_9\right]$	10	6	132	206	208	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_{10}\right]$	10	6	149	248	284	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_{20}\right]$	10	6	527	888	1315	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_{30}\right]$	10	6	1127	1928	3352	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_{40}\right]$	10	6	1951	3368	6513	5
$M^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$	10	6	3009	5208	11042	5
$M^*\left[\mathcal{P}_1^{(4)}, Q_3\right]$	15	7	28	39	1338	9
$M^*\left[\mathcal{P}_1^{(4)}, Q_4\right]$	15	7	45	57	6820	9
$M^*\left[\mathcal{P}_1^{(4)}, Q_5\right]$	15	7	66	79	21716	9
$M^*\left[\mathcal{P}_1^{(4)}, Q_6\right]$	15	7	91	105	48896	9
$M^*\left[\mathcal{P}_1^{(4)}, Q_{10}\right]$	15	7	231	249	460924	9
$M^*\left[\mathcal{P}_1^{(4)}, Q_{15}\right]$	15	7	496	519	2486284	9
$M^*\left[\mathcal{P}_1^{(5)}, Q_4\right]$	21	8	45	58	525213	14

Table 5.3: $M^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$



(a) $x - y + (x - 1/2)^2 - (y - 1/2)^2$



(b) $x - y - (x - 1/2)^2 + (y - 1/2)^2$

Figure 5.1: $\mathcal{V}^*\left[\mathcal{P}_1^{(2)}, Q_*\right]$

and based on this matrix we state the following conjecture, in which the supposed functions in $\mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$ are described as linear combinations of the functions corresponding to the vectors in $\mathcal{R}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$.

Conjecture 5.1. *There exist $\lambda_i \in \mathbb{R}$ for $i = 1, \cdots, 5$ such that if*

$$\begin{aligned}\mu_{\mathbf{c}}(x,y) \;=\; & \lambda_1\big((x-1/2)-(y-1/2)\big) \;+ \\ & \lambda_2\big((x-1/2)^2-(y-1/2)^2\big) \;+ \\ & \lambda_3\big((x-1/2)^3-(y-1/2)^3\big) \;+ \\ & \lambda_4\big((x-1/2)^2(y-1/2)-(y-1/2)^3\big) \;+ \\ & \lambda_5\big((x-1/2)(y-1/2)^2-(y-1/2)^3\big),\end{aligned}$$

then $\mu_{\mathbf{c}} \in \mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$.

We now assume this representation of $\mu_{\mathbf{c}} \in \mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$ and rewrite the axioms in this 5-dimensional subspace rather than in the original 10-dimensional space. For a given $q \in \mathbb{Z}_+$ and for each $(x,y) \in Q_q$, we have

$$\lambda_1 + \frac{1}{4}\lambda_3 + \frac{1}{4}\lambda_5 = 1 \tag{5.21}$$

$$\lambda_1 + 2(x-\frac{1}{2})\lambda_2 + 3(x-\frac{1}{2})^2\lambda_3 + 2(x-\frac{1}{2})(y-\frac{1}{2})\lambda_4 + (y-\frac{1}{2})^2\lambda_5 \geq 0 \tag{5.22}$$

$$\begin{aligned}-\lambda_1 - 2(y-\frac{1}{2})\lambda_2 - 3(y-\frac{1}{2})^2\lambda_3 + ((x-\frac{1}{2})^2 - 3(y-\frac{1}{2})^2)\lambda_4 \\ +\big(2(x-\frac{1}{2})(y-\frac{1}{2}) - 3(y-\frac{1}{2})^2\big)\lambda_5 \leq 0\end{aligned} \tag{5.23}$$

with (5.21) corresponding to (A1-$\mathcal{P}_\eta^{(k)}$), and (5.22) and (5.23) corresponding to (A4-$\mathcal{P}_\eta^{(k)}$) implemented as

$$\frac{\partial \mu_{\mathbf{c}}}{\partial x} \geq 0 \text{ and } \frac{\partial \mu_{\mathbf{c}}}{\partial y} \leq 0$$

respectively. Continuing to write $\mu_{\mathbf{c}}$ as in Conjecture 5.4, we note that $\mu_{\mathbf{c}}(0,1) = -\lambda_1 - \frac{1}{4}\lambda_3 - \frac{1}{4}\lambda_5 = -1$ which is redundant given (5.21), and therefore no equation corresponding to (A2-$\mathcal{P}_\eta^{(k)}$) is required. In addition, $\mu_{\mathbf{c}}(x,x) = 0$ for every $x \in Q_q$, and therefore no equations corresponding to (A3-$\mathcal{P}_\eta^{(k)}$) are required.

The results for computations with *lrs* in the 5-dimensional subspace were similar to those for $M^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ in the 10-dimensional subspace that are shown in Table 5.3. The differences are that for the 5-dimensional subspace, *Variables* is obviously 5 rather than 10, *Non-redundant Inequalities* is 1 rather than 6, and *All Equations & Inequalities* is 5 less than that for the 10-dimensional subspace problem. We also note that the non-zero rows in the matrix $\mathcal{R}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ in the 5-dimensional subspace correspond to the 5×5 identity matrix. Since the number of vertices in $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}, Q_q\right]$ again increased with q, as was expected, these results did not directly suggest a vertex set that defines $M^*\left[\mathcal{P}_1^{(3)}\right]$.

We conclude our discussion of $M^*\left[\mathcal{P}_1^{(3)}\right]$ with a few observations about the set $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$. We first see that (5.21) requires that all vertices lie on the plane

$$\lambda_1 + \frac{1}{4}\lambda_3 + \frac{1}{4}\lambda_5 = 1.$$

Second, we note that the "center" $M^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$ approximately corresponds to the vector

$$\lambda_1 = \frac{1}{4},\ \lambda_2 = 0,\ \lambda_3 = \frac{3}{2},\ \lambda_4 = -\frac{3}{2},\ \lambda_5 = \frac{3}{2}$$

in the 5-dimensional subspace, and the associated function is shown in Figure 5.2. Next, in an effort to explore the geometry of this non-polyhedral set we used the GGobi data visualization system (see e.g. [74]), which provides a variety of tools for visualizing high dimensional data. An interesting subset of $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}\right]$ that was identified using this tool is the set of vertices which satisfy

$$\lambda_1 = \lambda_2 = 0$$
$$\lambda_3 + \lambda_5 = 4$$
$$\lambda_4 < 0.$$

This set is shown in Figure 5.3 and contains nearly half of the 11042 vertices in $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$.

As shown in Table 5.3 in experiments related to $M^*\left[\mathcal{P}_1^{(4)}, Q_q\right]$ and $M^*\left[\mathcal{P}_1^{(5)}, Q_q\right]$, the number of vertices in $\mathcal{V}^*\left[\mathcal{P}_1^{(4)}, Q_q\right]$ and $\mathcal{V}^*\left[\mathcal{P}_1^{(5)}, Q_q\right]$ respectively, as was expected,

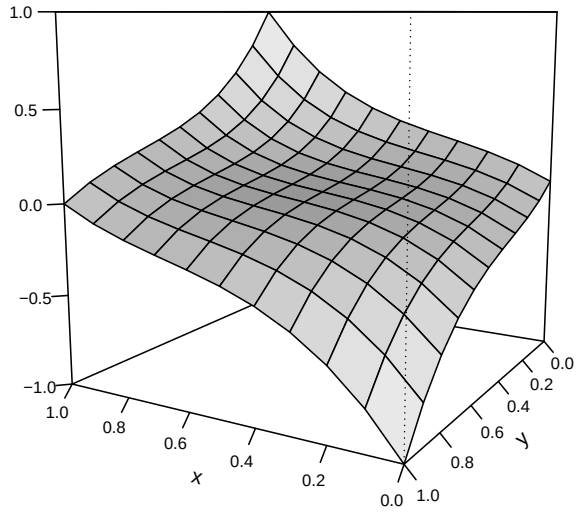
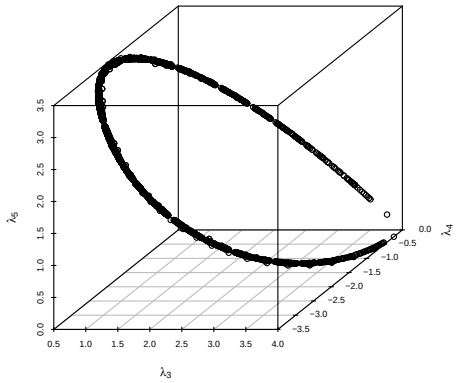
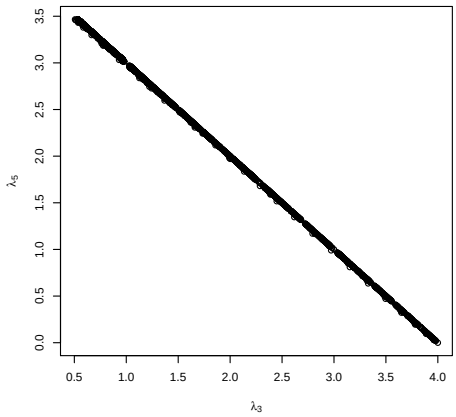


Figure 5.2: Center of $M^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$

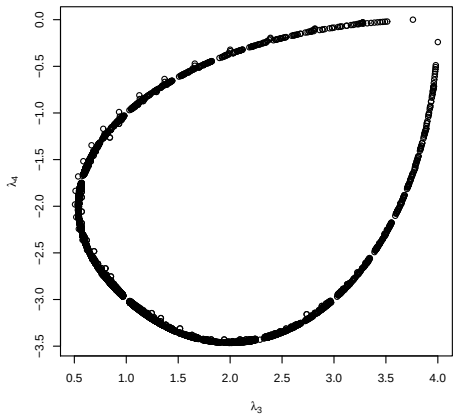
increased with q and therefore did not suggest a vertex set that defines $M^*\left[\mathcal{P}_1^{(4)}\right]$ or $M^*\left[\mathcal{P}_1^{(5)}\right]$. However, in these experiments the matrices $\mathcal{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$ and $\mathcal{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$ shown in Table 5.4 and Table 5.5 were identified. These results could be used to perform analyses similar to the one done for $M^*\left[\mathcal{P}_1^{(3)}\right]$, but we will not pursue this.



(a)



(b)



(c)

Figure 5.3: $\mathcal{V}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ for $\lambda_1 \approx \lambda_2 \approx 0$

$$\mathcal{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Table 5.4: $\mathcal{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$

$$\mathcal{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right] = \left[\begin{array}{cccccccccccccccccccccc}
c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} & c_{5,0} & c_{4,1} & c_{3,2} & c_{2,3} & c_{1,4} & c_{0,5} \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1
\end{array}\right]$$

Table 5.5: $\mathcal{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$

5.5.2 $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$

A summary of the results for $\hat{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ are shown in Table 5.6. Notable differences in comparison to the results for $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ include the reduction in *All Equalities & Inequalities* which is due to the fact that constraints corresponding to (A4-$\mathcal{P}_\eta^{(k)}$) are only generated for $x \geq y$ for $(x, y) \in [0, 1]^2$, as well as the substantial decrease in *Vertices*. In experiments related to $\hat{M}^*\big[\mathcal{P}_1^{(2)}, Q_q\big]$ we found that $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(2)}, Q_*\big] = \mathcal{V}^*\big[\mathcal{P}_1^{(2)}, Q_*\big]$ and $\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(2)}, Q_*\big] = \mathcal{R}^*\big[\mathcal{P}_1^{(2)}, Q_*\big]$ and therefore conclude that $\hat{M}^*\big[\mathcal{P}_1^{(2)}\big] = M^*\big[\mathcal{P}_1^{(2)}\big]$.

As shown in Table 5.6 in experiments related to $\hat{M}^*\big[\mathcal{P}_1^{(3)}, Q_q\big]$, the number of vertices in $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_q\big]$, as was expected, increased with q and therefore did not suggest a vertex set that defines $\hat{M}^*\big[\mathcal{P}_1^{(3)}\big]$. However, in these experiments the following matrix was identified

$$\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

and based on this matrix we state the following conjecture, in which the supposed functions in $\mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$ are described as linear combinations of the functions corresponding to the vectors in $\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$.

Conjecture 5.2. *There exist $\lambda_i \in \mathbb{R}$ for $i = 1, \cdots, 4$ such that if*

$$\begin{aligned} \mu_{\mathbf{c}}(x, y) \;=\; & \lambda_1\big((x - 1/2) - (y - 1/2)\big) \;+ \\ & \lambda_2\big((x - 1/2)^2 - (y - 1/2)^2\big) \;+ \\ & \lambda_3\big((x - 1/2)^3 - (y - 1/2)^3\big) \;+ \\ & \lambda_4\big((x - 1/2)^2(y - 1/2) - (x - 1/2)(y - 1/2)^2\big) \end{aligned}$$

then $\mu_{\mathbf{c}} \in \hat{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}\big]$.

Following the same approach used in the analysis of $\mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$, we now assume this

Class	Variables	Non-redundant Equations	Non-redundant Inequalities	All Equations & Inequalities	Vertices	Rank of Vertex Matrix
$\hat{M}^*\left[\mathcal{P}_1^{(1)}, Q_1\right]$	3	3	0	12	1	1
$\hat{M}^*\left[\mathcal{P}_1^{(2)}, Q_1\right]$	6	5	2	15	2	2
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_2\right]$	10	7	6	24	5	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_3\right]$	10	7	8	32	6	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_4\right]$	10	7	10	42	8	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_5\right]$	10	7	12	54	9	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_6\right]$	10	7	14	68	11	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_7\right]$	10	7	16	84	12	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_8\right]$	10	7	18	102	14	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_9\right]$	10	7	20	122	15	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_{10}\right]$	10	7	22	144	17	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_{20}\right]$	10	7	42	474	32	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_{30}\right]$	10	7	62	1004	47	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_{40}\right]$	10	7	82	1734	62	4
$\hat{M}^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$	10	7	102	2664	77	4
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_3\right]$	15	10	14	36	16	6
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_{10}\right]$	15	10	73	148	398	6
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_{20}\right]$	15	10	213	478	1988	6
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_{30}\right]$	15	10	419	1008	4792	6
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_{40}\right]$	15	10	693	1738	8980	6
$\hat{M}^*\left[\mathcal{P}_1^{(4)}, Q_{50}\right]$	15	10	1033	2668	14490	6
$\hat{M}^*\left[\mathcal{P}_1^{(5)}, Q_4\right]$	21	13	25	50	284	9
$\hat{M}^*\left[\mathcal{P}_1^{(5)}, Q_{10}\right]$	21	13	121	152	92430	9

Table 5.6: $\hat{M}^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$

representation of $\mu_{\mathbf{c}} \in \hat{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}\big]$ and rewrite the axioms in this 4-dimensional subspace rather than in the original 10-dimensional space. For a given $q \in \mathbb{Z}_+$ and for each $(x, y) \in Q_q$, we have

$$\begin{aligned}
\lambda_1 + \frac{1}{4}\lambda_3 - \frac{1}{4}\lambda_4 &= 1 \\
\lambda_1 + 2(x - \frac{1}{2})\lambda_2 + 3(x - \frac{1}{2})^2\lambda_3 + \big(2(x - \frac{1}{2})(y - \frac{1}{2}) - (y - \frac{1}{2})^2\big)\lambda_4 &\geq 0 \\
-\lambda_1 - 2(y - \frac{1}{2})\lambda_2 - 3(y - \frac{1}{2})^2\lambda_3 + \big((x - \frac{1}{2})^2 - 2(x - \frac{1}{2})(y - \frac{1}{2})\big)\lambda_4 &\leq 0.
\end{aligned}$$

As was the case with our analysis of $M^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$, the results of computations for $\hat{M}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ with *lrs* in the 4-dimensional subspace were similar to those for $\hat{M}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ in the 10-dimensional subspace that are shown in Table 5.6. Working in the 4-dimensional subspace facilitated the visualization of $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ and substantially reduced the time required by *lrs* thereby allowing us to utilize larger values of q.

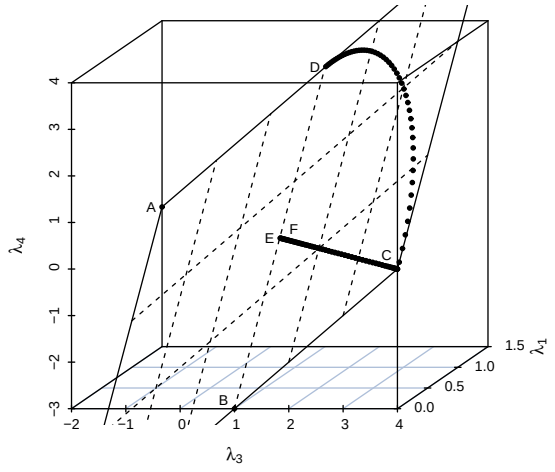
Figure 5.4 shows two projections of the set $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$, which contains 152 vertices, onto three dimensions, with Figure 5.4a depicting the fact that all vertices lie in the plane

$$\lambda_1 + \frac{1}{4}\lambda_3 - \frac{1}{4}\lambda_4 = 1.$$

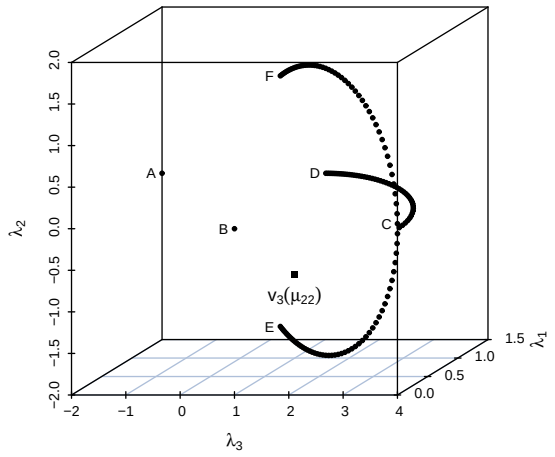
The values, after some minor rounding, of the labeled vertices are provided in Table 5.7 and the corresponding functions are shown in Figures 5.5 and Figure 5.6. In the reduced subspace we can write the following convex combinations

$$\begin{aligned}
v_5(\mu_{9,\varrho=1/2}(x, y)) &= \frac{1}{2}A + \frac{1}{3}C + \frac{1}{6}D, \\
v_3(\mu_{22}(x, y)) &= \frac{1}{3}D + \frac{2}{3}E, \text{ and} \\
(2/3)v_3(\mu_{24}(x, y)) &= \frac{4}{9}A + \frac{2}{9}B + \frac{1}{3}C,
\end{aligned}$$

however, we note that $(1/3)\tau_3(\mu_{9,\lim \varrho \to 0})$ cannot be represented as a convex combination in this subspace. It is interesting to note that the vectors AB and AD as well as the vectors CB and CD are orthogonal.

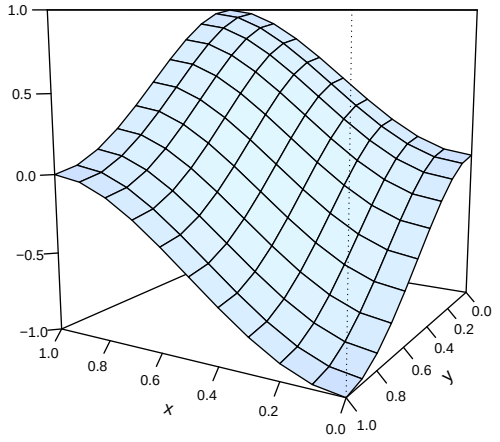


(a)

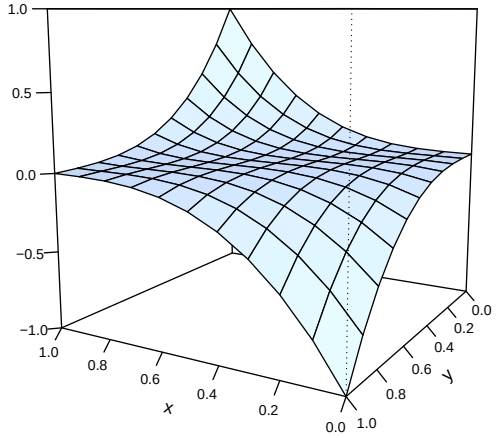


(b)

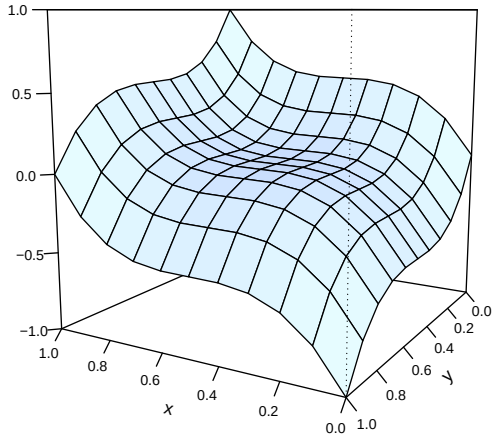
Figure 5.4: $\hat{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_{100}\right]$ in Reduced Dimension Space



(a) μ_A

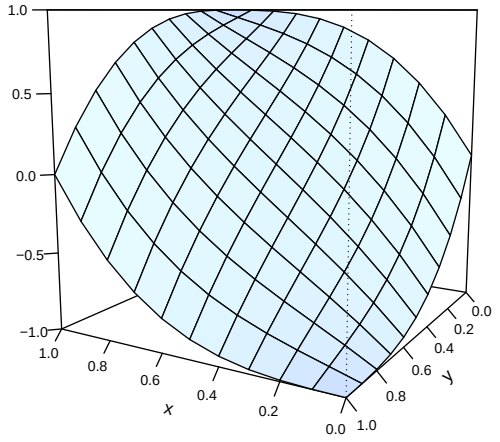


(b) μ_B

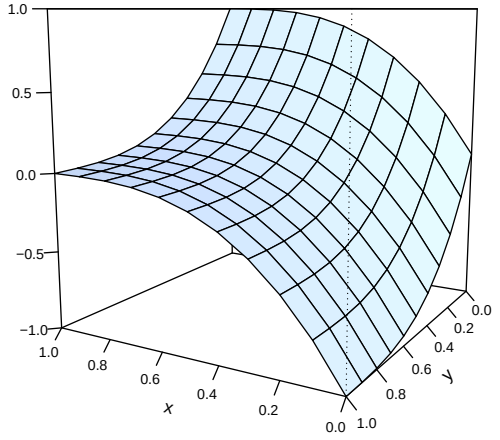


(c) μ_C

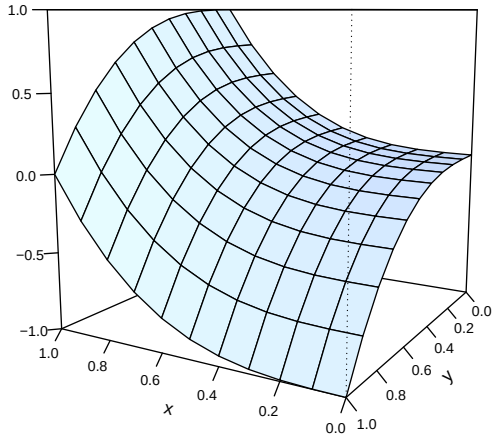
Figure 5.5: Three selected functions from $\hat{\mathcal{M}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$



(a) μ_D



(b) μ_E



(c) μ_F

Figure 5.6: Three additional functions from $\hat{\mathcal{M}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$

Vertex	λ_1	λ_2	λ_3	λ_4	μ
A	3/2	0	-2	0	$3(x^2-y^2)-2(x^3-y^3)$
B	0	0	1	-3	$(x^3-y^3)-3(x^2y-y^2x)$
C	0	0	4	0	$3(x-y)-6(x^2-y^2)+4(x^3-y^3)$
D	3/2	0	1	3	$3(x-y)-3(x^2-y^2)+(x^3-y^3)+3(x^2y-y^2x)$
E	3/4	-3/2	1	0	$3(x-y)-3(x^2-y^2)+(x^3-y^3)$
F	3/4	3/2	1	0	(x^3-y^3)

Table 5.7: Selected elements of $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ and functions from $\hat{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$

As shown in Table 5.6 in experiments related to $\hat{M}^*\big[\mathcal{P}_1^{(4)}, Q_q\big]$ and $\hat{M}^*\big[\mathcal{P}_1^{(5)}, Q_q\big]$, the number of vertices in $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(4)}, Q_q\big]$ and $\hat{\mathcal{V}}^*\big[\mathcal{P}_1^{(5)}, Q_q\big]$ respectively, as was expected, increased with q and therefore did not suggest a vertex set that defines $\hat{M}^*\big[\mathcal{P}_1^{(4)}\big]$ or $\hat{M}^*\big[\mathcal{P}_1^{(5)}\big]$. However, in these experiments the matrices $\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(4)}, Q_*\big]$ and $\hat{\mathcal{R}}^*\big[\mathcal{P}_1^{(5)}, Q_*\big]$ shown in Table 5.8 and Table 5.9 were identified. These results could be used to perform analyses similar to the one done for $\hat{M}^*\big[\mathcal{P}_1^{(3)}\big]$, but we will not pursue this.

$$\hat{\mathcal{R}}^*\left[\mathcal{P}_1^{(4)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

Table 5.8: $\widehat{R}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$

$$\hat{\mathcal{R}}^*\left[\mathcal{P}_1^{(5)}, Q_*\right] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} & c_{5,0} & c_{4,1} & c_{3,2} & c_{2,3} & c_{1,4} & c_{0,5} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}$$

Table 5.9: $\widehat{R}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$

5.5.3 $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$

A summary of the results for $\tilde{M}^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ are shown in Table 5.10. Notable differences in comparison to the results for $M^*\big[\mathcal{P}_1^{(k)}, Q_q\big]$ include the reduction in *All Equalities & Inequalities* which is due to the fact that constraints corresponding to (A4-$\mathcal{P}_\eta^{(k)}$) are only generated for $x \geq y$ and $x \leq 1 - y$ for $(x, y) \in [0, 1]^2$, as well as the substantial decrease in *Vertices*. In experiments related to $\tilde{M}^*\big[\mathcal{P}_1^{(2)}, Q_q\big]$ we found that $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(2)}, Q_1\big] = \mathcal{V}^*\big[\mathcal{P}_1^{(2)}, Q_*\big]$ and $\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(2)}, Q_1\big] = \mathcal{R}^*\big[\mathcal{P}_1^{(2)}, Q_*\big]$, but that for $q > 1$

$$\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(2)}, Q_q\big] = \tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(2)}, Q_q\big] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} \\ 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

which motivates the following conjecture.

Conjecture 5.3. $\tilde{M}^*\big[\mathcal{P}_1^{(2)}\big] = \big\{\, [0, 1, -1, 0, 0, 0] \,\big\}$ *and* $\mathcal{M}\big[\mathcal{P}_1^{(2)}\big] = \big\{x - y\big\}$.

As shown in Table 5.10 in experiments related to $\tilde{M}^*\big[\mathcal{P}_1^{(3)}, Q_q\big]$, the number of vertices in $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_q\big]$, as was expected, increased with q and therefore did not suggest a vertex set that defines $\tilde{M}^*\big[\mathcal{P}_1^{(3)}\big]$. However, in these experiments the following matrix was identified

$$\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

and based on this matrix we state the following conjecture, in which the supposed functions in $\tilde{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}\big]$ are described as linear combinations of the functions corresponding to the vectors in $\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$.

Class	Variables	Non-redundant Equations	Non-redundant Inequalities	All Equations & Inequalities	Vertices	Rank of Vertex Matrix
$\tilde{M}^*\left[\mathcal{P}_1^{(1)}, Q_1\right]$	3	3	0	12	1	1
$\tilde{M}^*\left[\mathcal{P}_1^{(2)}, Q_1\right]$	6	5	2	17	2	2
$\tilde{M}^*\left[\mathcal{P}_1^{(2)}, Q_2\right]$	6	6	0	26	1	1
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_2\right]$	10	8	4	29	4	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_3\right]$	10	8	4	40	4	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_4\right]$	10	8	5	55	5	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_5\right]$	10	8	5	72	5	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_6\right]$	10	8	6	93	6	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_7\right]$	10	8	6	116	6	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_8\right]$	10	8	7	143	7	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_9\right]$	10	8	7	157	7	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_{10}\right]$	10	8	8	205	8	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_{20}\right]$	10	8	13	695	13	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_{30}\right]$	10	8	18	1485	18	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_{40}\right]$	10	8	23	2575	23	3
$\tilde{M}^*\left[\mathcal{P}_1^{(3)}, Q_{50}\right]$	10	8	28	3965	28	3
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_3\right]$	15	12	7	44	8	4
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_{10}\right]$	15	13	8	209	8	3
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_{20}\right]$	15	13	13	699	13	3
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_{30}\right]$	15	13	18	1489	18	3
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_{40}\right]$	15	13	23	2579	23	3
$\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_{50}\right]$	15	13	28	3969	28	3
$\tilde{M}^*\left[\mathcal{P}_1^{(5)}, Q_4\right]$	21	16	13	63	34	6
$\tilde{M}^*\left[\mathcal{P}_1^{(5)}, Q_{10}\right]$	21	16	61	213	529	6
$\tilde{M}^*\left[\mathcal{P}_1^{(5)}, Q_{20}\right]$	21	16	221	703	4336	6

Table 5.10: $\tilde{M}^*\left[\mathcal{P}_1^{(*)}, Q_*\right]$

Conjecture 5.4. *There exist* $\lambda_i \in \mathbb{R}$ *for* $i = 1, \cdots, 3$ *such that if*

$$\begin{aligned}\mu_{\mathbf{c}}(x,y) \;=\; & \lambda_1\big((x-1/2)-(y-1/2)\big) \;+ \\ & \lambda_2\big((x-1/2)^3-(y-1/2)^3\big) \;+ \\ & \lambda_3\big((x-1/2)^2(y-1/2)-(x-1/2)(y-1/2)^2\big)\end{aligned}$$

then $\mu_{\mathbf{c}} \in \tilde{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}\big]$.

Following the same approach used is the analysis of $\mathcal{M}^*\big[\mathcal{P}_1^{(3)}\big]$, we now assume this representation of $\mu_{\mathbf{c}} \in \tilde{\mathcal{M}}^*\big[\mathcal{P}_1^{(3)}\big]$ and rewrite the axioms in this 3-dimensional subspace rather than in the original 10-dimensional space. For a given $q \in \mathbb{Z}_+$ and for each $(x,y) \in Q_q$, we have

$$\begin{aligned}\lambda_1 + \frac{\lambda_2}{4} - \frac{\lambda_3}{4} \;&=\; 1 \\ \lambda_1 + 3\big(x-\frac{1}{2}\big)^2\lambda_2 + \big(2(x-\frac{1}{2})(y-\frac{1}{2}) - (y-\frac{1}{2})^2\big)\lambda_3 \;&\geq\; 0 \\ -\lambda_1 - 3\big(y-\frac{1}{2}\big)^2\lambda_2 + \big((x-\frac{1}{2})^2 - 2(x-\frac{1}{2})(y-\frac{1}{2})\big)\lambda_3 \;&\leq\; 0.\end{aligned}$$

As was the case with our analysis of $M^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$, the results of computations for $\tilde{M}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ with *lrs* in the 3-dimensional subspace were similar to those for $\tilde{M}^*\big[\mathcal{P}_1^{(3)}, Q_*\big]$ in the 10-dimensional subspace that are shown in Table 5.10. Working in the 3-dimensional rather than the 10-dimensional subspace, again facilitated visualization and reduced the time required by *lrs* thereby allowing us to utilize larger values of q.

Figure 5.7 shows the set $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$ which has 53 vertices, with Figure 5.7a depicting the fact that all vertices in $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$ lie in the plane

$$\lambda_1 + \frac{\lambda_2}{4} - \frac{\lambda_3}{4} \;=\; 1,$$

and Figure 5.7b providing a hypothesized depiction of $\tilde{M}^*\big[\mathcal{P}_1^{(3)}\big]$ assuming it is the convex hull of $\tilde{\mathcal{V}}^*\big[\mathcal{P}_1^{(3)}, Q_{100}\big]$. A comparison of Figure 5.4 and Figure 5.7 shows the relationship between the basis functions of the sets $\hat{M}^*\big[\mathcal{P}_1^{(3)}\big]$ and $\tilde{M}^*\big[\mathcal{P}_1^{(3)}\big]$, and we

note the absence of $\left((x-\frac{1}{2})^2-(y-\frac{1}{2})^2\right)$ from the basis set of $\tilde{M}^*\left[\mathcal{P}_1^{(3)}\right]$.

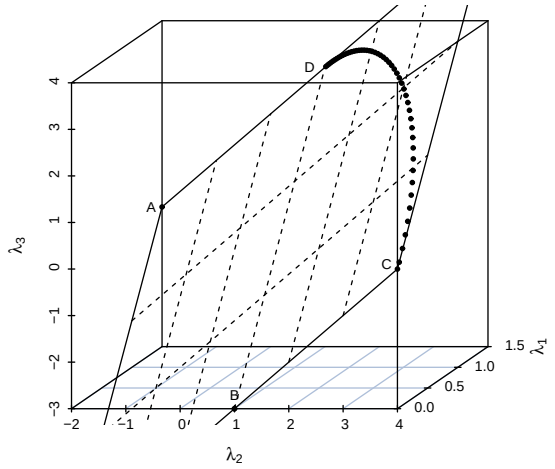
The values, after some minor rounding, of the labeled vertices are provided in Table 5.11 which can be seen to correspond to the vertices in Table 5.7 where $\lambda_2 = 0$ and the corresponding functions are shown as μ_A, μ_B, μ_C, and μ_D in Figures 5.5 and

Vertex	λ_1	λ_2	λ_3	μ
A	3/2	-2	0	$3(x^2-y^2)-2(x^3-y^3)$
B	0	1	-3	$(x^3-y^3)-3(x^2y-y^2x)$
C	0	4	0	$3(x-y)-6(x^2-y^2)+4(x^3-y^3)$
D	3/2	1	3	$3(x-y)-3(x^2-y^2)+(x^3-y^3)+3(x^2y-y^2x)$

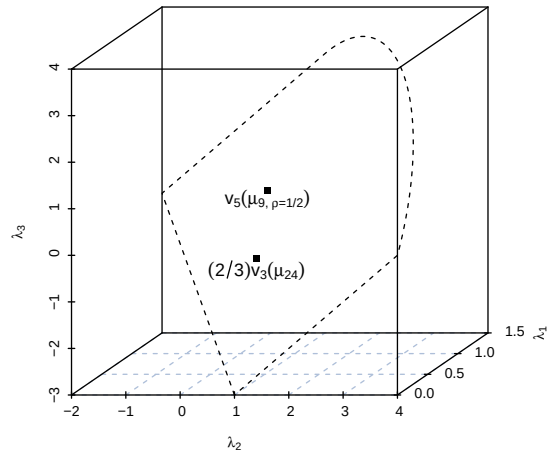
Table 5.11: Selected elements of $\tilde{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$ and functions in $\tilde{\mathcal{M}}^*\left[\mathcal{P}_1^{(3)}, Q_*\right]$

Figure 5.6. Similar to $\hat{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_{100}\right]$ we note that the vectors AB and AD as well as the vectors CB and CD are orthogonal.

As shown in Table 5.10 in experiments related to $\tilde{M}^*\left[\mathcal{P}_1^{(4)}, Q_q\right]$ and $\tilde{M}^*\left[\mathcal{P}_1^{(5)}, Q_q\right]$, the number of vertices in $\tilde{\mathcal{V}}^*\left[\mathcal{P}_1^{(4)}, Q_q\right]$ and $\tilde{\mathcal{V}}^*\left[\mathcal{P}_1^{(5)}, Q_q\right]$ respectively, as was expected, increased with q and therefore did not suggest a vertex set that defines $\tilde{M}^*\left[\mathcal{P}_1^{(4)}\right]$ or $\tilde{M}^*\left[\mathcal{P}_1^{(5)}\right]$. However, in these experiments the matrices $\tilde{\mathcal{R}}^*\left[\mathcal{P}_1^{(4)}, Q_*\right]$ and $\tilde{\mathcal{R}}^*\left[\mathcal{P}_1^{(5)}, Q_*\right]$ shown in Table 5.12 and Table 5.13 were identified. These results could be used to perform analyses similar to the one done for $\tilde{M}^*\left[\mathcal{P}_1^{(3)}\right]$, but we will not pursue this.



(a)



(b)

Figure 5.7: $\tilde{\mathcal{V}}^*\left[\mathcal{P}_1^{(3)}, Q_{100}\right]$ in Reduced Dimension Space

$$\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(4)}, Q_*\big] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 5.12: $\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(4)}, Q_*\big]$

$$\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(5)}, Q_*\big] = \begin{bmatrix} c_{0,0} & c_{1,0} & c_{0,1} & c_{2,0} & c_{1,1} & c_{0,2} & c_{3,0} & c_{2,1} & c_{1,2} & c_{0,3} & c_{4,0} & c_{3,1} & c_{2,2} & c_{1,3} & c_{0,4} & c_{5,0} & c_{4,1} & c_{3,2} & c_{2,3} & c_{1,4} & c_{0,5} \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}$$

Table 5.13: $\tilde{\mathcal{R}}^*\big[\mathcal{P}_1^{(5)}, Q_*\big]$

Chapter 6

Extensions

In this chapter we consider extensions of several of the ideas which we discussed in earlier chapters.

6.1 Real-Valued Feature Ranking Functions

We begin this section by introducing real-valued feature ranking functions. We then present some functions from the literature, and draw comparisons with some Boolean feature ranking functions. We conclude with a discussion of the results of an empirical study of the performance of real-valued ranking algorithms based on the functions we discussed.

6.1.1 Ranking Functions and Algorithms

If the vector $\lambda \in \mathbb{R}^m$ is the column of a real-valued document-term matrix A, then a *real-valued feature ranking function* is a function $\pi\colon \mathbb{R}^{|T|} \times \mathbb{R}^{|F|} \mapsto \mathbb{R}$ that for each $j \in V$ maps the vectors $\lambda[W_T]$, $\lambda[W_F]$ to the set of real numbers. The notation π will be used for a generic real-valued feature ranking function and specific real-valued feature ranking functions will be denoted as π_i for some integer i. When there is an obvious relationship between a Boolean and a real-valued feature ranking function, we will use the same value of i. We shall denote the set of all real-valued feature ranking functions as Π.

Just as with ranking algorithms based on Boolean feature ranking functions, each $\pi \in \Pi$ defines a different ranking algorithm and the set Π defines an entire class of ranking algorithms. We denote the ranking algorithm for a particular $\pi \in \Pi$ as π-RANKING, with the definition of this algorithm following that provided in §3.1.

We now present fourteen real-valued feature ranking functions. Again, let $\lambda \in \mathbb{R}^m$ be the column of a real-valued document-term matrix that corresponds to some feature in V. We will refer to $\lambda[W_T]$ and $\lambda[W_F]$ respectively as the relevant and irrelevant score vectors, and will let $\bar{\lambda}_T$ and $\bar{\lambda}_F$ denote the mean of $\lambda[W_T]$ and $\lambda[W_F]$. We have added 1 to the denominator of several of the functions in order to avoid division by zero errors.

π_0. The function, π_0 simply returns a random number sampled from the continuous uniform distribution U(0,1). It is included as a baseline for comparison with other functions.

π_1. The function, π_1 is the ratio of the mean of the relevant score vector to the mean of irrelevant score vector. It is defined as

$$\pi_1(\lambda[W_T], \lambda[W_F]) = \frac{\bar{\lambda}_T}{\bar{\lambda}_F + 1} \tag{6.1}$$

π_4. The function, π_4 is the difference of the mean of the relevant score vector and the mean of irrelevant score vector. It is defined as

$$\pi_4(\lambda[W_T], \lambda[W_F]) = \bar{\lambda}_T - \bar{\lambda}_F \tag{6.2}$$

We will also consider $|\pi_4|$.

π_8. The function, π_8 it is defined as

$$\pi_8(\lambda[W_T], \lambda[W_F]) = \sum_{i=1}^{|T|} \sum_{j=1}^{|F|} \Big| \lambda[W_T]_i - \lambda[W_F]_j \Big|. \tag{6.3}$$

It is the L_1 distance between the relevant score vector and the irrelevant score vector.

π_9. The function π_9 is the *point biserial correlation coefficient*. It is the *Pearson Product Moment Correlation* between a real-valued and a Boolean vector. In this case the real-valued vector is λ and the Boolean vector is the vector $[1^{|T|}, 0^{|F|}]$ that indicates relevance or irrelevance of each document. It is defined as

$$\pi_9(\lambda[W_T], \lambda[W_F]) = \frac{\bar{\lambda}_T - \bar{\lambda}_F}{s_\lambda} \sqrt{\frac{|T||F|}{(|T| + |F|)^2}} \tag{6.4}$$

where s_λ is the standard deviation of the vector λ. We will also consider the function $|\pi_9|$.

$\boldsymbol{\pi_{18}}$. The function, π_{18} is defined as

$$\pi_{18}(\lambda[W_T], \lambda[W_F]) = \bar{\lambda}_T. \tag{6.5}$$

It is the mean of the relevant score vector.

$\boldsymbol{\pi_{22}}$. The function, π_{22} is the ratio of the difference of the relevant score and irrelevant score vectors, to the sum of the relevant score and irrelevant score vectors. It is defined as

$$\pi_{22}(\lambda[W_T], \lambda[W_F]) = \frac{\bar{\lambda}_T - \bar{\lambda}_F}{\bar{\lambda}_T + \bar{\lambda}_F + 1}. \tag{6.6}$$

We will also consider the function $|\pi_{22}|$.

$\boldsymbol{\pi_{23}}$. The function, π_{23} is Fisher's Linear Discriminant. It is defined as

$$\pi_{23}(\lambda[W_T], \lambda[W_F]) = \frac{(\bar{\lambda}_T - \bar{\lambda}_F)^2}{v_{\lambda[W_T]} + v_{\lambda[W_T]}} \tag{6.7}$$

where $v_{\lambda[W_T]}$ and $v_{\lambda[W_F]}$ are the variances of the vectors $\lambda[W_T]$ and $\lambda[W_F]$ respectively.

$\boldsymbol{\pi_{24}}$. The function, π_{24} is a variant of Fisher's Linear Discriminant. It is defined as

$$\pi_{24}(\lambda[W_T], \lambda[W_F]) = \frac{\bar{\lambda}_T - \bar{\lambda}_F}{s_{\lambda[W_T]} + s_{\lambda[W_T]}} \tag{6.8}$$

where $s_{\lambda[W_T]}$ and $s_{\lambda[W_F]}$ are the standard deviations of the vectors $\lambda[W_T]$ and $\lambda[W_F]$ respectively. We will also consider the function $|\pi_{24}|$.

$\boldsymbol{\pi_{25}}$. The function, π_{25} is the $\max\{\text{AUC}, 1 - \text{AUC}\}$ where AUC is the area under the ROC curve for the vector λ.

$\boldsymbol{\pi_{26}}$. The function, π_{26} is the $\max\{\text{AUC}, 1 - \text{AUC}\}$ where conv. hull(AUC) is the area under the convex hull of the ROC curve for the vector λ.

6.1.2 Real-Valued Feature Ranking Function Properties

In this section we will show how many of the concepts we covered in our discussions of Boolean features and Boolean feature ranking functions translate to real-valued features and real-values ranking functions. We will continue to utilize the notation from §6.1.1, however, we will now assume that λ was generated by a real-valued classifier such as f^+ in §1.3 for which larger scores are considered indicative of relevant documents, or f^- in §1.3 for which larger scores are considered indicative of irrelevant documents. This assumption will allow us to closely follow §1.3, and as such we will consider AUC_+, AUC_-, U_+, and U_- and related notation to be as defined there. Many of the parallels that we discuss are based on the observation that AUC_+ and AUC_-, and U_+ and U_-, as well as $\bar{\lambda}_T$ and $\bar{\lambda}_F$, in the real-valued model play a role similar to that of $x = a/|T|$ and $y = b/|F|$ in the Boolean model. That $\bar{\lambda}_T$ and $\bar{\lambda}_F$ are included here is justified by the fact that while actually a test of the more general hypothesis that $\lambda[W_T]$ and $\lambda[W_F]$ come from the same population, the Wilcoxon test, which is frequently referred to as a non-parametric version of the two-sample t-test, can be viewed as testing that the means of two independent samples are different. (see e.g. [36] and [64], p.132-134).

Single Feature Classifier Model. Let λ_T and λ_F be randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively. If $\text{P}(\lambda_T > \lambda_F) = 1$, or equivalently if $\text{AUC}_+ = 1$, or equivalently if $U_+ = |T||F|$, or equivalently if $\bar{\lambda}_T > 0$ and $\bar{\lambda}_F = 0$ then there exists a binary classifier that correctly classifies all documents and λ can be viewed as containing "perfect information". Similarly, if $\text{P}(\lambda_F > \lambda_T) = 1$, or equivalently if $\text{AUC}_- = 1$, or equivalently if $U_- = |T||F|$, or equivalently if $\bar{\lambda}_T = 0$ and $\bar{\lambda}_F > 0$ then also there exists a binary classifier that correctly classifies all documents and λ can again be viewed as containing "perfect information". Note that when $\bar{\lambda}_F = 0$ the corresponding feature only appears in relevant documents and when $\bar{\lambda}_T = 0$ the corresponding feature only appears in irrelevant documents.

Positive and Negative Features. Let λ_T and λ_F be randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively, then if $\text{P}(\lambda_T > \lambda_F) - \text{P}(\lambda_F > \lambda_T) > 0$, or equivalently if $\text{AUC}_+ - \text{AUC}_- > 0$, or equivalently if $U_+ - U_- > 0$, or equivalently if $\bar{\lambda}_T - \bar{\lambda}_F > 0$, then

the presence of the corresponding feature in a document is evidence that the document is relevant and the feature is positive. Similarly, if $\mathrm{P}(\lambda_T > \lambda_F) - \mathrm{P}(\lambda_F > \lambda_T) < 0$, or equivalently if $\mathrm{AUC}_+ - \mathrm{AUC}_- < 0$, or equivalently if $U_+ - U_- < 0$, or equivalently if $\bar{\lambda}_T - \bar{\lambda}_F < 0$, then the presence of the corresponding feature in a document is evidence that the document is irrelevant and the feature is negative.

Monotone Feature Principle. We shall refer to a feature that appears in $\omega(V, K, \bar{\lambda}_T\text{-RANKING}, \uparrow)$, $\omega(V, K, \mathrm{AUC}_+\text{-RANKING}, \uparrow)$, $\omega(V, K, U_+\text{-RANKING}, \uparrow)$, or $\omega(V, K, \bar{\lambda}_T - \bar{\lambda}_F\text{-RANKING}, \uparrow)$ as a highly ranked positive feature, and a feature that appears in $\omega(V, K, \bar{\lambda}_F\text{-RANKING}, \uparrow)$, $\omega(V, K, \mathrm{AUC}_-\text{-RANKING}, \uparrow)$, $\omega(V, K, U_-\text{-RANKING}, \uparrow)$, or $\omega(V, K, \bar{\lambda}_F - \bar{\lambda}_T\text{-RANKING}, \uparrow)$ as a a highly ranked negative feature. Just as in the Boolean model, The Monotone Feature Principle for real-valued features says that, *non-noise features in textual data sets, that provide substantial separation, are highly ranked positive features.*

Class Separation and Noise. Let λ_T and λ_F be randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively, then if $\mathrm{P}(\lambda_T > \lambda_F) = \mathrm{P}(\lambda_F > \lambda_T) = 1/2$, or equivalently if $\mathrm{AUC}_+ = \mathrm{AUC}_- = 1/2$, or equivalently if $U_+/|T||F| = U_-/|T||F| = 1/2$, or equivalently if $\bar{\lambda}_T - \bar{\lambda}_F = 0$, then the corresponding feature is a class noise feature. In this case, the performance of the associated classifier is the same as if it had randomly assigned scores to documents. Similar to the Boolean case, in the distance of the point $(\bar{\lambda}_T, \bar{\lambda}_F)$ from the line of class noise $x = y$ in ROC space, provides a measure of separation.

Collection Noise. The variance and standard deviations are classic measures of noise, with smaller values considered to indicate less noise and larger values considered to indicate more noise. The denominator of π_9 is the standard deviation of λ, but since it only differs in a monotonic transformation from the variance, we will now consider the later. Since the variance of λ can be written as $\mathrm{Var}(\lambda) = \mathrm{E}(\lambda^2) - \mathrm{E}(\lambda)^2$ and noting that

$$\mathrm{E}(\lambda) = \frac{1}{|T| + |F|}\left(|T|\bar{\lambda}_T + |F|\bar{\lambda}_F\right)$$

we have

$$\mathrm{Var}(\lambda) = \frac{1}{|T| + |F|}\sum_{i=1}^{m}\lambda_i^2 - \frac{1}{\left(|T| + |F|\right)^2}\left(|T|\bar{\lambda}_T + |F|\bar{\lambda}_F\right)^2. \tag{6.9}$$

Considering $\mathrm{Var}(\lambda)$ as a quadratic function of $\bar{\lambda}_T$ and $\bar{\lambda}_F$ it can be shown that

$$\mathrm{argmax}\,\mathrm{Var}(\lambda)_{(\bar{\lambda}_T,\bar{\lambda}_F)} = \big\{(\bar{\lambda}_T,\bar{\lambda}_F) : \varrho\bar{\lambda}_T + (1-\varrho)\bar{\lambda}_F = 0\big\}.$$

So similar to the Boolean case in (4.5), $\mathrm{Var}(\lambda)$ takes its maximum along the line

$$\varrho\bar{\lambda}_T + (1-\varrho)\bar{\lambda}_F = 0 \tag{6.10}$$

and we shall refer to this *line of collection noise* and to features that appear on this line as *collection noise features*.

Strong Collection Noise. Motivated by the fact that the denominator of $\mathrm{den}(\pi_{23}) = \mathrm{Var}(\lambda[W_T])+\mathrm{Var}(\lambda[W_F])$ and the denominator of $\mathrm{den}(\pi_{24}) = \mathrm{Std}(\lambda[W_T])+\mathrm{Std}(\lambda[W_F])$, we now consider $\mathrm{den}(\pi_{23})$. The variance of $\lambda[W_T]$ can be written as

$$\mathrm{Var}(\lambda[W_T]) = \mathrm{E}(\lambda[W_T]^2) + \mathrm{E}(\lambda[W_F])^2 = \frac{1}{|T|}\sum_{i\in W_T}\lambda[W_T]_i^2 - \bar{\lambda}_T^2.$$

After proceeding similarly for $\mathrm{Var}(\lambda[W_F])$ we can write

$$\mathrm{den}(\pi_{23}) = \frac{1}{|T|}\sum_{i\in W_T}^{m}\lambda[W_T]_i^2 + \frac{1}{|F|}\sum_{i\in W_F}\lambda[W_F]_i^2 - \left(\bar{\lambda}_T^2 + \bar{\lambda}_F^2\right).$$

Considering $\mathrm{den}(\pi_{23})$ to be a quadratic function of $\bar{\lambda}_T$ and $\bar{\lambda}_F$ it can be shown that $\mathrm{den}(\pi_{23})$ takes its maximum when $\bar{\lambda}_T = \bar{\lambda}_F = 0$. So similar to the Boolean case in Proposition 4.7, $\mathrm{den}(\pi_{23})$ and $\mathrm{den}(\pi_{24})$ take their maximum at the point $(\bar{\lambda}_T, \bar{\lambda}_F) = (0,0)$ and we shall refer to this point as the *point of strong collection noise*.

Separation to Noise Ratios. We now remark that three of the real-valued ranking functions introduced in §6.1.1, namely π_9, π_{23} and π_{24} can be viewed as ratios of separation functions to noise functions.

Real-Valued Feature Ranking Function Axioms. We now state a set of axioms for real-valued feature ranking function that are similar to those introduced for Boolean feature ranking functions §5.1. We consider a real-valued feature ranking function,

$\pi\colon \mathbb{R}^{|T|} \times \mathbb{R}^{|F|} \mapsto \mathbb{R}$ to be *desirable* if it satisfies the following axioms

$$\bar{\lambda}_T > 0 \text{ and } \bar{\lambda}_F = 0 \Rightarrow \pi(\lambda[W_T], \lambda[W_F]) = 1 \tag{A1}$$

$$\bar{\lambda}_T = 0 \text{ and } \bar{\lambda}_F > 0 \Rightarrow \pi(\lambda[W_T], \lambda[W_F]) = -1 \tag{A2}$$

$$\bar{\lambda}_T - \bar{\lambda}_F = 1/2 \Rightarrow \pi(\lambda[W_T], \lambda[W_F]) = 0 \tag{A3}$$

$$\bar{\lambda}_T \geq \bar{\lambda}'_T \text{ and } \bar{\lambda}_F \leq \bar{\lambda}'_F \Rightarrow \pi(\lambda[W_T], \lambda[W_F]) \geq \pi(\lambda'[W_T], \lambda'[W_F]) \tag{A4}$$

Obviously, based on our discussions earlier in this section, these axioms could have been stated in terms of the AUC, the Wilcoxon statistics, or the $\mathrm{P}(\lambda_T > \lambda_F)$ where λ_T and λ_F are randomly selected scores from $\lambda[W_T]$ and $\lambda[W_F]$ respectively.

6.1.3 Real-valued Feature Ranking Results

In this section, we use the methodology presented in §2 to evaluate the relative performance of π-RANKING for each of the feature ranking functions discussed in §6.1.1. The detailed non-discounted results are provided in Appendix P and the detailed discounted results are provided in Appendix Q. The non-discounted results are summarized in Table 6.1 and the discounted results are summarized in Table 6.2. In addition, Figure 6.1, Figure 6.2, and Figure 6.3 depict the relationship between σ_K, θ_K and φ_K, and ν_K. Based on this data we make the following observations.

- The top of the lex-max-min, lex, avg and $\bar{\varphi}$ rankings, for both the non-discounted and discounted experiments included the classic feature ranking functions, i.e. the correlation coefficient and Fisher's linear discriminant and its variant.
- There seems to be a relatively strong correlation between the rankings of the feature ranking functions by $\widehat{\nu}$ and by $\bar{\varphi}$
- The feature ranking functions ranked highest in terms of $\widehat{\sigma}$, notably π_8, the L_1 distance, were among the worst performing features ranking functions as measured by $\widehat{\nu}$.
- The performance of feature ranking functions and their absolute values was not

substantially different.

- Both σ_K and θ_K increased with ν_K, while φ_K decreased with ν_K.

It is interesting to note that there are obviously quite a few similarities between these results and the results of the experiments using Boolean feature ranking functions. We will discuss this observation further in §7.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
π_{23}	$\lvert\pi_9\rvert$	$\lvert\pi_{24}\rvert$	π_9	0.01	$\lvert\pi_4\rvert$	0.98	π_8	0.51	$\lvert\pi_9\rvert$	0.89
$\lvert\pi_{24}\rvert$	π_9	π_{23}	$\lvert\pi_9\rvert$	0.01	π_8	0.98	$\lvert\pi_4\rvert$	0.47	π_9	0.88
$\lvert\pi_{22}\rvert$	π_{24}	π_{24}	π_{24}	0.02	$\lvert\pi_{22}\rvert$	0.98	$\lvert\pi_{22}\rvert$	0.46	$\lvert\pi_{24}\rvert$	0.88
π_{22}	$\lvert\pi_{24}\rvert$	$\lvert\pi_9\rvert$	π_0	0.04	π_{23}	0.98	π_{25}	0.46	π_{23}	0.87
π_{24}	π_0	$\lvert\pi_{22}\rvert$	$\lvert\pi_{24}\rvert$	0.04	$\lvert\pi_{24}\rvert$	0.98	π_{26}	0.45	π_{24}	0.87
$\lvert\pi_4\rvert$	π_{23}	$\lvert\pi_4\rvert$	π_{23}	0.07	π_{25}	0.98	π_4	0.43	$\lvert\pi_4\rvert$	0.86
π_4	$\lvert\pi_{22}\rvert$	π_9	$\lvert\pi_{22}\rvert$	0.12	π_{26}	0.98	π_{22}	0.43	$\lvert\pi_{22}\rvert$	0.86
π_{25}	$\lvert\pi_4\rvert$	π_{25}	$\lvert\pi_4\rvert$	0.13	π_4	0.97	π_{23}	0.43	π_{25}	0.86
π_9	π_{22}	π_{22}	π_{22}	0.13	π_{22}	0.97	π_1	0.41	π_4	0.84
$\lvert\pi_9\rvert$	π_4	π_4	π_4	0.14	π_{24}	0.97	π_{18}	0.40	π_{22}	0.84
π_{26}	π_{25}	π_{26}	π_{25}	0.15	π_9	0.96	$\lvert\pi_{24}\rvert$	0.40	π_{26}	0.83
π_8	π_{26}	π_8	π_{26}	0.20	$\lvert\pi_9\rvert$	0.96	π_{24}	0.38	π_8	0.78
π_1	π_8	π_1	π_8	0.48	π_1	0.92	π_9	0.35	π_1	0.72
π_{18}	π_1	π_{18}	π_1	0.53	π_{18}	0.90	$\lvert\pi_9\rvert$	0.35	π_{18}	0.69
π_0	π_{18}	π_0	π_{18}	0.54	π_0	0.16	π_0	0.01	π_0	0.06

Table 6.1: π-RANKING Not Discounted Stopwords Included

6.2 Boolean Greedy Algorithms

Given a Boolean feature ranking function $\mu \in \mathcal{M}$, a Boolean document-term matrix A, *two* features $i, j \in V$, and a Boolean function $f\colon \mathbb{B} \times \mathbb{B} \mapsto \mathbb{B}$ we can compute

$$\mu\big(f(i,j)\big) = \mu(a_{f(i,j)}, b_{f(i,j)}, c_{f(i,j)}, d_{f(i,j)})$$

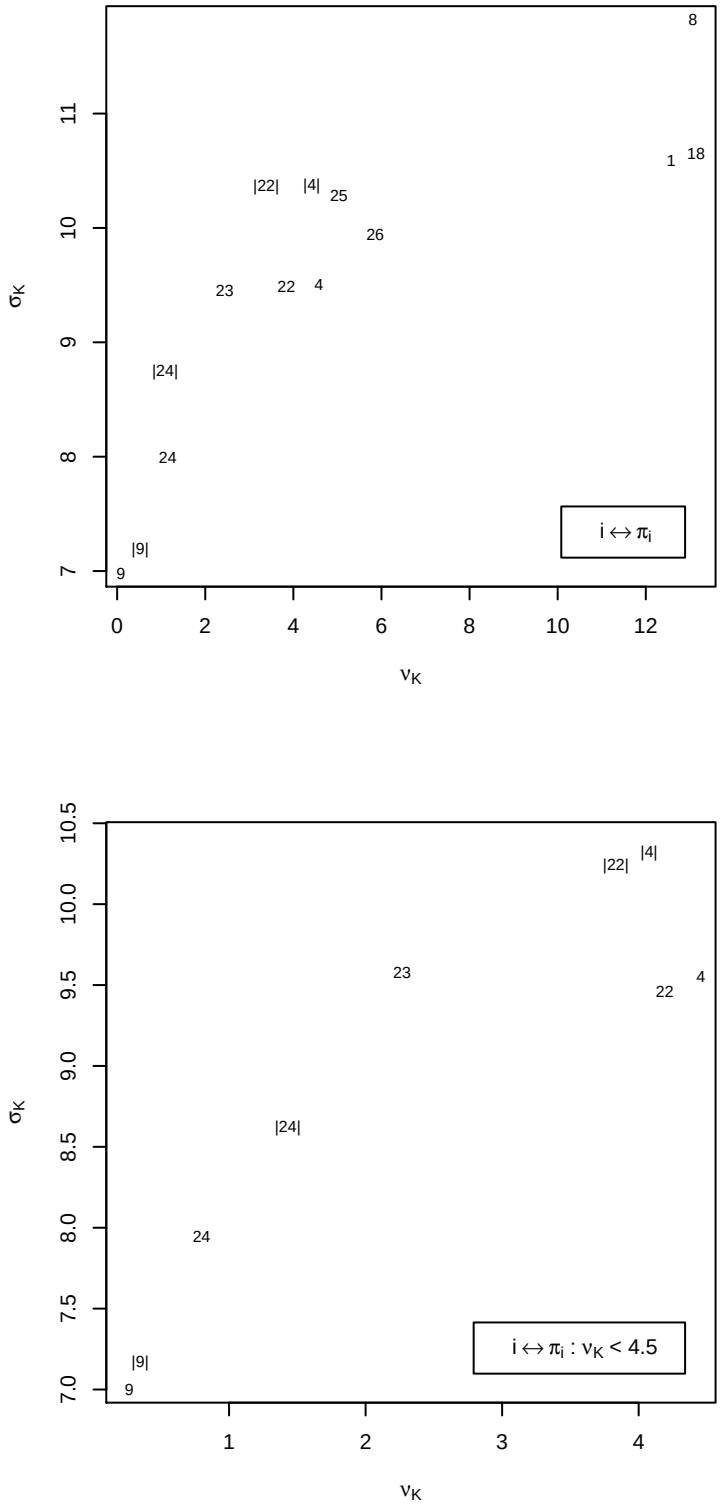


Figure 6.1: π-RANKING Stopwords Included Not Discounted: σ_K vs ν_K

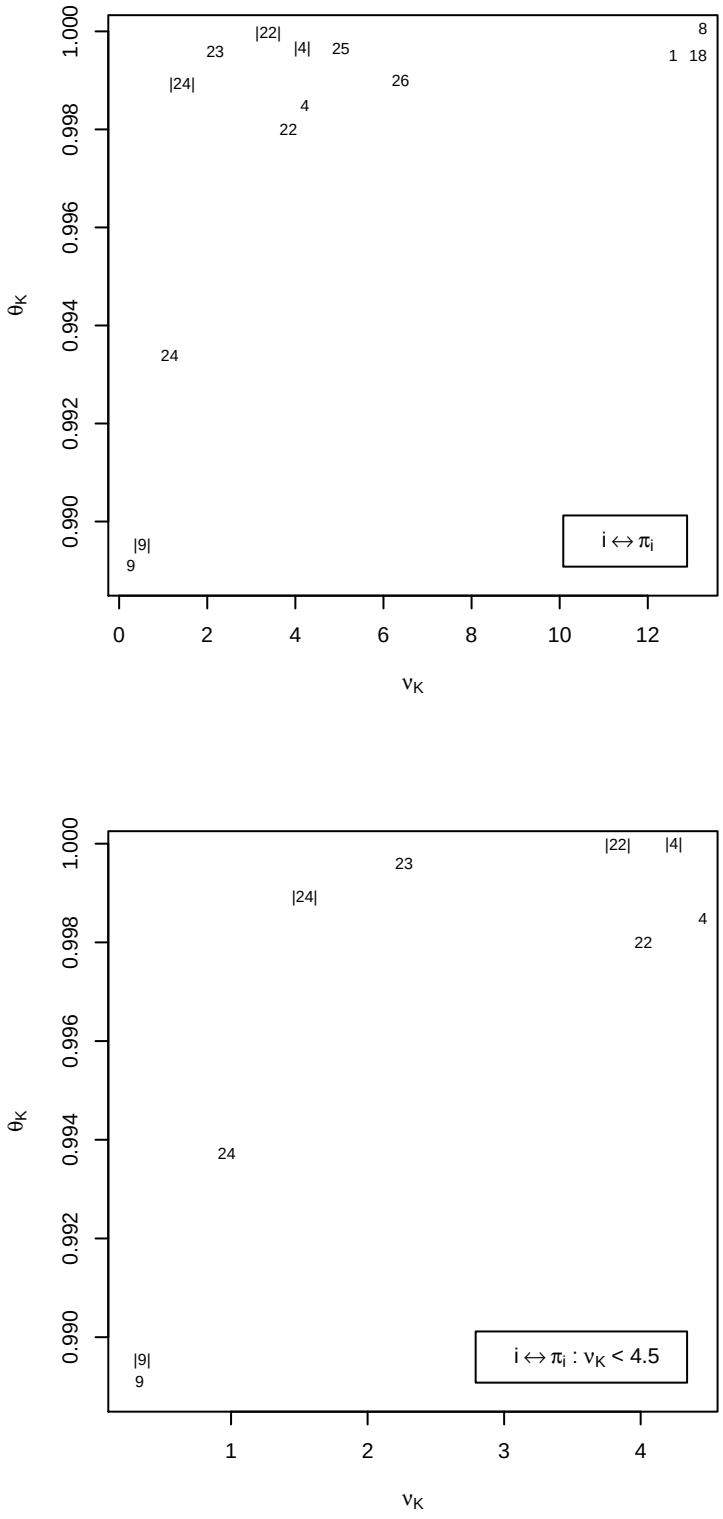


Figure 6.2: π-RANKING Stopwords Included Not Discounted: θ_K vs ν_K

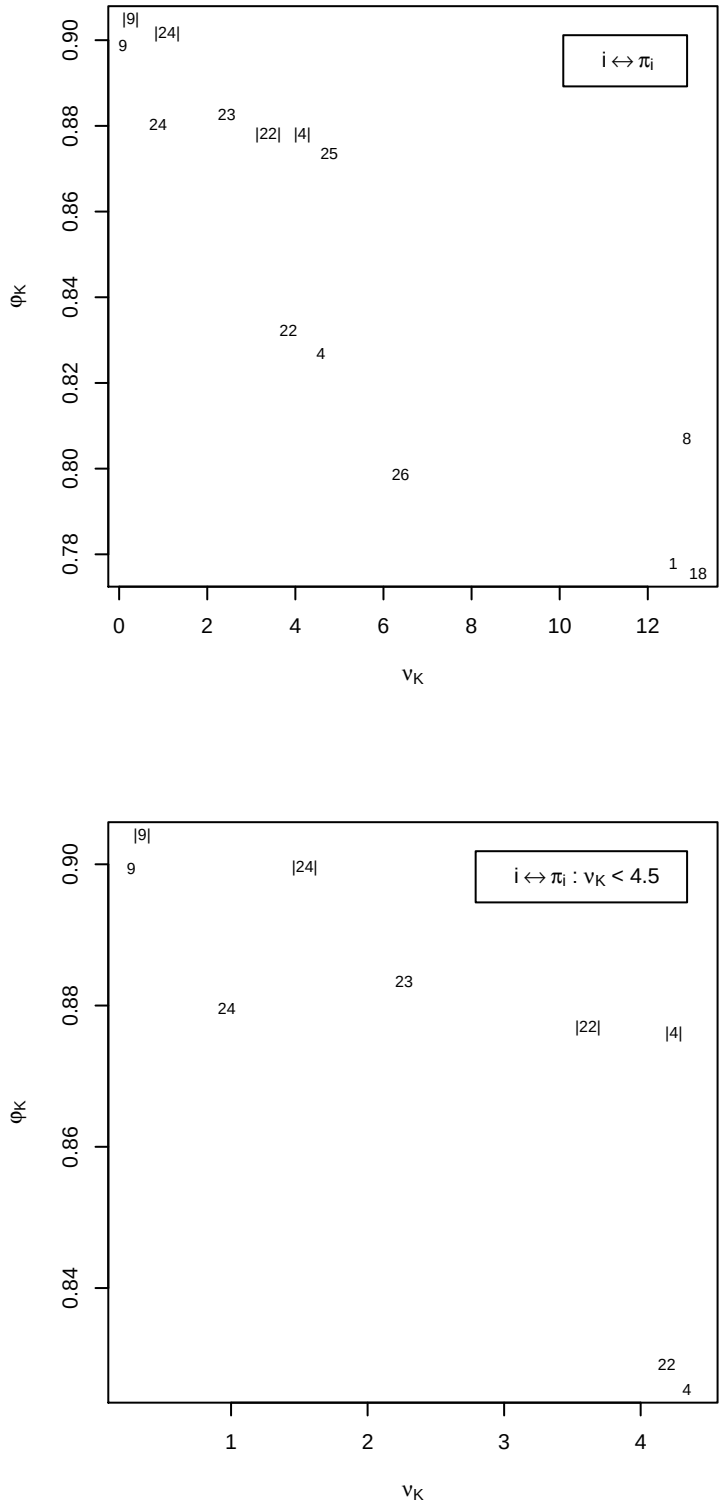


Figure 6.3: π-RANKING Stopwords Included Not Discounted: φ_K vs ν_K

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}'$		$\widehat{\sigma}'$		$\bar{\varphi}$	
$\lvert\pi_{24}\rvert$	$\lvert\pi_9\rvert$	$\lvert\pi_{24}\rvert$	π_9	0.01	$\lvert\pi_4\rvert$	0.98	$\lvert\pi_{22}\rvert$	0.41	$\lvert\pi_9\rvert$	0.89
π_{24}	π_9	$\lvert\pi_9\rvert$	$\lvert\pi_9\rvert$	0.01	$\lvert\pi_{22}\rvert$	0.98	$\lvert\pi_4\rvert$	0.40	π_9	0.88
π_{23}	π_{24}	π_{24}	π_{24}	0.02	π_{23}	0.98	π_{23}	0.40	$\lvert\pi_{24}\rvert$	0.88
π_9	$\lvert\pi_{24}\rvert$	π_{23}	π_0	0.04	π_{25}	0.98	$\lvert\pi_{24}\rvert$	0.39	π_{23}	0.87
$\lvert\pi_9\rvert$	π_0	π_9	$\lvert\pi_{24}\rvert$	0.04	π_4	0.97	π_{25}	0.39	π_{24}	0.87
$\lvert\pi_{22}\rvert$	π_{23}	$\lvert\pi_{22}\rvert$	π_{23}	0.07	π_8	0.97	π_{22}	0.37	$\lvert\pi_4\rvert$	0.86
π_{22}	$\lvert\pi_{22}\rvert$	$\lvert\pi_4\rvert$	$\lvert\pi_{22}\rvert$	0.12	π_{22}	0.97	π_{24}	0.37	π_{22}	0.86
$\lvert\pi_4\rvert$	$\lvert\pi_4\rvert$	π_{25}	$\lvert\pi_4\rvert$	0.13	π_{24}	0.97	π_4	0.36	$\lvert\pi_{22}\rvert$	0.86
π_4	π_{22}	π_{22}	π_{22}	0.13	$\lvert\pi_{24}\rvert$	0.97	π_9	0.35	π_{25}	0.86
π_{25}	π_4	π_4	π_4	0.14	π_{26}	0.97	$\lvert\pi_9\rvert$	0.35	π_4	0.85
π_{26}	π_{25}	π_{26}	π_{25}	0.15	π_9	0.96	π_{26}	0.35	π_{26}	0.85
π_8	π_{26}	π_8	π_{26}	0.20	$\lvert\pi_9\rvert$	0.96	π_8	0.30	π_8	0.80
π_1	π_8	π_1	π_8	0.48	π_1	0.86	π_1	0.23	π_1	0.74
π_{18}	π_1	π_{18}	π_1	0.53	π_{18}	0.84	π_{18}	0.22	π_{18}	0.72
π_0	π_{18}	π_0	π_{18}	0.54	π_0	0.13	π_0	0.01	π_0	0.06

Table 6.2: π-RANKING Discounted Stopwords Included

where

$$
\begin{aligned}
a_{f(i,j)} &\triangleq \text{the number of relevant documents for which } f(i,j) = 1 \\
b_{f(i,j)} &\triangleq \text{the number of irrelevant documents for which } f(i,j) = 1 \\
c_{f(i,j)} &\triangleq \text{the number of relevant documents for which } f(i,j) = 0 \\
d_{f(i,j)} &\triangleq \text{the number of irrelevant documents for which } f(i,j) = 0
\end{aligned}
$$

and $\mu\big(f(i,j)\big)$ can be viewed as the value of μ for the "feature" $f(i,j)$. For example, if f is the AND function, the "feature" $f(i,j)$ is the *composite* feature i AND j, and $a_{f(i,j)}$ is the number of relevant documents that contain feature i *and* feature j. Returning to the *fuel* topic we discussed in §4 one might imagine that the feature "crude" AND "prices" might be an interesting feature. Notice that, the information in the contingency tables $\boxplus_j$ for all $j \in V$ is not sufficient to compute $\mu\big(f(i,j)\big)$. In order to calculate $a_{f(i,j)}$ for any f, we must know *which* relevant documents contain feature i and *which* relevant documents contain feature j. The vectors $A[W_T, i]$ and $A[W_T, j]$ contain exactly this information, whereas the contingency tables $\boxplus_i$ and $\boxplus_j$ only tell us *how many* relevant

documents contain each feature. For example to compute $a_{f(i,j)}$ where f is the AND function, simply compute the vector

$$f_{W_T}(i,j) = A[W_T, \{i\}] \text{ AND } A[W_T, \{j\}]$$

and then count the number of 1s that it contains. Clearly, the calculation of the vector $f_{W_T}(i,j)$ can be consider an intermediate step for subsequent computations. For example, we could use it to compute $\mu\big(f(f(i,j),k)\big)$ for $i,j,k \in V$.

The ability to compute the value of a feature ranking function $\mu \in \mathcal{M}$ for a composite feature can be used to implement a greedy feature selection algorithm as shown below.

μ-GREEDY

Input: The set V, a Boolean document-term matrix A, a Boolean function f, a function $\mu \in \mathcal{M}$, a sort order $\updownarrow \in \{\downarrow, \uparrow\}$ and an integer K.

Step 1: Set $k := 0$, $\lambda = 0^{|T|+|F|}$, $z = 0$.

Step 2: Set $s_k := j^*$, $z := \mu\big(f(\lambda, j^*)\big)$, and $\lambda := f(\lambda, j^*)$, where $j^* := \text{argmax}_{\,j \in V \setminus S}\ \mu\big(f(\lambda, \{j\})\big)$.

Step 3: If $k := K$ then set $(s_k) := s_1, s_2, \cdots, s_K$ and goto **Output**, otherwise set $k := k + 1$ and goto **Step 2**.

Output: Output (s_k).

At each iteration a typical greedy algorithm selects the feature which will result in the largest increase in the objective function, which in this case is the feature ranking function μ, and it terminates when the objective function cannot be increased. It should be mentioned that μ-GREEDY differs in that at each iteration, it adds the feature that results in the largest value of z, even if that value is does not represent an increase in z, until K features have been selected. The reason for this algorithmic design is that our feature set evaluation methodology requires each set to contain exactly K features.

While μ-GREEDY can be implemented for any Boolean function f, we only used the Boolean OR in the experiments discussed in §6.2.1.

6.2.1 μ-GREEDY Results

In this section, we use the methodology presented in §2 to evaluate the relative performance of μ-GREEDY for each of the Boolean feature ranking functions discussed in §3.8. The detailed non-discounted results are provided in Appendix M and the detailed discounted results are provided in Appendix N. The non-discounted results are summarized in Table 6.3 and the discounted results are summarized in Table 6.4. In addition, Figure 6.4, Figure 6.5, and Figure 6.6 depict the relationship between σ_K, θ_K and φ_K, and ν_K. Based on this data we make the following observations.

- The feature ranking functions ranked highest in terms of $\widehat{\sigma}$, were among the worst performing features ranking functions as measured by $\widehat{\nu}$.
- Interestingly, μ_8 selected relatively little noise but achieved the best separation of those functions that selected relatively little noise. However, other feature ranking functions such as μ_{18} and μ_{19} that selected a large amount of noise when used in μ-RANKING, continued to do so in μ-GREEDY.
- Further details related to this observation are provided in Appendix O where in Table O.1 and Figure O.2 where we note the relatively low values of $\sigma_K[J]$ and $\dot{\sigma}[J]$ associated with μ_8, and the continued high values associated with μ_{18} and μ_{19}.
- The value of $\widehat{\sigma}$ achieved by functions that selected relatively little noise was lower than in the μ-RANKING results.
- Except for the functions that selected a relatively large number of noise features, there was little variation in the values of φ_K.

Further support for this suggestion is offered by the data in Appendix O where in Table O.1 and Figure O.2 we see that μ_8 had the largest value of $\sigma_K[J]$, and along with μ_{19} had the second largest value of $\dot{\sigma}[J]$.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
μ_8	μ_4	μ_2	μ_1	0.00	μ_2	0.98	$\lvert\mu_5\rvert$	0.41	μ_2	0.91
μ_5	μ_{15}	μ_4	μ_3	0.00	μ_4	0.98	μ_{18}	0.41	μ_7	0.91
μ_6	μ_{24}	μ_{13}	μ_4	0.00	$\lvert\mu_4\rvert$	0.98	$\lvert\mu_6\rvert$	0.40	$\lvert\mu_7\rvert$	0.91
μ_{13}	μ_{23}	μ_{15}	μ_9	0.00	μ_8	0.98	μ_{19}	0.36	μ_{17}	0.91
$\lvert\mu_4\rvert$	μ_{12}	μ_{24}	μ_{10}	0.00	μ_{12}	0.98	μ_5	0.28	μ_3	0.90
μ_{15}	μ_{21}	μ_{12}	μ_{11}	0.00	μ_{13}	0.98	μ_6	0.28	μ_9	0.90
μ_4	μ_{10}	μ_{21}	μ_{12}	0.00	μ_{15}	0.98	μ_8	0.27	$\lvert\mu_9\rvert$	0.90
μ_{23}	μ_{16}	μ_{23}	μ_{15}	0.00	μ_{21}	0.98	μ_{13}	0.23	μ_{10}	0.90
μ_2	μ_9	$\lvert\mu_{24}\rvert$	μ_{16}	0.00	μ_{23}	0.98	μ_4	0.22	μ_{11}	0.90
μ_{12}	μ_{11}	$\lvert\mu_4\rvert$	μ_{20}	0.00	μ_{24}	0.98	$\lvert\mu_4\rvert$	0.22	μ_{12}	0.90
μ_{21}	μ_3	μ_8	$\lvert\mu_{20}\rvert$	0.00	$\lvert\mu_{24}\rvert$	0.98	μ_2	0.21	μ_{15}	0.90
μ_{24}	μ_{20}	μ_{10}	μ_{21}	0.00	μ_7	0.97	μ_{14}	0.21	μ_{16}	0.90
$\lvert\mu_{24}\rvert$	$\lvert\mu_{20}\rvert$	μ_{16}	μ_{22}	0.00	$\lvert\mu_7\rvert$	0.97	μ_{15}	0.21	μ_{20}	0.90
μ_9	μ_1	μ_7	$\lvert\mu_{22}\rvert$	0.00	μ_9	0.97	μ_{23}	0.21	$\lvert\mu_{20}\rvert$	0.90
μ_{10}	μ_{22}	$\lvert\mu_7\rvert$	μ_{23}	0.00	$\lvert\mu_9\rvert$	0.97	μ_{24}	0.21	μ_{21}	0.90
μ_{11}	$\lvert\mu_{22}\rvert$	μ_9	μ_{24}	0.00	μ_{10}	0.97	$\lvert\mu_{24}\rvert$	0.21	μ_{24}	0.90
μ_{16}	μ_{13}	μ_{11}	μ_2	0.01	μ_{11}	0.97	μ_{10}	0.20	$\lvert\mu_{24}\rvert$	0.90
$\lvert\mu_9\rvert$	$\lvert\mu_4\rvert$	$\lvert\mu_9\rvert$	$\lvert\mu_4\rvert$	0.01	μ_{16}	0.97	μ_{12}	0.20	μ_1	0.89
μ_7	μ_2	μ_{17}	μ_7	0.01	μ_{17}	0.96	μ_{16}	0.20	μ_4	0.89
$\lvert\mu_7\rvert$	$\lvert\mu_{24}\rvert$	μ_3	$\lvert\mu_7\rvert$	0.01	μ_5	0.95	μ_{21}	0.20	μ_{13}	0.89
μ_{17}	μ_7	μ_{20}	$\lvert\mu_9\rvert$	0.01	μ_6	0.95	μ_7	0.19	μ_{22}	0.89
μ_3	$\lvert\mu_7\rvert$	$\lvert\mu_{20}\rvert$	μ_{13}	0.01	μ_3	0.92	$\lvert\mu_7\rvert$	0.19	μ_{23}	0.89
μ_{20}	$\lvert\mu_9\rvert$	μ_5	μ_{17}	0.01	μ_{14}	0.92	μ_9	0.19	$\lvert\mu_4\rvert$	0.88
$\lvert\mu_{20}\rvert$	μ_{17}	μ_6	$\lvert\mu_{24}\rvert$	0.01	μ_{20}	0.92	$\lvert\mu_9\rvert$	0.19	$\lvert\mu_{22}\rvert$	0.88
μ_1	μ_8	μ_1	μ_8	0.02	$\lvert\mu_{20}\rvert$	0.92	μ_{11}	0.19	μ_5	0.84
μ_{22}	μ_0	μ_{22}	μ_0	0.04	$\lvert\mu_6\rvert$	0.89	μ_{17}	0.19	μ_6	0.84
$\lvert\mu_{22}\rvert$	μ_5	$\lvert\mu_{22}\rvert$	μ_5	0.09	$\lvert\mu_5\rvert$	0.88	μ_3	0.18	μ_8	0.84
$\lvert\mu_6\rvert$	μ_6	$\lvert\mu_5\rvert$	μ_6	0.09	μ_{18}	0.88	μ_{20}	0.18	$\lvert\mu_5\rvert$	0.73
$\lvert\mu_5\rvert$	μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	0.12	μ_1	0.86	$\lvert\mu_{20}\rvert$	0.18	$\lvert\mu_6\rvert$	0.70
μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	$\lvert\mu_6\rvert$	0.40	μ_{22}	0.86	μ_1	0.15	μ_{18}	0.67
μ_{18}	$\lvert\mu_5\rvert$	μ_{18}	$\lvert\mu_5\rvert$	0.41	$\lvert\mu_{22}\rvert$	0.85	μ_{22}	0.15	μ_{14}	0.40
μ_{19}	μ_{18}	μ_{19}	μ_{18}	0.59	μ_{19}	0.84	$\lvert\mu_{22}\rvert$	0.15	μ_{19}	0.27
μ_0	μ_{19}	μ_0	μ_{19}	0.59	μ_0	0.15	μ_0	0.01	μ_0	0.06

Table 6.3: μ-GREEDY Not Discounted Stopwords Included

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}'$		$\widehat{\sigma}'$		$\bar{\varphi}$	
μ_4	μ_4	μ_2	μ_1	0.00	μ_2	0.98	$\lvert\mu_5\rvert$	0.27	μ_2	0.91
μ_{13}	μ_{15}	μ_4	μ_3	0.00	μ_4	0.98	$\lvert\mu_6\rvert$	0.27	μ_7	0.91
$\lvert\mu_4\rvert$	μ_{24}	μ_{15}	μ_4	0.00	$\lvert\mu_4\rvert$	0.98	μ_8	0.27	$\lvert\mu_7\rvert$	0.91
μ_2	μ_{23}	μ_{24}	μ_9	0.00	μ_{12}	0.98	μ_5	0.25	μ_{17}	0.91
μ_{15}	μ_{12}	μ_{12}	μ_{10}	0.00	μ_{13}	0.98	μ_6	0.25	μ_3	0.90
μ_{24}	μ_{21}	μ_{13}	μ_{11}	0.00	μ_{15}	0.98	μ_4	0.22	μ_9	0.90
$\lvert\mu_{24}\rvert$	μ_{10}	μ_{21}	μ_{12}	0.00	μ_{21}	0.98	$\lvert\mu_4\rvert$	0.22	$\lvert\mu_9\rvert$	0.90
μ_{23}	μ_{16}	μ_{23}	μ_{15}	0.00	μ_{23}	0.98	μ_{13}	0.22	μ_{10}	0.90
μ_{12}	μ_9	$\lvert\mu_{24}\rvert$	μ_{16}	0.00	μ_{24}	0.98	μ_2	0.21	μ_{11}	0.90
μ_{21}	μ_{11}	$\lvert\mu_4\rvert$	μ_{20}	0.00	$\lvert\mu_{24}\rvert$	0.98	μ_{15}	0.21	μ_{12}	0.90
μ_{10}	μ_3	μ_{10}	$\lvert\mu_{20}\rvert$	0.00	μ_7	0.97	μ_{23}	0.21	μ_{15}	0.90
μ_{16}	μ_{20}	μ_{16}	μ_{21}	0.00	$\lvert\mu_7\rvert$	0.97	μ_{24}	0.21	μ_{16}	0.90
μ_8	$\lvert\mu_{20}\rvert$	μ_7	μ_{22}	0.00	μ_8	0.97	$\lvert\mu_{24}\rvert$	0.21	μ_{20}	0.90
μ_7	μ_1	$\lvert\mu_7\rvert$	$\lvert\mu_{22}\rvert$	0.00	μ_9	0.97	μ_{10}	0.20	$\lvert\mu_{20}\rvert$	0.90
$\lvert\mu_7\rvert$	μ_{22}	μ_8	μ_{23}	0.00	$\lvert\mu_9\rvert$	0.97	μ_{12}	0.20	μ_{21}	0.90
μ_9	$\lvert\mu_{22}\rvert$	μ_9	μ_{24}	0.00	μ_{10}	0.97	μ_{16}	0.20	μ_{24}	0.90
μ_{11}	μ_{13}	μ_{11}	μ_2	0.01	μ_{11}	0.97	μ_{18}	0.20	$\lvert\mu_{24}\rvert$	0.90
$\lvert\mu_9\rvert$	$\lvert\mu_4\rvert$	$\lvert\mu_9\rvert$	$\lvert\mu_4\rvert$	0.01	μ_{16}	0.97	μ_{21}	0.20	μ_1	0.89
μ_{17}	μ_2	μ_{17}	μ_7	0.01	μ_{17}	0.96	μ_7	0.19	μ_4	0.89
μ_3	$\lvert\mu_{24}\rvert$	μ_3	$\lvert\mu_7\rvert$	0.01	μ_5	0.94	$\lvert\mu_7\rvert$	0.19	μ_{13}	0.89
μ_{20}	μ_7	μ_{20}	$\lvert\mu_9\rvert$	0.01	μ_6	0.94	μ_9	0.19	μ_{22}	0.89
$\lvert\mu_{20}\rvert$	$\lvert\mu_7\rvert$	$\lvert\mu_{20}\rvert$	μ_{13}	0.01	μ_3	0.92	$\lvert\mu_9\rvert$	0.19	μ_{23}	0.89
μ_5	$\lvert\mu_9\rvert$	μ_5	μ_{17}	0.01	μ_{20}	0.92	μ_{11}	0.19	$\lvert\mu_4\rvert$	0.88
μ_6	μ_{17}	μ_6	$\lvert\mu_{24}\rvert$	0.01	$\lvert\mu_{20}\rvert$	0.91	μ_{17}	0.19	$\lvert\mu_{22}\rvert$	0.88
μ_1	μ_8	μ_1	μ_8	0.02	μ_{14}	0.89	μ_3	0.18	μ_5	0.85
μ_{22}	μ_0	μ_{22}	μ_0	0.04	μ_1	0.86	μ_{14}	0.18	μ_6	0.85
$\lvert\mu_{22}\rvert$	μ_5	$\lvert\mu_{22}\rvert$	μ_5	0.09	$\lvert\mu_5\rvert$	0.86	μ_{20}	0.18	μ_8	0.83
$\lvert\mu_6\rvert$	μ_6	$\lvert\mu_5\rvert$	μ_6	0.09	μ_{22}	0.86	$\lvert\mu_{20}\rvert$	0.18	$\lvert\mu_5\rvert$	0.74
$\lvert\mu_5\rvert$	μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	0.12	$\lvert\mu_6\rvert$	0.85	μ_1	0.15	$\lvert\mu_6\rvert$	0.71
μ_{14}	$\lvert\mu_6\rvert$	μ_{14}	$\lvert\mu_6\rvert$	0.40	$\lvert\mu_{22}\rvert$	0.85	μ_{19}	0.15	μ_{18}	0.71
μ_{18}	$\lvert\mu_5\rvert$	μ_{18}	$\lvert\mu_5\rvert$	0.41	μ_{18}	0.82	μ_{22}	0.15	μ_{14}	0.32
μ_{19}	μ_{18}	μ_{19}	μ_{18}	0.59	μ_{19}	0.75	$\lvert\mu_{22}\rvert$	0.15	μ_{19}	0.17
μ_0	μ_{19}	μ_0	μ_{19}	0.59	μ_0	0.13	μ_0	0.01	μ_0	0.05

Table 6.4: μ-GREEDY Discounted Stopwords Included

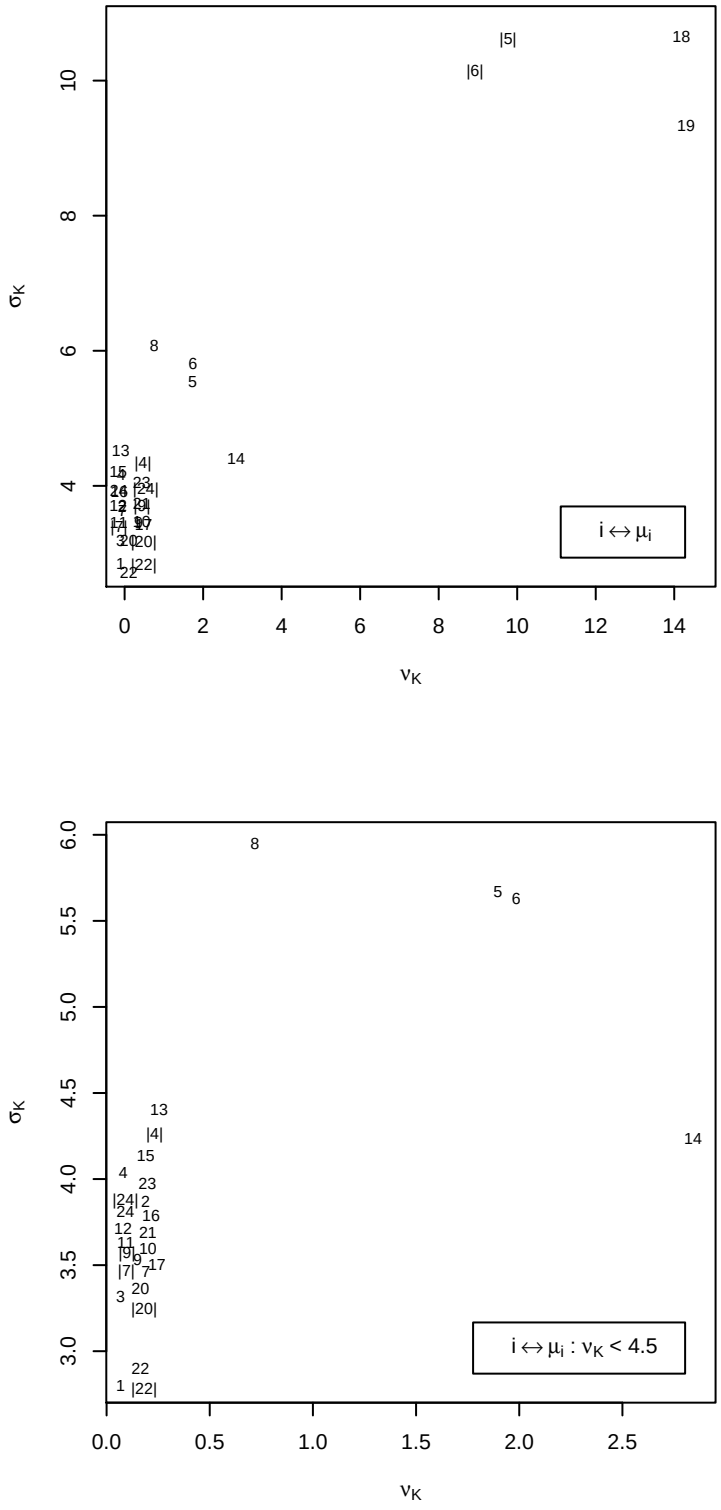


Figure 6.4: μ-GREEDY Stopwords Included Not Discounted: σ_K vs ν_K

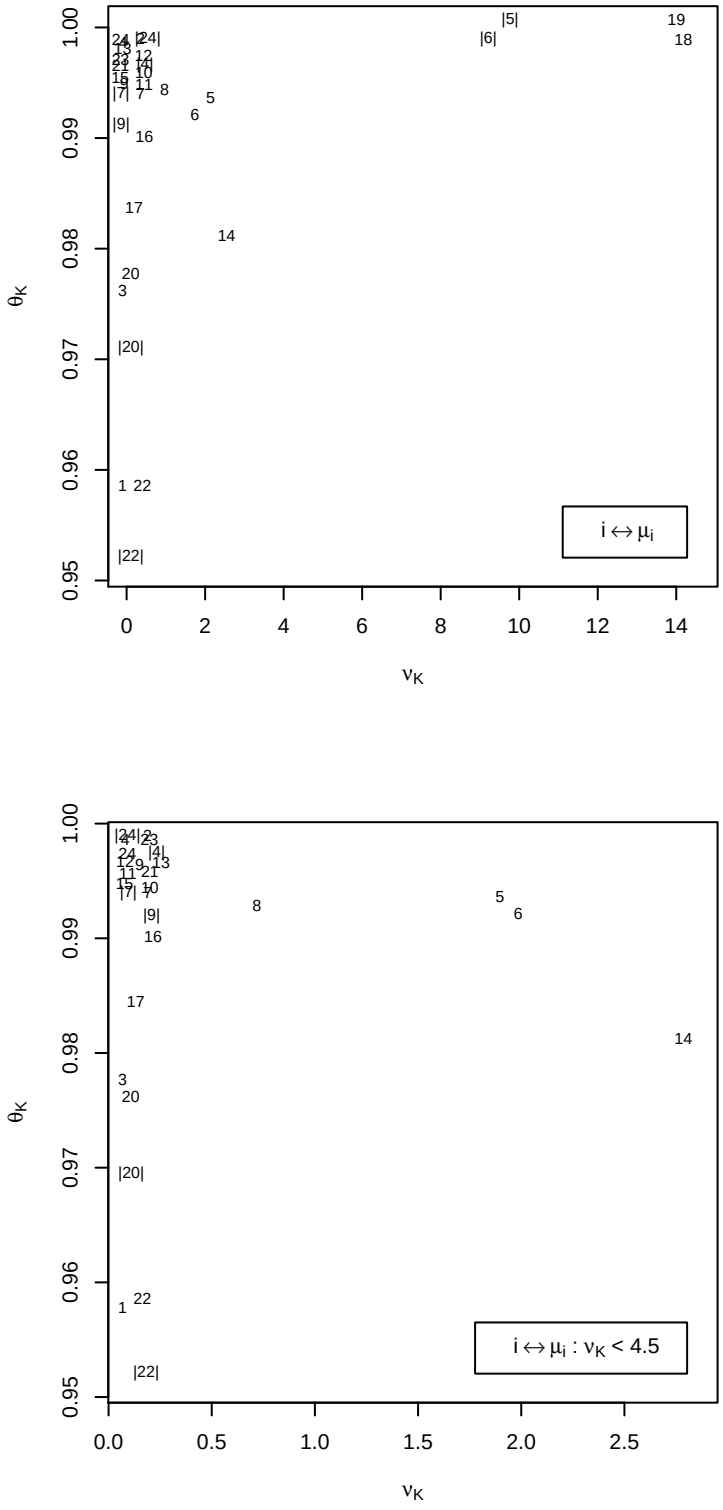


Figure 6.5: μ-GREEDY Stopwords Included Not Discounted: θ_K vs ν_K

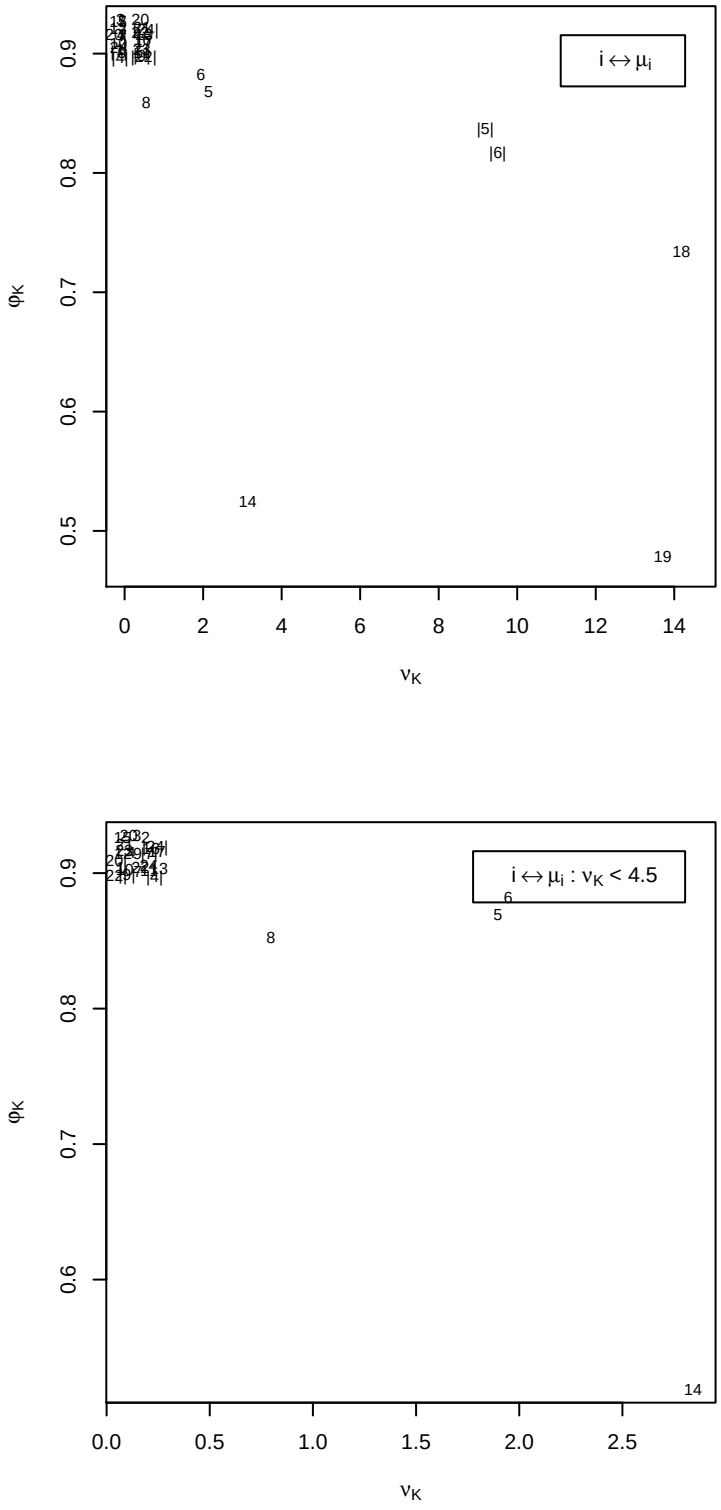


Figure 6.6: μ-GREEDY Stopwords Included Not Discounted: φ_K vs ν_K

6.3 Lexicographic Greedy Algorithms

We have seen that while using μ_8 in μ-RANKING in selects the feature set with the maximum average Hamming distance, it also selects a large number of noise features. By contrast, the use of μ_4 with μ-RANKING is in concert with the Feature Monotonicity Principle, and does not have this problem. In this section we investigate an approach for using this observation to develop a set of extensions the lexicographic Hamming distance ϕ_α and then using these extensions to implement a class of greedy algorithms.

If $u, v \in \mathbb{B}^n$ then we shall let

$$\hat{d_y}(u, v) = \hat{d}_y(u[S], v[S]) = \sum_{j \in S} u_j - v_j$$

where $y = \chi^S$ for some $S \subseteq V$. While we will view $\hat{d_y}(u, v)$ as a measure of the "distance" between the vectors u and v, it clearly is not a metric, e.g. it may be negative, and owing to this fact will refer it as the *Hamming difference*. Just as the feature ranking function μ_8 is based on the Hamming distance, the feature ranking function μ_4 can be seen to be based on the Hamming difference. Now, recalling that the vector $\mathbf{h}(y)$ is based on the Hamming distance, and that the lexicographic Hamming distance, $\phi_\alpha(y)$ is a function of $\mathbf{h}(y)$, we now show how to define a variant of $\mathbf{h}(y)$ that is based on the Hamming difference, and show how it can be used to define a variant of $\phi_\alpha(y)$.

If again $y = \chi^S$ for some $S \subseteq V$, then we define the vector $\hat{\mathbf{h}}(y) = \left[\hat{\mathbf{h}}^+(y), \hat{\mathbf{h}}^-(y)\right]$ where

$$\hat{\mathbf{h}}^+(y) = \left[\hat{h}_o^+(y), \hat{h}_1^+(y), \cdots, \hat{h}_n^+(y)\right]$$

and

$$\hat{h}_k^+(y) = \left| \left\{(u, v) \in T \times F \colon \hat{d}_y(u, v) = k\right\} \right|$$

for $k = 0, 1, \ldots, n$, and

$$\hat{\mathbf{h}}^-(y) = \left[\hat{h}_1^-(y), \cdots, \hat{h}_n^-(y)\right]$$

and

$$\hat{h}_k^-(y) = \Big|\, \{(u,v) \in T \times F \colon \hat{d}_y^-(u,v) = -k\} \,\Big|$$

for $k = 1, \ldots, n$. The components $\hat{h}_k^+(y)$ for $k = 0, 1, \ldots, n$ are the number of pairs in $T[S] \times F[S]$ that are exactly at Hamming difference k with $\hat{h}_k^+(y) = 0$ for any $k > |S|$, and the components $\hat{h}_k^-(y)$ for $k = 1, \ldots, n$ are the number of pairs in $T[S] \times F[S]$ that are exactly at Hamming difference $-k$. Now, using $\widehat{\mathbf{h}}$ we define

$$\phi_\alpha^\Delta(y) = \sum_{k=0}^{n} \big(\hat{h}_k^+(y) - \hat{h}_k^-(y)\big)\big(1 - \alpha^k\big).$$

Just as $\phi_\alpha(y)$ can be viewed as a weighted version of the average Hamming distance, $\phi_\alpha^\Delta(y)$ can be viewed as a weighted version of the average Hamming difference. It is also possible to develop variants to the function θ and ρ. Noting that $h_0(y)$ is the number of relevant-irrelevant pairs that are *not* separated by a set $S \subseteq V$ we can write

$$\theta(y) = \sum_{k=1}^{n} h_k(y)$$

and will define

$$\theta^\Delta(y) = \sum_{k=1}^{n} \hat{h}_k^+(y) + \sum_{k=1}^{n} \hat{h}_k^-(y).$$

We note that we can write

$$\rho(y) = \min\{k \colon 1 \le k \le K \text{ and } h_k(y) > 0\}$$

and will define

$$\rho^\Delta(y) = \min\{k \colon 1 \le k \le K \text{ and } h_k^+(y) > 0\}.$$

In order to use $\phi_\alpha^\Delta(y)$ or other functions based on $\mathbf{h}(y)$ or $\hat{\mathbf{h}}(y)$ in a greedy algorithm we now introduce two matrices. Following [13], we shall let $A_{(T,F)} = [a_{rj}]$ denote the $|T||F| \times n$ matrix in which $a_{rj} = |u_j - v_j|$ where $u = A[r_T, \cdot]$ and $v = A[r_F, \cdot]$ for all $r = (r_T, r_F) \in W_T \times W_F$ and all $j \in V$. That is, the rows of $A_{(T,F)}$ are constructed by computing the absolute value of the difference of each $u \in T$ and $v \in F$. We will refer

to this matrix as the *Hamming distance matrix*.

In addition, we will let $\hat{A}_{(T,F)} = [a_{rj}]$ denote the $|T||F| \times n$ matrix in which $a_{rj} = u_j - v_j$ where $u = A[r_T, \cdot]$ and $v = A[r_F, \cdot]$ for all $r = (r_T, r_F) \in W_T \times W_F$ and all $j \in V$, i.e.

$$\hat{a}_{rj} = \begin{cases} 1 & \text{if } u_j = 1 \text{ and } v_j = 0, \\ 0 & \text{if } u_j = v_j, \text{ and} \\ -1 & \text{if } u_j = 0 \text{ and } v_j = 1, \end{cases}$$

and $A_{(T,F)} = |\hat{A}_{(T,F)}|$. We will refer to this matrix as the *Hamming difference matrix*.

Informally, $A_{(T,F)}$ supports a model of the differences between relevant and irrelevant documents that is based on μ_8, while the model employed by $\hat{A}_{(T,F)}$ is based on μ_4. In fact, if $j \in V$, a_j is a column in $A_{(T,F)}$, $\hat{a}_j$ is a column in $\hat{A}_{(T,F)}$, and $\boxplus_j \in \boxplus$ then

$$\mu_8(\boxplus_j) = \frac{1}{|T||F|} \sum_{i=1}^{|T||F|} a_{i,j} \quad \text{and} \quad \mu_4(\boxplus_j) = \frac{1}{|T||F|} \sum_{i=1}^{|T||F|} \hat{a}_{i,j}.$$

However, since both $A_{(T,F)}$ and $\hat{A}_{(T,F)}$ contain the distance between each relevant-irrelevant document pair, they can be used to calculate much more sophisticated distance measures $\phi_\alpha(y)$ and $\phi_\alpha^\Delta(y)$ that are based the vector $\mathbf{h}(y)$ and $\hat{\mathbf{h}}(y)$ respectively and can be used to update these vectors as the set S changes.

If $j \in V$ and $S = \{j\}$, and $y = \chi^S$, then the column vectors $A_{(T,F)}[\cdot, \{j\}]$ and $\hat{A}_{(T,F)}[\cdot, \{j\}]$ respectively contain all of the information required to compute $\mathbf{h}(y)$ and $\hat{\mathbf{h}}(y)$. Further, if $S' = \{j, j'\}$ and $y' = \chi^{S'}$ for $j' \in V$, then the column vectors $A_{(T,F)}[\cdot, \{j\}] + A_{(T,F)}[\cdot, \{j'\}]$ and $\hat{A}_{(T,F)}[\cdot, \{j\}] + \hat{A}_{(T,F)}[\cdot, \{j'\}]$ contain all of the information required to compute $\mathbf{h}(y')$ and $\hat{\mathbf{h}}(y')$. This observation can be used in an obvious way to implement a class of greedy feature selection algorithms as shown below.

h_k-GREEDY

Input: The set V, a Hamming distance matrix A, a function ϕ_α, and an integer K.

Step 1: Set $k := 0$, $\lambda = 0^{|T|+|F|}$, $z = 0$.

Step 2: Set $s_k := j^*$, and $z := \phi_\alpha(\lambda + a_j^*)$, and $\lambda := \lambda + a_j^*$, where $j^* := \text{argmax}_{j \in V \setminus S} \ \phi_\alpha\big(\lambda + a_j\big)$, and a_j, a_{j^*} are columns in A.

Step 3: If $k := K$ then set $(s_k) := s_1, s_2, \cdots, s_K$ and goto **Output**, otherwise set $k := k + 1$ and goto **Step 2**.

Output: Output (s_k).

Notice there is a slight abuse of notation in our definition of h_k-GREEDY in that we write ϕ_α as function of a column vector rather than a subset of V. Also notice that h_k-GREEDY can clearly be adapted to the situation where A is a Hamming difference matrix and a function such as ϕ_α^Δ, θ^Δ, or ρ^Δ is used. It should also be mentioned that the remarks regarding the stopping criteria of μ-GREEDY also apply to h_k-GREEDY, i.e. h_k-GREEDY always selects K features regardless of the value of the objective function.

6.3.1 h_k-GREEDY Results

In this section, we use the methodology presented in §2 to evaluate the relative performance of h_k-GREEDY for each of the feature ranking functions discussed in §6.3. The detailed non-discounted results are provided in Appendix R and the detailed discounted results are provided in Appendix S. The non-discounted results are summarized in Table 6.5 and the discounted results are summarized in Table 6.6. In addition, Figure 6.7, Figure 6.8, and Figure 6.9 depict the relationship between σ_K, θ_K and φ_K, and ν_K. Based on this data we make the following observations.

- The algorithms based on $\hat{\mathbf{h}}(y)$ were superior to those based on $\mathbf{h}(y)$ on all criteria and all rankings except for $\widehat{\theta}$ and $\widehat{\sigma}$.
- The superiority of the algorithms based on $\mathbf{h}(y)$ on the $\widehat{\theta}$ and $\widehat{\sigma}$ involved the retrieval of a large number of noise features.

- It is interesting to note that the number of noise features selected as well as the separation increased with α for both algorithms based on $\mathbf{h}(y)$ and those based on $\hat{\mathbf{h}}(y)$.

- In ϕ_{α}^{Δ}-GREEDY, varying α provides a means of selecting points on the σ_K vs ν_K or $\widehat{\sigma}$ vs $\widehat{\nu}$ "efficient frontier".

These results serve to validate the approach of extending the use of Hamming difference based functions from their use in ranking algorithms to greedy algorithms based on $\hat{\mathbf{h}}(y)$.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
$\phi^{\Delta}_{\alpha=0.99}$	$\phi^{\Delta}_{\alpha=\epsilon}$	$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=\epsilon}$	0.01	$\phi_{\alpha=\epsilon}$	0.99	$\phi_{\alpha=0.99}$	0.54	$\phi^{\Delta}_{\alpha=0.25}$	0.91
ρ^{Δ}	$\phi^{\Delta}_{\alpha=0.25}$	$\phi^{\Delta}_{\alpha=0.75}$	$\phi^{\Delta}_{\alpha=0.25}$	0.02	$\phi_{\alpha=0.25}$	0.99	$\phi_{\alpha=0.5}$	0.53	$\phi^{\Delta}_{\alpha=0.5}$	0.90
$\phi^{\Delta}_{\alpha=0.75}$	θ^{Δ}	$\phi^{\Delta}_{\alpha=0.25}$	θ^{Δ}	0.02	$\phi_{\alpha=0.5}$	0.99	$\phi_{\alpha=0.75}$	0.53	$\phi^{\Delta}_{\alpha=\epsilon}$	0.88
$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=0.99}$	$\phi^{\Delta}_{\alpha=0.5}$	0.03	$\phi_{\alpha=0.75}$	0.99	ρ	0.53	$\phi^{\Delta}_{\alpha=0.75}$	0.88
$\phi^{\Delta}_{\alpha=0.25}$	$\phi^{\Delta}_{\alpha=0.75}$	ρ^{Δ}	$\phi^{\Delta}_{\alpha=0.75}$	0.06	$\phi_{\alpha=0.99}$	0.99	$\phi_{\alpha=0.25}$	0.52	θ^{Δ}	0.85
θ^{Δ}	ρ^{Δ}	$\phi^{\Delta}_{\alpha=\epsilon}$	ρ^{Δ}	0.13	θ	0.99	θ	0.51	$\phi^{\Delta}_{\alpha=0.99}$	0.84
$\phi^{\Delta}_{\alpha=\epsilon}$	$\phi^{\Delta}_{\alpha=0.99}$	θ^{Δ}	$\phi^{\Delta}_{\alpha=0.99}$	0.14	ρ	0.99	$\phi_{\alpha=\epsilon}$	0.47	ρ^{Δ}	0.84
$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	0.43	$\phi^{\Delta}_{\alpha=0.25}$	0.98	$\phi^{\Delta}_{\alpha=0.99}$	0.44	$\phi_{\alpha=\epsilon}$	0.83
θ	θ	$\phi_{\alpha=0.25}$	θ	0.46	$\phi^{\Delta}_{\alpha=0.5}$	0.98	ρ^{Δ}	0.41	$\phi_{\alpha=0.25}$	0.82
$\phi_{\alpha=0.25}$	$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.25}$	0.48	$\phi^{\Delta}_{\alpha=0.75}$	0.98	$\phi^{\Delta}_{\alpha=0.75}$	0.39	$\phi_{\alpha=0.5}$	0.81
$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.25}$	θ	$\phi_{\alpha=0.5}$	0.48	$\phi^{\Delta}_{\alpha=0.99}$	0.98	$\phi^{\Delta}_{\alpha=0.5}$	0.35	θ	0.81
$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	0.49	ρ^{Δ}	0.98	$\phi^{\Delta}_{\alpha=0.25}$	0.31	$\phi_{\alpha=0.75}$	0.79
ρ	ρ	$\phi_{\alpha=0.99}$	ρ	0.49	$\phi^{\Delta}_{\alpha=\epsilon}$	0.86	θ^{Δ}	0.22	$\phi_{\alpha=0.99}$	0.78
$\phi_{\alpha=0.99}$	$\phi_{\alpha=0.99}$	ρ	$\phi_{\alpha=0.99}$	0.50	θ^{Δ}	0.82	$\phi^{\Delta}_{\alpha=\epsilon}$	0.19	ρ	0.78

Table 6.5: h_k-GREEDY Not Discounted Stopwords Included

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}'$		$\widehat{\sigma}'$		$\bar{\varphi}$	
$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=\epsilon}$	$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=\epsilon}$	0.01	$\phi_{\alpha=\epsilon}$	0.98	$\phi^{\Delta}_{\alpha=0.99}$	0.37	$\phi^{\Delta}_{\alpha=0.25}$	0.91
$\phi^{\Delta}_{\alpha=0.75}$	$\phi^{\Delta}_{\alpha=0.25}$	$\phi^{\Delta}_{\alpha=0.25}$	$\phi^{\Delta}_{\alpha=0.25}$	0.02	$\phi_{\alpha=0.25}$	0.98	$\phi^{\Delta}_{\alpha=0.75}$	0.36	$\phi^{\Delta}_{\alpha=0.5}$	0.90
$\phi^{\Delta}_{\alpha=0.25}$	θ^{Δ}	$\phi^{\Delta}_{\alpha=0.75}$	θ^{Δ}	0.02	$\phi_{\alpha=0.5}$	0.98	ρ^{Δ}	0.35	$\phi^{\Delta}_{\alpha=\epsilon}$	0.88
ρ^{Δ}	$\phi^{\Delta}_{\alpha=0.5}$	$\phi^{\Delta}_{\alpha=0.99}$	$\phi^{\Delta}_{\alpha=0.5}$	0.03	$\phi_{\alpha=0.75}$	0.98	$\phi^{\Delta}_{\alpha=0.5}$	0.34	$\phi^{\Delta}_{\alpha=0.75}$	0.88
$\phi^{\Delta}_{\alpha=0.99}$	$\phi^{\Delta}_{\alpha=0.75}$	ρ^{Δ}	$\phi^{\Delta}_{\alpha=0.75}$	0.06	$\phi_{\alpha=0.99}$	0.98	$\phi^{\Delta}_{\alpha=0.25}$	0.31	$\phi^{\Delta}_{\alpha=0.99}$	0.86
θ^{Δ}	ρ^{Δ}	$\phi^{\Delta}_{\alpha=\epsilon}$	ρ^{Δ}	0.13	$\phi^{\Delta}_{\alpha=0.25}$	0.98	$\phi_{\alpha=0.25}$	0.29	ρ^{Δ}	0.86
$\phi^{\Delta}_{\alpha=\epsilon}$	$\phi^{\Delta}_{\alpha=0.99}$	θ^{Δ}	$\phi^{\Delta}_{\alpha=0.99}$	0.14	$\phi^{\Delta}_{\alpha=0.5}$	0.98	$\phi_{\alpha=0.5}$	0.29	θ^{Δ}	0.85
$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	$\phi_{\alpha=\epsilon}$	0.43	$\phi^{\Delta}_{\alpha=0.75}$	0.98	$\phi_{\alpha=0.75}$	0.29	$\phi_{\alpha=\epsilon}$	0.84
θ	θ	θ	θ	0.46	θ	0.98	$\phi_{\alpha=0.99}$	0.29	$\phi_{\alpha=0.25}$	0.83
$\phi_{\alpha=0.25}$	$\phi_{\alpha=0.25}$	$\phi_{\alpha=0.25}$	$\phi_{\alpha=0.25}$	0.48	ρ	0.98	θ	0.29	$\phi_{\alpha=0.5}$	0.83
$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.5}$	$\phi_{\alpha=0.5}$	0.48	$\phi^{\Delta}_{\alpha=0.99}$	0.97	ρ	0.29	θ	0.83
$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	$\phi_{\alpha=0.75}$	0.49	ρ^{Δ}	0.97	$\phi_{\alpha=\epsilon}$	0.28	$\phi_{\alpha=0.75}$	0.82
ρ	ρ	ρ	ρ	0.49	$\phi^{\Delta}_{\alpha=\epsilon}$	0.86	θ^{Δ}	0.22	ρ	0.82
$\phi_{\alpha=0.99}$	$\phi_{\alpha=0.99}$	$\phi_{\alpha=0.99}$	$\phi_{\alpha=0.99}$	0.50	θ^{Δ}	0.82	$\phi^{\Delta}_{\alpha=\epsilon}$	0.19	$\phi_{\alpha=0.99}$	0.81

Table 6.6: h_k-GREEDY Discounted Stopwords Included

6.4 Ensemble Methods

In §4.1 and §4.2 we concluded that noise and separation, noise and size (as measured by θ), and noise and robustness are all pairs of *competing criteria*. This conclusion suggests that any optimization problem constructed to perform feature selection on textual data, should be formulated as a multicriteria optimization problem. In §4.7, we observed that many Boolean feature ranking functions are ratios of separating functions to noise functions and therefore are implicitly pursuing a strategy to address this issue. The following result shows, however, that this strategy may yield a solution that is not Pareto optimal.

Proposition 6.1. *If $\psi \in \Psi$, $\eta \in \aleph$, and $j \in V$, then*

$$\big(V, \langle \psi(x_j, y_j), \eta(x_j, y_j) \rangle, \mathbb{R}^2\big) \,/\, \frac{\psi(x_j, y_j)}{\eta(x_j, y_j)} \,/\, \big(\mathbb{R}, \leq\big) \tag{6.11}$$

is not efficient.

Proof. Note that since (6.11) creates a total order, it identifies a unique optimum feature which we denote by $j \in V$. If $\psi(x_j, y_j) \geq \psi(x_i, y_i)$ (or $\eta(x_j, y_j) \leq \eta(x_i, y_i)$) and $\eta(x_j, y_j) < \eta(x_i, y_i)$ (or $\psi(x_j, y_j) > \psi(x_i, y_i)$) for each $i \in V$, $i \neq j$, then by definition feature j is efficient and is correctly identified by (6.11). However, it is possible for $j \in V$ to be the unique optimum for (6.11), but for there to exist an $i \in V$ such that $\psi(x_j, y_j) \geq \psi(x_i, y_i)$ (or $\eta(x_j, y_j) \leq \eta(x_i, y_i)$) and $\eta(x_i, y_i) \leq \eta(x_j, y_j)$ (or $\psi(x_i, y_i) \geq \eta(x_i, y_i)$), in which case the set of efficient features contains both i and j. ■

As an example of the case where the efficient set contains multiple features, let $V = \{1, 2, 3\}$ with $\big(\psi(x_1, y_1), \eta(x_1, y_1)\big) = (0.8, 0.1)$, $\big(\psi(x_2, y_2), \eta(x_2, y_2)\big) = (0.9, 0.2)$, and $\big(\psi(x_3, y_3), \eta(x_3, y_3)\big) = (0.7, 0.3)$. Since $0.8/0.1 > 0.9/0.2 > 0.7/0.3$, feature 1 will be the unique optimum found by (6.11), but the efficient set contains both feature 1 and feature 2. So, if there is a unique efficient feature, then (6.11) will identify it, but if there are multiple efficient features, then it "artificially" ranks one above the other.

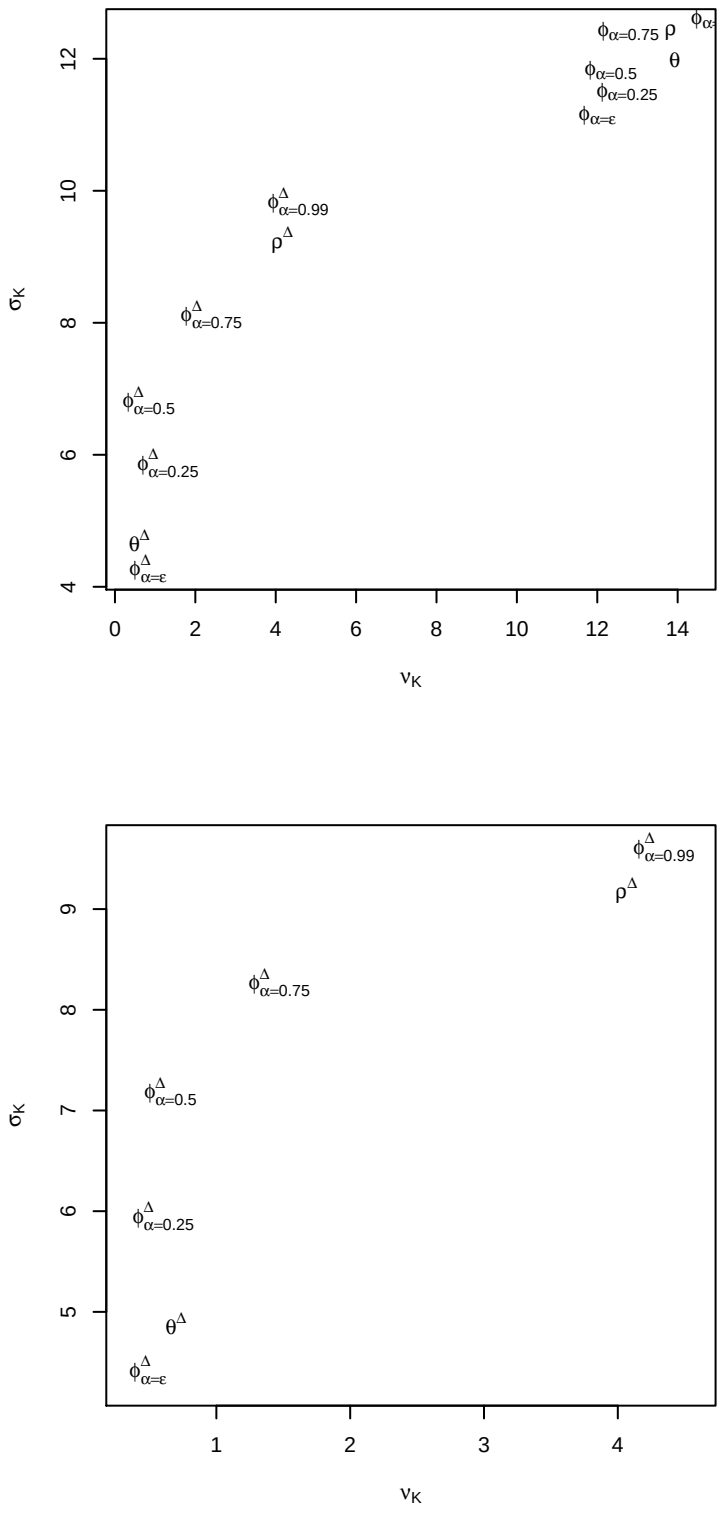


Figure 6.7: h_k-GREEDY Stopwords Included Not Discounted: σ_K vs ν_K

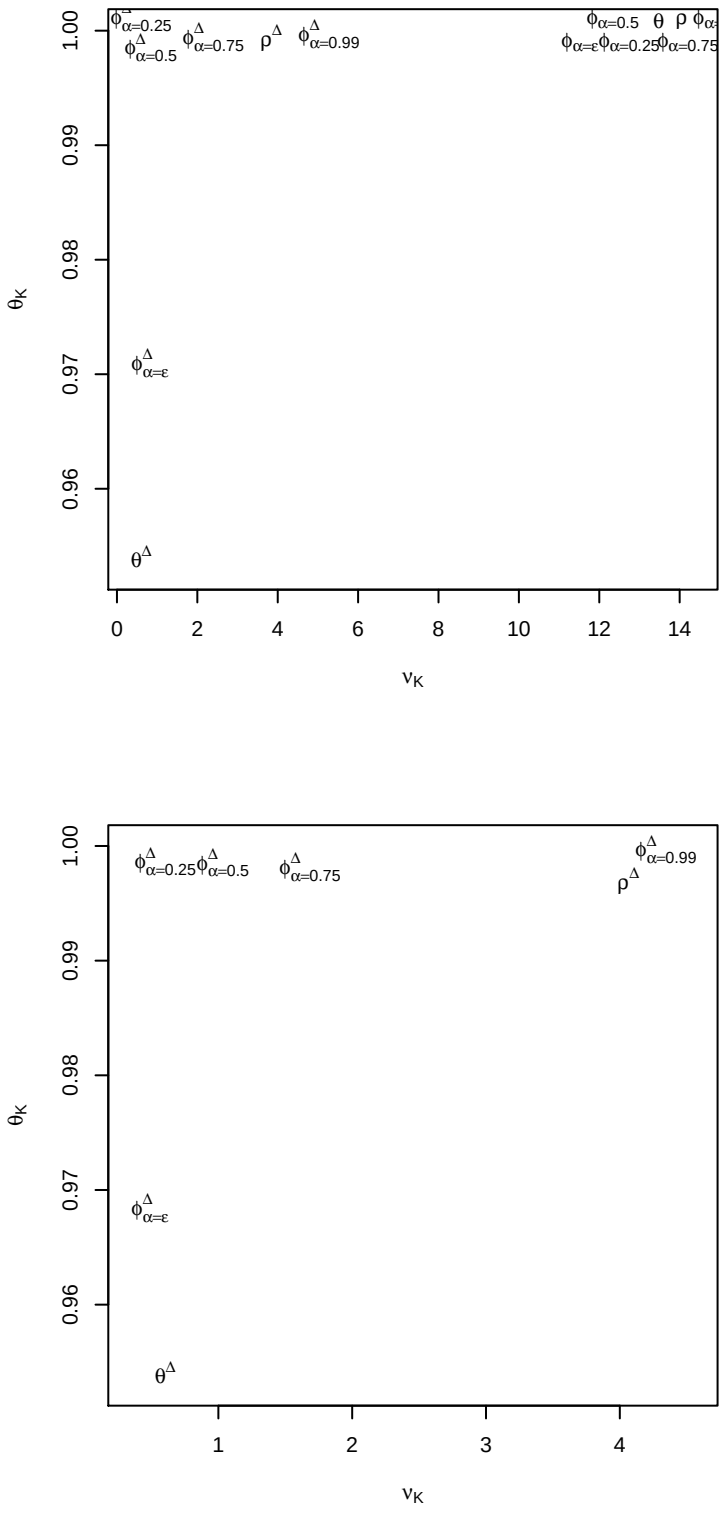


Figure 6.8: h_k-GREEDY Stopwords Included Not Discounted: θ_K vs ν_K

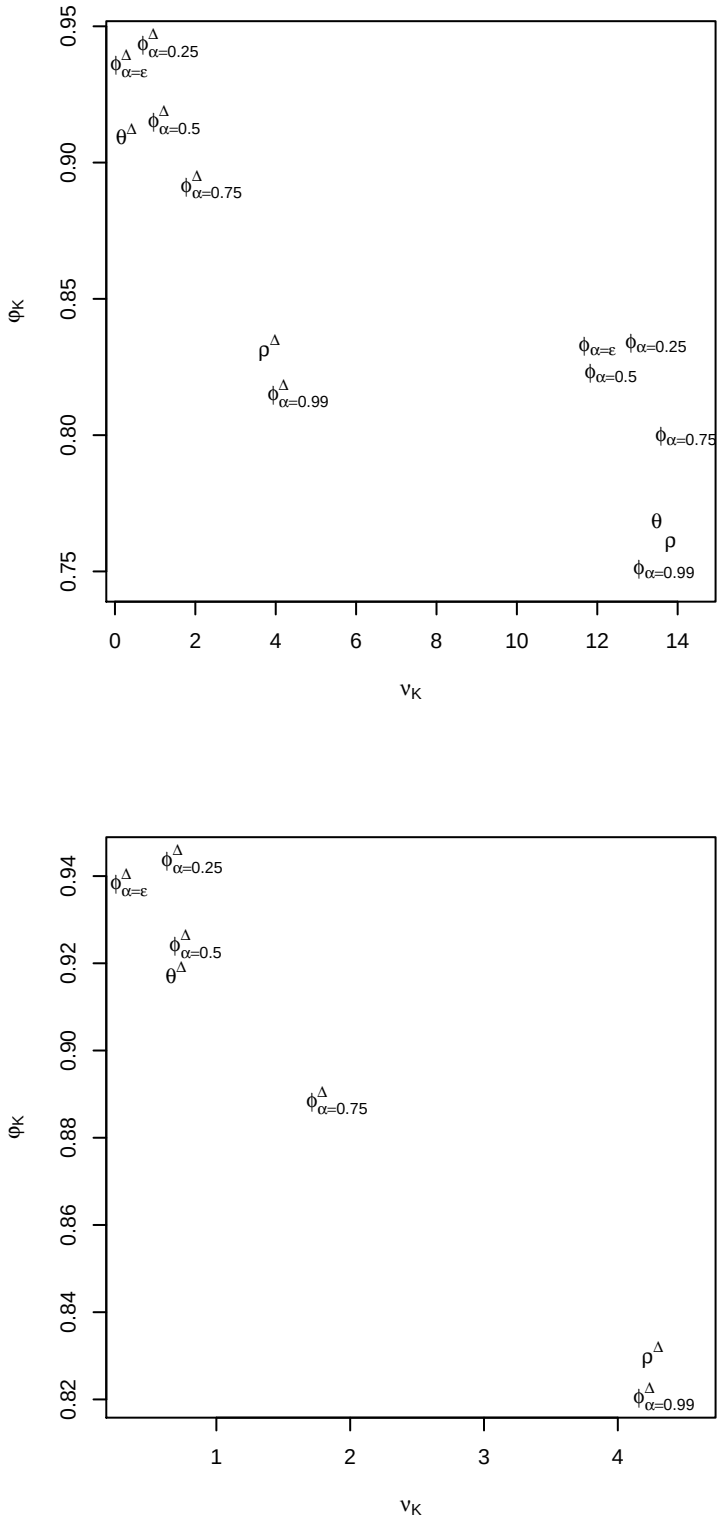


Figure 6.9: h_k-GREEDY Stopwords Included Not Discounted: φ_K vs ν_K

By contrast lex-max-min[1], as shown in Table 6.7, finds that the efficient set consists of both feature 1 and feature 2.

Feature	ψ	$1-\eta$	1^{rst} Worst	2^{nd} Worst
1	0.8	0.9	0.8	0.9
2	0.9	0.8	0.8	0.9
3	0.7	0.7	0.7	0.7

Table 6.7: Lex-Max-Min Solution

In order to address the fact that feature selection involves multiple competing criteria we will consider two families of feature selection algorithms; one based on lex-max-min and the other based on the weighted sum method. Let $\varepsilon_i \subseteq \mathcal{M} \cup \Psi \cup \aleph$ for some index i. The feature selection algorithm ℓ_i-ENSEMBLE uses a similar implementation of lex-max-min to that defined in §2.14, with the criteria in ε_i, to rank V. The feature selection algorithm α_i-ENSEMBLE ranks V by the average of the criteria in ε_i. Since each subset of $\varepsilon_i \subseteq \mathcal{M} \cup \Psi \cup \aleph$ defines a different ranking algorithm we have defined two classes of ranking algorithms, and owing to the relationship between the sets of criteria and the Boolean feature ranking functions, we will collectively refer to these algorithms as μ-ENSEMBLE algorithms.

We will limit ourselves to considering the performance of implementations of μ-ENSEMBLE algorithms based on the following sets of criteria.

- $\varepsilon_1 = \{\ \mu_4,\ \eta_1\ \}$
- $\varepsilon_2 = \{\ \mu_4,\ \eta_2\ \}$
- $\varepsilon_3 = \{\ \mu_4,\ \eta_1,\ \eta_2^{(\varrho)}\ \}$
- $\varepsilon_4 = \{\ \mu_9,\ \mu_{11},\ \mu_{12},\ \mu_{16},\ \mu_{17},\ \mu_{24}\ \}$
- $\varepsilon_5 = \{\ \mu_9,\ |\mu_9|,\ \mu_{10},\ \mu_{11},\ \mu_{12},\ \mu_{16},\ \mu_{17},\ \mu_{21},\ \mu_{23},\ \mu_{24}\ , |\mu_{24}|\ \}$

[1]Note that we use $1-\eta$ since smaller values are preferred and all values of η are in [0,1].

We recall that η_1 was given in (4.3) and corresponds to den(μ_9), η_2 was given in (4.9) and corresponds to den(μ_{23}). We now introduce

$$\eta_2^{(\varrho)} = \varrho x(1-x) + (1-\varrho)y(1-y)$$

which is a variant of η_2 weighted by the collection weight factors ϱ and $1-\varrho$. The sets ℓ_1, ℓ_2, and ℓ_3, combine the prototypical separating function μ_4 with a different collection of noise functions. These sets will allow us to compare the traditional approach of using the ratio of a separating function to a noise function to rank features with the use of ensemble algorithms where the criteria include separating functions and noise functions and ranking is accomplished either using lex-max-min or the weighted sum method.

6.4.1 μ-ENSEMBLE Results

In this section, we use the methodology presented in §2 to evaluate the relative performance of each variant of μ-ENSEMBLE that was introduced in §6.4. The detailed non-discounted results are provided in Appendix T and the detailed discounted results are provided in Appendix U. The non-discounted results are summarized in Table 6.8 and the discounted results are summarized in Table 6.9. In addition, Figure 6.10, Figure 6.11, and Figure 6.12 depict the relationship between σ_K, θ_K and φ_K, and ν_K. Based on this data we make the following observations.

- The performance of μ_9-RANKING and α_1-ENSEMBLE was similar on both the non-discounted and discounted measures, with μ_9-RANKING having slightly lower values of $\hat{\nu}$ and slightly larger values of φ, and α_1-ENSEMBLE having slightly larger values of $\hat{\theta}$ and $\hat{\sigma}$.

- The performance of μ_{24}-RANKING and α_2-ENSEMBLE was also similar, as was the performance of μ_9-RANKING, μ_{24}-RANKING, and α_3-ENSEMBLE.

- These observations about α_i-ENSEMBLE are interesting in that they suggest that if $\psi \in \Psi$ and $\eta \in \aleph$, it is possible for an implementation of μ-RANKING

using a *polynomial* function such as $(1/2)\big(1 - \eta(x,y) + \psi(x,y)\big)$ to achieve performance similar an implementation of μ-RANKING using the *rational* function $\psi(x,y)/\eta(x,y)$.

- As shown in the plots in Appendix T and Appendix U, the ℓ_i-ENSEMBLE algorithms, with the exception of ℓ_4-ENSEMBLE had noticeably higher variance on all measures than the corresponding α_i-ENSEMBLE algorithms.

- Of the ensembles ε_4 and ε_5 which contained feature ranking functions from $\mathcal{M}$, the performance of ℓ_4-ENSEMBLE was clearly better than that of ℓ_5-ENSEMBLE, while the performance of α_4-ENSEMBLE and α_5-ENSEMBLE was similar.

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}$		$\widehat{\sigma}$		$\bar{\varphi}$	
α_1	α_2	ℓ_4	α_2	0.01	ℓ_1	0.97	ℓ_1	0.43	α_5	0.89
ℓ_4	α_3	α_1	α_3	0.01	ℓ_3	0.97	ℓ_3	0.43	ℓ_4	0.88
α_2	α_4	α_2	α_4	0.01	ℓ_4	0.97	ℓ_2	0.42	α_4	0.88
α_3	α_5	α_3	α_5	0.01	α_1	0.97	α_1	0.39	α_1	0.87
α_4	α_1	α_4	ℓ_4	0.02	α_2	0.97	ℓ_4	0.38	α_2	0.87
α_5	ℓ_4	α_5	α_1	0.02	α_3	0.97	α_2	0.37	α_3	0.87
ℓ_2	ℓ_5	ℓ_2	ℓ_5	0.05	ℓ_2	0.96	α_3	0.37	ℓ_2	0.85
ℓ_5	ℓ_2	ℓ_1	ℓ_2	0.11	α_4	0.96	α_4	0.36	ℓ_5	0.85
ℓ_1	ℓ_1	ℓ_3	ℓ_1	0.13	α_5	0.96	α_5	0.35	ℓ_1	0.84
ℓ_3	ℓ_3	ℓ_5	ℓ_3	0.13	ℓ_5	0.95	ℓ_5	0.31	ℓ_3	0.84

Table 6.8: μ-ENSEMBLE Not Discounted Stopwords Included

Lex-Max-Min	Lex	Avg	$\widehat{\nu}$		$\widehat{\theta}'$		$\widehat{\sigma}'$		$\bar{\varphi}$	
ℓ_4	α_2	ℓ_4	α_2	0.01	ℓ_1	0.97	α_1	0.38	α_5	0.89
α_2	α_3	α_1	α_3	0.01	ℓ_3	0.97	ℓ_1	0.37	ℓ_4	0.88
α_3	α_5	α_2	α_4	0.01	ℓ_4	0.97	ℓ_2	0.37	α_2	0.88
α_1	α_4	α_3	α_5	0.01	α_1	0.97	ℓ_3	0.37	α_4	0.88
α_5	α_1	α_5	ℓ_4	0.02	α_2	0.97	ℓ_4	0.37	α_1	0.87
α_4	ℓ_4	α_4	α_1	0.02	α_3	0.97	α_2	0.37	α_3	0.87
ℓ_5	ℓ_5	ℓ_2	ℓ_5	0.05	ℓ_2	0.96	α_3	0.37	ℓ_1	0.86
ℓ_2	ℓ_2	ℓ_1	ℓ_2	0.11	α_4	0.96	α_4	0.35	ℓ_2	0.86
ℓ_1	ℓ_1	ℓ_3	ℓ_1	0.13	α_5	0.96	α_5	0.35	ℓ_3	0.86
ℓ_3	ℓ_3	ℓ_5	ℓ_3	0.13	ℓ_5	0.95	ℓ_5	0.29	ℓ_5	0.85

Table 6.9: μ-ENSEMBLE Discounted Stopwords Included

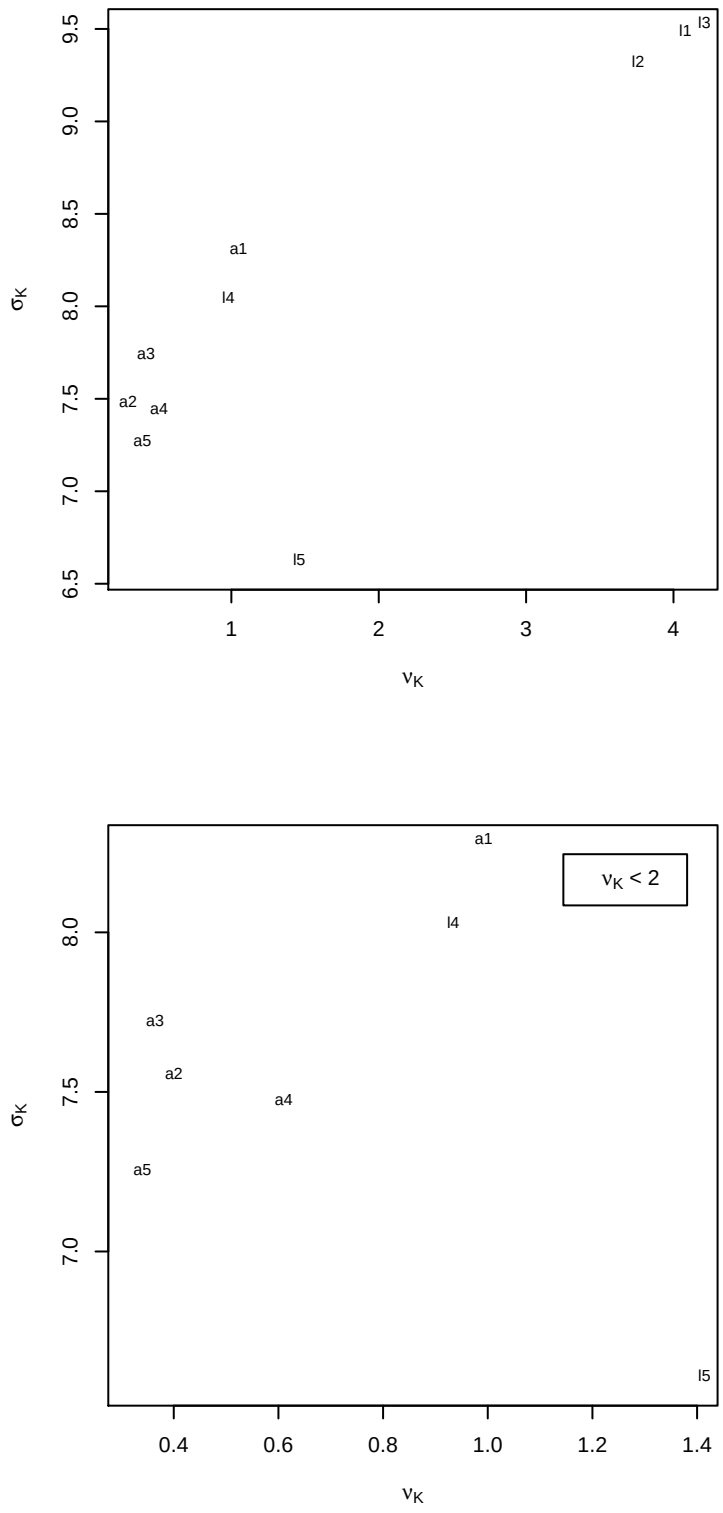


Figure 6.10: μ-ENSEMBLE Stopwords Included Not Discounted: σ_K vs ν_K

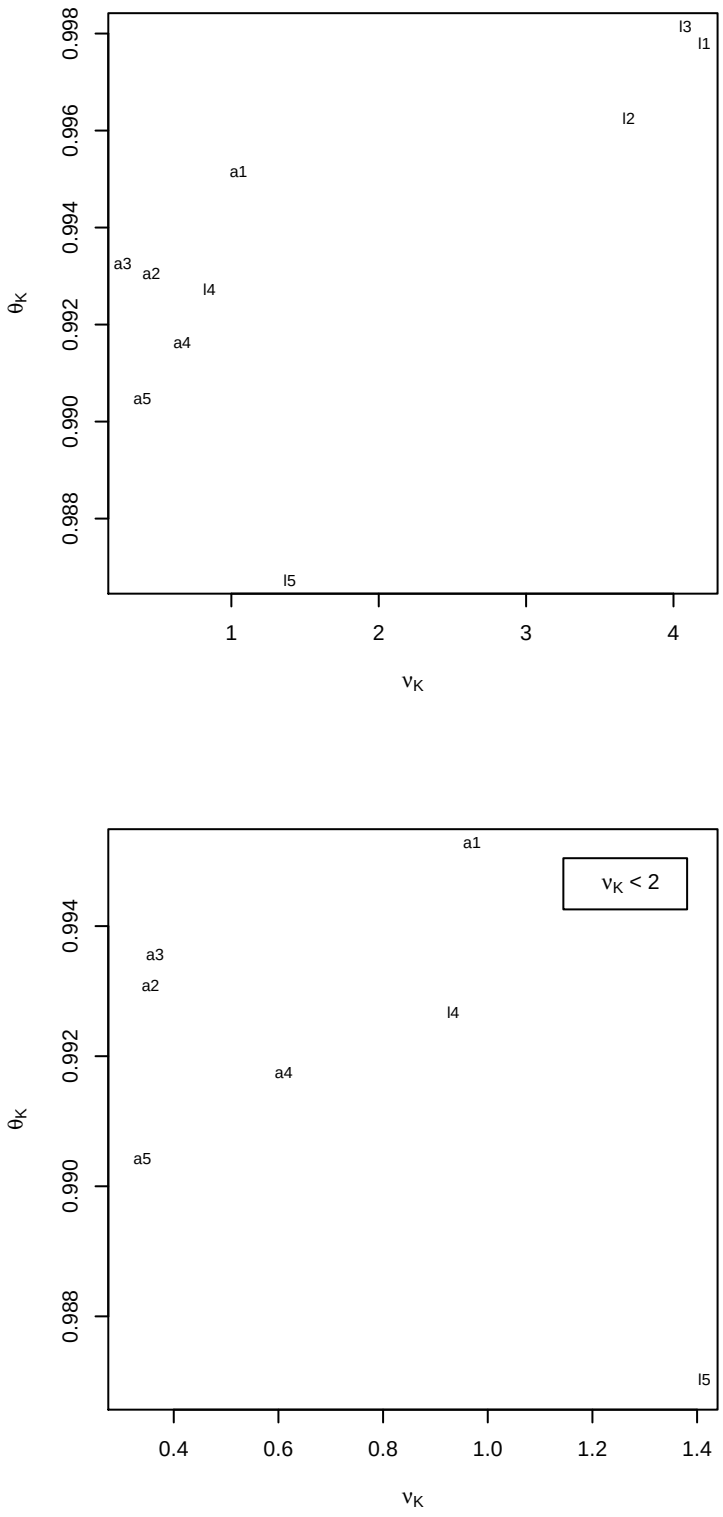


Figure 6.11: μ-ENSEMBLE Stopwords Included Not Discounted: θ_K vs ν_K

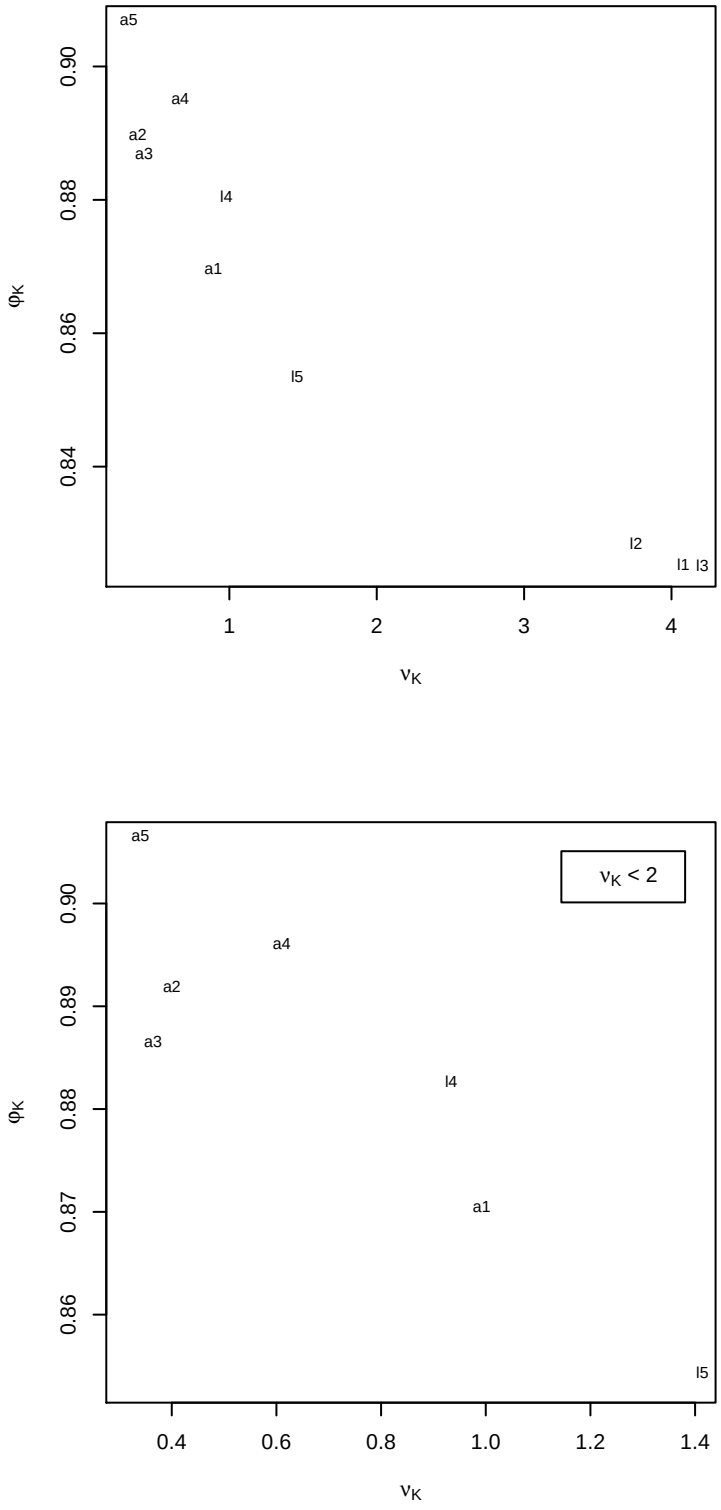


Figure 6.12: μ-ENSEMBLE Stopwords Included Not Discounted: φ_K vs ν_K

Chapter 7

Conclusions

In this dissertation we used a combination of empirical and theoretical tools to analyze feature selection selection algorithms. We now present an overview of the material.

- Typically the evaluation of feature selection algorithms involves comparison of the performance of one or more classification algorithms on a data set before and after feature selection has been performed. By contrast, to evaluate the performance of feature selection algorithms we introduced a set of criteria that measure properties of the feature set itself. Included in this set of criteria are direct measures of *separation* and *noise* as well as indirect measures of *size* and *quality*. When a measure is monotonically non-decreasing in the number of features selected, we used the AUC of the measure as a function of the number of features selected, and when this was not the case we simply computed the measure for different size feature sets and used the average of these values. In order to understand the performance of algorithms as a function of these multiple criteria we ranked them using the lex-max-min ordering.

- We described five models which form the basis of many Boolean feature ranking functions. Specifically, we described the probabilistic model, the term frequency model, the separation model, the information retrieval model, and the single feature classifier model, and presented a review of thirty-two functions in terms of these models. Typically such functions are viewed as mapping the four elements of a 2×2 contingency table to the set of real numbers, but we found it useful to view them to be functions in ROC space. Most of these functions were from the literature, but we also introduced a new function which we called the *rareness*. From the single feature classifier model and ROC curves we also developed a

definition of positive and negative features. Following this review we showed that several classic feature ranking functions are actually monotonic transformations of each other.

- Using the Reuters-21578 text categorization test collection, we ran a series of experiments and used our feature set evaluation methodology to assess the performance of the Boolean feature ranking functions that we studied. Not surprisingly, many of the classic feature ranking functions such as the correlation coefficient, the entropy, F_α, Fisher's linear discriminant and its variants, and the Gini criteria, exhibited some of the best performance. However, the results of these experiments also showed that paradoxically the functions which had the largest separation also selected the most noise, i.e. they indicated that *noise separates*.

- Motivated by the idea that noise separates, we performed several additional experiments to better understand this phenomenon. We used variants of our measures that were discounted when noise features were selected. Using these discounted measures we were able to confirm that stopwords, which are known noise features, are strong separators. We also noted with interest that the new *rareness* function performed especially well on the discounted measures. In addition, in experiments in which stopwords were eliminated *a priori*, we were able to conclude that there are many noise features that are not stopwords. We also noticed in plots of separation versus noise that many of the functions which appeared on what we loosely considered to be the "efficient frontier", and especially those which selected very little noise, and achieved moderate separation, were rational functions in which the numerator and denominator were each members of a small set of functions.

- Viewed in ROC space, both the numerators and denominators of these rational functions measured the distance from an arbitrary feature to either a special feature, or a special set of features. The numerators measure the distance from what we refer to as the *line of class noise*, provide a measure of separation, and are either the difference of the positive and negative feature frequencies or the square of this difference. The denominators measure the distance from what we

refer to as the *line of collection noise* or from the *point of strong collection noise* and provide a measure of noise. By assuming that textual data follows a Zipf distribution, we were also able to show the equivalence of two characterizations of a noise feature. We also observed what we refer to as the *monotone feature principle*, which states that in contrast to many non-textual data sets, in textual data sets, non-noise features that provide substantial separation are highly ranked positive features.

- In an effort to broaden our understanding of the set of *desirable* Boolean feature ranking functions we defined a set of linear axioms which we believe that such functions should satisfy. Then we considered the set of all functions that satisfy these axioms and can be represented as a linear combination of a set of finite basis functions. The basis we used allowed us to consider functions that are finite degree two-dimensional power series about the point $(1/2, 1/2)$ in ROC space. We were able to establish several properties of the set of functions which satisfy these axioms, including that it is convex, that it has an infinite number of vertices, and that it included functions from the literature. We also used the *lrs* software tools to identify the vertex sets for several low dimensional representations, and the GGobi software tools to visualize these vertex sets.

- We presented a natural extension of the axioms for Boolean feature ranking functions to real-valued feature ranking functions. In addition to those based on real-valued feature ranking functions, we also considered several other feature ranking algorithms including two greedy algorithms and one that employs an ensemble of ranking functions. By repeating many of our experiments with these new algorithms we were able to see that the conclusion that separation should be measured by the difference of the positive and negative feature frequencies extends beyond the Boolean model.

Future work related to this dissertation could include investigating the performance of new feature ranking functions that were identified as satisfying our axioms. In addition, we note that the approach we took to study of set of functions satisfying our

axioms is very general, and many variants could be pursued using a different set of axioms and different basis functions.

References

[1] H. Almuallim and T. G. Dietterich. Efficient algorithms for identifying relevant features. In *Proceedings of the Ninth Canadian Conference on Artificial Intelligence*, pages 38–45. Morgan Kaufmann, May 1992.

[2] Hussein Almuallim and Thomas G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69:279–305, 1994.

[3] Andrei Anghelescu, Endre Boros, David Lewis, Vladimir Menkov, David Neu, and Paul Kantor. Rutgers filtering work at trec 2002: Adaptive and batch. In E. M. Voorhees and Lori P. Buckland, editors, *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002)*. Department of Commerce, National Institute of Standards and Technology, 2002.

[4] D. Avis. lrs: A revised implementation of the reverse search vertex enumeration algorithm. In G. Kalai and G. Ziegler, editors, *Polytopes - Combinatorics and Computation*, pages 177–198. Birkhauser-Verlag, 2000.

[5] Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[6] Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2010. R package version 0.999375-44.

[7] Edwin Beckenbach and Richard Bellman. *An Introduction to Inequalities*. Random House, 1961.

[8] F. A. Behringer. On optimal decisions under complete ignorance: A new criterion stronger than both pareto and maxmin. *European Journal of Operational Research*, 1(5):295–306, September 1977.

[9] D. A. Bell and H. Wang. A formalism for relevance and its application in feature sebset selection. *Machine Learning*, 41:175–195, 2000.

[10] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 67:245–285, 1997.

[11] Endre Boros, Peter L. Hammer, Toshihide Ibaraki, Alexander Kogan, Eddy Mayoraz, and Ilya Muchnik. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12:292–306, 2000.

[12] Endre Boros, T. Horiyama, T. Ibaraki, K. Makino, M. Yagiura, P.B. Kantor, and D.J. Neu. Feature selection and information retrieval. RDLDL Seminar, April 2001.

[13] Endre Boros, Takashi Horiyama, Toshihide Ibaraki, Kazuhisa Makino, and Mutsunori Yagiura. Finding essential attributes from binary data. *Annals of Mathematics and Artificial Intelligence*, 39(3):223–257, 2003.

[14] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Chapman and HallCRC, 1984.

[15] Chris Buckley. The importance of proper weighting methods. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 349–352, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[16] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 28–36. Morgan Kaufmann, 1994.

[17] Bernard C. K. Choi. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, 148(11):1127–1132, 1998.

[18] C.K. Chow. Boolean functions realizable with single threshold devices. In *Proc. IRE,* **49**, pages 370–371, 1961.

[19] V. Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233235, 1979.

[20] Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49, 2008.

[21] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[22] Wikipedia contributors. Cross-validation. http://en.wikipedia.org/wiki/Cross-validation.

[23] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *Advances In Neural Information Processing Systems (NIPS 2003)*, volume 16. MIT Press, 2004.

[24] Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships in partially defined boolean functions. *Annals of Operations Research*, 16:299–326, 1988.

[25] David L. Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality, 2000.

[26] Edward R. Dougherty and Charles R. Giardina. *Mathematical methods for artificial intelligence and autonomous systems*. Prentice-Hall, Inc., 1988.

[27] Matthias Ehrgott. *Multicriteria Optimization*. Springer Berlin Heidelberg, second edition, 2010.

[28] Paul F. Evangelista, Mark J. Embrechts, and Boleslaw K. Szymanski. Taming the curse of dimensionality in kernels and novelty detection. In *In Ajith Abraham, Bernard de Baets, Mario Kppen, and Bertam Nickolay, editors, Applied Soft Computing Technologies: The Challenge of Complexity*, pages 431–444. Springer Verlag, 2006.

[29] B.S. Everitt. *The Analysis of Contingency Tables*. Halsted Press, 1977.

[30] Stefan Evert and Marco Baroni. zipfr: Word frequency distributions in r. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session, Prague, Czech Republic, 2007.

[31] Susana Eyheramendy, David D. Lewis, and David Madigan. On the naive bayes model for text classification. In C.M. Bishop and B.J. Frey, editors, *Proceedings of The Ninth International Workshop on Artificial Intelligence and Statistics*, pages 332–339, 2003.

[32] Susana Eyheramendy and David Madigan. A novel feature selection score for text categorization. In *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics*, pages 1–8. SIAM, April 2005.

[33] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. `http://home.comcast.net/~tom.fawcett/public_html/papers/`, March 2004.

[34] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[35] Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68:97–106, July 2007.

[36] Michael P. Fay and Michael A. Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.

[37] Ingo Feinerer. *tm: Text Mining Package*, 2009. R package version 0.5-1.

[38] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5), March 2008. `http://www.jstatsoft.org/v25/i05/`.

[39] Peter A. Flach and Shaomin Wu. Repairing concavities in roc curves. In *Proceeding of the 2003 UK Workshop on Computational Intelligence*, pages 38–44, 2003.

[40] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[41] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall PTR, 1992.

[42] David G. Garson. Dichotomous measures from statnotes: topics in multivariate analysis. Retrieved from `http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm`, August 2010.

[43] Isabelle Guyon and Andr Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[44] Isabelle Guyon, Asa Ben Hur, Steve Gunn, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT Press, 2004.

[45] A.B. Hammer, P.L. Hammer, and I. Muchnik. Logical analysis of chinese labor productivity patterns. *Annals of Operations Research*, 87(0):165–176, 1999.

[46] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, April 1982.

[47] R. J. Herrnstein, Donald H. Loveland, and Cynthia Cable. Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2(4):285–302, 1976.

[48] F.J. Banzaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

[49] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129. Morgan Kaufmann Publisher, 1994.

[50] K. Kira and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 129–134, Menlo Park, 1992. AAAI Press/The MIT Press.

[51] Donald E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley, 1969.

[52] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[53] M. Kostreva and W. Ogryczak. Linear optimization with multiple equitable criteria. *RAIRO Operations Research*, 33:275–297, 1999.

[54] D. D. Lewis. Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 212–217, San Mateo, California, 1992. Morgan Kaufmann.

[55] David D. Lewis. Reuters-21578 text categorization test collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[56] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[57] SMART Stopword List. ftp://ftp.cs.cornell.edu/pub/smart/english.stop.

[58] Irina I. Lozina. *Composite Boolean Separators for Data Analysis with Applications in Computed Tomography and Gene Expression Microarray Data*. PhD thesis, Rutgers University, May 2007.

[59] Saturnino Luz. Xml-encoded version of reuters-21578. `http://modnlp.berlios.de/reuters21578.html`.

[60] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[61] Christopher D. Manning and Hinrich Schütze. *Foundataions of Statistical Natural Language Processing*. MIT Press, 2002.

[62] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[63] Fabian Model, Pter Adorjn, Alexander Olek, and Christian Piepenbrock. Feature selection for dna methylation based cancer classification. *Bioinformatics*, 17, June 2001.

[64] Gottfried E. Noether. *Introduction to Statistics: The Nonparametric Way*. Springer-Verlag, 1991.

[65] Peter Norvig. How to write a spelling corrector. `http://norvig.com/spell-correct.html`.

[66] W. Ogryczak and Tomasz Sliwinski. On direct methods for lexicographic min-max optimization. In *Computational Science and Its Applications - ICCSA 2006*, volume 3982 of *Lecture Notes in Computer Science*, pages 802–811. Springer, 2006.

[67] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[68] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[69] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[70] Peter Seibel. *Practical Common Lisp*. Apress, 2005.

[71] L.S. Shapley and M. Shubik. A method for evaluating the distribution of power in a committee system. *Amer. Polit. Sci. Rev.*, 48:787–792, 1954.

[72] Rosie Shier. The mann-whitney u test. `http://mlsc.lboro.ac.uk/resources/statistics/Mannwhitney.pdf`, 2004.

[73] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Research and Development in Information Retrieval*, pages 21–29, 1996.

[74] Deborah F. Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. *Comput. Stat. Data Anal.*, 43(4):423–444, August 2003.

[75] Lynne Truss. *Eats, Shoot & Leaves: The Zero Tolerance Approach to Punctuation*. Gotham Books, 2004.

[76] E. M. Voorhees and Lori P. Buckland, editors. *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002)*. Department of Commerce, National Institute of Standards and Technology, 2002.

[77] Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li. A feature selection method based on improved fishers discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38, 2001.

[78] Edward J. Wegman and Jeffrey L. Solka. Short course on statistical data mining, part 10. Eighth U.S. Army Conference On Applied Statistics, October 2002. http://www.armyconference.org/.

[79] Wikipedia. Mann-whitney u — Wikipedia, the free encyclopedia, 2010. [Online; accessed 26-October-2010].

[80] R.O. Winder. Chow parameters in threshold logic. *J. ACM*, 18:265–289, 1971.

[81] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[82] Eunseog Youn and Myong K. Jeong. Class dependent feature scaling method using naive bayes classifier for text datamining. *Pattern Recognition Letters*, 30(5):477–485, April 2009.

Appendix A

Correlation Coefficient

This appendix contains the derivation of the correlation coefficient which is denoted μ_9 and introduced in (3.79). For a given feature t and a given document, let X and Y be random variables with

$$X = \begin{cases} 1 & \text{if feature } t \text{ appears in the document,} \\ 0 & \text{if feature } t \text{ appears in the document.} \end{cases}$$

and

$$Y = \begin{cases} 1 & \text{if the document is relevant,} \\ 0 & \text{if the document is irrelevant.} \end{cases}$$

The formula for the correlation coefficient is

$$r(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{(E[X^2] - E[X]^2)(E[Y^2] - E[Y]^2)}} \tag{A.1}$$

Using the notation introduced in §1.2 and

$$\begin{aligned} \rho &\triangleq \text{encoding for a relevant document} \\ \overline{\rho} &\triangleq \text{encoding for an irrelevant document} \\ \tau &\triangleq \text{encoding that a feature is present} \\ \overline{\tau} &\triangleq \text{encoding that a feature is not present} \end{aligned}$$

we have

$$E[X] = \tau \frac{a+b}{a+b+c+d} + \overline{\tau} \frac{c+d}{a+b+c+d} \tag{A.2}$$

$$E[X^2] = \tau^2 \frac{a+b}{a+b+c+d} + \overline{\tau}^2 \frac{c+d}{a+b+c+d} \tag{A.3}$$

$$E[Y] = \rho\frac{a+c}{a+b+c+d} + \overline{\rho}\frac{b+d}{a+b+c+d} \tag{A.4}$$

$$E[Y^2] = \rho^2\frac{a+c}{a+b+c+d} + \overline{\rho}^2\frac{b+d}{a+b+c+d} \tag{A.5}$$

$$\begin{aligned} E[XY] &= \tau\rho\frac{a}{a+b+c+d} + \tau\overline{\rho}\frac{b}{a+b+c+d} && \text{(A.6)} \\ &+ \overline{\tau}\rho\frac{c}{a+b+c+d} + \overline{\tau\rho}\frac{d}{a+b+c+d} && \text{(A.7)} \end{aligned}$$

$$r(X,Y) = \frac{\tau\rho(ad-bc) + \overline{\tau\rho}(ad-bc) - \overline{\tau}\rho(ad-bc) - \tau\overline{\rho}(ad-bc)}{(\tau-\overline{\tau})(\rho-\overline{\rho})\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{A.8}$$

Now we remark that the correlation coefficient is invariant under any linear transformation of X and Y, that is,

$$r(X,Y) = r(\alpha_1 X + \beta, \alpha_2 Y + \beta).$$

This observation follows easily from the fact that for any random variable Z

$$E[\alpha Z + \beta] = \alpha E[Z] + \beta$$

and

$$V[\alpha Z + \beta] = \alpha^2 V[Z]$$

Therefore, the correlation is independent under a linear transformation of the encoding $(\tau, \overline{\tau}, \rho, \overline{\rho})$.

For example, the linear transformation $2X-1$ maps $(0,1)$ to $(-1,1)$. For $(\tau, \overline{\tau}, \rho, \overline{\rho}) = (1,0,1,0)$ and any linear transformation of $(\tau, \overline{\tau}, \rho, \overline{\rho})$, for example $(-1,1,-1,1)$ we have

$$r(X,Y) = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{A.9}$$

In general, any pair of integers (u, v) can be mapped to the pair $(0, 1)$ using the linear transformation $\frac{1}{v-u}(x-u)$ where $x \in \{u, v\}$.

Appendix B

Gini Split Criterion

This appendix contains the derivation of the *Gini split criterion* which is denoted μ_{11} and introduced in (3.83). The *Gini split criterion*, which is based on the *Gini impurity*, is commonly utilized as a "split criterion" in the construction of decision trees in the context of classification problems. In a classification problem with n classes the Gini impurity on a training set (or subset of a training set) S with $|S| = N$ is defined as

$$\text{gini}(S) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the probability that an arbitrary $s \in S$ is a member of class j. That is, p_j is the relative frequency of class j in S. The information retrieval problem discussed in this paper can be viewed as a two class classification problem with class 1 being the set of relevant documents and class 2 being the set of irrelevant documents. The Gini impurity therefore becomes

$$\text{gini}(S) = 1 - p_1^2 - p_2^2.$$

When constructing a decision tree on a data set with binary (or binarized) features, S is recursively split into two subsets, based on the value of one feature. The subset S_1 contains all points in which the feature takes the value 1 and the subset S_2 other contains all points in which the feature takes the value 0. When the Gini impurity is utilized as the basis for a split criterion, the feature which has the smallest $\text{gini}_{split}(S)$ is used to create the split, where

$$\text{gini}_{split}(S) = \frac{|S_1|}{N}\ \text{gini}(S_1) + \frac{|S_2|}{N}\ \text{gini}(S_2).$$

To utilize the Gini split criterion for feature selection we rank features in increasing order of $\text{gini}_{split}(S)$. To compute $\text{gini}_{split}(S)$ for a feature t we let S_1 be the set of all documents which contain t and S_2 be the set of all documents which do not contain t and write

$$\text{gini}_{split}(S) = \frac{a+b}{m}\ \text{gini}(S_1) + \frac{c+d}{m}\ \text{gini}(S_2).$$

We have

$$\text{gini}(S_1) = 1 - P(R|S_1) - P(I|S_1) = 1 - \left(\frac{a}{a+b}\right)^2 - \left(\frac{b}{a+b}\right)^2$$

where $P(R|S_1)$ is the probability that a document is relevant given that it is in S_1, that is, that it contains feature t, and $P(I|S_1)$ is the probability that a document is irrelevant given that it is in S_1, that is, that it contains feature t. Similarly,

$$\text{gini}(S_2) = 1 - P(R|S_2) - P(I|S_2) = 1 - \left(\frac{c}{c+d}\right)^2 - \left(\frac{d}{c+d}\right)^2$$

where $P(R|S_2)$ is the probability that a document is relevant given that it is in S_2, that is, that it does not contain feature t, and $P(I|S_2)$ is the probability that a document is irrelevant given that it is in S_2, that is, that it does not contain feature t. Substituting $\text{gini}(S_1)$ and $\text{gini}(S_2)$ into $\text{gini}_{split}(S)$ yields

$$\begin{aligned}
\text{gini}_{split}(S) &= \frac{a+b}{m}\left(1 - \left(\frac{a}{a+b}\right)^2 - \left(\frac{b}{a+b}\right)^2\right) \\
&+ \frac{c+d}{m}\left(1 - \left(\frac{c}{c+d}\right)^2 - \left(\frac{d}{c+d}\right)^2\right) \\
&= \frac{a+b}{m}\left(1 - \frac{a^2+b^2}{(a+b)^2}\right) + \frac{c+d}{m}\left(1 - \frac{c^2+d^2}{(c+d)^2}\right) \\
&= \frac{a+b}{m}\left(\frac{(a+b)^2}{(a+b)^2} - \frac{a^2+b^2}{(a+b)^2}\right) + \frac{c+d}{m}\left(\frac{(c+d)^2}{(c+d)^2} - \frac{c^2+d^2}{(c+d)^2}\right) \\
&= \frac{a+b}{m}\left(\frac{a^2+2ab+b^2}{(a+b)^2} - \frac{a^2+b^2}{(a+b)^2}\right) + \frac{c+d}{m}\left(\frac{c^2+2cd+d^2}{(c+d)^2} - \frac{c^2+d^2}{(c+d)^2}\right) \\
&= \frac{a+b}{m}\left(\frac{2ab}{(a+b)^2}\right) + \frac{c+d}{m}\left(\frac{2cd}{(c+d)^2}\right) \\
&= \frac{2}{m}\left(\frac{ab}{a+b} + \frac{cd}{c+d}\right).
\end{aligned}$$

Writing the Gini split criterion as

$$\text{gini}_{split}(S) = \frac{2}{m}\left[\frac{ad(b+c)+bc(a+d)}{(a+b)(c+d)}\right] \tag{B.1}$$

exposes the factors ad and bc. It now becomes clear that $\text{gini}_{split}(S)$ takes on its smallest (i.e. best) value of 0, in the "perfect information" cases mentioned in §3.6, when either $a = d = 0$ and b and c are large, specifically $b = |F|$ and $c = |T|$, or when $b = c = 0$ and a and d are large, specifically, $a = |T|$ and $d = |F|$. Further, $\text{gini}_{split}(S)$ takes on its largest (i.e. worst) value of $\frac{1}{2}$ when $ad = bc$ which given that $a + c$ and $b + d$ are constants occurs if and only if $a = b$ and $c = d$. Since the feature appears in the same number of relevant as irrelevant documents and is absent from the same number of relevant as irrelevant documents, it appears in the same fraction of relevant and irrelevant documents and therefore can be viewed as providing "no information" as was mentioned in §3.6.

Appendix C

Information Gain

This appendix contains the derivation of the *information gain* which is denoted μ_{12} and introduced in (3.86). Suppose we are given a classification problem, with class variable Y having k values, a decision variable X having n values, and a training set (or subset of a training set) S. The *entropy* of Y in S is defined as

$$H(Y) = \sum_{j=1}^{k} -P(Y = y_j) \log_2 P(Y = y_j)$$

where $P(Y = y_j)$ is the probability that an arbitrary $s \in S$ is a member of class y_j. That is, $P(Y = y_j)$ is the relative frequency of class y_j in S.

Since the information retrieval problem discussed in this paper can be viewed as a two class classification problem with class y_1 being the set of relevant documents and class y_2 being the set of irrelevant documents, we have

$$P(Y = y_1) = \frac{a+c}{m} \text{ and } P(Y = y_2) = \frac{b+d}{m}$$

and for the entropy we have

$$H(Y) = -\frac{a+c}{m} \log_2 \frac{a+c}{m} - \frac{b+d}{m} \log_2 \frac{b+d}{m}.$$

Entropy can be seen to provide a measure of the homogeneity of the set S. A set S in which all elements are from one class is completely homogeneous, and $H(Y)$ will have the entropy function's minimal value of 0, while a set S in which the elements are equally distributed among the classes, $H(Y)$ will have the entropy function's maximum value of 1.

Many decision tree construction algorithms utilize this notion of homogeneity. At each iteration they select a variable on which to partition the set S, with the goal of increasing the homogeneity of the subsets resulting from the partition. For example, the ID3 algorithm selects the variable X with the largest value of information gain, which is defined as

$$I(Y|X) = H(Y) - H(Y|X)$$

where the conditional entropy of Y given X is

$$H(Y|X) = \sum_{i=1}^{n} P(X = x_i) H(Y|X = x_i)$$

and $H(Y|X = x_i)$ is the entropy function computed on the subset of S where $X = x_i$.

Since the entropy, $H(Y)$ is the same for all variables, the variable with the largest information gain, can be seen to have the smallest conditional entropy. That is, rather than splitting on the variable with the largest value of information gain, ID3 could equivalently split on the variable with the smallest conditional entropy.

For our information retrieval problem, letting x_1 correspond to the case where a feature is present in a document, and x_2 correspond to the case where a feature is absent from a document yields

$$P(X = x_1) = \frac{a+b}{m} \text{ and } P(X = x_2) = \frac{c+d}{m}$$

which yields

$$H(Y|X = x_1) = -\frac{a}{a+b}\log_2\frac{a}{a+b} - \frac{b}{a+b}\log_2\frac{b}{a+b}$$

$$H(Y|X = x_2) = -\frac{c}{c+d}\log_2\frac{c}{c+d} - \frac{d}{c+d}\log_2\frac{d}{c+d}$$

$$H(Y|X) = \frac{a+b}{m}H(Y|X = x_1) + \frac{c+d}{m}H(Y|X = x_2)$$

and finally

$$\begin{aligned} I(Y|X) &= -\frac{a+c}{m}\log_2\frac{a+c}{m} - \frac{b+d}{m}\log_2\frac{b+d}{m} \\ &\quad -\left[\frac{a+b}{m}\left(-\frac{a}{a+b}\log_2\frac{a}{a+b} - \frac{b}{a+b}\log_2\frac{b}{a+b}\right)\right. \\ &\quad \left. +\frac{c+d}{m}\left(-\frac{c}{c+d}\log_2\frac{c}{c+d} - \frac{d}{c+d}\log_2\frac{d}{c+d}\right)\right]. \end{aligned}$$

In the "perfect information" cases mentioned in §3.6, that is, where $b = c = 0$ and where $a = d = 0$, elementary calculations show that $H(Y|X) = 0$ and therefore $I(Y|X) = H(Y)$, i.e. the maximal value. In the "no information" cases mentioned in §3.6, that is, where $a = b = 0$ (using $c + d = m$) and where $c = d = 0$ (using $a + b = m$) elementary calculations show that we have $I(Y|X) = 0$, i.e. the minimal value.

Appendix D

Rareness

In [13] it was shown that for a given K, that unless $|T|$ and $|F|$ exceed some determined threshold which is a function of K, that *many* support sets exist. It was then reasoned that for a given T and F that if K does exceed this threshold that the support sets that exist are just those that randomly exist in the data set, but when K lies below the threshold, the support sets are those (desirable ones that are) associated with the underlying phenomenon of the data set. While our problem formulation differs from the assumptions used to prove this result, we are nevertheless motivated by the idea that sets of features that are in someway unusual or *rare* are desirable. In [13] it was the size of a support set that determined its rareness. In this section we introduce a measure of the rareness of a feature that is based on the probability of a feature occurring, given an assumption about the manner in which the collection is generated.

Assume that relevant and irrelevant documents are generated by two independent stochastic processes that choose features from V. The process that generates relevant documents chooses $|T|$ features while the process that generates irrelevant documents chooses $|F|$ features. For a given feature $j \in V$ we can consider the relevant document generating process as performing $|T|$ Bernoulli trials in which success is choosing feature j, failure is choosing any feature other than feature j, the probability of success is

$$p_j = \frac{a_j + b_j}{m}$$

and the probability of failure is

$$q_j = \frac{c_j + d_j}{m}.$$

Similarly, for the feature $j \in V$ we can consider the irrelevant document generating

process as performing $|F|$ Bernoulli trials with the same definitions of success, failure and their associated probabilities. Let the random variable X_j^T be the number of relevant documents containing feature j and let the random variable X_j^F be the number of irrelevant documents containing feature j. Clearly X_j^T and X_j^F follow a binomial distribution with

$$\begin{aligned}
\mathrm{P}(X_j^T = a_j) &= \binom{|T|}{a_j} \left(\frac{a_j + b_j}{m}\right)^{a_j} \left(\frac{c_j + d_j}{m}\right)^{|T|-a_j} \\
&= \frac{|T|!}{a_j!(|T| - a_j)!} \left(\frac{a_j + b_j}{m}\right)^{a_j} \left(\frac{c_j + d_j}{m}\right)^{|T|-a_j} \\
&= \frac{|T|!}{a_j!\ c_j!} \left(\frac{a_j + b_j}{m}\right)^{a_j} \left(\frac{c_j + d_j}{m}\right)^{c_j}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(X_j^F = b_j) &= \binom{|F|}{b_j} \left(\frac{a_j + b_j}{m}\right)^{b_j} \left(\frac{c_j + d_j}{m}\right)^{|F|-b_j} \\
&= \frac{|F|!}{b_j!\ (|F| - b_j)!} \left(\frac{a_j + b_j}{m}\right)^{b_j} \left(\frac{c_j + d_j}{m}\right)^{|F|-b_j} \\
&= \frac{|F|!}{b_j!\ d_j!} \left(\frac{a_j + b_j}{m}\right)^{b_j} \left(\frac{c_j + d_j}{m}\right)^{d_j}.
\end{aligned}$$

Given this model of document generation, a natural definition of the *rareness* of a feature is the probability that the feature will occur in a relevant documents and in b irrelevant documents, with features having *lower* values of this probability being "rarer" and therefore more desirable. The rareness of a feature therefore is a function $\zeta\colon \boxplus \mapsto [0, 1]$ given by

$$\zeta(a_j, b_j, c_j, d_j) = \mathrm{P}(X_j^T = a_j,\ X_j^F = b_j)$$

and since X_j^T and X_j^F are independent random variables we have

$$\begin{aligned}
\zeta(a_j, b_j, c_j, d_j) &= \mathrm{P}(X_j^T = a_j)\ \mathrm{P}(X_j^F = b_j) \\
&= \frac{|T|!\ |F|!}{a_j!\ b_j!\ c_j!\ d_j!} \left(\frac{a_j + b_j}{m}\right)^{a_j+b_j} \left(\frac{c_j + d_j}{m}\right)^{c_j+d_j}.
\end{aligned}$$

Since lower values are considered more desirable, when rareness is used as a feature ranking function, features are ranked in ascending order.

Appendix E

Fisher's Linear Discriminant

This appendix contains the derivation of Fisher's linear discriminant. For a given feature t and a relevant document, let X be a random variable with

$$X = \begin{cases} 1 & \text{if feature } t \text{ appears in the document,} \\ 0 & \text{if feature } t \text{ does not appear in the document,} \end{cases}$$

and for an irrelevant document, let Y be a random variable with

$$Y = \begin{cases} 1 & \text{if feature } t \text{ appears in the document,} \\ 0 & \text{if feature } t \text{ does not appear in the document.} \end{cases}$$

The formula for Fisher's linear discriminant is

$$\mathcal{L}(X, Y) = \frac{[\mathrm{E}(X) - \mathrm{E}(Y)]^2}{\mathrm{Var}(X) + \mathrm{Var}(Y)}$$

Using the notation introduced in §1.2 we have

$$\mathrm{E}(X) = 1\frac{a}{|T|} + 0\frac{c}{|T|} = \frac{a}{|T|},$$

$$\mathrm{E}(Y) = 1\frac{b}{|F|} + 0\frac{d}{|F|} = \frac{b}{|F|},$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 1^2\frac{a}{|T|} + 0^2\frac{c}{|T|} - \left(\frac{a}{|T|}\right)^2 = \frac{a}{|T|} - \left(\frac{a}{|T|}\right)^2,$$

and

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = 1^2\frac{b}{|F|} + 0^2\frac{d}{|F|} - \left(\frac{b}{|F|}\right)^2 = \frac{b}{|F|} - \left(\frac{b}{|F|}\right)^2.$$

Alternatively, we could have noted that X and Y are Bernoulli random variables with probability of occurrence $a/|T|$ and $b/|F|$ respectively, and recalled that any Bernoulli random variable, with probability of occurrence p has expected value equal to p and variance equal to $p(1-p)$.

Therefore,

$$\begin{aligned}\mathcal{L}(X,Y) &= \frac{\left(\frac{a}{|T|}-\frac{b}{|F|}\right)^2}{\frac{a}{|T|}-\left(\frac{a}{|T|}\right)^2+\frac{b}{|F|}-\left(\frac{b}{|F|}\right)^2}, \\ &= \frac{\left(\frac{a}{|T|}-\frac{b}{|F|}\right)^2}{\frac{a}{|T|}\left(1-\frac{a}{|T|}\right)+\frac{b}{|F|}\left(1-\frac{b}{|F|}\right)}.\end{aligned}$$

Fisher's linear discriminant for Boolean text data was discussed in [77] and here we offer a closed form formula for it. We shall also use a variant of Fisher's linear discriminant, that is defined as

$$\begin{aligned}\mathcal{L}_G(X,Y) &= \frac{|\mathrm{E}(X)-\mathrm{E}(Y)|}{\sqrt{\mathrm{Var}(X)}+\sqrt{\mathrm{Var}(Y)}} \\ &= \frac{\left|\frac{a}{|T|}-\frac{b}{|F|}\right|}{\sqrt{\frac{a}{|T|}\left(1-\frac{a}{|T|}\right)}+\sqrt{\frac{b}{|F|}\left(1-\frac{b}{|F|}\right)}}\end{aligned}$$

as well as a version of it, in which the absolute value is not applied

$$\begin{aligned}\widetilde{\mathcal{L}}_G(X,Y) &= \frac{\mathrm{E}(X)-\mathrm{E}(Y)}{\sqrt{\mathrm{Var}(X)}+\sqrt{\mathrm{Var}(Y)}} \\ &= \frac{\frac{a}{|T|}-\frac{b}{|F|}}{\sqrt{\frac{a}{|T|}\left(1-\frac{a}{|T|}\right)}+\sqrt{\frac{b}{|F|}\left(1-\frac{b}{|F|}\right)}}\end{aligned}$$

(see e.g. [63]).

Appendix F

Proof that Function Sets are Closed

Corollary F.1. *The sets of functions $\mathcal{M}^*[\mathcal{F}]$, $\hat{\mathcal{M}}^*[\mathcal{F}]$, and $\tilde{\mathcal{M}}^*[\mathcal{F}]$ are closed under convex combination.*

Proof. Given $\mu_{\mathbf{c}}$, $\mu_{\mathbf{d}} \in \tilde{\mathcal{M}}^*[\mathcal{F}]$, i.e. both vectors satisfy axioms (A1)–(A6), it is RTP that each convex combination of these two vectors also satisfies these axioms. For each axiom we will show that if $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ satisfy the axiom, then $(\mu_{\mathbf{c}} + \mu_{\mathbf{d}})/2$ satisfies the axiom.

Since $\mu_{\mathbf{c}}, \mu_{\mathbf{d}} \in \tilde{\mathcal{M}}^*[\mathcal{F}]$, they satisfy (A1), i.e.

$$\sum_{f \in \mathcal{F}} c_f f(1,0) = 1$$

and

$$\sum_{f \in \mathcal{F}} d_f f(1,0) = 1.$$

Since these equations are linear, adding them and dividing the result by 2 yields

$$\sum_{f \in \mathcal{F}} \frac{(c_f + d_f) f(1,0)}{2} = 1$$

and we see that the convex combination of $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ also satisfies (A1). That the convex combination of $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ also satisfies (A2) and A3 follows similarly.

Since $\mu_{\mathbf{c}}, \mu_{\mathbf{d}} \in \tilde{\mathcal{M}}^*[\mathcal{F}]$, they satisfy (A4), i.e.

$$\sum_{f \in \mathcal{F}} c_f f(x,y) \geq \sum_{f \in \mathcal{F}} c_f f(x',y')$$

and

$$\sum_{f\in\mathcal{F}} d_f f(x,y) \geq \sum_{f\in\mathcal{F}} d_f f(x',y').$$

Since these inequalities are linear, adding them and dividing the result by 2 yields

$$\sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(x,y)}{2} \geq \sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(x',y')}{2}$$

and we see that the convex combination of $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ also satisfies (A4).

Since $\mu_{\mathbf{c}}, \mu_{\mathbf{d}} \in \tilde{\mathcal{M}}^*[\mathcal{F}]$, they satisfy (A5), i.e.

$$\sum_{f\in\mathcal{F}} c_f f(x,y) = -\sum_{f\in\mathcal{F}} c_f f(y,x)$$

and

$$\sum_{f\in\mathcal{F}} d_f f(x,y) = -\sum_{f\in\mathcal{F}} d_f f(y,x).$$

Since these inequalities are linear, adding them and dividing the result by 2 yields

$$\sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(x,y)}{2} = -\sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(y,x)}{2}$$

and we see that the convex combination of $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ also satisfies (A5).

Since $\mu_{\mathbf{c}}, \mu_{\mathbf{d}} \in \tilde{\mathcal{M}}^*[\mathcal{F}]$, they satisfy (A6), i.e.

$$\sum_{f\in\mathcal{F}} c_f f(x,y) = \sum_{f\in\mathcal{F}} c_f f(1-y,1-x)$$

and

$$\sum_{f\in\mathcal{F}} d_f f(x,y) = \sum_{f\in\mathcal{F}} d_f f(1-y,1-x).$$

Since these inequalities are linear, adding them and dividing the result by 2 yields

$$\sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(x,y)}{2} = \sum_{f\in\mathcal{F}} \frac{(c_f+d_f)f(1-y,1-x)}{2}$$

and we see that the convex combination of $\mu_{\mathbf{c}}$ and $\mu_{\mathbf{d}}$ also satisfies (A6). ■

Appendix G

SMART Stopword List

a	a's	able	about	above	according
accordingly	across	actually	after	afterwards	again
against	ain't	all	allow	allows	almost
alone	along	already	also	although	always
am	among	amongst	an	and	another
any	anybody	anyhow	anyone	anything	anyway
anyways	anywhere	apart	appear	appreciate	appropriate
are	aren't	around	as	aside	ask
asking	associated	at	available	away	awfully
b	be	became	because	become	becomes
becoming	been	before	beforehand	behind	being
believe	below	beside	besides	best	better
between	beyond	both	brief	but	by
c	c'mon	c's	came	can	can't
cannot	cant	cause	causes	certain	certainly
changes	clearly	co	com	come	comes
concerning	consequently	consider	considering	contain	containing
contains	corresponding	could	couldn't	course	currently
d	definitely	described	despite	did	didn't
different	do	does	doesn't	doing	don't
done	down	downwards	during	e	each
edu	eg	eight	either	else	elsewhere
enough	entirely	especially	et	etc	even

ever	every	everybody	everyone	everything	everywhere
ex	exactly	example	except	f	far
few	fifth	first	five	followed	following
follows	for	former	formerly	forth	four
from	further	furthermore	g	get	gets
getting	given	gives	go	goes	going
gone	got	gotten	greetings	h	had
hadn't	happens	hardly	has	hasn't	have
haven't	having	he	he's	hello	help
hence	her	here	here's	hereafter	hereby
herein	hereupon	hers	herself	hi	him
himself	his	hither	hopefully	how	howbeit
however	i	i'd	i'll	i'm	i've
ie	if	ignored	immediate	in	inasmuch
inc	indeed	indicate	indicated	indicates	inner
insofar	instead	into	inward	is	isn't
it	it'd	it'll	it's	its	itself
j	just	k	keep	keeps	kept
know	knows	known	l	last	lately
later	latter	latterly	least	less	lest
let	let's	like	liked	likely	little
look	looking	looks	ltd	m	mainly
many	may	maybe	me	mean	meanwhile
merely	might	more	moreover	most	mostly
much	must	my	myself	n	name
namely	nd	near	nearly	necessary	need
needs	neither	never	nevertheless	new	next
nine	no	nobody	non	none	noone
nor	normally	not	nothing	novel	now
nowhere	o	obviously	of	off	often

oh	ok	okay	old	on	once
one	ones	only	onto	or	other
others	otherwise	ought	our	ours	ourselves
out	outside	over	overall	own	p
particular	particularly	per	perhaps	placed	please
plus	possible	presumably	probably	provides	q
que	quite	qv	r	rather	rd
re	really	reasonably	regarding	regardless	regards
relatively	respectively	right	s	said	same
saw	say	saying	says	second	secondly
see	seeing	seem	seemed	seeming	seems
seen	self	selves	sensible	sent	serious
seriously	seven	several	shall	she	should
shouldn't	since	six	so	some	somebody
somehow	someone	something	sometime	sometimes	somewhat
somewhere	soon	sorry	specified	specify	specifying
still	sub	such	sup	sure	t
t's	take	taken	tell	tends	th
than	thank	thanks	thanx	that	that's
thats	the	their	theirs	them	themselves
then	thence	there	there's	thereafter	thereby
therefore	therein	theres	thereupon	these	they
they'd	they'll	they're	they've	think	third
this	thorough	thoroughly	those	though	three
through	throughout	thru	thus	to	together
too	took	toward	towards	tried	tries
truly	try	trying	twice	two	u
un	under	unfortunately	unless	unlikely	until
unto	up	upon	us	use	used
useful	uses	using	usually	uucp	v

value	various	very	via	viz	vs
w	want	wants	was	wasn't	way
we	we'd	we'll	we're	we've	welcome
well	went	were	weren't	what	what's
whatever	when	whence	whenever	where	where's
whereafter	whereas	whereby	wherein	whereupon	wherever
whether	which	while	whither	who	who's
whoever	whole	whom	whose	why	will
willing	wish	with	within	without	won't
wonder	would	would	wouldn't	x	y
yes	yet	you	you'd	you'll	you're
you've	your	yours	yourself	yourselves	z
zero					

Appendix H

μ-RANKING Results Stopwords Included Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_0	Train	0.15	0.28	0.01	0.34	0.04	0.91	0.06	0.10
μ_0	Test	0.14	0.26	0.01	0.32	0.04	0.91	0.03	0.06
μ_0	Δ%	-6.21	-5.37	-7.01	-6.53	0.00	0.00	-41.81	-39.86
μ_1	Train	0.84	0.93	0.20	3.93	0.00	0.13	0.87	0.90
μ_1	Test	0.81	0.89	0.18	3.29	0.00	0.13	0.81	0.83
μ_1	Δ%	-3.22	-3.89	-12.08	-16.31	0.00	0.00	-6.59	-8.46
μ_2	Train	0.95	0.98	0.29	5.55	0.00	0.09	0.90	0.92
μ_2	Test	0.94	0.97	0.27	4.94	0.00	0.09	0.87	0.88
μ_2	Δ%	-0.88	-1.23	-8.31	-10.94	0.00	0.00	-3.23	-4.33
μ_3	Train	0.92	0.97	0.24	4.62	0.00	0.11	0.89	0.92
μ_3	Test	0.91	0.95	0.21	3.98	0.00	0.11	0.85	0.87
μ_3	Δ%	-1.18	-1.62	-10.29	-13.85	0.00	0.00	-4.21	-5.58
μ_4	Train	0.97	1.00	0.43	9.58	0.14	4.42	0.84	0.82
μ_4	Test	0.97	1.00	0.41	8.90	0.14	4.42	0.81	0.79
μ_4	Δ%	-0.32	-0.10	-6.22	-7.07	0.00	0.00	-3.08	-3.16
$\lvert\mu_4\rvert$	Train	0.98	1.00	0.46	10.37	0.12	3.97	0.86	0.87
$\lvert\mu_4\rvert$	Test	0.98	1.00	0.44	9.87	0.12	3.97	0.84	0.84
$\lvert\mu_4\rvert$	Δ%	-0.21	-0.01	-4.78	-4.89	0.00	0.00	-2.57	-2.93
μ_5	Train	0.96	1.00	0.45	10.38	0.38	8.98	0.77	0.77
μ_5	Test	0.96	1.00	0.43	9.84	0.38	8.98	0.75	0.74
μ_5	Δ%	-0.05	-0.01	-4.71	-5.28	0.00	0.00	-2.97	-3.76
$\lvert\mu_5\rvert$	Train	0.89	1.00	0.41	10.48	0.41	9.48	0.74	0.84
$\lvert\mu_5\rvert$	Test	0.89	1.00	0.39	10.08	0.41	9.48	0.72	0.81
$\lvert\mu_5\rvert$	Δ%	-0.14	0.00	-3.26	-3.74	0.00	0.00	-2.95	-2.81
μ_6	Train	0.96	1.00	0.45	10.39	0.38	9.00	0.77	0.77
μ_6	Test	0.96	1.00	0.43	9.84	0.38	9.00	0.75	0.74
μ_6	Δ%	-0.04	-0.01	-4.71	-5.28	0.00	0.00	-2.98	-3.78
$\lvert\mu_6\rvert$	Train	0.89	1.00	0.41	10.48	0.41	9.46	0.73	0.84
$\lvert\mu_6\rvert$	Test	0.89	1.00	0.40	10.09	0.41	9.46	0.71	0.82
$\lvert\mu_6\rvert$	Δ%	-0.14	0.00	-3.23	-3.68	0.00	0.00	-2.91	-2.79
μ_7	Train	0.95	0.98	0.29	5.55	0.01	0.13	0.90	0.92
μ_7	Test	0.94	0.97	0.26	4.93	0.01	0.13	0.86	0.87
μ_7	Δ%	-1.03	-1.01	-8.45	-11.26	0.00	0.00	-3.71	-4.92
$\lvert\mu_7\rvert$	Train	0.95	0.98	0.29	5.55	0.01	0.13	0.90	0.92

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_7\rvert$	Test	0.94	0.97	0.26	4.93	0.01	0.13	0.86	0.87
$\lvert\mu_7\rvert$	Δ%	-1.03	-1.01	-8.45	-11.26	0.00	0.00	-3.71	-4.92
μ_8	Train	0.99	1.00	0.54	12.43	0.50	14.41	0.78	0.74
μ_8	Test	0.98	1.00	0.53	12.14	0.50	14.41	0.76	0.72
μ_8	Δ%	-0.12	0.00	-2.36	-2.31	0.00	0.00	-2.66	-3.12
μ_9	Train	0.96	0.99	0.34	6.93	0.01	0.23	0.89	0.91
μ_9	Test	0.95	0.98	0.32	6.30	0.01	0.23	0.87	0.88
μ_9	Δ%	-0.61	-0.65	-7.10	-9.07	0.00	0.00	-2.69	-3.33
$\lvert\mu_9\rvert$	Train	0.96	0.99	0.34	7.01	0.01	0.27	0.89	0.91
$\lvert\mu_9\rvert$	Test	0.95	0.98	0.32	6.38	0.01	0.27	0.87	0.88
$\lvert\mu_9\rvert$	Δ%	-0.61	-0.64	-7.12	-8.97	0.00	0.00	-2.68	-3.18
μ_{10}	Train	0.96	0.99	0.35	7.11	0.01	0.29	0.89	0.91
μ_{10}	Test	0.96	0.98	0.32	6.49	0.01	0.29	0.87	0.88
μ_{10}	Δ%	-0.60	-0.59	-7.00	-8.78	0.00	0.00	-2.71	-3.24
μ_{11}	Train	0.96	0.99	0.34	7.01	0.01	0.26	0.89	0.91
μ_{11}	Test	0.95	0.98	0.32	6.38	0.01	0.26	0.87	0.88
μ_{11}	Δ%	-0.61	-0.64	-7.11	-8.92	0.00	0.00	-2.68	-3.18
μ_{12}	Train	0.97	0.99	0.37	7.90	0.01	0.52	0.89	0.90
μ_{12}	Test	0.96	0.99	0.35	7.29	0.01	0.52	0.86	0.88
μ_{12}	Δ%	-0.50	-0.20	-6.32	-7.69	0.00	0.00	-2.58	-2.59
μ_{13}	Train	0.98	1.00	0.49	11.11	0.37	10.65	0.79	0.74
μ_{13}	Test	0.98	1.00	0.46	10.57	0.37	10.65	0.77	0.71
μ_{13}	Δ%	-0.21	-0.01	-4.89	-4.86	0.00	0.00	-2.95	-4.04
μ_{14}	Train	0.96	1.00	0.40	9.02	0.32	9.22	0.35	0.43
μ_{14}	Test	0.96	1.00	0.40	9.08	0.32	9.22	0.31	0.39
μ_{14}	Δ%	-0.08	-0.00	0.20	0.61	0.00	0.00	-8.88	-9.61
μ_{15}	Train	0.97	1.00	0.41	9.00	0.13	4.22	0.86	0.84
μ_{15}	Test	0.96	1.00	0.38	8.34	0.13	4.22	0.83	0.82
μ_{15}	Δ%	-0.44	0.02	-6.15	-7.30	0.00	0.00	-2.56	-2.50
μ_{16}	Train	0.97	0.99	0.37	7.94	0.05	1.64	0.88	0.88
μ_{16}	Test	0.96	0.99	0.35	7.32	0.05	1.64	0.85	0.86
μ_{16}	Δ%	-0.52	-0.12	-6.52	-7.79	0.00	0.00	-2.71	-2.73
μ_{17}	Train	0.96	0.99	0.33	6.88	0.01	0.51	0.89	0.90
μ_{17}	Test	0.95	0.98	0.31	6.26	0.01	0.51	0.87	0.87
μ_{17}	Δ%	-0.60	-0.61	-7.05	-9.06	0.00	0.00	-2.65	-3.40
μ_{18}	Train	0.88	1.00	0.41	10.78	0.59	14.48	0.68	0.74
μ_{18}	Test	0.88	1.00	0.40	10.48	0.59	14.48	0.66	0.72
μ_{18}	Δ%	0.07	0.02	-2.33	-2.82	0.00	0.00	-2.57	-2.58
μ_{19}	Train	0.84	1.00	0.36	9.46	0.59	14.00	0.27	0.47
μ_{19}	Test	0.84	1.00	0.36	9.47	0.59	14.00	0.23	0.42
μ_{19}	Δ%	0.03	0.00	0.02	0.05	0.00	0.00	-13.43	-10.88
μ_{20}	Train	0.92	0.97	0.24	4.62	0.00	0.11	0.89	0.92
μ_{20}	Test	0.91	0.95	0.21	3.98	0.00	0.11	0.85	0.87
μ_{20}	Δ%	-1.18	-1.62	-10.29	-13.85	0.00	0.00	-4.21	-5.58

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_{20}\rvert$	Train	0.92	0.97	0.24	4.64	0.00	0.11	0.89	0.92
$\lvert\mu_{20}\rvert$	Test	0.91	0.96	0.22	4.00	0.00	0.11	0.85	0.87
$\lvert\mu_{20}\rvert$	$\Delta\%$	-1.15	-1.59	-10.26	-13.80	0.00	0.00	-4.22	-5.68
μ_{21}	Train	0.97	0.99	0.38	8.10	0.02	0.64	0.88	0.90
μ_{21}	Test	0.96	0.99	0.36	7.49	0.02	0.64	0.86	0.87
μ_{21}	$\Delta\%$	-0.44	-0.16	-6.22	-7.48	0.00	0.00	-2.63	-2.82
μ_{22}	Train	0.84	0.93	0.20	3.93	0.00	0.13	0.87	0.90
μ_{22}	Test	0.81	0.89	0.18	3.29	0.00	0.13	0.81	0.83
μ_{22}	$\Delta\%$	-3.21	-3.79	-12.03	-16.22	0.00	0.00	-6.57	-8.31
$\lvert\mu_{22}\rvert$	Train	0.84	0.93	0.20	3.94	0.00	0.14	0.87	0.90
$\lvert\mu_{22}\rvert$	Test	0.82	0.89	0.18	3.30	0.00	0.14	0.81	0.83
$\lvert\mu_{22}\rvert$	$\Delta\%$	-3.21	-3.77	-12.02	-16.18	0.00	0.00	-6.58	-8.40
μ_{23}	Train	0.98	1.00	0.43	9.37	0.05	1.81	0.88	0.89
μ_{23}	Test	0.97	1.00	0.40	8.84	0.05	1.81	0.85	0.87
μ_{23}	$\Delta\%$	-0.22	-0.03	-5.12	-5.59	0.00	0.00	-2.57	-2.18
μ_{24}	Train	0.97	0.99	0.38	8.04	0.02	0.89	0.87	0.88
μ_{24}	Test	0.96	0.99	0.36	7.41	0.02	0.89	0.85	0.86
μ_{24}	$\Delta\%$	-0.48	-0.20	-6.45	-7.83	0.00	0.00	-2.65	-2.67
$\lvert\mu_{24}\rvert$	Train	0.97	1.00	0.40	8.72	0.03	1.12	0.88	0.90
$\lvert\mu_{24}\rvert$	Test	0.97	1.00	0.38	8.17	0.03	1.12	0.86	0.88
$\lvert\mu_{24}\rvert$	$\Delta\%$	-0.33	-0.06	-5.51	-6.27	0.00	0.00	-2.54	-2.39

Table H.1: μ-RANKING Stopwords Included Not Discounted

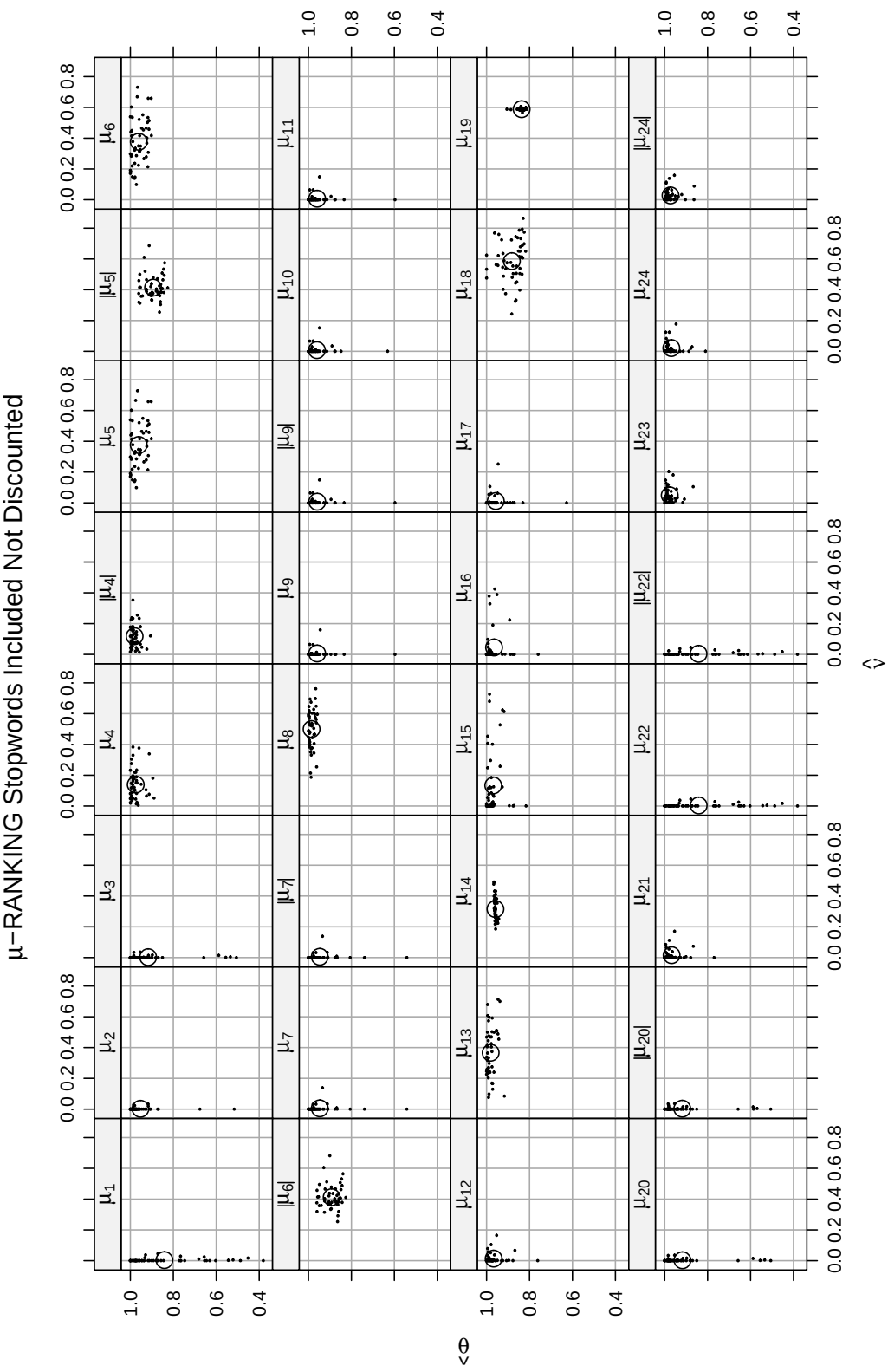

μ−RANKING Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ6|
μ7
|μ7|
μ8
μ12
μ13
μ14
μ15
μ20
|μ20|
μ21
μ22
0.0 0.2 0.4 0.6 0.8
1.0
0.8
0.6
0.4
θ̂
V̂

Figure H.1

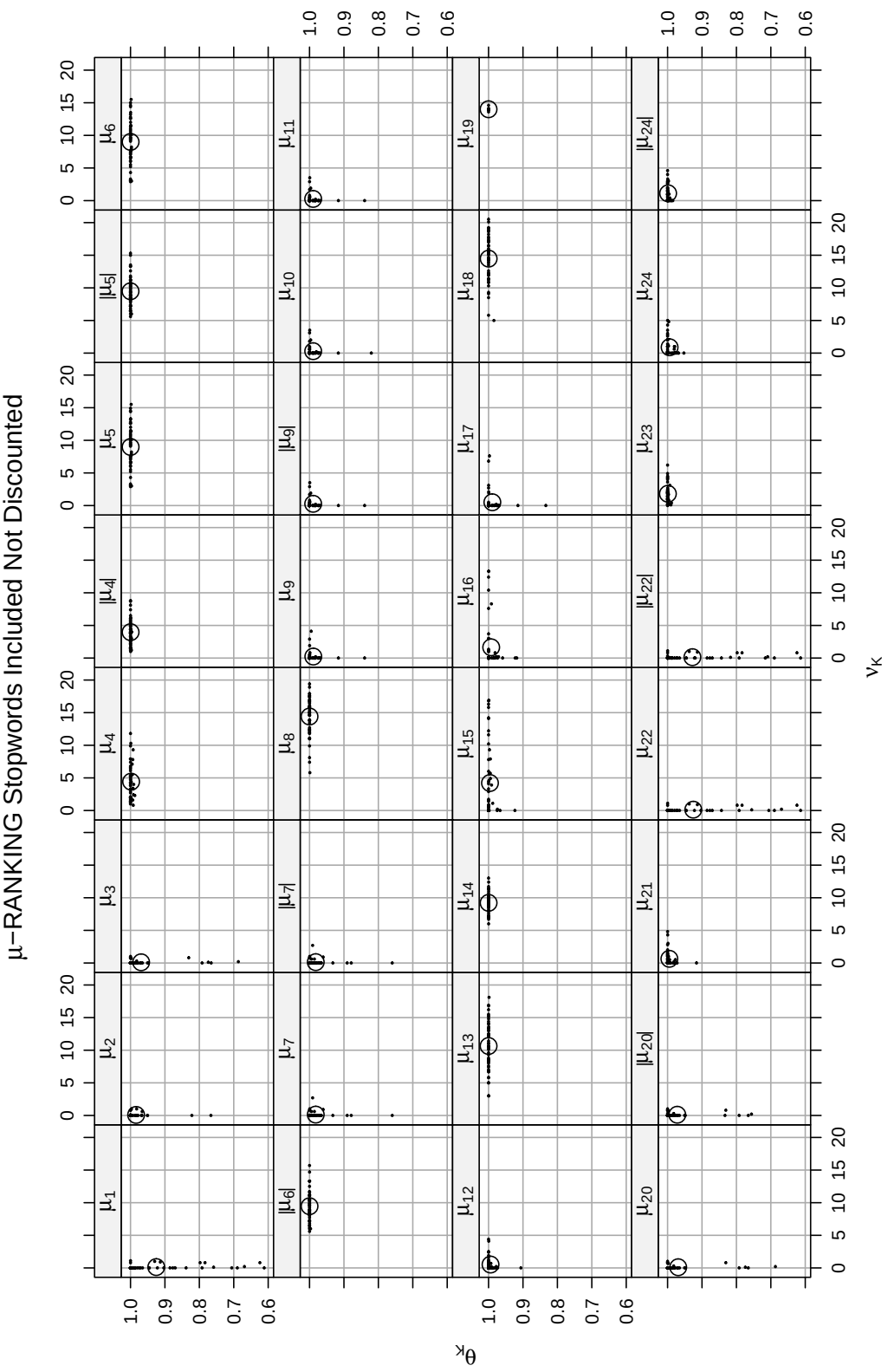


Figure H.2

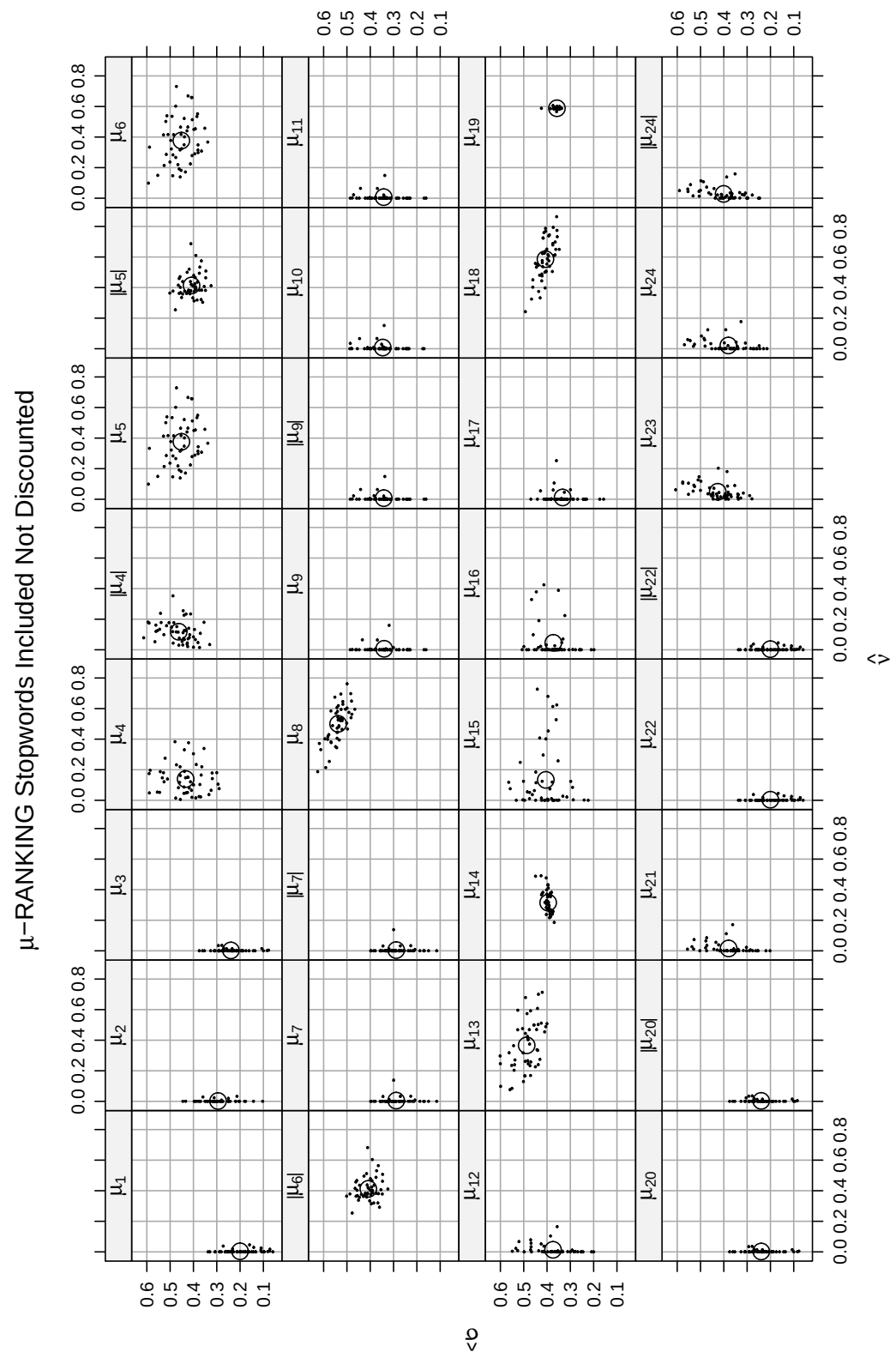


Figure H.3

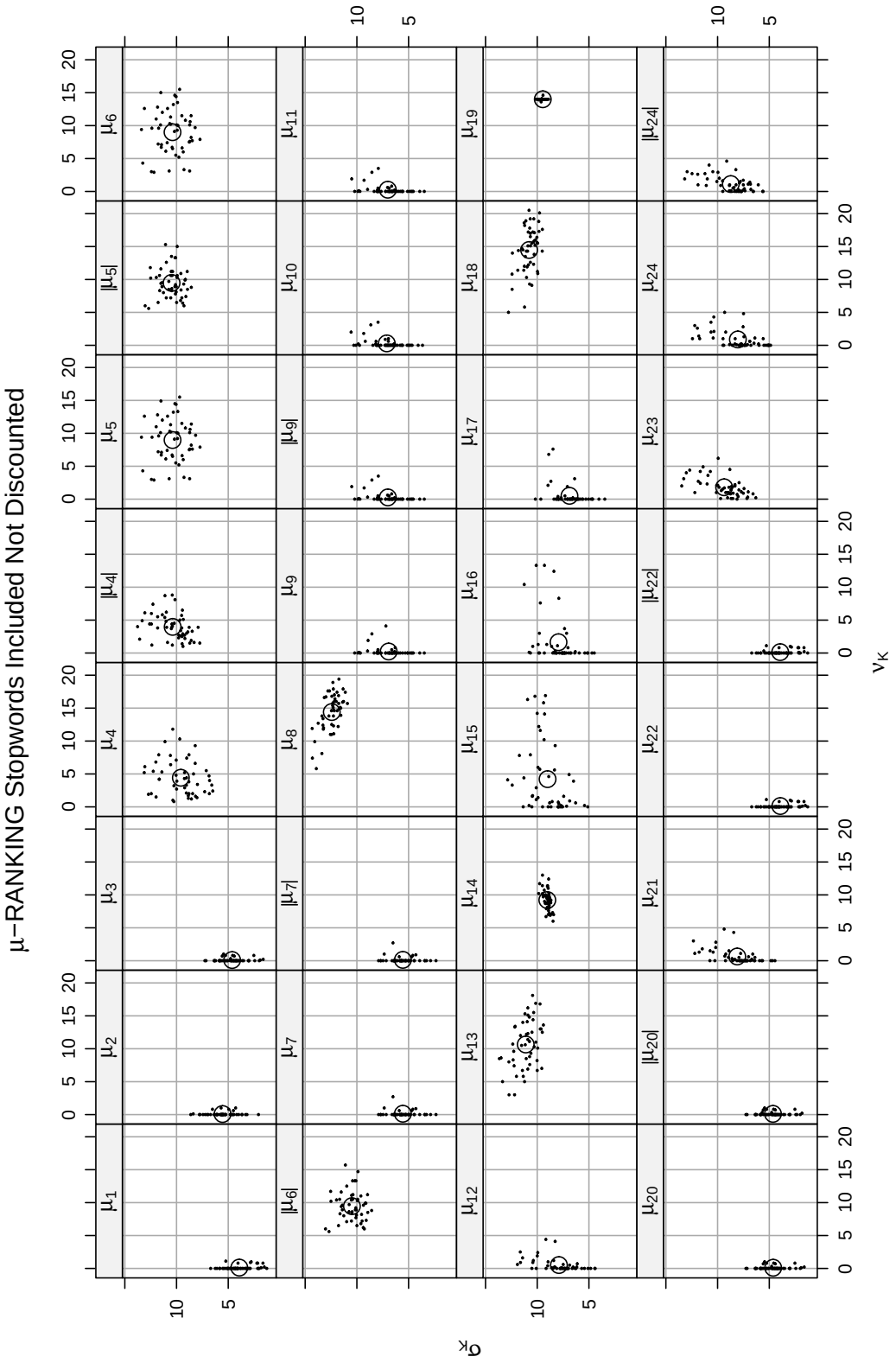

μ−RANKING Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
σκ
νκ

Figure H.4

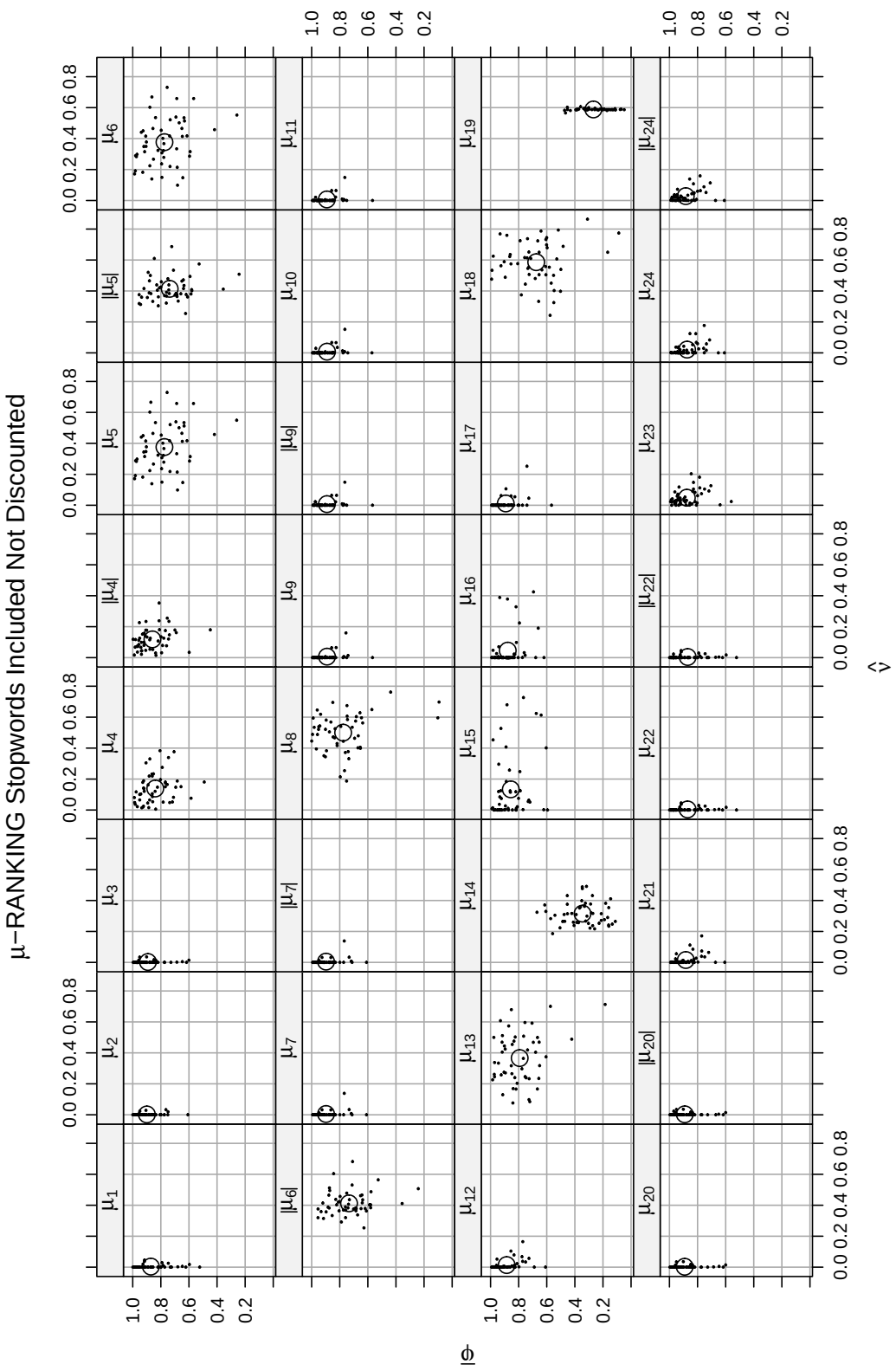

μ−RANKING Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
0.0 0.2 0.4 0.6 0.8
1.0 0.8 0.6 0.4 0.2
Φ̂
V̂

Figure H.5

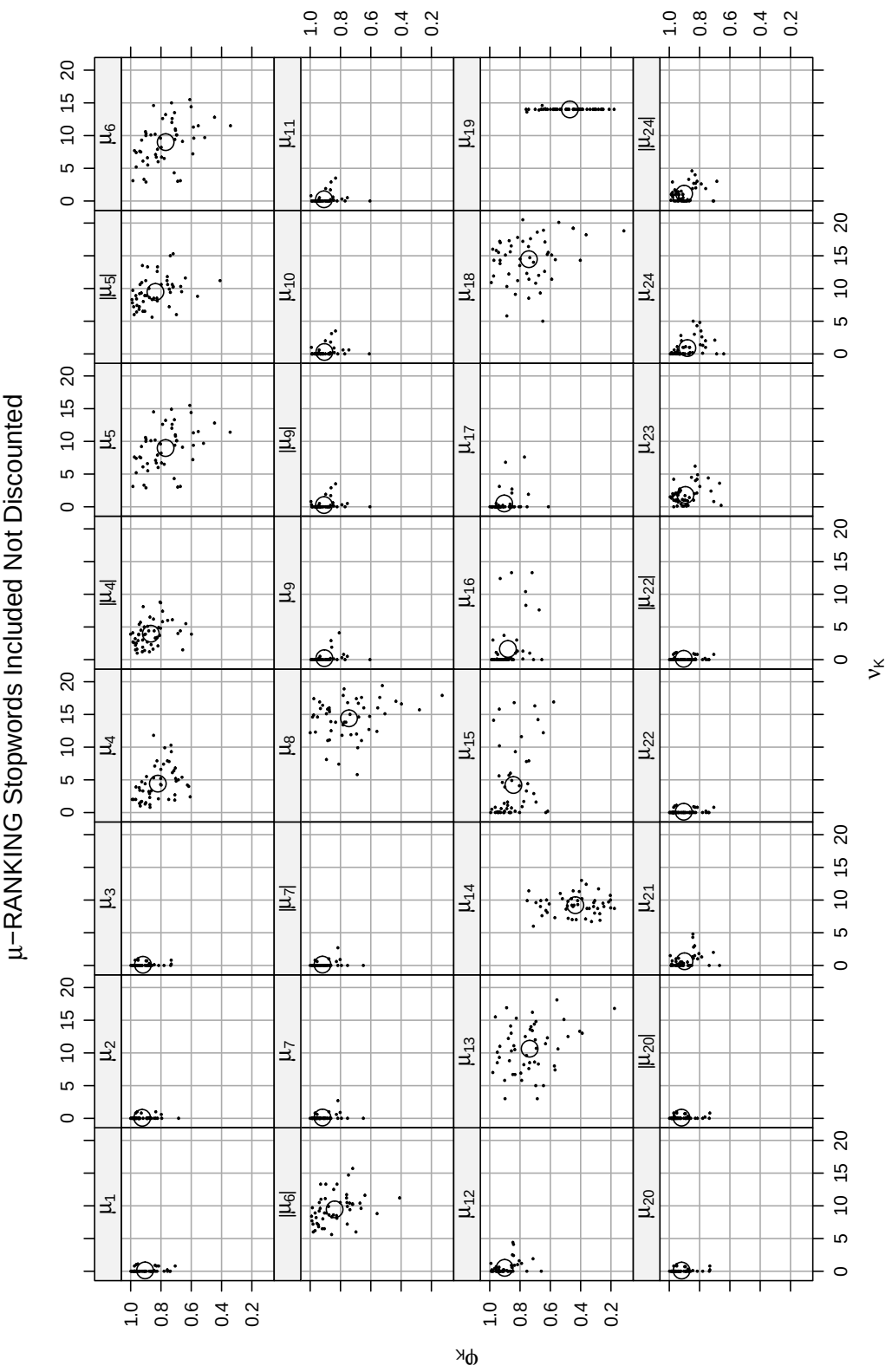

μ−RANKING Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
φK
νK

Figure H.6

Appendix I

μ-RANKING Results Stopwords Included Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_0	Train	0.12	0.22	0.01	0.26	0.04	0.91	0.05	0.09
μ_0	Test	0.11	0.21	0.01	0.24	0.04	0.91	0.03	0.05
μ_0	Δ%	-7.89	-7.07	-9.00	-8.56	0.00	0.00	-43.60	-42.58
μ_1	Train	0.84	0.92	0.20	3.92	0.00	0.13	0.87	0.90
μ_1	Test	0.81	0.89	0.18	3.28	0.00	0.13	0.81	0.83
μ_1	Δ%	-3.21	-3.97	-12.02	-16.28	0.00	0.00	-6.57	-8.35
μ_2	Train	0.95	0.98	0.29	5.54	0.00	0.09	0.90	0.92
μ_2	Test	0.94	0.97	0.27	4.93	0.00	0.09	0.87	0.88
μ_2	Δ%	-0.89	-1.23	-8.27	-10.89	0.00	0.00	-3.22	-4.31
μ_3	Train	0.92	0.97	0.24	4.62	0.00	0.11	0.89	0.92
μ_3	Test	0.91	0.95	0.21	3.98	0.00	0.11	0.85	0.87
μ_3	Δ%	-1.18	-1.62	-10.24	-13.77	0.00	0.00	-4.19	-5.56
μ_4	Train	0.97	0.99	0.37	7.54	0.14	4.42	0.86	0.86
μ_4	Test	0.96	0.99	0.34	7.01	0.14	4.42	0.83	0.84
μ_4	Δ%	-0.53	-0.35	-6.00	-7.04	0.00	0.00	-2.80	-2.88
$\lvert\mu_4\rvert$	Train	0.98	1.00	0.41	8.52	0.12	3.97	0.87	0.89
$\lvert\mu_4\rvert$	Test	0.98	1.00	0.39	8.11	0.12	3.97	0.85	0.86
$\lvert\mu_4\rvert$	Δ%	-0.22	-0.04	-4.55	-4.75	0.00	0.00	-2.41	-2.49
μ_5	Train	0.93	0.99	0.30	6.55	0.38	8.98	0.82	0.86
μ_5	Test	0.93	0.99	0.28	6.16	0.38	8.98	0.80	0.83
μ_5	Δ%	-0.48	-0.36	-5.06	-6.09	0.00	0.00	-2.48	-3.11
$\lvert\mu_5\rvert$	Train	0.87	1.00	0.28	6.89	0.41	9.48	0.75	0.87
$\lvert\mu_5\rvert$	Test	0.87	1.00	0.27	6.64	0.41	9.48	0.73	0.85
$\lvert\mu_5\rvert$	Δ%	-0.21	-0.03	-3.42	-3.69	0.00	0.00	-2.35	-2.04
μ_6	Train	0.93	0.99	0.30	6.55	0.38	9.00	0.82	0.86
μ_6	Test	0.93	0.99	0.28	6.15	0.38	9.00	0.80	0.83
μ_6	Δ%	-0.48	-0.36	-5.06	-6.10	0.00	0.00	-2.48	-3.12
$\lvert\mu_6\rvert$	Train	0.87	1.00	0.28	6.90	0.41	9.46	0.74	0.87
$\lvert\mu_6\rvert$	Test	0.86	1.00	0.27	6.65	0.41	9.46	0.73	0.85
$\lvert\mu_6\rvert$	Δ%	-0.21	-0.03	-3.39	-3.60	0.00	0.00	-2.30	-2.06
μ_7	Train	0.95	0.98	0.29	5.52	0.01	0.13	0.90	0.92
μ_7	Test	0.94	0.97	0.26	4.90	0.01	0.13	0.86	0.87
μ_7	Δ%	-1.03	-1.01	-8.44	-11.24	0.00	0.00	-3.68	-4.89
$\lvert\mu_7\rvert$	Train	0.95	0.98	0.29	5.52	0.01	0.13	0.90	0.92

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_7\rvert$	Test	0.94	0.97	0.26	4.90	0.01	0.13	0.86	0.87
$\lvert\mu_7\rvert$	Δ%	-1.03	-1.01	-8.44	-11.24	0.00	0.00	-3.68	-4.89
μ_8	Train	0.97	1.00	0.29	5.71	0.50	14.41	0.81	0.85
μ_8	Test	0.97	1.00	0.29	5.55	0.50	14.41	0.80	0.82
μ_8	Δ%	-0.18	-0.04	-2.58	-2.74	0.00	0.00	-2.12	-2.62
μ_9	Train	0.96	0.99	0.34	6.84	0.01	0.23	0.89	0.91
μ_9	Test	0.95	0.98	0.31	6.22	0.01	0.23	0.87	0.88
μ_9	Δ%	-0.61	-0.65	-7.09	-9.04	0.00	0.00	-2.66	-3.34
$\lvert\mu_9\rvert$	Train	0.96	0.99	0.34	6.90	0.01	0.27	0.89	0.91
$\lvert\mu_9\rvert$	Test	0.95	0.98	0.32	6.28	0.01	0.27	0.87	0.88
$\lvert\mu_9\rvert$	Δ%	-0.61	-0.64	-7.06	-8.93	0.00	0.00	-2.66	-3.21
μ_{10}	Train	0.96	0.99	0.34	6.99	0.01	0.29	0.89	0.91
μ_{10}	Test	0.95	0.98	0.32	6.37	0.01	0.29	0.87	0.88
μ_{10}	Δ%	-0.60	-0.59	-6.95	-8.79	0.00	0.00	-2.68	-3.28
μ_{11}	Train	0.96	0.99	0.34	6.90	0.01	0.26	0.89	0.91
μ_{11}	Test	0.95	0.98	0.32	6.29	0.01	0.26	0.87	0.88
μ_{11}	Δ%	-0.61	-0.64	-7.05	-8.87	0.00	0.00	-2.65	-3.21
μ_{12}	Train	0.97	0.99	0.37	7.66	0.01	0.52	0.89	0.90
μ_{12}	Test	0.96	0.99	0.35	7.08	0.01	0.52	0.86	0.88
μ_{12}	Δ%	-0.50	-0.20	-6.28	-7.60	0.00	0.00	-2.54	-2.72
μ_{13}	Train	0.96	0.99	0.31	6.27	0.37	10.65	0.85	0.85
μ_{13}	Test	0.96	0.99	0.30	5.92	0.37	10.65	0.83	0.83
μ_{13}	Δ%	-0.48	-0.30	-5.10	-5.69	0.00	0.00	-2.36	-2.95
μ_{14}	Train	0.93	0.99	0.26	5.25	0.32	9.22	0.29	0.40
μ_{14}	Test	0.93	0.99	0.26	5.25	0.32	9.22	0.27	0.37
μ_{14}	Δ%	-0.02	0.02	-0.09	-0.02	0.00	0.00	-7.21	-8.29
μ_{15}	Train	0.96	0.99	0.35	7.24	0.13	4.22	0.87	0.87
μ_{15}	Test	0.96	0.99	0.33	6.65	0.13	4.22	0.85	0.85
μ_{15}	Δ%	-0.51	-0.02	-6.57	-8.16	0.00	0.00	-2.42	-2.42
μ_{16}	Train	0.96	0.99	0.35	7.27	0.05	1.64	0.88	0.89
μ_{16}	Test	0.96	0.99	0.33	6.68	0.05	1.64	0.86	0.86
μ_{16}	Δ%	-0.55	-0.14	-6.69	-8.15	0.00	0.00	-2.61	-2.71
μ_{17}	Train	0.96	0.99	0.33	6.68	0.01	0.51	0.89	0.91
μ_{17}	Test	0.95	0.98	0.31	6.07	0.01	0.51	0.87	0.88
μ_{17}	Δ%	-0.60	-0.60	-7.07	-9.16	0.00	0.00	-2.63	-3.30
μ_{18}	Train	0.82	0.99	0.20	4.97	0.59	14.48	0.71	0.85
μ_{18}	Test	0.82	0.99	0.19	4.80	0.59	14.48	0.69	0.83
μ_{18}	Δ%	-0.31	-0.09	-2.92	-3.46	0.00	0.00	-2.09	-2.44
μ_{19}	Train	0.75	0.99	0.15	4.03	0.59	14.00	0.17	0.39
μ_{19}	Test	0.75	0.99	0.15	4.03	0.59	14.00	0.15	0.36
μ_{19}	Δ%	0.00	0.03	-0.12	-0.01	0.00	0.00	-8.25	-7.82
μ_{20}	Train	0.92	0.97	0.24	4.62	0.00	0.11	0.89	0.92
μ_{20}	Test	0.91	0.95	0.21	3.98	0.00	0.11	0.85	0.87
μ_{20}	Δ%	-1.18	-1.62	-10.24	-13.77	0.00	0.00	-4.19	-5.56

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_{20}\rvert$	Train	0.92	0.97	0.24	4.64	0.00	0.11	0.89	0.92
$\lvert\mu_{20}\rvert$	Test	0.91	0.96	0.22	4.00	0.00	0.11	0.85	0.87
$\lvert\mu_{20}\rvert$	$\Delta\%$	-1.15	-1.59	-10.23	-13.74	0.00	0.00	-4.21	-5.65
μ_{21}	Train	0.97	0.99	0.37	7.79	0.02	0.64	0.88	0.90
μ_{21}	Test	0.96	0.99	0.35	7.21	0.02	0.64	0.86	0.88
μ_{21}	$\Delta\%$	-0.43	-0.15	-6.18	-7.45	0.00	0.00	-2.58	-2.83
μ_{22}	Train	0.84	0.93	0.20	3.92	0.00	0.13	0.87	0.90
μ_{22}	Test	0.81	0.89	0.18	3.29	0.00	0.13	0.81	0.83
μ_{22}	$\Delta\%$	-3.20	-3.88	-11.97	-16.19	0.00	0.00	-6.55	-8.22
$\lvert\mu_{22}\rvert$	Train	0.84	0.93	0.20	3.94	0.00	0.14	0.87	0.90
$\lvert\mu_{22}\rvert$	Test	0.82	0.89	0.18	3.30	0.00	0.14	0.81	0.83
$\lvert\mu_{22}\rvert$	$\Delta\%$	-3.20	-3.85	-11.98	-16.16	0.00	0.00	-6.57	-8.30
μ_{23}	Train	0.98	1.00	0.40	8.56	0.05	1.81	0.88	0.90
μ_{23}	Test	0.97	1.00	0.38	8.09	0.05	1.81	0.86	0.88
μ_{23}	$\Delta\%$	-0.22	-0.03	-5.02	-5.52	0.00	0.00	-2.47	-2.52
μ_{24}	Train	0.97	0.99	0.37	7.60	0.02	0.89	0.88	0.89
μ_{24}	Test	0.96	0.99	0.35	7.00	0.02	0.89	0.85	0.86
μ_{24}	$\Delta\%$	-0.47	-0.20	-6.42	-7.89	0.00	0.00	-2.62	-2.65
$\lvert\mu_{24}\rvert$	Train	0.97	1.00	0.39	8.23	0.03	1.12	0.88	0.91
$\lvert\mu_{24}\rvert$	Test	0.97	1.00	0.37	7.71	0.03	1.12	0.86	0.88
$\lvert\mu_{24}\rvert$	$\Delta\%$	-0.32	-0.05	-5.49	-6.38	0.00	0.00	-2.47	-2.45

Table I.1: μ-RANKING Stopwords Included Discounted

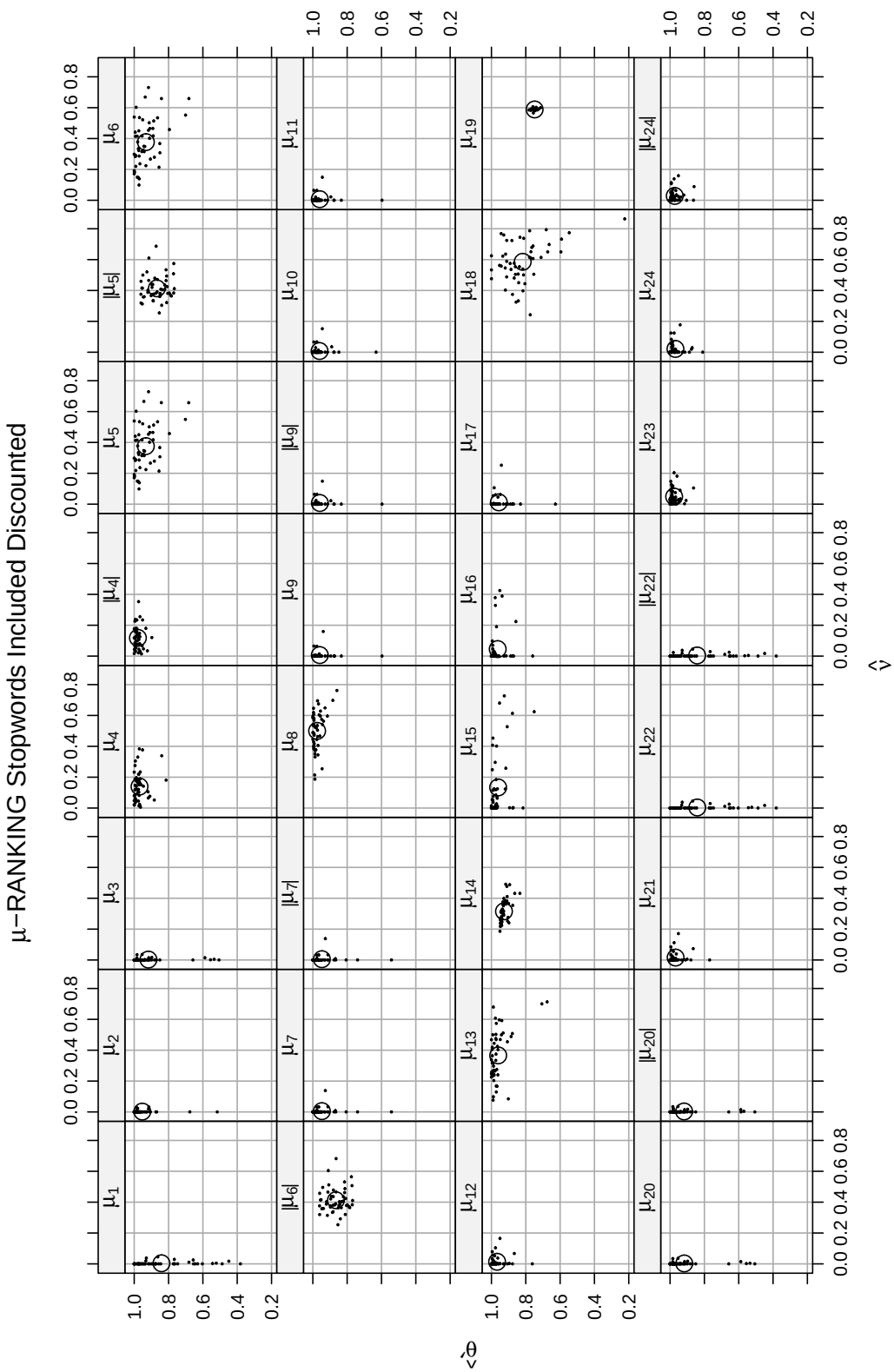

μ−RANKING Stopwords Included Discounted
$\hat{\theta}$
$\hat{V}$
μ1
μ2
μ3
μ4
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
0.0 0.2 0.4 0.6 0.8
0.2 0.4 0.6 0.8 1.0

Figure I.1

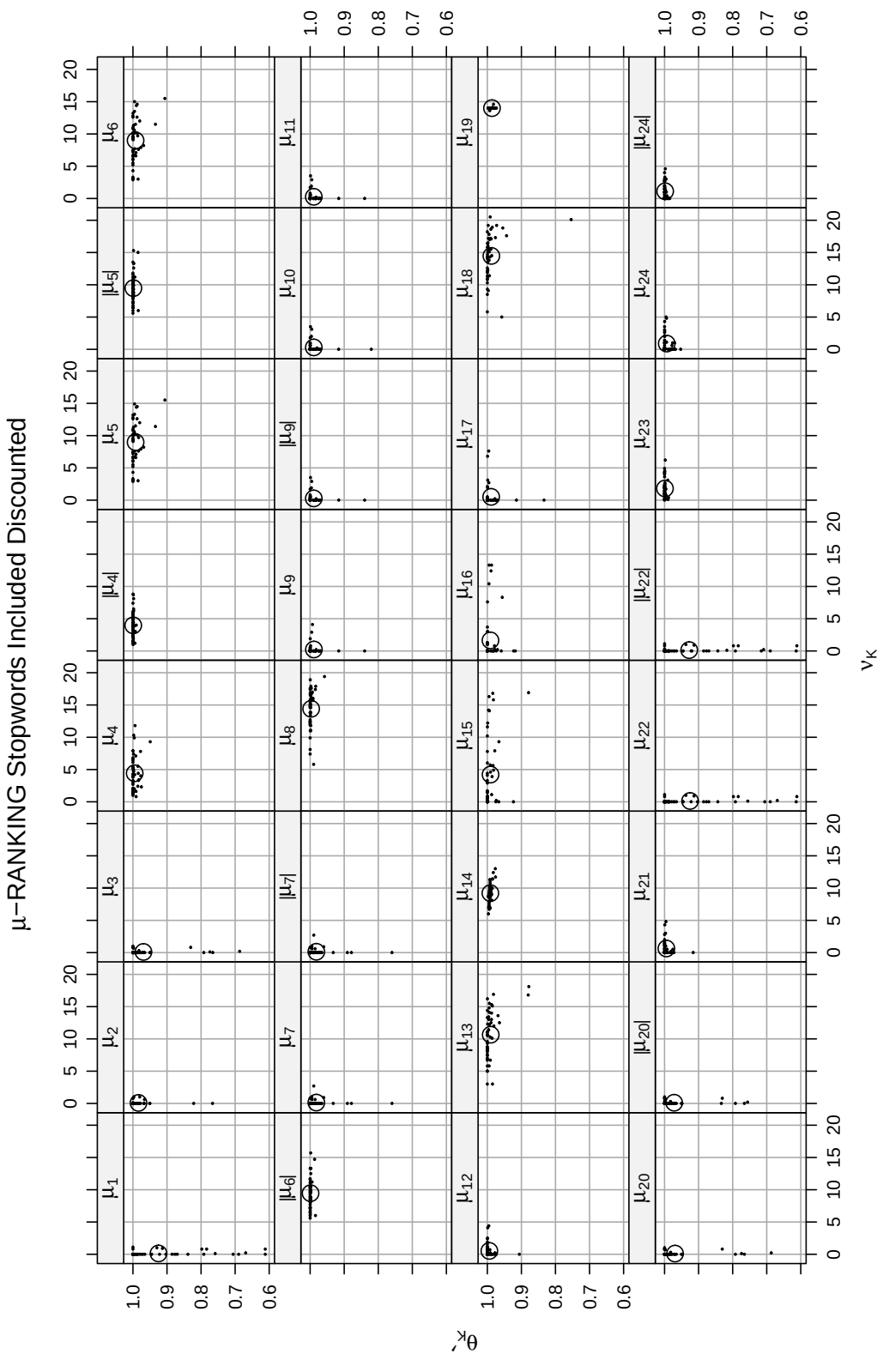

μ−RANKING Stopwords Included Discounted
θκ'
νκ
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|

Figure I.2

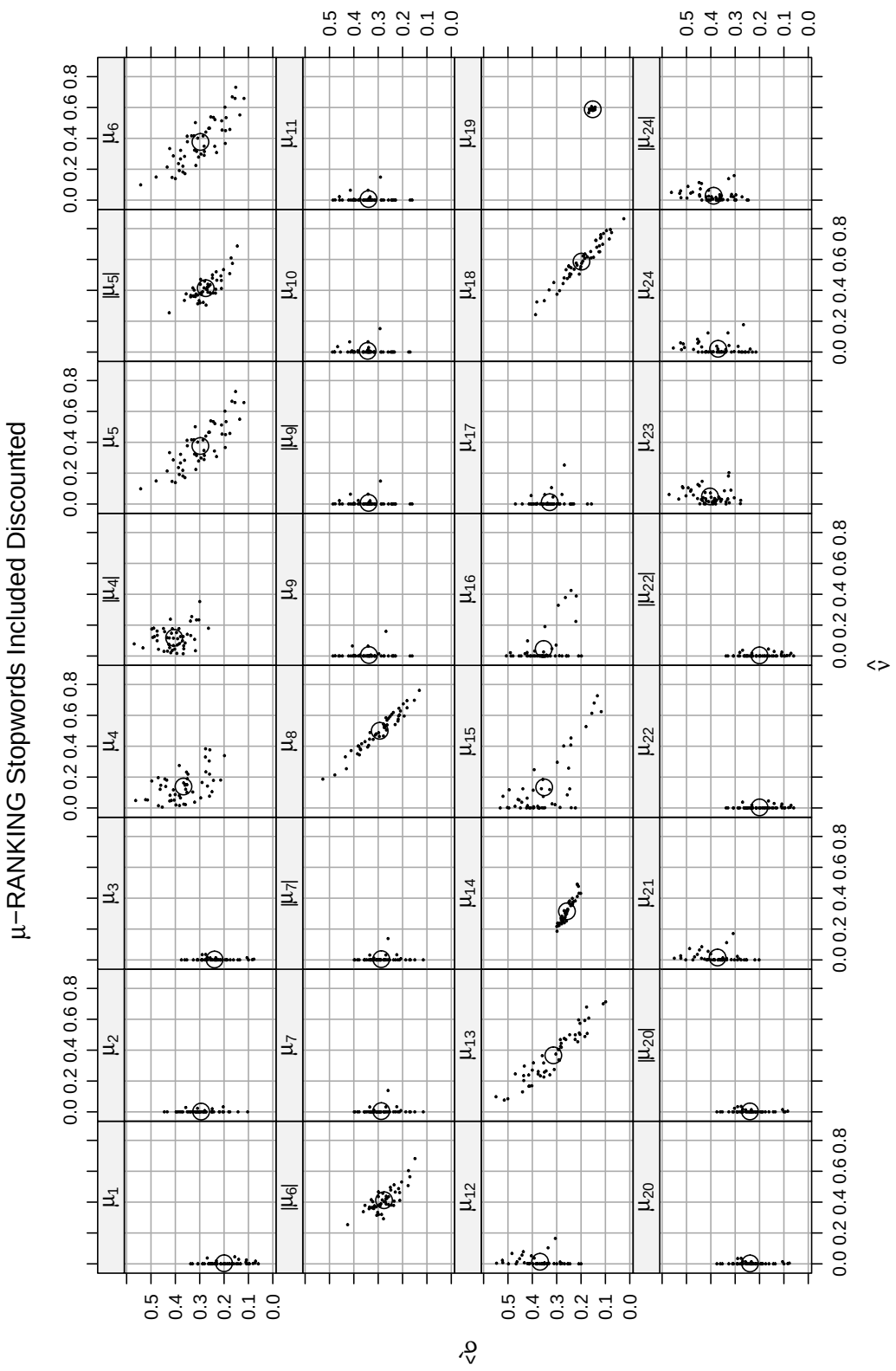

μ−RANKING Stopwords Included Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
0.0 0.2 0.4 0.6 0.8
0.0 0.1 0.2 0.3 0.4 0.5
σ²
V̂

Figure I.3

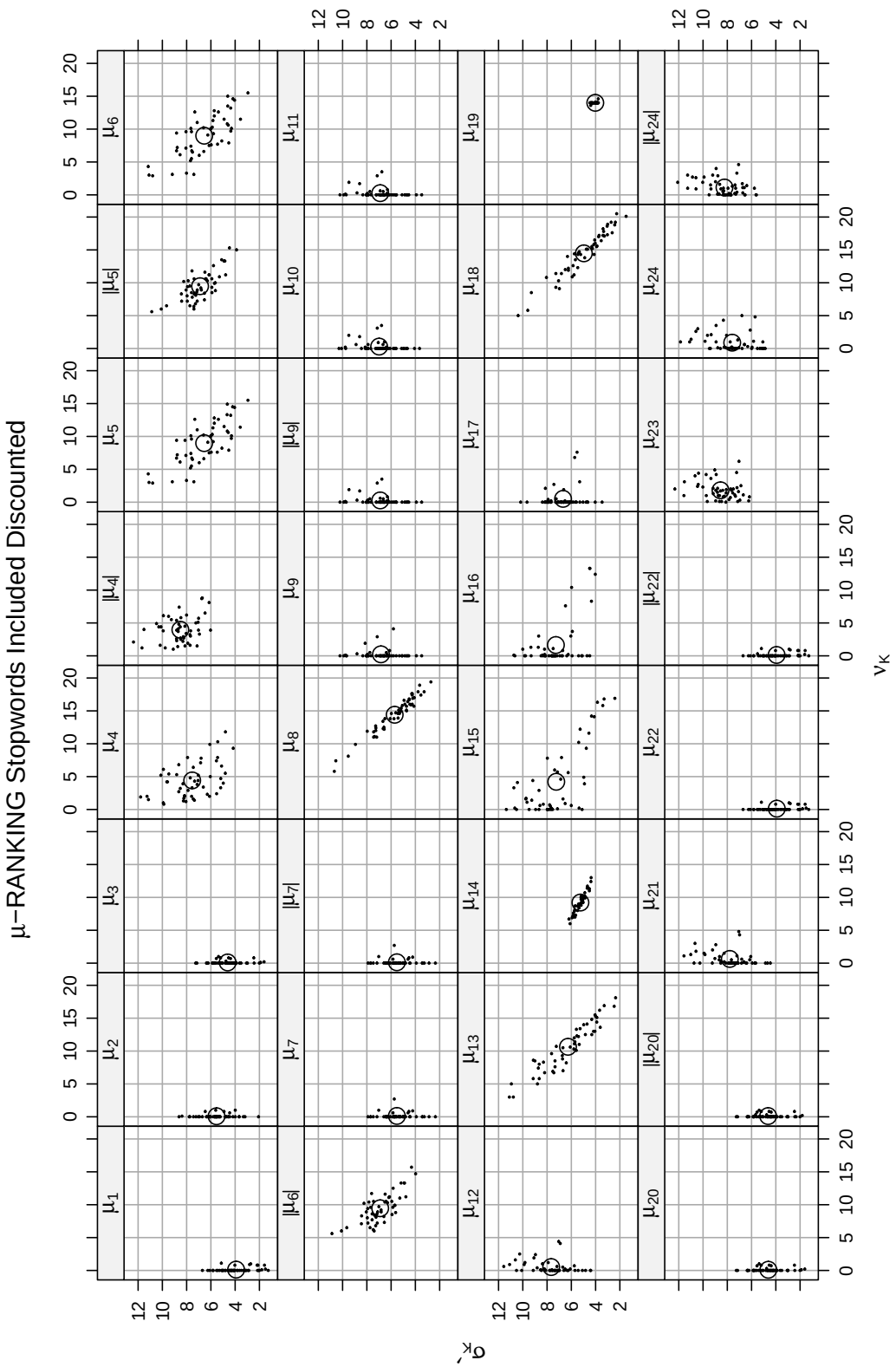


Figure I.4

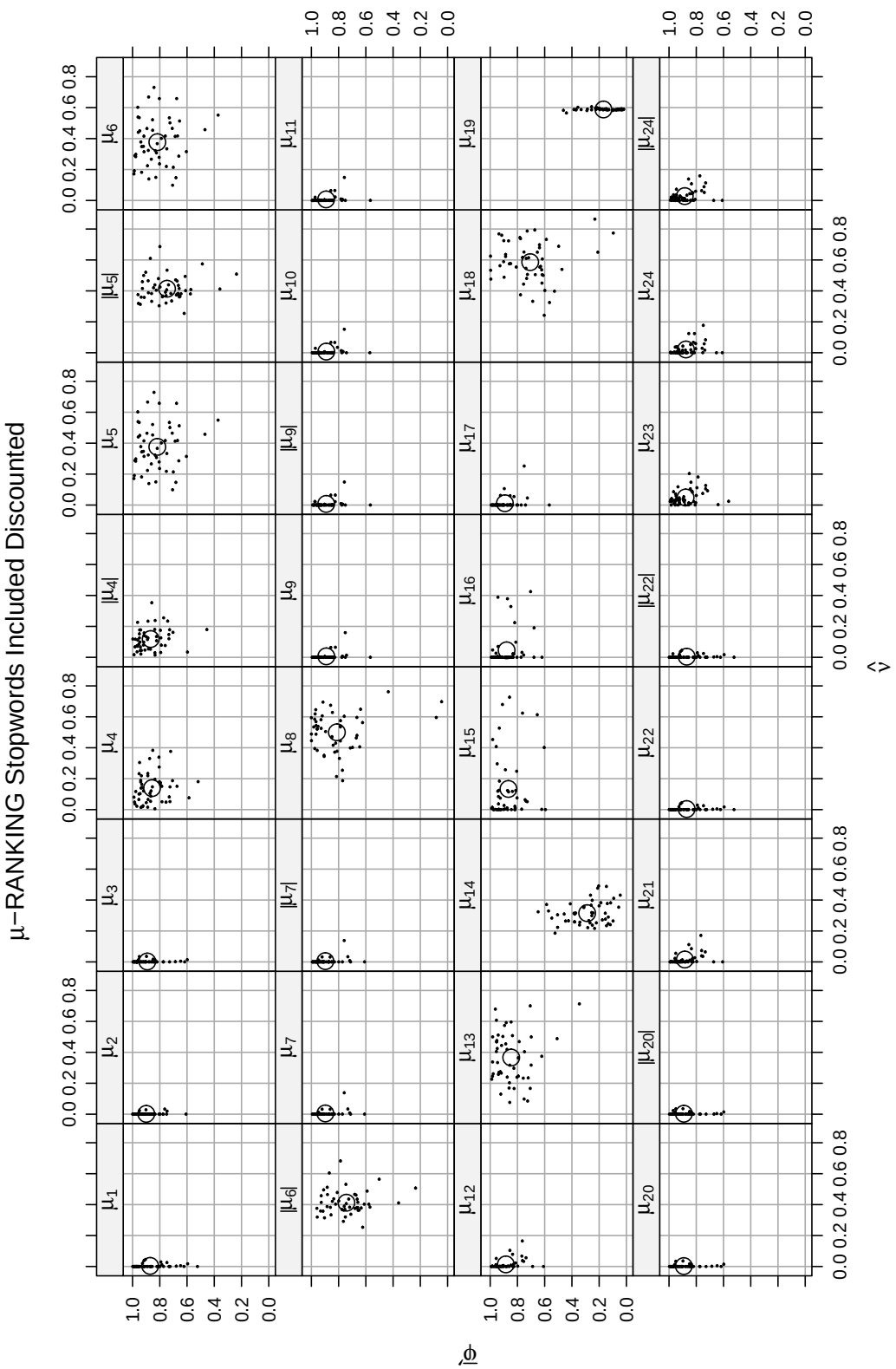


Figure I.5

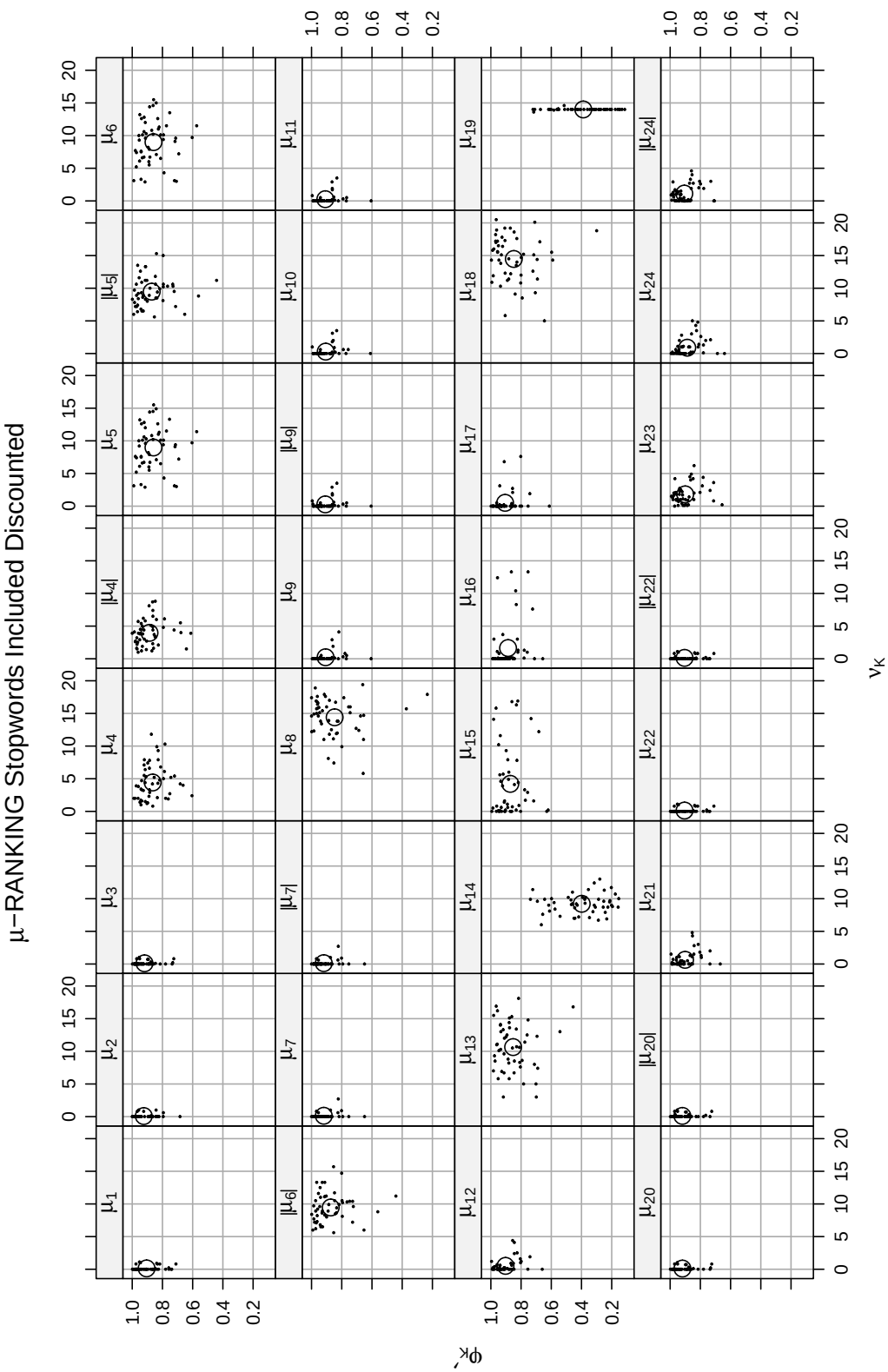

μ−RANKING Stopwords Included Discounted
φK'
νK
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|

Figure I.6

Appendix J

μ-RANKING Results Stopwords Excluded Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_0	Train	0.13	0.24	0.01	0.28	0	0	0.06	0.09
μ_0	Test	0.12	0.22	0.01	0.25	0	0	0.03	0.06
μ_0	Δ%	-10.51	-8.33	-11.80	-10.08			-39.56	-39.55
μ_1	Train	0.84	0.93	0.20	3.93	0	0	0.87	0.90
μ_1	Test	0.81	0.89	0.18	3.29	0	0	0.81	0.83
μ_1	Δ%	-3.20	-3.99	-12.04	-16.34			-6.55	-8.36
μ_2	Train	0.95	0.98	0.29	5.54	0	0	0.90	0.92
μ_2	Test	0.94	0.97	0.27	4.94	0	0	0.87	0.88
μ_2	Δ%	-0.88	-1.22	-8.28	-10.89			-3.22	-4.32
μ_3	Train	0.92	0.97	0.24	4.62	0	0	0.89	0.92
μ_3	Test	0.91	0.95	0.21	3.99	0	0	0.85	0.87
μ_3	Δ%	-1.18	-1.62	-10.25	-13.82			-4.19	-5.62
μ_4	Train	0.97	1.00	0.40	8.60	0	0	0.86	0.86
μ_4	Test	0.97	1.00	0.38	7.93	0	0	0.83	0.84
μ_4	Δ%	-0.47	-0.10	-6.35	-7.75			-2.80	-2.55
$\lvert\mu_4\rvert$	Train	0.98	1.00	0.44	9.63	0	0	0.87	0.89
$\lvert\mu_4\rvert$	Test	0.98	1.00	0.42	9.17	0	0	0.85	0.87
$\lvert\mu_4\rvert$	Δ%	-0.20	-0.03	-4.58	-4.86			-2.46	-2.39
μ_5	Train	0.97	1.00	0.41	8.80	0	0	0.84	0.85
μ_5	Test	0.97	1.00	0.39	8.18	0	0	0.82	0.83
μ_5	Δ%	-0.39	-0.09	-6.05	-7.07			-2.87	-2.92
$\lvert\mu_5\rvert$	Train	0.94	1.00	0.43	9.81	0	0	0.82	0.88
$\lvert\mu_5\rvert$	Test	0.94	1.00	0.42	9.38	0	0	0.80	0.86
$\lvert\mu_5\rvert$	Δ%	-0.13	-0.01	-3.89	-4.34			-2.46	-2.68
μ_6	Train	0.97	1.00	0.41	8.81	0	0	0.84	0.85
μ_6	Test	0.97	1.00	0.39	8.18	0	0	0.82	0.83
μ_6	Δ%	-0.39	-0.09	-6.04	-7.07			-2.87	-2.93
$\lvert\mu_6\rvert$	Train	0.94	1.00	0.43	9.81	0	0	0.82	0.88
$\lvert\mu_6\rvert$	Test	0.94	1.00	0.42	9.39	0	0	0.80	0.86
$\lvert\mu_6\rvert$	Δ%	-0.13	-0.01	-3.85	-4.28			-2.42	-2.51
μ_7	Train	0.95	0.98	0.29	5.54	0	0	0.90	0.92
μ_7	Test	0.94	0.97	0.26	4.92	0	0	0.86	0.87

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_7	Δ%	-1.03	-1.01	-8.44	-11.22			-3.69	-4.92
$\lvert\mu_7\rvert$	Train	0.95	0.98	0.29	5.54	0	0	0.90	0.92
$\lvert\mu_7\rvert$	Test	0.94	0.97	0.26	4.92	0	0	0.86	0.87
$\lvert\mu_7\rvert$	Δ%	-1.03	-1.01	-8.44	-11.22			-3.69	-4.92
μ_8	Train	0.98	1.00	0.48	10.52	0	0	0.84	0.87
μ_8	Test	0.98	1.00	0.46	10.13	0	0	0.82	0.85
μ_8	Δ%	-0.12	-0.00	-3.32	-3.67			-2.25	-2.02
μ_9	Train	0.96	0.99	0.34	6.88	0	0	0.89	0.91
μ_9	Test	0.95	0.98	0.32	6.26	0	0	0.87	0.88
μ_9	Δ%	-0.61	-0.64	-7.08	-9.05			-2.67	-3.30
$\lvert\mu_9\rvert$	Train	0.96	0.99	0.34	6.94	0	0	0.89	0.91
$\lvert\mu_9\rvert$	Test	0.95	0.98	0.32	6.32	0	0	0.87	0.88
$\lvert\mu_9\rvert$	Δ%	-0.61	-0.63	-7.04	-8.94			-2.65	-3.20
μ_{10}	Train	0.96	0.99	0.35	7.04	0	0	0.89	0.91
μ_{10}	Test	0.96	0.98	0.32	6.42	0	0	0.87	0.88
μ_{10}	Δ%	-0.60	-0.59	-6.93	-8.89			-2.68	-3.28
μ_{11}	Train	0.96	0.99	0.34	6.94	0	0	0.89	0.91
μ_{11}	Test	0.95	0.98	0.32	6.33	0	0	0.87	0.88
μ_{11}	Δ%	-0.61	-0.63	-7.03	-8.89			-2.65	-3.20
μ_{12}	Train	0.97	0.99	0.37	7.78	0	0	0.89	0.90
μ_{12}	Test	0.96	0.99	0.35	7.18	0	0	0.86	0.88
μ_{12}	Δ%	-0.50	-0.20	-6.27	-7.60			-2.54	-2.79
μ_{13}	Train	0.98	1.00	0.42	9.10	0	0	0.85	0.84
μ_{13}	Test	0.97	1.00	0.40	8.45	0	0	0.82	0.82
μ_{13}	Δ%	-0.38	-0.05	-6.01	-7.18			-2.94	-3.32
μ_{14}	Train	0.95	1.00	0.34	7.28	0	0	0.32	0.40
μ_{14}	Test	0.95	1.00	0.34	7.29	0	0	0.29	0.37
μ_{14}	Δ%	-0.01	0.01	0.05	0.08			-7.35	-7.70
μ_{15}	Train	0.97	0.99	0.39	8.34	0	0	0.87	0.87
μ_{15}	Test	0.96	0.99	0.37	7.68	0	0	0.85	0.85
μ_{15}	Δ%	-0.47	-0.05	-6.37	-7.85			-2.45	-2.20
μ_{16}	Train	0.96	0.99	0.37	7.67	0	0	0.88	0.89
μ_{16}	Test	0.96	0.99	0.34	7.05	0	0	0.86	0.86
μ_{16}	Δ%	-0.53	-0.11	-6.65	-8.07			-2.63	-2.72
μ_{17}	Train	0.96	0.99	0.33	6.80	0	0	0.89	0.91
μ_{17}	Test	0.95	0.98	0.31	6.17	0	0	0.87	0.87
μ_{17}	Δ%	-0.60	-0.61	-7.06	-9.19			-2.64	-3.42
μ_{18}	Train	0.94	1.00	0.42	9.55	0	0	0.80	0.85
μ_{18}	Test	0.94	1.00	0.40	9.01	0	0	0.78	0.82
μ_{18}	Δ%	-0.13	-0.06	-4.58	-5.66			-2.77	-3.68
μ_{19}	Train	0.92	1.00	0.34	7.66	0	0	0.33	0.45
μ_{19}	Test	0.92	1.00	0.34	7.63	0	0	0.31	0.42
μ_{19}	Δ%	0.00	-0.00	-0.19	-0.30			-7.25	-6.92
μ_{20}	Train	0.92	0.97	0.24	4.62	0	0	0.89	0.92

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_{20}	Test	0.91	0.95	0.21	3.99	0	0	0.85	0.87
μ_{20}	Δ%	-1.18	-1.62	-10.25	-13.82			-4.19	-5.62
$\lvert\mu_{20}\rvert$	Train	0.92	0.97	0.24	4.64	0	0	0.89	0.92
$\lvert\mu_{20}\rvert$	Test	0.91	0.96	0.22	4.00	0	0	0.85	0.87
$\lvert\mu_{20}\rvert$	Δ%	-1.15	-1.60	-10.24	-13.80			-4.21	-5.72
μ_{21}	Train	0.97	0.99	0.38	7.95	0	0	0.88	0.90
μ_{21}	Test	0.96	0.99	0.35	7.36	0	0	0.86	0.88
μ_{21}	Δ%	-0.44	-0.16	-6.14	-7.44			-2.58	-2.94
μ_{22}	Train	0.84	0.93	0.20	3.93	0	0	0.87	0.90
μ_{22}	Test	0.81	0.89	0.18	3.29	0	0	0.81	0.83
μ_{22}	Δ%	-3.19	-3.90	-11.98	-16.24			-6.54	-8.21
$\lvert\mu_{22}\rvert$	Train	0.84	0.93	0.20	3.94	0	0	0.87	0.90
$\lvert\mu_{22}\rvert$	Test	0.82	0.89	0.18	3.30	0	0	0.81	0.83
$\lvert\mu_{22}\rvert$	Δ%	-3.19	-3.87	-11.99	-16.23			-6.55	-8.30
μ_{23}	Train	0.98	1.00	0.42	9.04	0	0	0.88	0.90
μ_{23}	Test	0.97	1.00	0.40	8.51	0	0	0.86	0.88
μ_{23}	Δ%	-0.22	-0.02	-5.02	-5.83			-2.48	-2.24
μ_{24}	Train	0.97	0.99	0.38	7.80	0	0	0.88	0.89
μ_{24}	Test	0.96	0.99	0.35	7.18	0	0	0.85	0.86
μ_{24}	Δ%	-0.47	-0.19	-6.40	-7.86			-2.62	-2.65
$\lvert\mu_{24}\rvert$	Train	0.97	1.00	0.40	8.52	0	0	0.89	0.91
$\lvert\mu_{24}\rvert$	Test	0.97	1.00	0.38	7.98	0	0	0.86	0.89
$\lvert\mu_{24}\rvert$	Δ%	-0.33	-0.04	-5.49	-6.40			-2.45	-2.12

Table J.1: μ-RANKING Stopwords Excluded Not Discounted

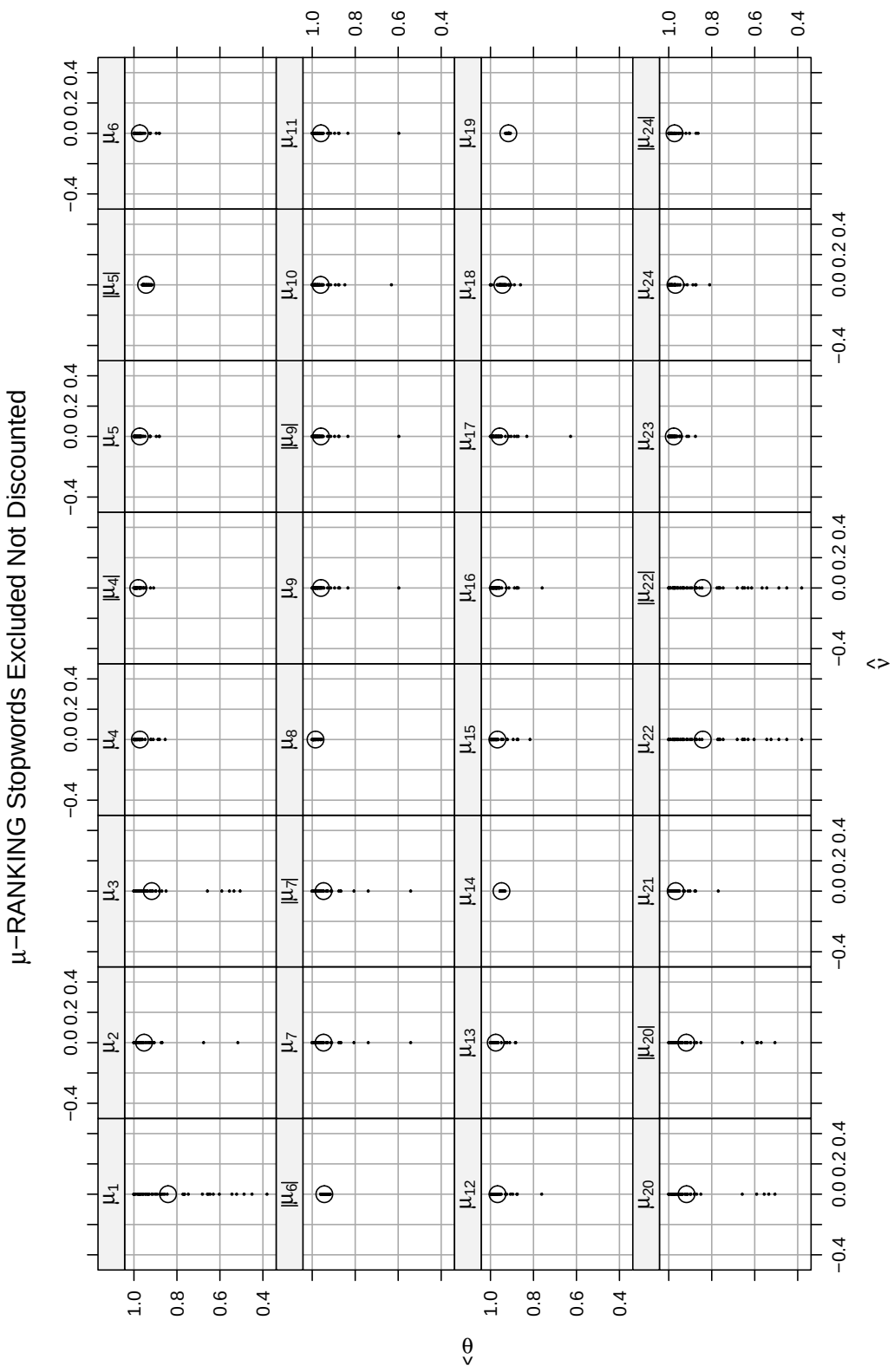

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
|μ6|
μ7
|μ7|
μ8
μ12
μ13
μ14
μ15
μ20
|μ20|
μ21
μ22
θ̂<
V̂<
−0.4
0.0
0.2
0.4
0.6
0.8
1.0

Figure J.1

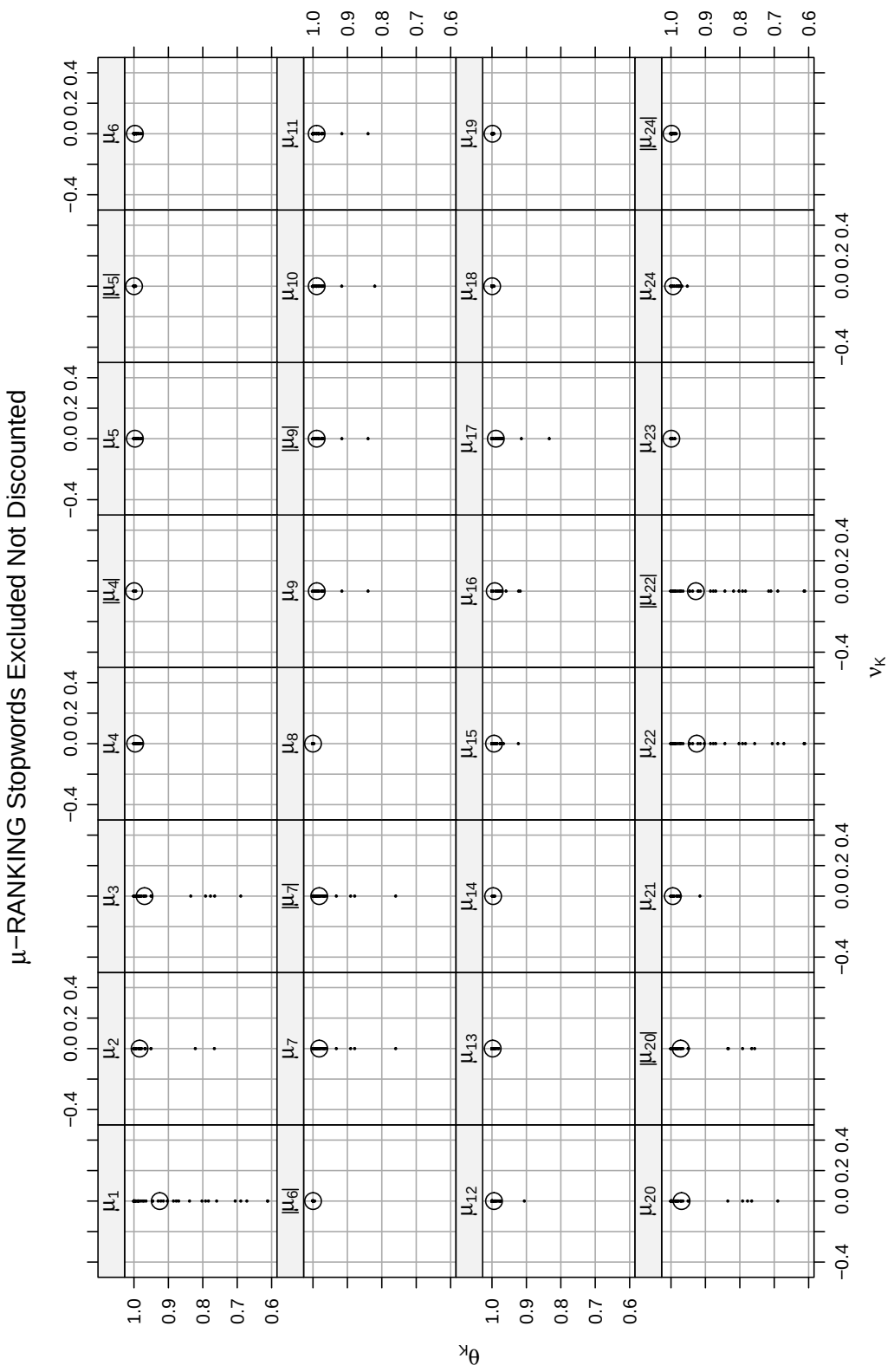

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
−0.4
0.0
0.2
0.4
0.6
0.7
0.8
0.9
1.0
θK
νK

Figure J.2

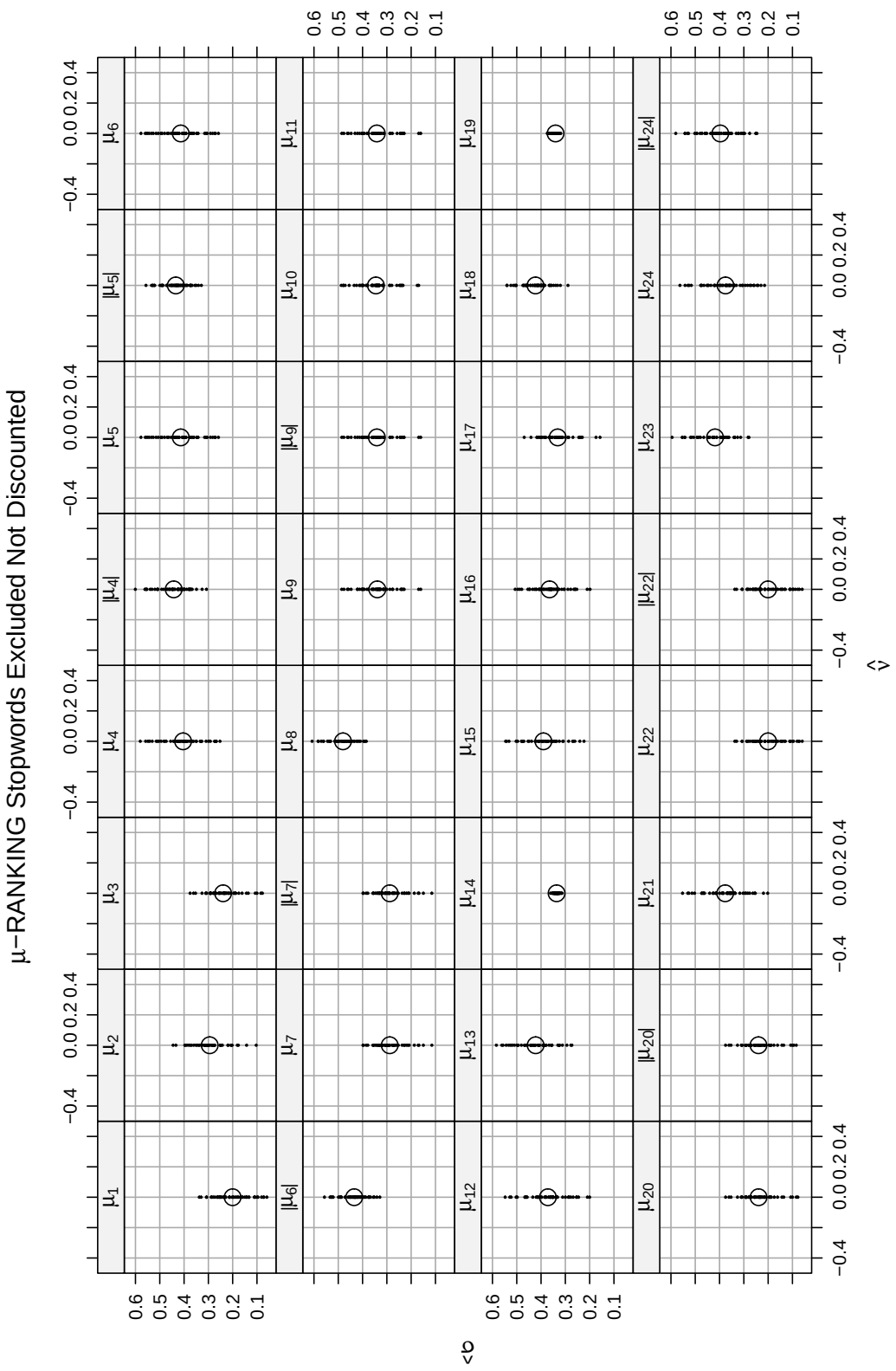

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
−0.4
0.0
0.2
0.4
0.6
0.5
0.4
0.3
0.2
0.1
σ̂
ν̂

Figure J.3

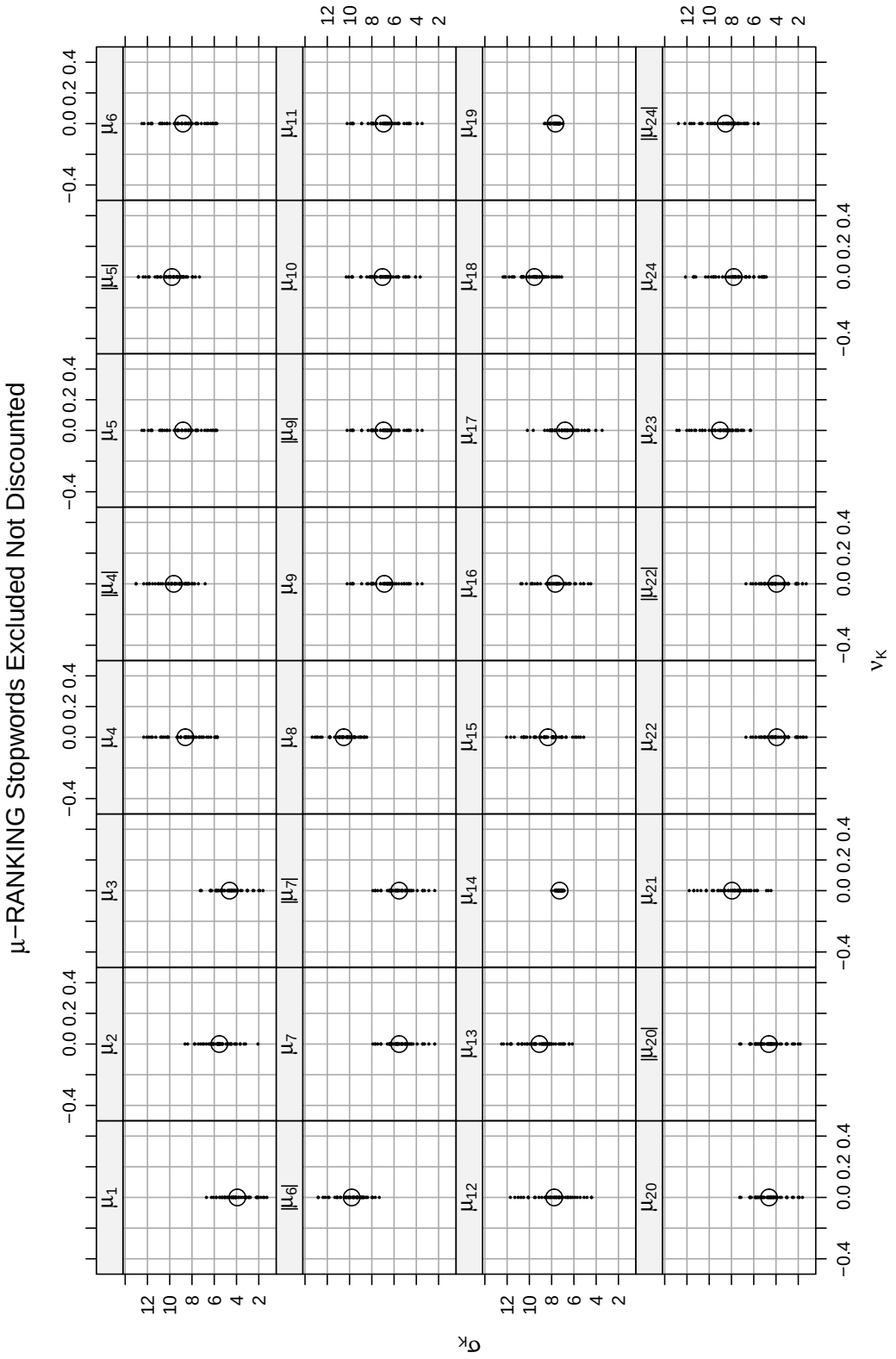

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
|μ6|
μ7
|μ7|
μ8
μ12
μ13
μ14
μ15
|μ20|
μ21
μ22
|μ22|
μ20
μ23
μ24
|μ24|
μ5
|μ5|
μ6
|μ4|
μ9
|μ9|
μ10
μ11
μ16
μ17
μ18
μ19
−0.4
0.0
0.2
0.4
2
4
6
8
10
12
σK
νK

Figure J.4

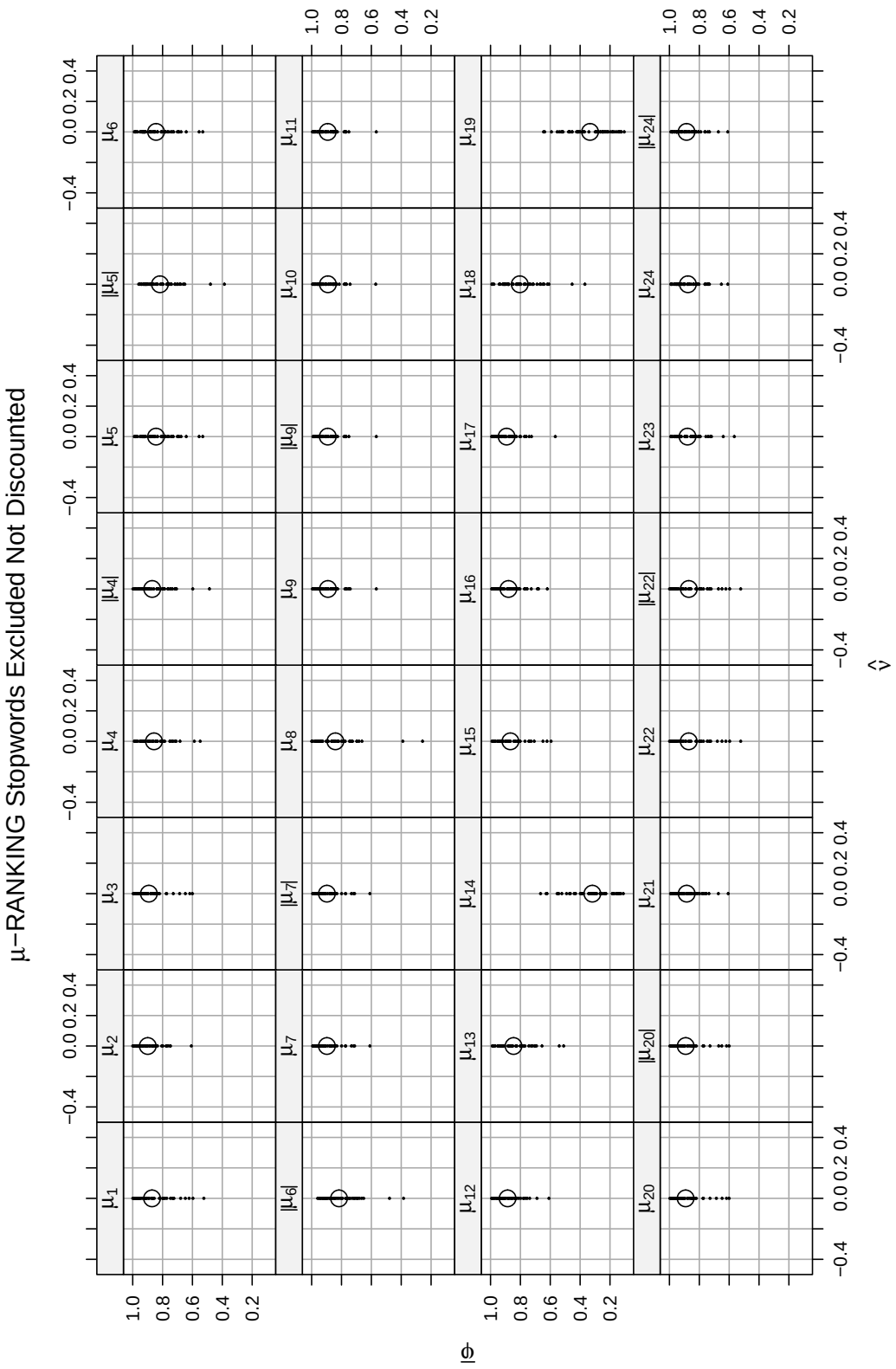

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
μ5
|μ6|
μ6
|μ5|
μ11
μ10
|μ9|
μ9
|μ8|
μ8
|μ7|
μ7
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
−0.4
0.0
0.2
0.4
0.2
0.4
0.6
0.8
1.0
Φ
V̂

Figure J.5

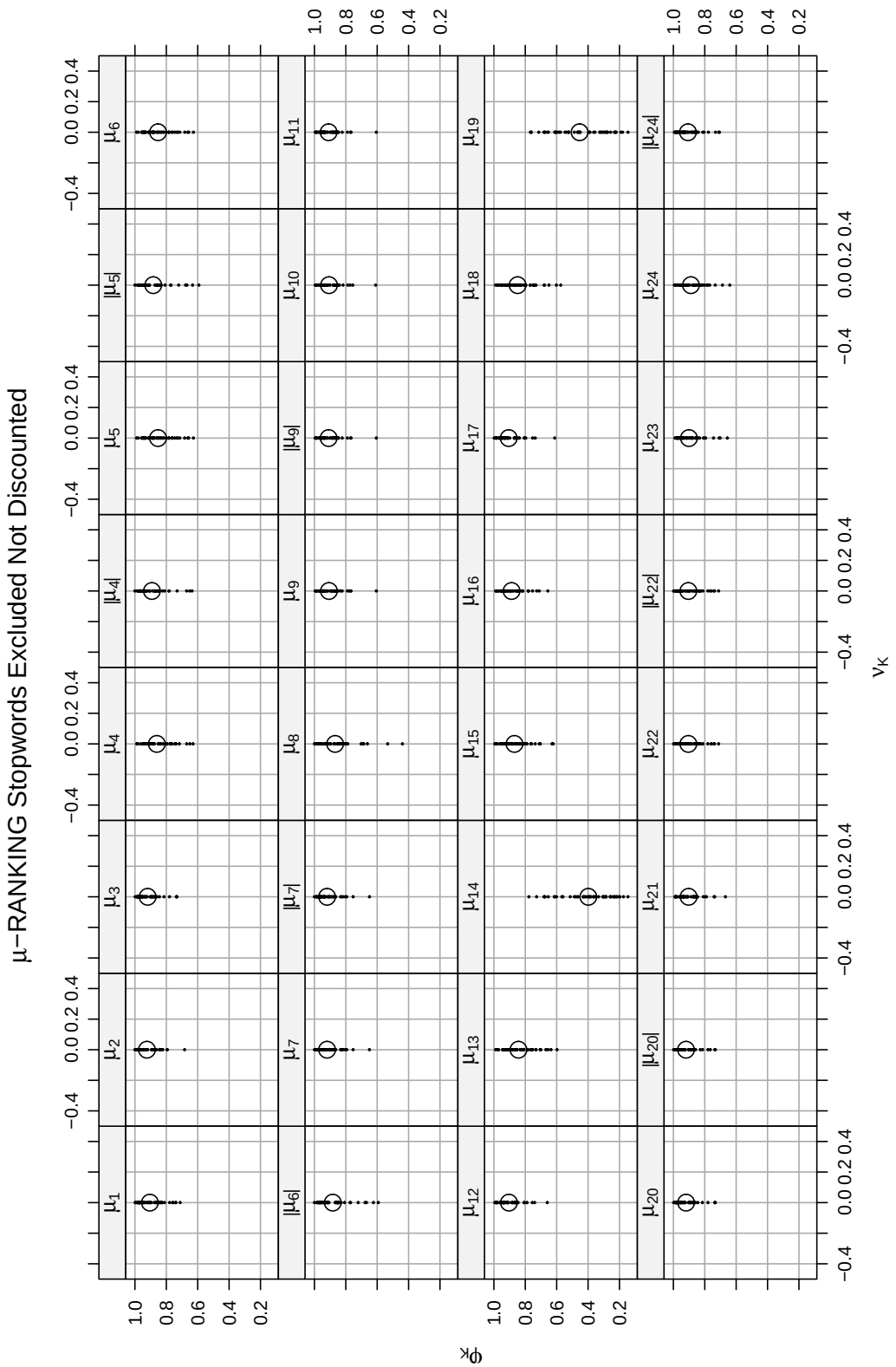
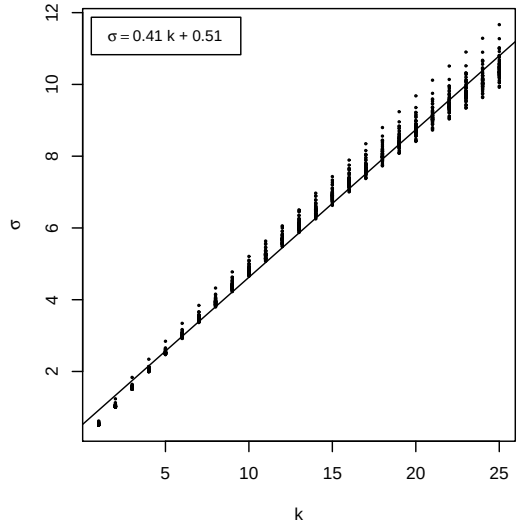

μ−RANKING Stopwords Excluded Not Discounted
μ1
μ2
μ3
μ4
μ5
μ6
|μ6|
μ7
|μ7|
μ8
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
−0.4
0.0
0.2
0.4
0.2
0.4
0.6
0.8
1.0
φK
νK

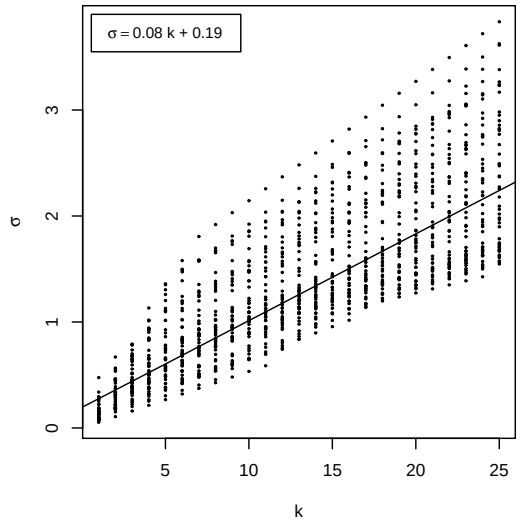
Figure J.6

Appendix K

μ-RANKING Results Noise Growth Rate



(a) Noise Growth Rate for μ_8 on Topic Fuel



(b) Noise Growth Rate for μ_9 on Topic Fuel

Figure K.1: μ-RANKING Noise Growth Rate Examples

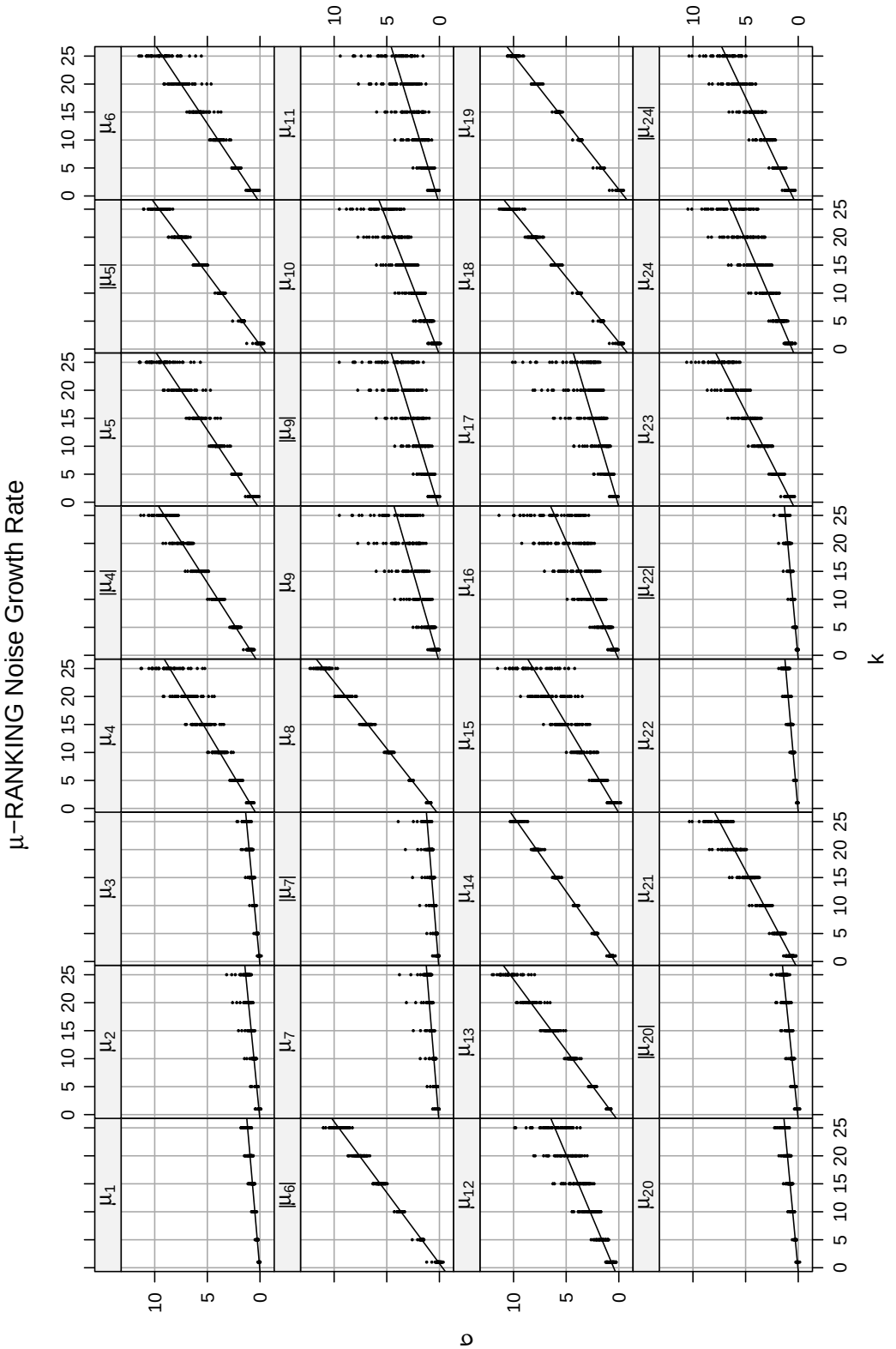

μ−RANKING Noise Growth Rate
σ
k
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|

Figure K.2

Parameter	Median Intercept	Median Slope	σ_K
μ_0	-0.04	0.11	2.65
μ_1	0.03	0.05	1.18
μ_2	0.05	0.05	1.35
μ_3	0.04	0.05	1.28
μ_4	0.65	0.32	8.57
$\lvert\mu_4\rvert$	0.61	0.34	9.09
μ_5	0.45	0.35	9.27
$\lvert\mu_5\rvert$	-0.30	0.40	9.58
μ_6	0.44	0.35	9.27
$\lvert\mu_6\rvert$	-0.30	0.39	9.57
μ_7	0.09	0.04	1.18
$\lvert\mu_7\rvert$	0.10	0.04	1.16
μ_8	0.53	0.42	11.00
μ_9	0.20	0.15	4.06
$\lvert\mu_9\rvert$	0.22	0.16	4.31
μ_{10}	0.20	0.21	5.41
μ_{11}	0.24	0.16	4.31
μ_{12}	0.49	0.22	6.08
μ_{13}	0.50	0.39	10.32
μ_{14}	0.33	0.37	9.67
μ_{15}	0.25	0.32	8.13
μ_{16}	0.18	0.24	6.09
μ_{17}	0.12	0.16	4.06
μ_{18}	-0.53	0.43	10.20
μ_{19}	-0.49	0.42	9.93
μ_{20}	0.03	0.05	1.28
$\lvert\mu_{20}\rvert$	0.05	0.05	1.41
μ_{21}	0.42	0.28	7.48
μ_{22}	0.03	0.05	1.19
$\lvert\mu_{22}\rvert$	0.04	0.05	1.23
μ_{23}	0.62	0.27	7.39
μ_{24}	0.58	0.23	6.28
$\lvert\mu_{24}\rvert$	0.56	0.25	6.86

Table K.1: μ-RANKING Noise Growth Rate

Appendix L

μ-RANKING Results Robustness

Parameter	Mean	Median	SD	IQR	CV	ν_K
μ_0	1.06	1.00	0.24	0.00	0.22	0.81
μ_1	10.26	3.23	13.66	13.00	1.35	0.16
μ_2	11.76	3.35	15.42	16.45	1.33	0.13
μ_3	10.74	3.31	14.31	14.03	1.36	0.14
μ_4	15.34	5.84	17.73	27.02	1.19	4.56
$\lvert\mu_4\rvert$	13.85	4.40	17.50	23.35	1.31	4.12
μ_5	16.68	6.70	18.27	31.33	1.12	8.84
$\lvert\mu_5\rvert$	17.10	6.96	19.28	33.84	1.17	9.61
μ_6	16.70	6.90	18.27	31.38	1.12	8.86
$\lvert\mu_6\rvert$	17.24	7.45	19.29	33.91	1.16	9.58
μ_7	12.33	3.50	15.87	18.37	1.31	0.18
$\lvert\mu_7\rvert$	12.33	3.50	15.87	18.37	1.31	0.18
μ_8	26.17	26.75	21.01	43.67	0.83	14.17
μ_9	13.69	3.87	17.12	22.44	1.27	0.30
$\lvert\mu_9\rvert$	11.90	2.62	16.34	16.72	1.40	0.36
μ_{10}	13.88	3.74	17.34	23.01	1.26	0.38
μ_{11}	13.64	3.77	17.11	22.41	1.27	0.32
μ_{12}	14.23	3.91	17.62	24.70	1.26	0.61
μ_{13}	17.86	8.18	18.79	33.53	1.09	10.16
μ_{14}	30.39	39.88	20.61	42.00	0.68	9.34
μ_{15}	16.14	7.86	17.95	27.50	1.18	4.10
μ_{16}	14.76	4.90	17.75	25.75	1.23	1.66
μ_{17}	13.58	3.84	17.07	22.27	1.28	0.54
μ_{18}	24.10	21.79	20.68	42.76	0.89	14.28
μ_{19}	44.15	50.00	14.04	2.92	0.33	13.99
μ_{20}	10.74	3.31	14.31	14.03	1.36	0.14
$\lvert\mu_{20}\rvert$	9.64	2.51	13.71	10.95	1.45	0.19
μ_{21}	14.48	4.31	17.69	25.09	1.24	0.80
μ_{22}	10.42	3.38	13.78	13.54	1.35	0.15
$\lvert\mu_{22}\rvert$	9.42	2.63	13.26	11.05	1.44	0.20
μ_{23}	15.45	5.22	18.00	27.91	1.20	1.94
μ_{24}	14.86	4.44	17.89	26.68	1.23	0.98
$\lvert\mu_{24}\rvert$	12.88	2.91	17.05	20.32	1.36	1.18

Table L.1: μ-RANKING Stopwords Included Robustness

Appendix M

μ-GREEDY Results Stopwords Included Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_0	Train	0.15	0.28	0.01	0.34	0.04	0.90	0.06	0.10
μ_0	Test	0.14	0.26	0.01	0.30	0.04	0.90	0.03	0.05
μ_0	Δ%	-8.32	-7.40	-9.40	-9.55	0.00	0.00	-46.25	-46.72
μ_1	Train	0.86	0.96	0.15	2.85	0.00	0.11	0.89	0.91
μ_1	Test	0.75	0.82	0.13	2.27	0.00	0.11	0.77	0.78
μ_1	Δ%	-12.64	-14.47	-15.24	-20.15	0.00	0.00	-13.03	-14.50
μ_2	Train	0.98	1.00	0.21	3.82	0.01	0.15	0.91	0.92
μ_2	Test	0.95	0.97	0.19	3.22	0.01	0.15	0.87	0.86
μ_2	Δ%	-3.04	-2.52	-11.06	-15.75	0.00	0.00	-4.45	-6.27
μ_3	Train	0.92	0.98	0.18	3.32	0.00	0.11	0.90	0.92
μ_3	Test	0.86	0.90	0.16	2.73	0.00	0.11	0.83	0.83
μ_3	Δ%	-6.81	-7.56	-12.63	-17.58	0.00	0.00	-8.10	-10.13
μ_4	Train	0.98	1.00	0.22	4.04	0.00	0.12	0.89	0.91
μ_4	Test	0.96	0.98	0.19	3.45	0.00	0.12	0.86	0.86
μ_4	Δ%	-2.49	-1.77	-10.64	-14.60	0.00	0.00	-3.70	-5.31
$\lvert\mu_4\rvert$	Train	0.98	1.00	0.22	4.21	0.01	0.18	0.88	0.90
$\lvert\mu_4\rvert$	Test	0.96	0.98	0.20	3.61	0.01	0.18	0.85	0.86
$\lvert\mu_4\rvert$	Δ%	-2.34	-1.62	-10.35	-14.15	0.00	0.00	-3.77	-5.22
μ_5	Train	0.95	0.99	0.28	5.67	0.09	1.94	0.84	0.88
μ_5	Test	0.94	0.99	0.26	5.04	0.09	1.94	0.81	0.83
μ_5	Δ%	-0.79	-0.63	-8.39	-11.17	0.00	0.00	-3.56	-4.89
$\lvert\mu_5\rvert$	Train	0.88	1.00	0.41	10.47	0.41	9.48	0.73	0.84
$\lvert\mu_5\rvert$	Test	0.88	1.00	0.39	10.08	0.41	9.48	0.71	0.81
$\lvert\mu_5\rvert$	Δ%	-0.13	0.00	-3.27	-3.75	0.00	0.00	-2.96	-2.82
μ_6	Train	0.95	0.99	0.28	5.68	0.09	1.95	0.84	0.88
μ_6	Test	0.94	0.99	0.26	5.05	0.09	1.95	0.81	0.83
μ_6	Δ%	-0.79	-0.63	-8.35	-11.12	0.00	0.00	-3.56	-4.88
$\lvert\mu_6\rvert$	Train	0.89	1.00	0.40	10.26	0.40	9.22	0.70	0.82
$\lvert\mu_6\rvert$	Test	0.88	1.00	0.39	9.86	0.40	9.22	0.68	0.80
$\lvert\mu_6\rvert$	Δ%	-0.14	-0.00	-3.38	-3.86	0.00	0.00	-3.48	-3.28
μ_7	Train	0.97	0.99	0.19	3.51	0.01	0.15	0.91	0.91
μ_7	Test	0.92	0.95	0.17	2.92	0.01	0.15	0.86	0.84
μ_7	Δ%	-4.79	-4.58	-11.88	-16.75	0.00	0.00	-5.32	-7.67
$\lvert\mu_7\rvert$	Train	0.97	0.99	0.19	3.51	0.01	0.15	0.91	0.91

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_7\rvert$	Test	0.92	0.95	0.17	2.92	0.01	0.15	0.86	0.84
$\lvert\mu_7\rvert$	Δ%	-4.79	-4.58	-11.88	-16.75	0.00	0.00	-5.32	-7.67
μ_8	Train	0.98	0.99	0.27	5.95	0.02	0.76	0.84	0.85
μ_8	Test	0.96	0.99	0.25	5.35	0.02	0.76	0.81	0.83
μ_8	Δ%	-1.25	-0.48	-8.10	-10.03	0.00	0.00	-3.32	-3.17
μ_9	Train	0.97	1.00	0.19	3.58	0.00	0.15	0.90	0.91
μ_9	Test	0.93	0.96	0.17	2.98	0.00	0.15	0.86	0.84
μ_9	Δ%	-4.34	-3.83	-12.02	-16.77	0.00	0.00	-5.10	-7.65
$\lvert\mu_9\rvert$	Train	0.97	0.99	0.19	3.57	0.01	0.15	0.90	0.90
$\lvert\mu_9\rvert$	Test	0.92	0.95	0.17	2.96	0.01	0.15	0.85	0.84
$\lvert\mu_9\rvert$	Δ%	-4.33	-3.79	-12.29	-17.11	0.00	0.00	-5.10	-7.63
μ_{10}	Train	0.97	1.00	0.20	3.60	0.00	0.14	0.90	0.91
μ_{10}	Test	0.93	0.96	0.17	3.00	0.00	0.14	0.86	0.84
μ_{10}	Δ%	-4.25	-3.69	-11.96	-16.69	0.00	0.00	-5.01	-7.44
μ_{11}	Train	0.97	1.00	0.19	3.59	0.00	0.15	0.90	0.91
μ_{11}	Test	0.93	0.96	0.17	2.98	0.00	0.15	0.86	0.84
μ_{11}	Δ%	-4.36	-3.85	-12.02	-16.79	0.00	0.00	-5.10	-7.65
μ_{12}	Train	0.98	1.00	0.20	3.72	0.00	0.14	0.90	0.91
μ_{12}	Test	0.94	0.97	0.18	3.12	0.00	0.14	0.86	0.86
μ_{12}	Δ%	-3.54	-2.96	-11.45	-16.15	0.00	0.00	-4.30	-6.02
μ_{13}	Train	0.98	1.00	0.23	4.40	0.01	0.20	0.89	0.91
μ_{13}	Test	0.96	0.98	0.20	3.77	0.01	0.20	0.86	0.86
μ_{13}	Δ%	-2.40	-1.55	-10.61	-14.25	0.00	0.00	-3.74	-5.39
μ_{14}	Train	0.92	0.98	0.21	4.29	0.12	2.84	0.40	0.53
μ_{14}	Test	0.91	0.98	0.22	4.45	0.12	2.84	0.33	0.44
μ_{14}	Δ%	-0.52	-0.44	2.60	3.76	0.00	0.00	-17.68	-15.71
μ_{15}	Train	0.98	1.00	0.21	4.09	0.00	0.13	0.90	0.92
μ_{15}	Test	0.95	0.98	0.19	3.48	0.00	0.13	0.87	0.87
μ_{15}	Δ%	-2.80	-2.00	-10.97	-14.85	0.00	0.00	-3.99	-5.16
μ_{16}	Train	0.97	0.99	0.20	3.79	0.00	0.16	0.90	0.91
μ_{16}	Test	0.93	0.96	0.18	3.18	0.00	0.16	0.86	0.85
μ_{16}	Δ%	-3.66	-2.97	-11.81	-16.09	0.00	0.00	-4.78	-6.39
μ_{17}	Train	0.96	0.98	0.19	3.55	0.01	0.19	0.91	0.91
μ_{17}	Test	0.92	0.95	0.17	2.96	0.01	0.19	0.86	0.85
μ_{17}	Δ%	-3.98	-3.58	-11.80	-16.73	0.00	0.00	-5.13	-7.09
μ_{18}	Train	0.88	1.00	0.41	10.78	0.59	14.48	0.67	0.74
μ_{18}	Test	0.88	1.00	0.40	10.48	0.59	14.48	0.66	0.72
μ_{18}	Δ%	0.07	0.02	-2.33	-2.82	0.00	0.00	-2.57	-2.58
μ_{19}	Train	0.84	1.00	0.36	9.46	0.59	14.00	0.27	0.47
μ_{19}	Test	0.84	1.00	0.36	9.47	0.59	14.00	0.23	0.42
μ_{19}	Δ%	0.03	0.00	0.02	0.05	0.00	0.00	-13.42	-10.88
μ_{20}	Train	0.92	0.98	0.18	3.32	0.00	0.11	0.90	0.92
μ_{20}	Test	0.86	0.90	0.16	2.73	0.00	0.11	0.83	0.83
μ_{20}	Δ%	-6.81	-7.56	-12.63	-17.58	0.00	0.00	-8.10	-10.13

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_{20}\rvert$	Train	0.92	0.97	0.18	3.29	0.00	0.11	0.90	0.92
$\lvert\mu_{20}\rvert$	Test	0.85	0.90	0.15	2.70	0.00	0.11	0.82	0.82
$\lvert\mu_{20}\rvert$	Δ%	-6.84	-7.57	-12.91	-17.93	0.00	0.00	-8.15	-10.12
μ_{21}	Train	0.98	1.00	0.20	3.74	0.00	0.14	0.90	0.92
μ_{21}	Test	0.95	0.97	0.18	3.14	0.00	0.14	0.86	0.86
μ_{21}	Δ%	-3.32	-2.73	-11.40	-16.02	0.00	0.00	-4.25	-5.79
μ_{22}	Train	0.86	0.96	0.15	2.85	0.00	0.11	0.89	0.91
μ_{22}	Test	0.75	0.82	0.13	2.27	0.00	0.11	0.77	0.78
μ_{22}	Δ%	-12.64	-14.47	-15.24	-20.15	0.00	0.00	-13.03	-14.50
$\lvert\mu_{22}\rvert$	Train	0.85	0.95	0.15	2.83	0.00	0.11	0.88	0.90
$\lvert\mu_{22}\rvert$	Test	0.75	0.81	0.12	2.24	0.00	0.11	0.76	0.77
$\lvert\mu_{22}\rvert$	Δ%	-12.71	-14.47	-15.63	-20.58	0.00	0.00	-13.14	-14.46
μ_{23}	Train	0.98	1.00	0.21	3.93	0.00	0.14	0.89	0.91
μ_{23}	Test	0.96	0.98	0.19	3.32	0.00	0.14	0.86	0.86
μ_{23}	Δ%	-2.58	-1.96	-11.05	-15.42	0.00	0.00	-3.82	-5.53
μ_{24}	Train	0.98	1.00	0.21	3.81	0.00	0.15	0.90	0.91
μ_{24}	Test	0.95	0.98	0.19	3.21	0.00	0.15	0.87	0.86
μ_{24}	Δ%	-2.80	-2.18	-11.16	-15.73	0.00	0.00	-3.73	-5.24
$\lvert\mu_{24}\rvert$	Train	0.98	1.00	0.21	3.83	0.01	0.16	0.90	0.91
$\lvert\mu_{24}\rvert$	Test	0.95	0.98	0.19	3.22	0.01	0.16	0.86	0.87
$\lvert\mu_{24}\rvert$	Δ%	-2.78	-2.17	-11.36	-15.82	0.00	0.00	-3.65	-5.13

Table M.1: μ-GREEDY Stopwords Included Not Discounted

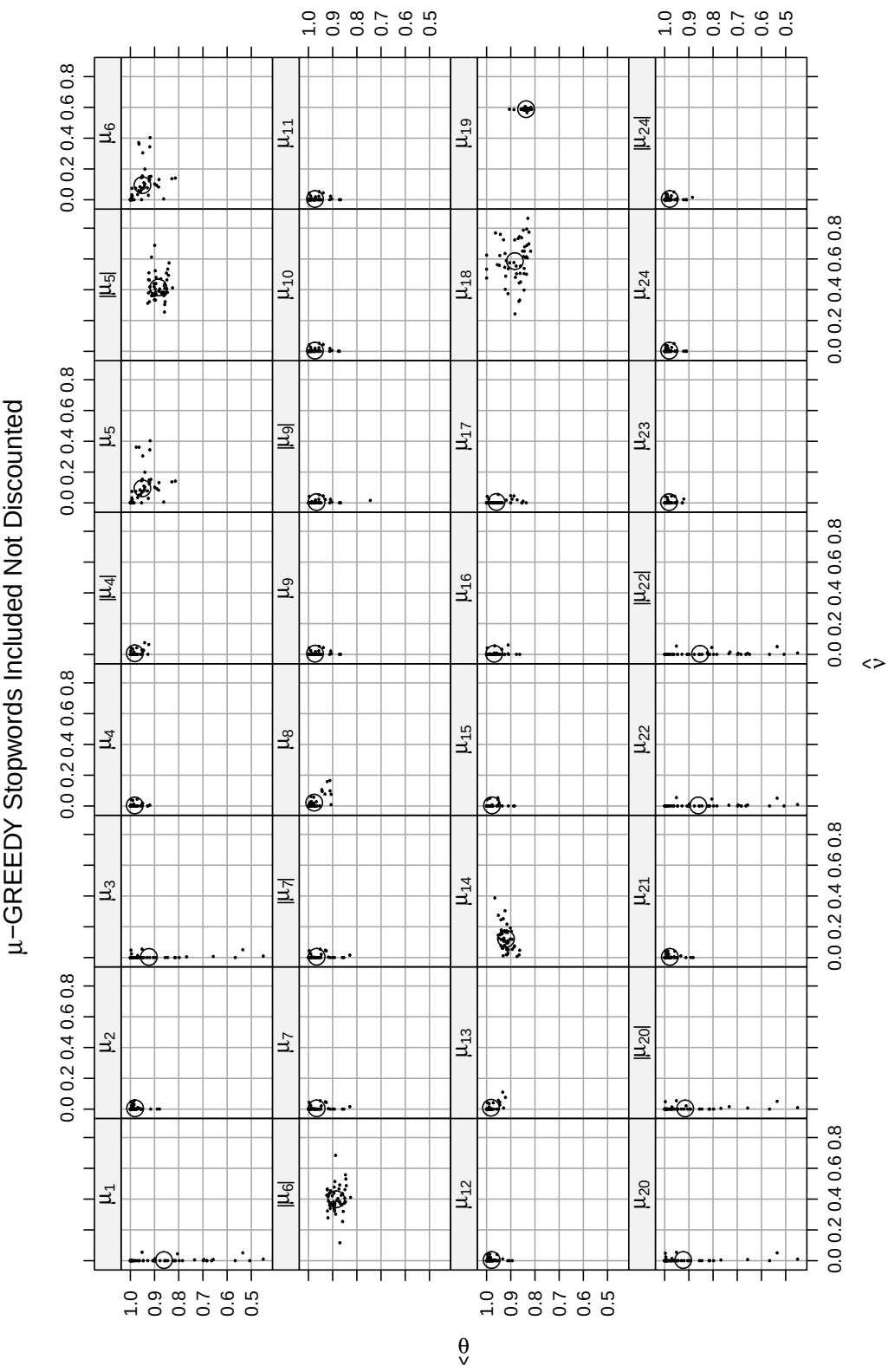


Figure M.1

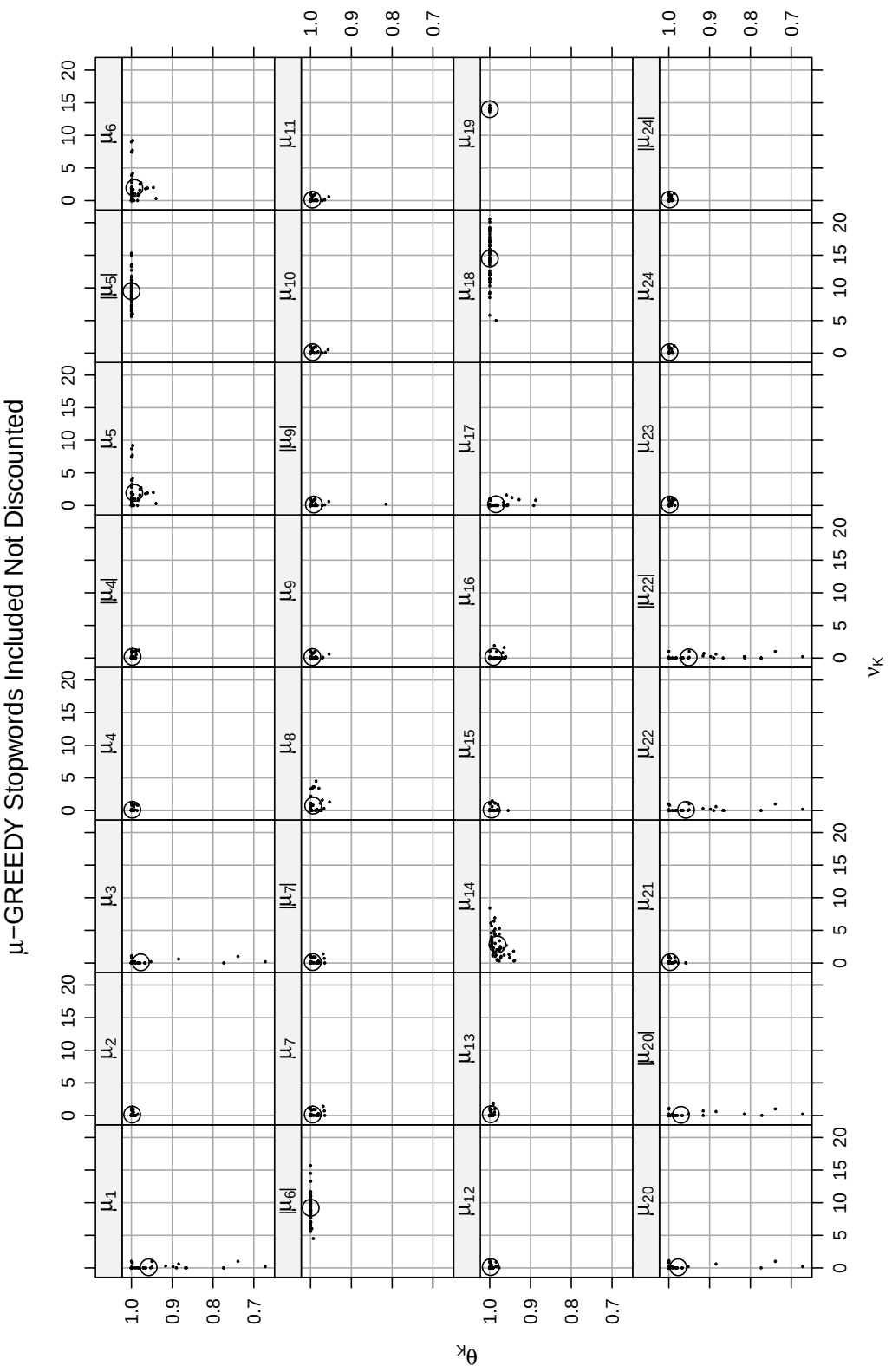

μ−GREEDY Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
θκ
νK

Figure M.2

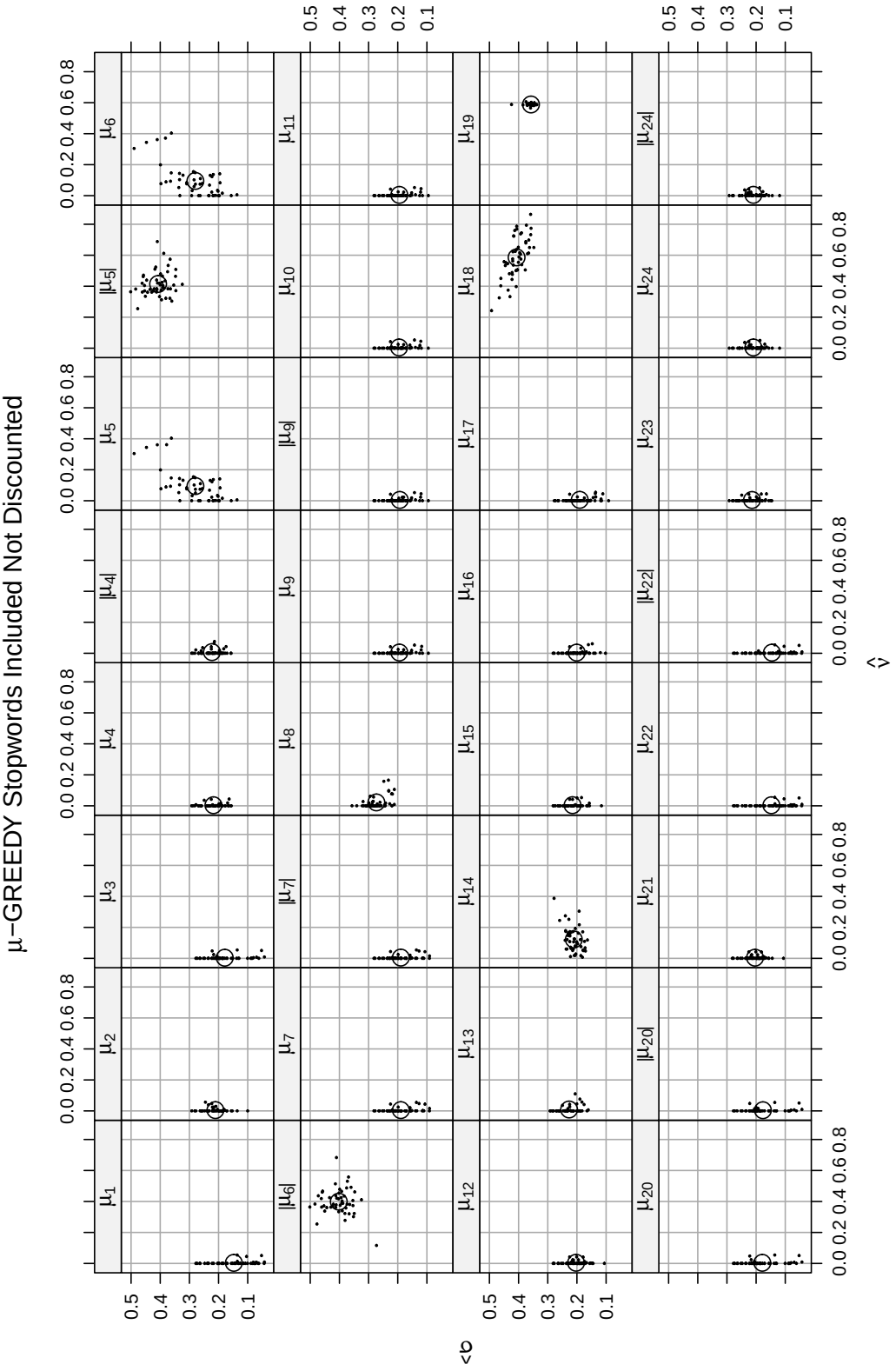

μ−GREEDY Stopwords Included Not Discounted
σ²
V̂
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|

Figure M.3

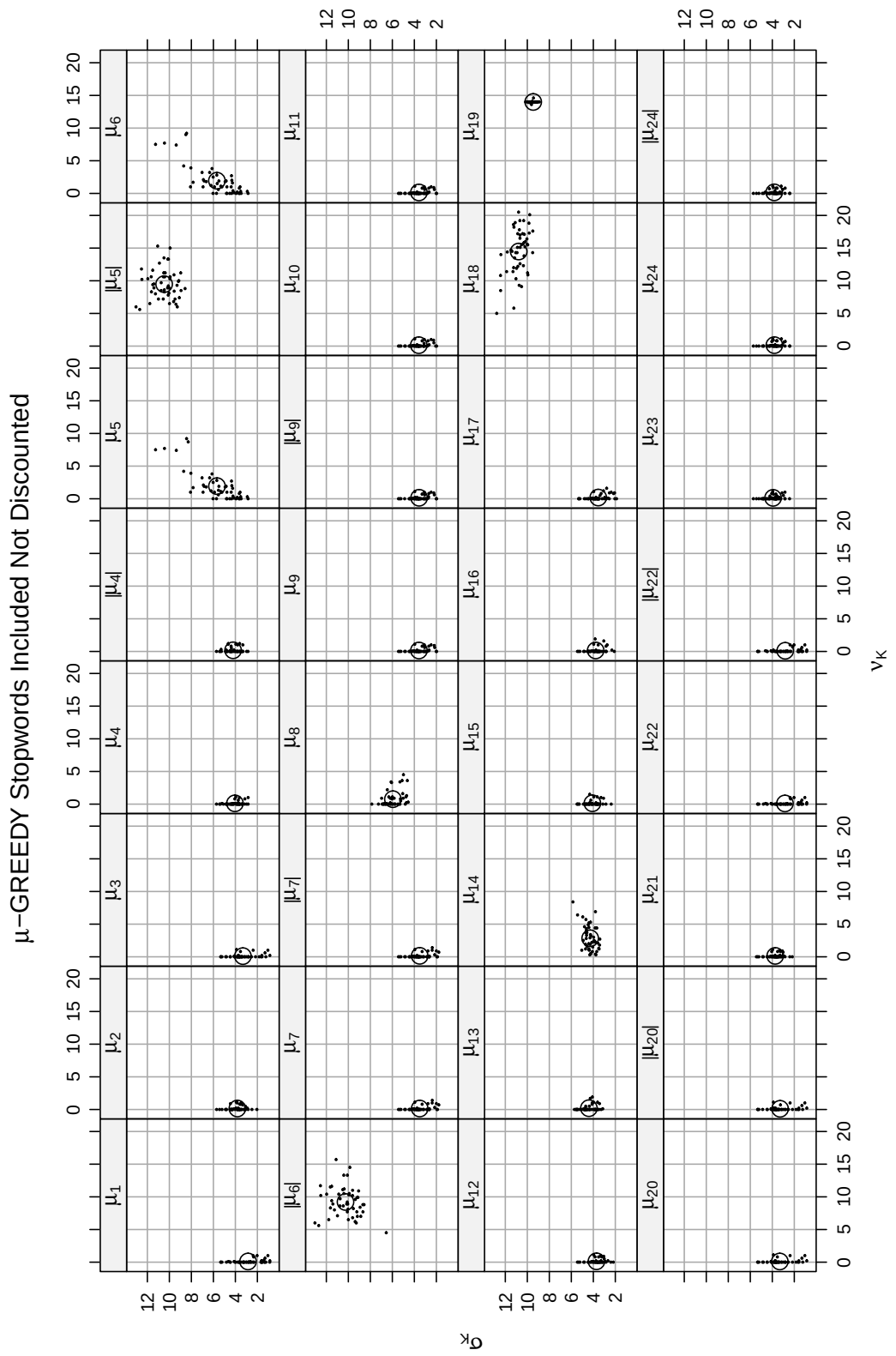

μ−GREEDY Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
σK
νK

Figure M.4

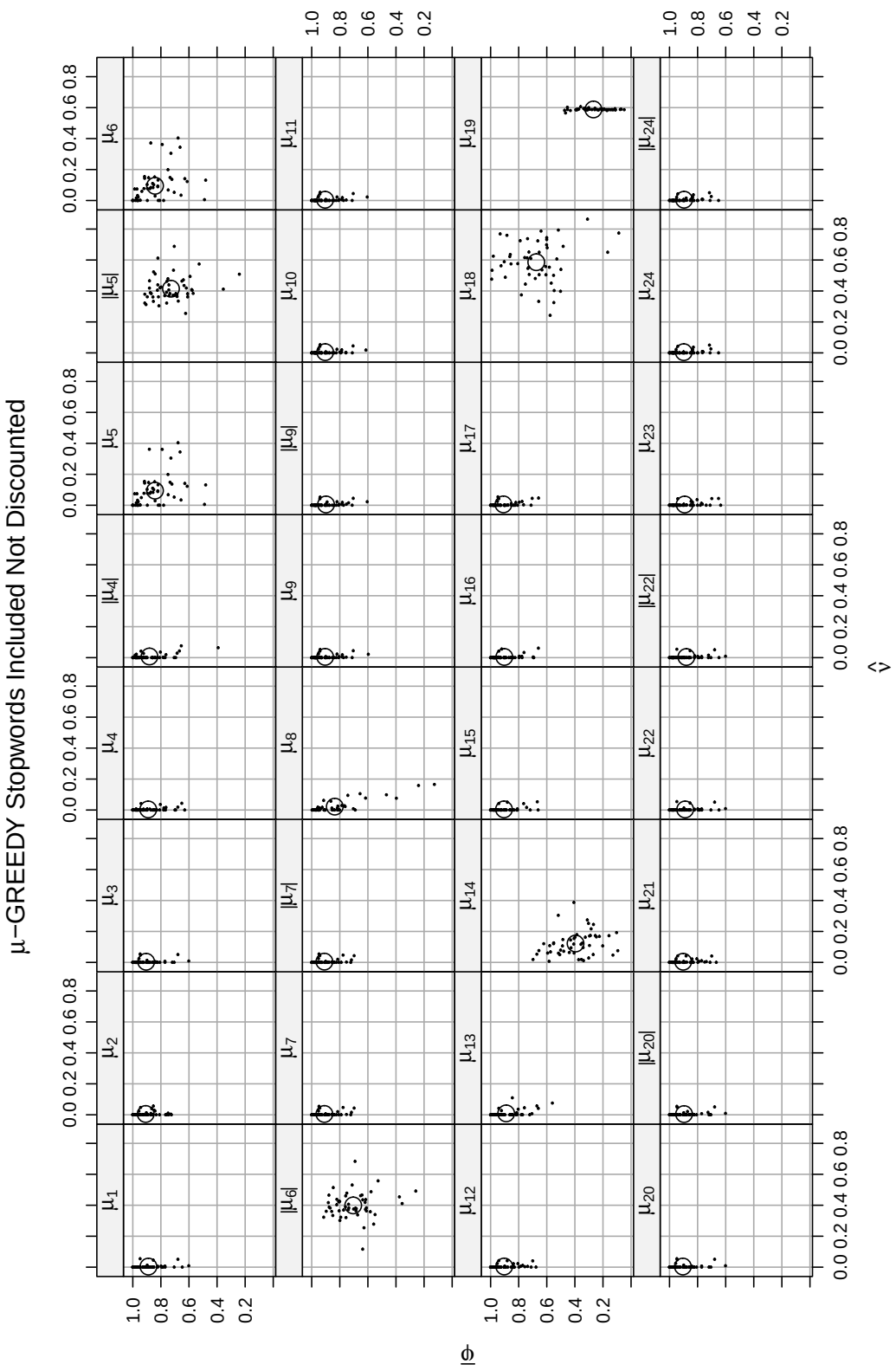


Figure M.5

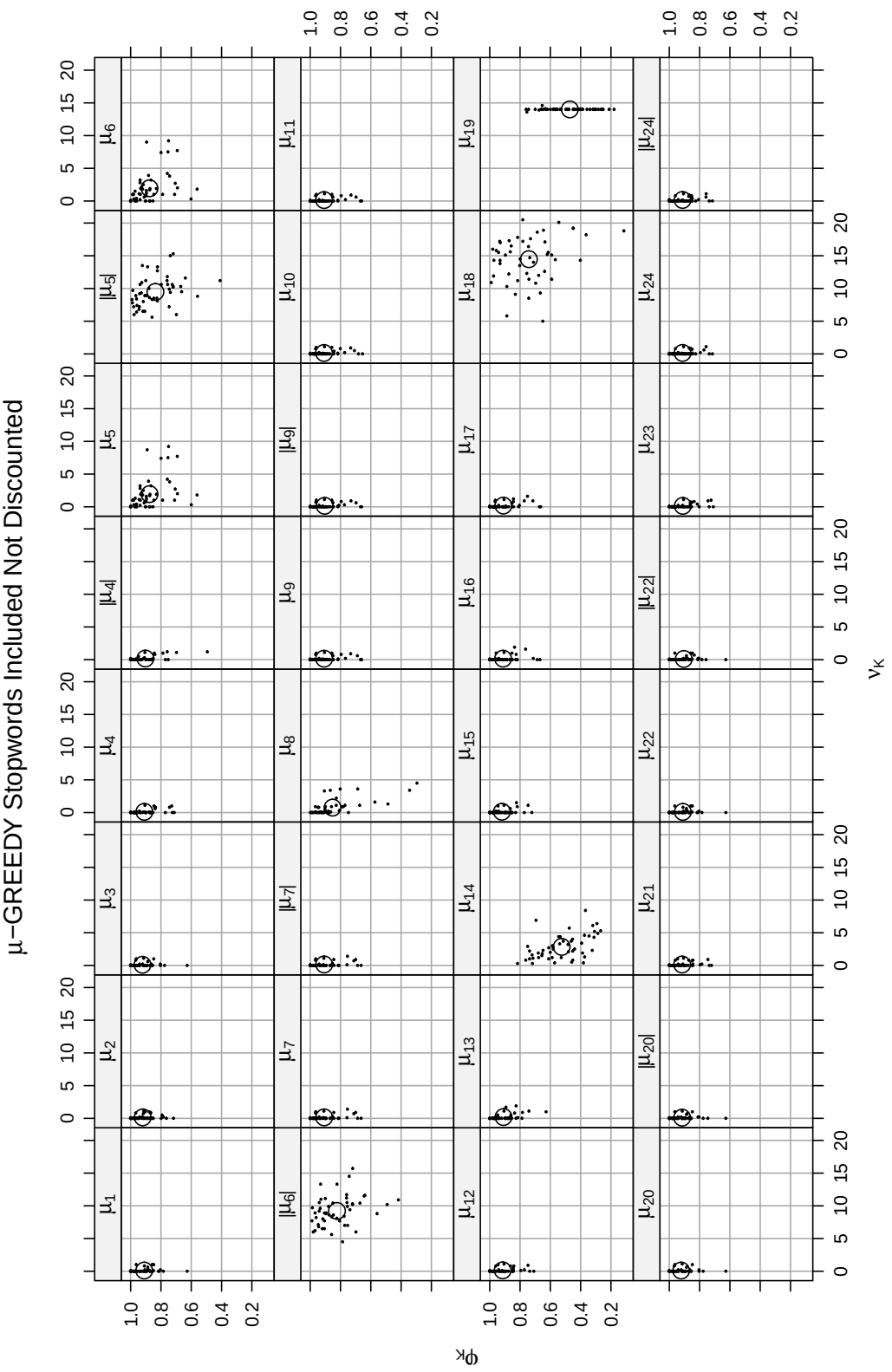

μ−GREEDY Stopwords Included Not Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
φK
νK

Figure M.6

Appendix N

μ-GREEDY Results Stopwords Included Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
μ_0	Train	0.13	0.23	0.01	0.27	0.04	0.90	0.05	0.10
μ_0	Test	0.11	0.21	0.01	0.24	0.04	0.90	0.03	0.05
μ_0	Δ%	-9.80	-9.49	-11.27	-11.62	0.00	0.00	-45.54	-46.24
μ_1	Train	0.86	0.96	0.15	2.84	0.00	0.11	0.89	0.91
μ_1	Test	0.75	0.82	0.13	2.27	0.00	0.11	0.77	0.78
μ_1	Δ%	-12.58	-14.52	-15.17	-20.15	0.00	0.00	-12.98	-14.44
μ_2	Train	0.98	1.00	0.21	3.81	0.01	0.15	0.91	0.92
μ_2	Test	0.95	0.97	0.19	3.21	0.01	0.15	0.87	0.86
μ_2	Δ%	-3.04	-2.60	-11.00	-15.75	0.00	0.00	-4.44	-6.29
μ_3	Train	0.92	0.98	0.18	3.31	0.00	0.11	0.90	0.92
μ_3	Test	0.86	0.90	0.16	2.73	0.00	0.11	0.83	0.83
μ_3	Δ%	-6.77	-7.61	-12.57	-17.58	0.00	0.00	-8.06	-10.05
μ_4	Train	0.98	1.00	0.22	4.03	0.00	0.12	0.89	0.91
μ_4	Test	0.96	0.98	0.19	3.44	0.00	0.12	0.86	0.86
μ_4	Δ%	-2.48	-1.76	-10.60	-14.55	0.00	0.00	-3.68	-5.28
$\lvert\mu_4\rvert$	Train	0.98	1.00	0.22	4.19	0.01	0.18	0.88	0.90
$\lvert\mu_4\rvert$	Test	0.96	0.98	0.20	3.59	0.01	0.18	0.85	0.85
$\lvert\mu_4\rvert$	Δ%	-2.32	-1.60	-10.34	-14.15	0.00	0.00	-3.76	-5.15
μ_5	Train	0.94	0.99	0.25	5.08	0.09	1.94	0.85	0.88
μ_5	Test	0.93	0.98	0.23	4.47	0.09	1.94	0.82	0.84
μ_5	Δ%	-1.14	-0.78	-9.07	-12.02	0.00	0.00	-3.46	-4.84
$\lvert\mu_5\rvert$	Train	0.86	1.00	0.27	6.89	0.41	9.48	0.74	0.87
$\lvert\mu_5\rvert$	Test	0.86	1.00	0.26	6.64	0.41	9.48	0.72	0.85
$\lvert\mu_5\rvert$	Δ%	-0.22	-0.03	-3.43	-3.70	0.00	0.00	-2.40	-2.04
μ_6	Train	0.94	0.99	0.25	5.08	0.09	1.95	0.85	0.88
μ_6	Test	0.93	0.98	0.23	4.47	0.09	1.95	0.82	0.84
μ_6	Δ%	-1.13	-0.78	-9.04	-11.97	0.00	0.00	-3.46	-4.83
$\lvert\mu_6\rvert$	Train	0.85	1.00	0.27	6.72	0.40	9.22	0.71	0.86
$\lvert\mu_6\rvert$	Test	0.85	1.00	0.26	6.46	0.40	9.22	0.69	0.83
$\lvert\mu_6\rvert$	Δ%	-0.26	-0.03	-3.57	-3.86	0.00	0.00	-2.86	-2.62
μ_7	Train	0.97	0.99	0.19	3.51	0.01	0.15	0.91	0.91
μ_7	Test	0.92	0.95	0.17	2.92	0.01	0.15	0.86	0.84
μ_7	Δ%	-4.74	-4.63	-11.80	-16.72	0.00	0.00	-5.29	-7.58
$\lvert\mu_7\rvert$	Train	0.97	0.99	0.19	3.51	0.01	0.15	0.91	0.91

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_7\rvert$	Test	0.92	0.95	0.17	2.92	0.01	0.15	0.86	0.84
$\lvert\mu_7\rvert$	Δ%	-4.74	-4.63	-11.80	-16.72	0.00	0.00	-5.29	-7.58
μ_8	Train	0.97	0.99	0.27	5.82	0.02	0.76	0.83	0.85
μ_8	Test	0.96	0.99	0.25	5.23	0.02	0.76	0.80	0.82
μ_8	Δ%	-1.23	-0.42	-8.14	-10.02	0.00	0.00	-3.24	-3.19
μ_9	Train	0.97	1.00	0.19	3.57	0.00	0.15	0.90	0.91
μ_9	Test	0.93	0.96	0.17	2.98	0.00	0.15	0.86	0.84
μ_9	Δ%	-4.31	-3.91	-11.95	-16.74	0.00	0.00	-5.04	-7.59
$\lvert\mu_9\rvert$	Train	0.97	0.99	0.19	3.56	0.01	0.15	0.90	0.90
$\lvert\mu_9\rvert$	Test	0.92	0.95	0.17	2.95	0.01	0.15	0.85	0.84
$\lvert\mu_9\rvert$	Δ%	-4.32	-3.87	-12.24	-17.09	0.00	0.00	-5.06	-7.58
μ_{10}	Train	0.97	0.99	0.20	3.59	0.00	0.14	0.90	0.91
μ_{10}	Test	0.93	0.96	0.17	2.99	0.00	0.14	0.86	0.84
μ_{10}	Δ%	-4.21	-3.77	-11.89	-16.67	0.00	0.00	-4.98	-7.41
μ_{11}	Train	0.97	1.00	0.19	3.58	0.00	0.15	0.90	0.91
μ_{11}	Test	0.93	0.96	0.17	2.98	0.00	0.15	0.86	0.84
μ_{11}	Δ%	-4.33	-3.92	-11.96	-16.76	0.00	0.00	-5.06	-7.60
μ_{12}	Train	0.98	1.00	0.20	3.71	0.00	0.14	0.90	0.91
μ_{12}	Test	0.94	0.97	0.18	3.11	0.00	0.14	0.86	0.86
μ_{12}	Δ%	-3.55	-3.06	-11.39	-16.12	0.00	0.00	-4.29	-6.03
μ_{13}	Train	0.98	1.00	0.22	4.37	0.01	0.20	0.89	0.91
μ_{13}	Test	0.95	0.98	0.20	3.74	0.01	0.20	0.86	0.86
μ_{13}	Δ%	-2.38	-1.59	-10.55	-14.23	0.00	0.00	-3.78	-5.56
μ_{14}	Train	0.89	0.97	0.18	3.72	0.12	2.84	0.32	0.49
μ_{14}	Test	0.88	0.96	0.19	3.83	0.12	2.84	0.27	0.41
μ_{14}	Δ%	-0.34	-0.25	2.12	2.85	0.00	0.00	-16.72	-16.30
μ_{15}	Train	0.98	1.00	0.21	4.07	0.00	0.13	0.90	0.92
μ_{15}	Test	0.95	0.97	0.19	3.47	0.00	0.13	0.87	0.87
μ_{15}	Δ%	-2.79	-2.10	-10.93	-14.87	0.00	0.00	-3.97	-5.14
μ_{16}	Train	0.97	0.99	0.20	3.77	0.00	0.16	0.90	0.91
μ_{16}	Test	0.93	0.96	0.18	3.17	0.00	0.16	0.86	0.85
μ_{16}	Δ%	-3.64	-3.06	-11.78	-16.07	0.00	0.00	-4.79	-6.34
μ_{17}	Train	0.96	0.98	0.19	3.54	0.01	0.19	0.91	0.91
μ_{17}	Test	0.92	0.95	0.17	2.95	0.01	0.19	0.86	0.84
μ_{17}	Δ%	-3.94	-3.65	-11.72	-16.69	0.00	0.00	-5.11	-7.16
μ_{18}	Train	0.82	0.99	0.20	4.97	0.59	14.48	0.71	0.85
μ_{18}	Test	0.82	0.99	0.19	4.80	0.59	14.48	0.69	0.83
μ_{18}	Δ%	-0.31	-0.09	-2.92	-3.46	0.00	0.00	-2.09	-2.44
μ_{19}	Train	0.75	0.99	0.15	4.03	0.59	14.00	0.17	0.39
μ_{19}	Test	0.75	0.99	0.15	4.03	0.59	14.00	0.15	0.36
μ_{19}	Δ%	0.00	0.03	-0.12	-0.01	0.00	0.00	-8.25	-7.82
μ_{20}	Train	0.92	0.98	0.18	3.31	0.00	0.11	0.90	0.92
μ_{20}	Test	0.86	0.90	0.16	2.73	0.00	0.11	0.83	0.83
μ_{20}	Δ%	-6.77	-7.61	-12.57	-17.58	0.00	0.00	-8.06	-10.05

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\lvert\mu_{20}\rvert$	Train	0.91	0.97	0.18	3.28	0.00	0.11	0.90	0.91
$\lvert\mu_{20}\rvert$	Test	0.85	0.90	0.15	2.69	0.00	0.11	0.82	0.82
$\lvert\mu_{20}\rvert$	$\Delta\%$	-6.83	-7.62	-12.88	-17.95	0.00	0.00	-8.11	-10.04
μ_{21}	Train	0.98	1.00	0.20	3.73	0.00	0.14	0.90	0.92
μ_{21}	Test	0.95	0.97	0.18	3.13	0.00	0.14	0.86	0.86
μ_{21}	$\Delta\%$	-3.33	-2.83	-11.34	-15.98	0.00	0.00	-4.24	-5.79
μ_{22}	Train	0.86	0.96	0.15	2.84	0.00	0.11	0.89	0.91
μ_{22}	Test	0.75	0.82	0.13	2.27	0.00	0.11	0.77	0.78
μ_{22}	$\Delta\%$	-12.58	-14.52	-15.17	-20.15	0.00	0.00	-12.98	-14.44
$\lvert\mu_{22}\rvert$	Train	0.85	0.95	0.15	2.82	0.00	0.11	0.88	0.90
$\lvert\mu_{22}\rvert$	Test	0.74	0.81	0.12	2.24	0.00	0.11	0.76	0.77
$\lvert\mu_{22}\rvert$	$\Delta\%$	-12.67	-14.53	-15.58	-20.60	0.00	0.00	-13.09	-14.41
μ_{23}	Train	0.98	1.00	0.21	3.92	0.00	0.14	0.89	0.91
μ_{23}	Test	0.96	0.98	0.19	3.31	0.00	0.14	0.86	0.86
μ_{23}	$\Delta\%$	-2.57	-1.95	-11.01	-15.37	0.00	0.00	-3.79	-5.51
μ_{24}	Train	0.98	1.00	0.21	3.80	0.00	0.15	0.90	0.91
μ_{24}	Test	0.95	0.98	0.19	3.20	0.00	0.15	0.87	0.86
μ_{24}	$\Delta\%$	-2.80	-2.27	-11.11	-15.70	0.00	0.00	-3.72	-5.28
$\lvert\mu_{24}\rvert$	Train	0.98	1.00	0.21	3.81	0.01	0.16	0.90	0.91
$\lvert\mu_{24}\rvert$	Test	0.95	0.98	0.19	3.21	0.01	0.16	0.86	0.87
$\lvert\mu_{24}\rvert$	$\Delta\%$	-2.79	-2.26	-11.32	-15.77	0.00	0.00	-3.65	-5.16

Table N.1: μ-GREEDY Stopwords Included Discounted

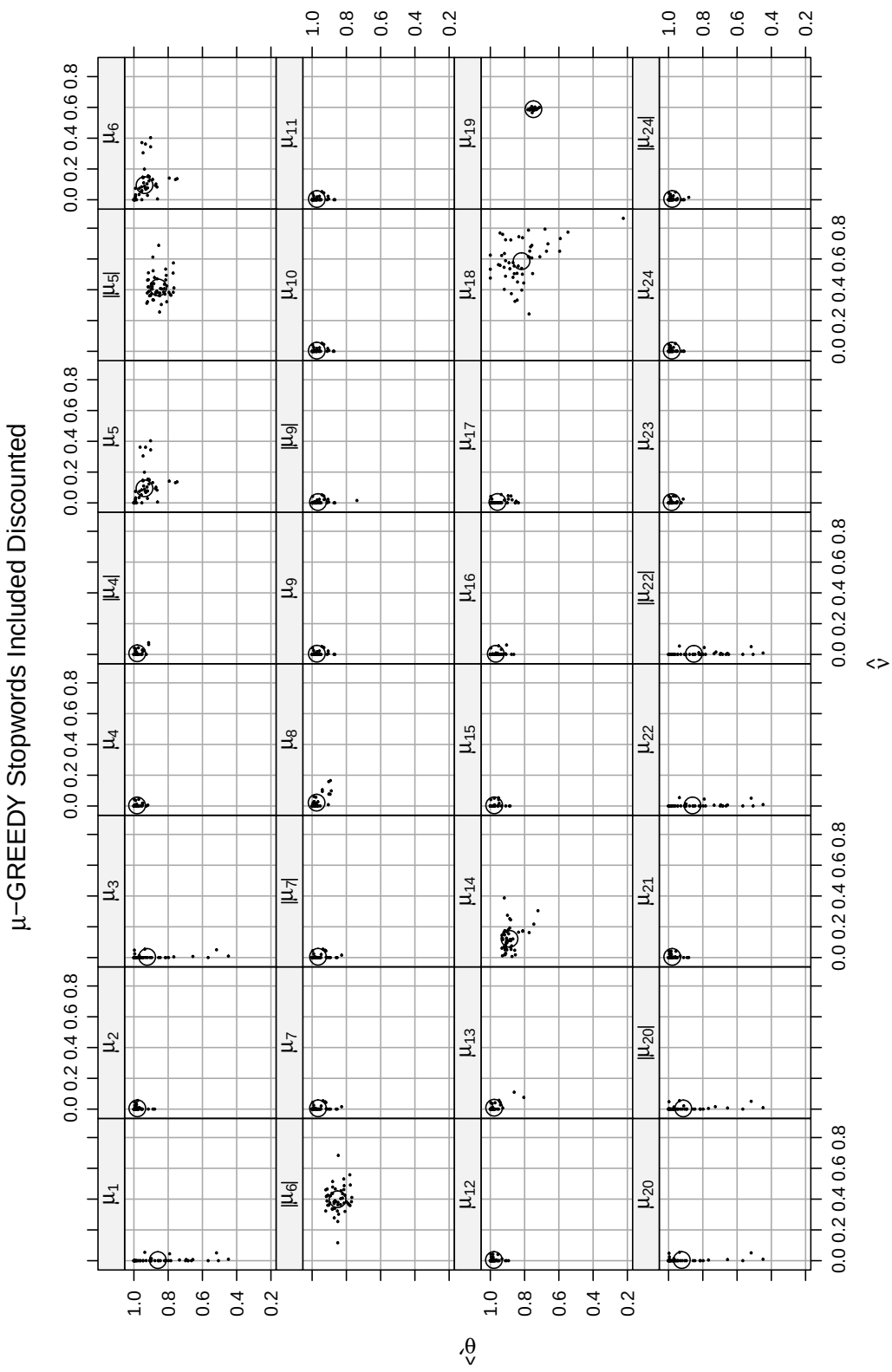

μ−GREEDY Stopwords Included Discounted
$\hat{\theta}$
$\hat{V}$
μ_1
μ_2
μ_3
μ_4
μ_5
μ_6
$|\mu_6|$
μ_7
$|\mu_7|$
μ_8
$|\mu_4|$
$|\mu_5|$
μ_9
$|\mu_9|$
μ_{10}
μ_{11}
μ_{12}
μ_{13}
μ_{14}
μ_{15}
μ_{16}
μ_{17}
μ_{18}
μ_{19}
μ_{20}
$|\mu_{20}|$
μ_{21}
μ_{22}
$|\mu_{22}|$
μ_{23}
μ_{24}
$|\mu_{24}|$
0.0 0.2 0.4 0.6 0.8
0.2 0.4 0.6 0.8 1.0

Figure N.1

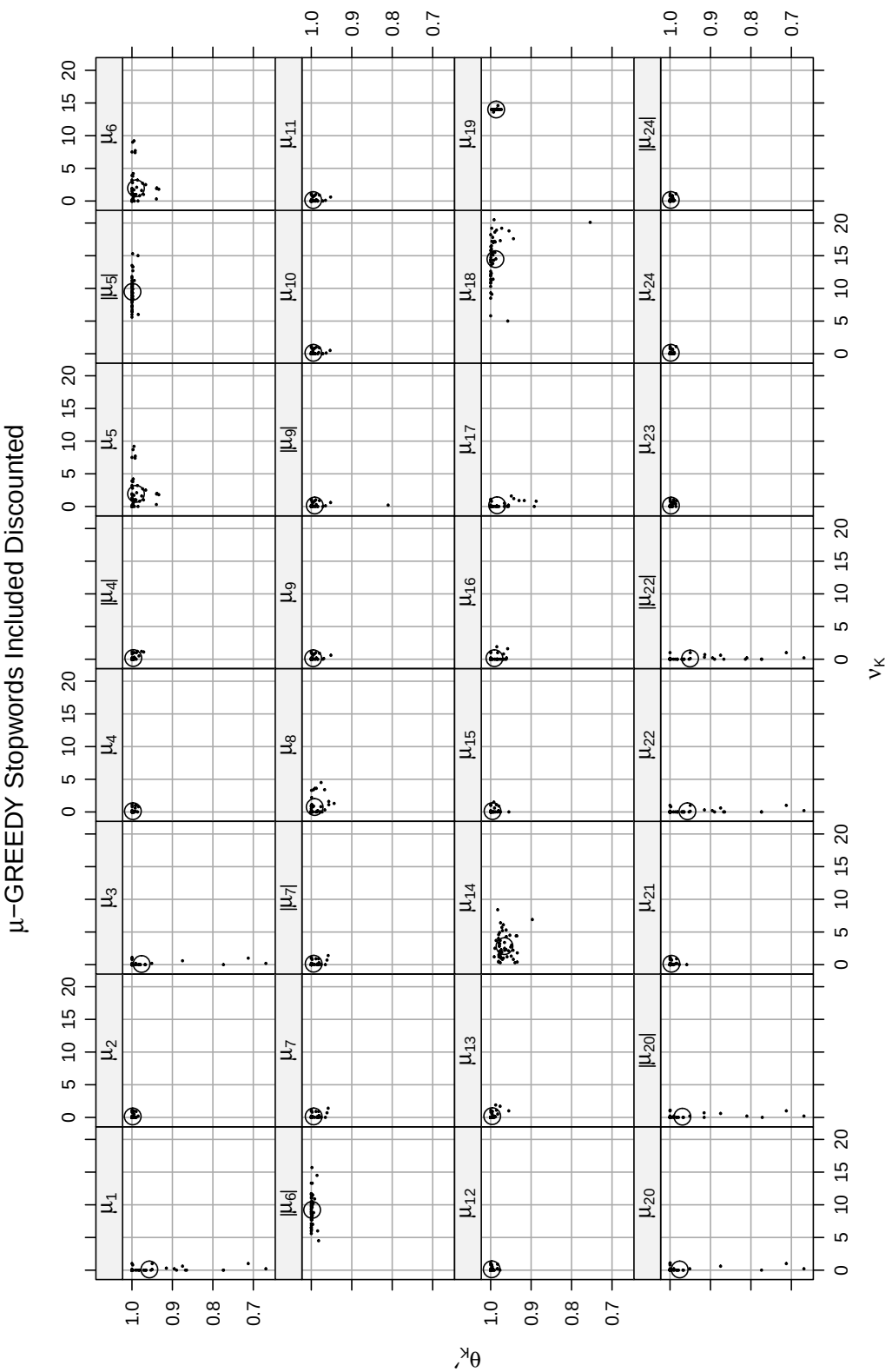

μ−GREEDY Stopwords Included Discounted
θK'
vK
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|

Figure N.2

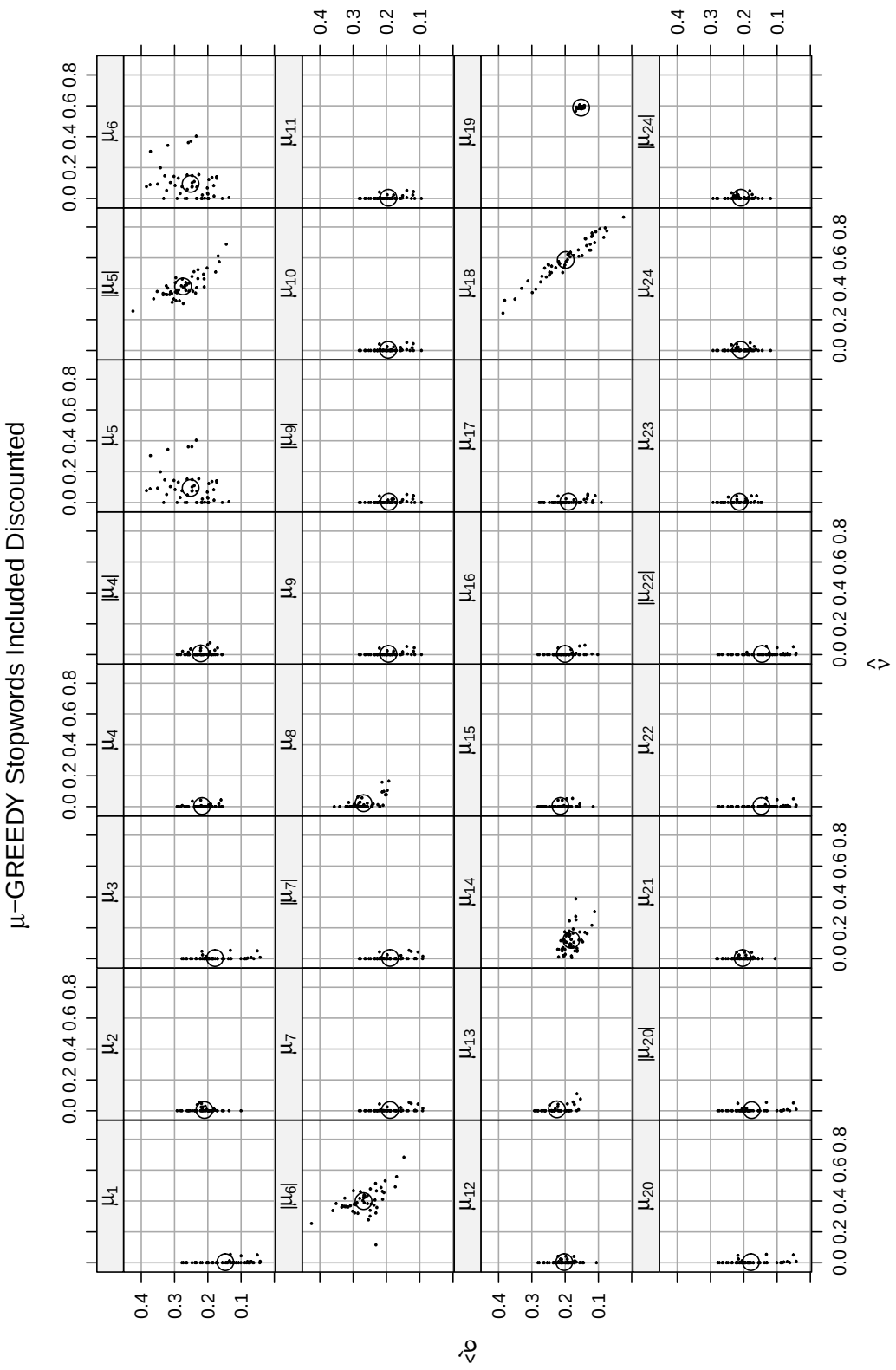

μ−GREEDY Stopwords Included Discounted
$\hat{\sigma}^2$
$\hat{v}$
μ_1
μ_2
μ_3
μ_4
$|\mu_4|$
μ_5
$|\mu_5|$
μ_6
$|\mu_6|$
μ_7
$|\mu_7|$
μ_8
μ_9
$|\mu_9|$
μ_{10}
μ_{11}
μ_{12}
μ_{13}
μ_{14}
μ_{15}
μ_{16}
μ_{17}
μ_{18}
μ_{19}
μ_{20}
$|\mu_{20}|$
μ_{21}
μ_{22}
$|\mu_{22}|$
μ_{23}
μ_{24}
$|\mu_{24}|$
0.0 0.2 0.4 0.6 0.8
0.1 0.2 0.3 0.4

Figure N.3

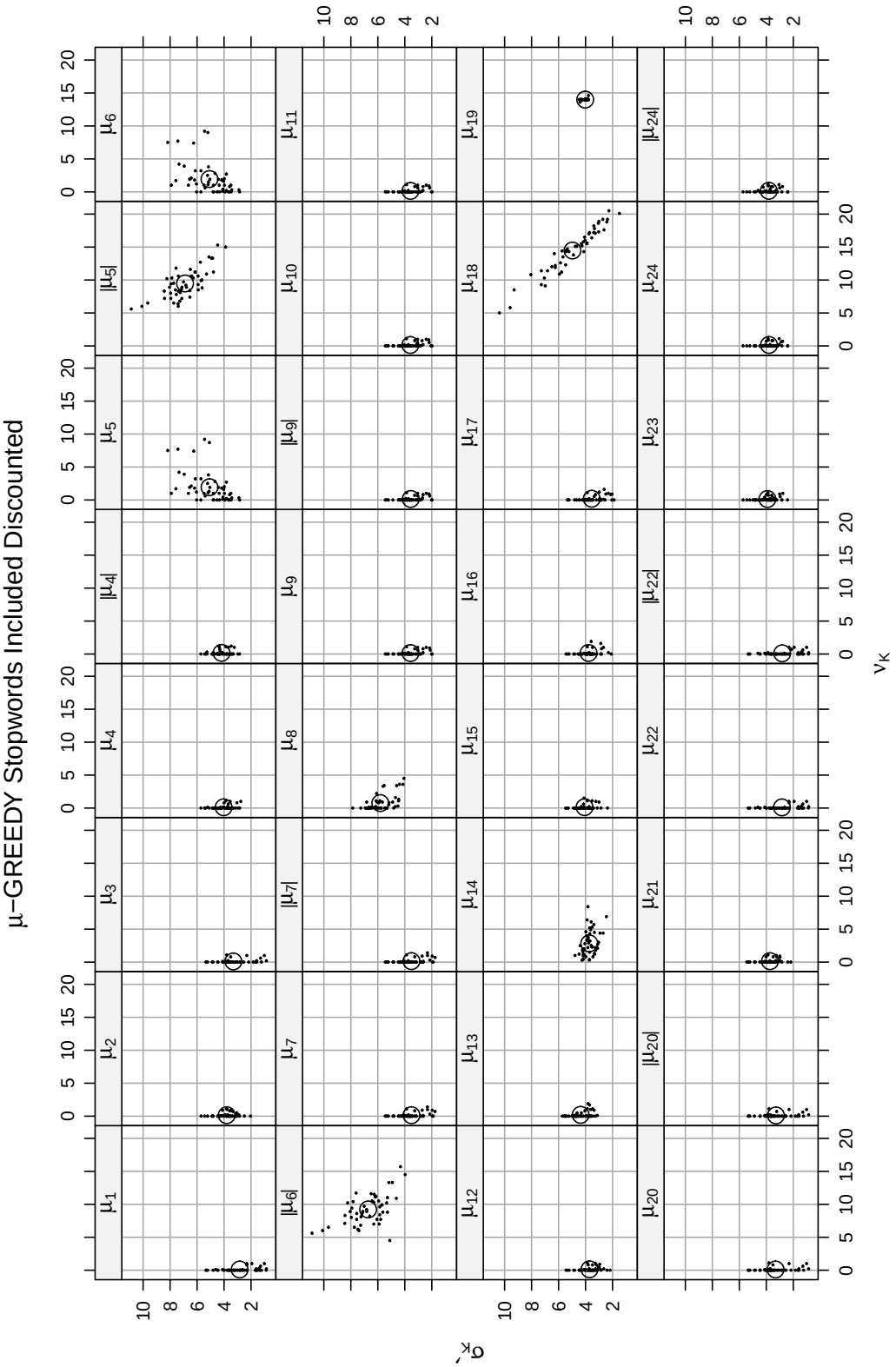

μ−GREEDY Stopwords Included Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
σK'
νK

Figure N.4

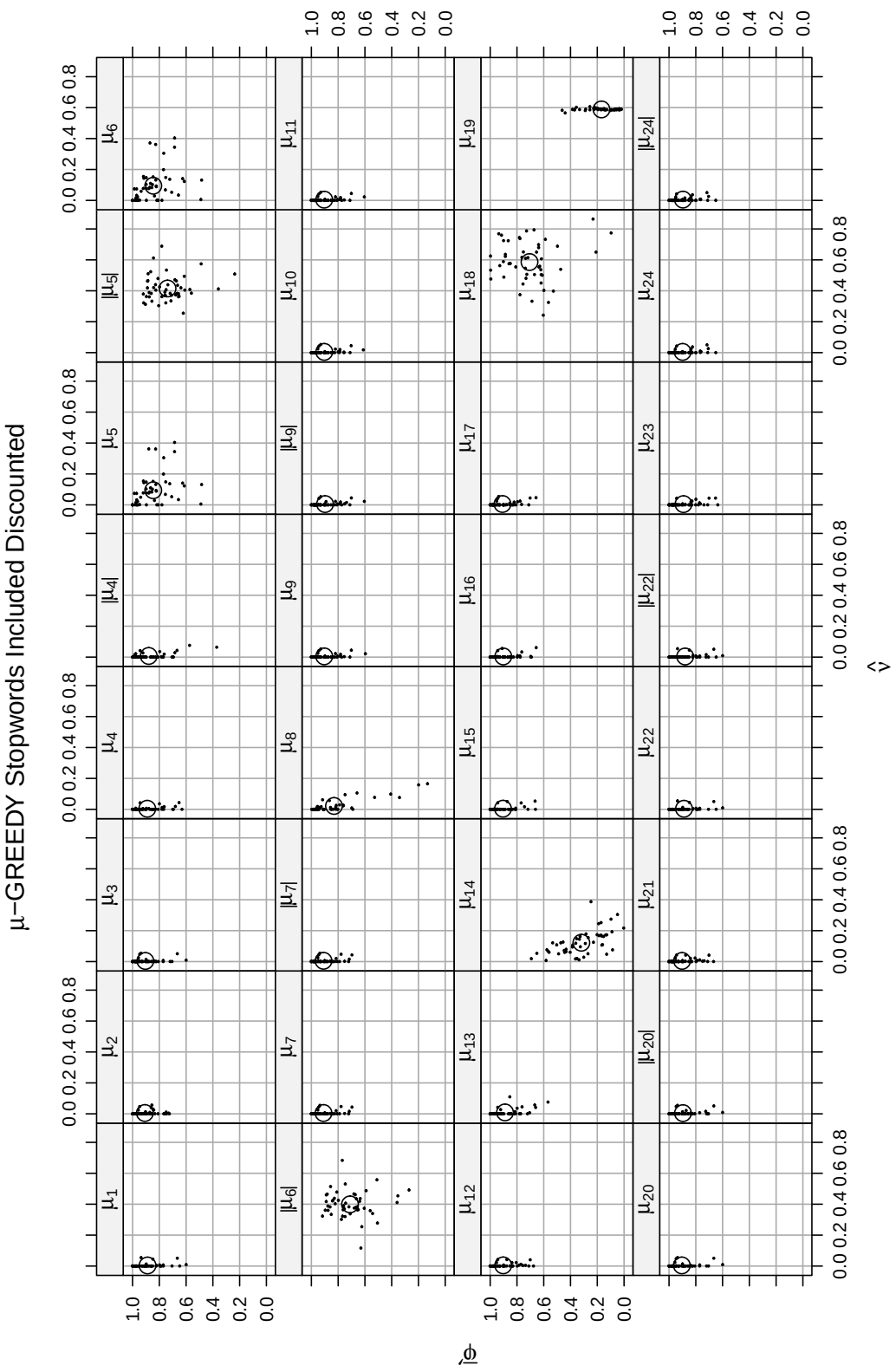

μ−GREEDY Stopwords Included Discounted
$\hat{\phi}'$
$\hat{v}$
μ_1
μ_2
μ_3
μ_4
μ_5
μ_6
μ_7
μ_8
μ_9
μ_{10}
μ_{11}
μ_{12}
μ_{13}
μ_{14}
μ_{15}
μ_{16}
μ_{17}
μ_{18}
μ_{19}
μ_{20}
μ_{21}
μ_{22}
μ_{23}
μ_{24}
$|\mu_4|$
$|\mu_5|$
$|\mu_6|$
$|\mu_7|$
$|\mu_9|$
$|\mu_{20}|$
$|\mu_{22}|$
$|\mu_{24}|$
0.0 0.2 0.4 0.6 0.8
1.0 0.8 0.6 0.4 0.2 0.0

Figure N.5

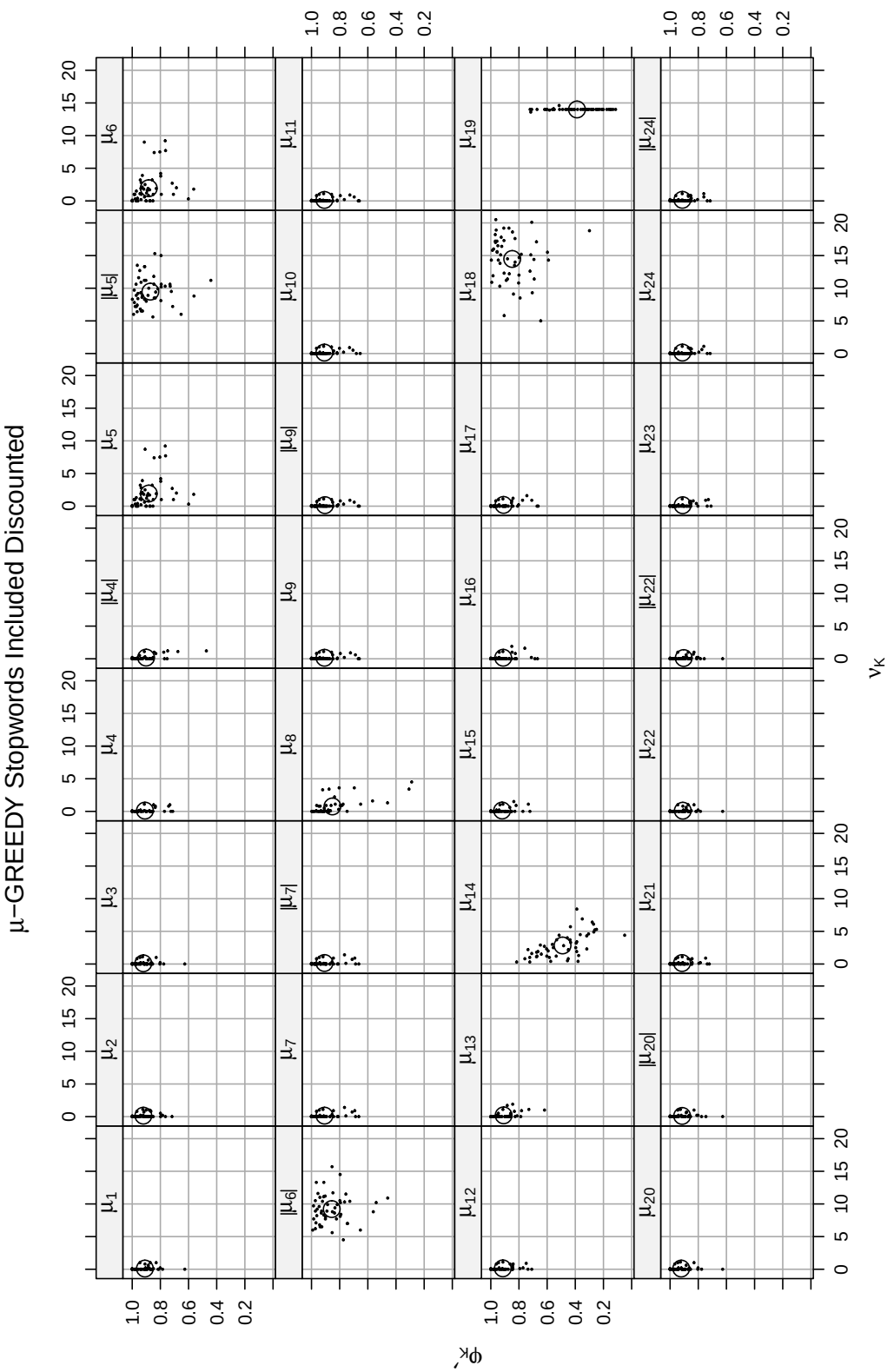
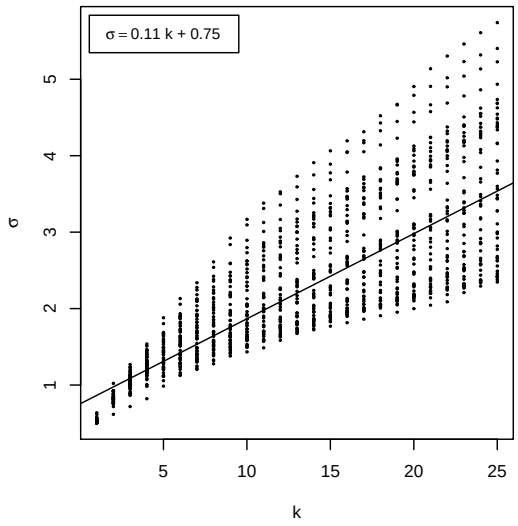

μ−GREEDY Stopwords Included Discounted
μ1
μ2
μ3
μ4
|μ4|
μ5
|μ5|
μ6
|μ6|
μ7
|μ7|
μ8
μ9
|μ9|
μ10
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
μ20
|μ20|
μ21
μ22
|μ22|
μ23
μ24
|μ24|
φk'
νk

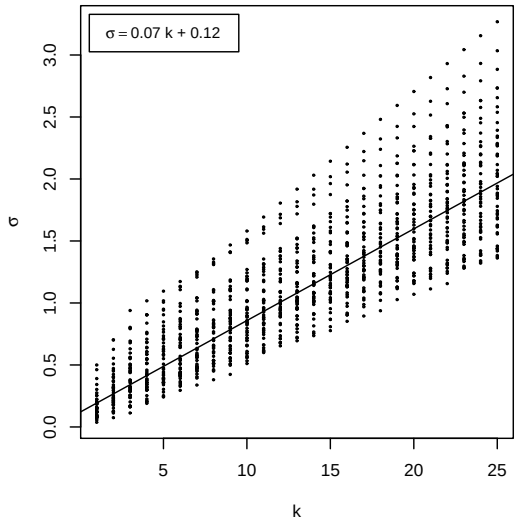
Figure N.6

Appendix O

μ-GREEDY Results Noise Growth Rate



(a) Noise Growth Rate for μ_8 on Topic Fuel



(b) Noise Growth Rate for μ_9 on Topic Fuel

Figure O.1: μ-GREEDY Noise Growth Rate Examples

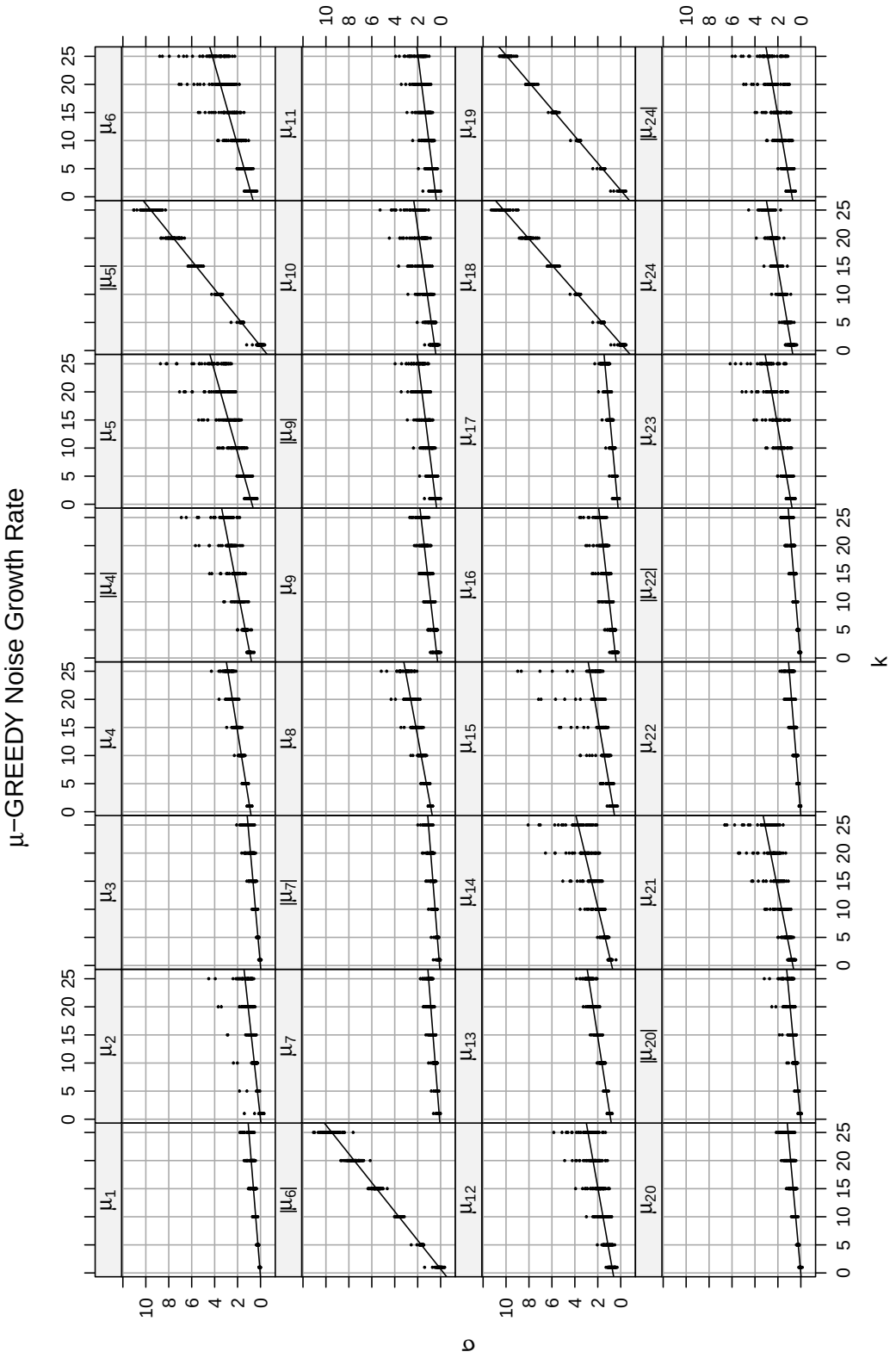

μ−GREEDY Noise Growth Rate
σ
k
μ1
μ2
μ3
μ4
μ5
μ6
|μ6|
μ7
|μ7|
μ8
|μ4|
μ9
|μ5|
μ10
|μ9|
μ11
μ12
μ13
μ14
μ15
μ16
μ17
μ18
μ19
|μ20|
μ21
|μ22|
μ22
μ23
|μ24|
μ24
μ20

Figure O.2

Parameter	Median Intercept	Median Slope	σ_K
μ_0	-0.04	0.11	2.67
μ_1	0.03	0.04	1.01
μ_2	0.00	0.05	1.35
μ_3	0.01	0.04	1.09
μ_4	0.88	0.08	2.85
$\lvert\mu_4\rvert$	0.85	0.10	3.23
μ_5	0.74	0.14	4.19
$\lvert\mu_5\rvert$	-0.31	0.40	9.57
μ_6	0.74	0.14	4.19
$\lvert\mu_6\rvert$	-0.27	0.39	9.48
μ_7	0.09	0.04	1.05
$\lvert\mu_7\rvert$	0.10	0.04	1.06
μ_8	0.79	0.09	3.06
μ_9	0.30	0.06	1.71
$\lvert\mu_9\rvert$	0.39	0.06	1.94
μ_{10}	0.46	0.07	2.23
μ_{11}	0.41	0.06	1.97
μ_{12}	0.68	0.09	2.85
μ_{13}	0.88	0.08	2.79
μ_{14}	0.78	0.12	3.72
μ_{15}	0.60	0.08	2.70
μ_{16}	0.44	0.06	1.83
μ_{17}	0.26	0.04	1.38
μ_{18}	-0.53	0.43	10.20
μ_{19}	-0.49	0.42	9.92
μ_{20}	0.01	0.04	1.11
$\lvert\mu_{20}\rvert$	0.01	0.05	1.16
μ_{21}	0.70	0.10	3.12
μ_{22}	0.03	0.04	1.01
$\lvert\mu_{22}\rvert$	0.03	0.04	1.04
μ_{23}	0.82	0.09	2.97
μ_{24}	0.76	0.08	2.86
$\lvert\mu_{24}\rvert$	0.76	0.09	2.90

Table O.1: μ-GREEDY Noise Growth Rate

Appendix P

π-RANKING Results Stopwords Included Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
π_0	Train	0.16	0.28	0.01	0.35	0.04	0.88	0.06	0.10
π_0	Test	0.15	0.26	0.01	0.32	0.04	0.88	0.04	0.07
π_0	Δ%	-8.59	-7.91	-8.22	-8.19	0.00	0.00	-34.94	-32.81
π_1	Train	0.92	1.00	0.41	10.59	0.53	12.76	0.72	0.78
π_1	Test	0.92	1.00	0.40	10.31	0.53	12.76	0.70	0.76
π_1	Δ%	0.01	0.02	-2.36	-2.68	0.00	0.00	-2.46	-2.80
π_4	Train	0.97	1.00	0.43	9.50	0.14	4.40	0.84	0.83
π_4	Test	0.97	1.00	0.41	8.88	0.14	4.40	0.81	0.80
π_4	Δ%	-0.39	-0.06	-5.53	-6.61	0.00	0.00	-2.78	-2.73
$\lvert\pi_4\rvert$	Train	0.98	1.00	0.47	10.37	0.13	4.16	0.86	0.88
$\lvert\pi_4\rvert$	Test	0.98	1.00	0.45	10.03	0.13	4.16	0.84	0.85
$\lvert\pi_4\rvert$	Δ%	-0.14	-0.00	-3.04	-3.28	0.00	0.00	-2.30	-2.49
π_8	Train	0.98	1.00	0.51	11.75	0.48	13.07	0.78	0.81
π_8	Test	0.98	1.00	0.50	11.50	0.48	13.07	0.76	0.78
π_8	Δ%	-0.08	0.00	-1.62	-2.18	0.00	0.00	-2.39	-2.89
π_9	Train	0.96	0.99	0.35	7.05	0.01	0.27	0.88	0.90
π_9	Test	0.95	0.98	0.32	6.45	0.01	0.27	0.86	0.87
π_9	Δ%	-0.63	-0.59	-6.74	-8.50	0.00	0.00	-2.87	-3.28
$\lvert\pi_9\rvert$	Train	0.96	0.99	0.35	7.12	0.01	0.27	0.89	0.90
$\lvert\pi_9\rvert$	Test	0.95	0.98	0.33	6.52	0.01	0.27	0.86	0.87
$\lvert\pi_9\rvert$	Δ%	-0.63	-0.59	-6.70	-8.40	0.00	0.00	-2.85	-3.13
π_{18}	Train	0.90	1.00	0.40	10.65	0.54	12.88	0.69	0.78
π_{18}	Test	0.90	1.00	0.40	10.37	0.54	12.88	0.68	0.76
π_{18}	Δ%	0.04	0.02	-2.20	-2.58	0.00	0.00	-2.49	-2.74
π_{22}	Train	0.97	1.00	0.43	9.41	0.13	4.11	0.84	0.83
π_{22}	Test	0.97	1.00	0.40	8.78	0.13	4.11	0.82	0.81
π_{22}	Δ%	-0.41	-0.07	-5.65	-6.68	0.00	0.00	-2.78	-2.63
$\lvert\pi_{22}\rvert$	Train	0.98	1.00	0.46	10.29	0.12	3.72	0.86	0.88
$\lvert\pi_{22}\rvert$	Test	0.98	1.00	0.45	9.94	0.12	3.72	0.84	0.86
$\lvert\pi_{22}\rvert$	Δ%	-0.16	-0.00	-3.14	-3.35	0.00	0.00	-2.32	-2.21
π_{23}	Train	0.98	1.00	0.43	9.53	0.07	2.18	0.87	0.88
π_{23}	Test	0.98	1.00	0.42	9.12	0.07	2.18	0.85	0.86
π_{23}	Δ%	-0.23	-0.02	-4.06	-4.25	0.00	0.00	-2.38	-2.34
π_{24}	Train	0.97	0.99	0.38	8.00	0.02	0.88	0.87	0.88

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
π_{24}	Test	0.96	0.99	0.36	7.40	0.02	0.88	0.85	0.86
π_{24}	$\Delta\%$	-0.41	-0.21	-6.36	-7.49	0.00	0.00	-2.59	-2.31
$\lvert\pi_{24}\rvert$	Train	0.98	1.00	0.40	8.67	0.04	1.43	0.88	0.90
$\lvert\pi_{24}\rvert$	Test	0.97	1.00	0.39	8.32	0.04	1.43	0.86	0.88
$\lvert\pi_{24}\rvert$	$\Delta\%$	-0.14	-0.04	-3.97	-4.01	0.00	0.00	-2.15	-1.95
π_{25}	Train	0.98	1.00	0.46	10.29	0.15	4.77	0.86	0.87
π_{25}	Test	0.98	1.00	0.44	9.84	0.15	4.77	0.84	0.85
π_{25}	$\Delta\%$	-0.21	0.00	-4.16	-4.32	0.00	0.00	-2.59	-2.93
π_{26}	Train	0.98	1.00	0.45	9.94	0.20	6.12	0.83	0.80
π_{26}	Test	0.97	1.00	0.42	9.30	0.20	6.12	0.80	0.77
π_{26}	$\Delta\%$	-0.30	-0.05	-5.86	-6.47	0.00	0.00	-3.02	-4.17

Table P.1: π-RANKING Stopwords Included Not Discounted

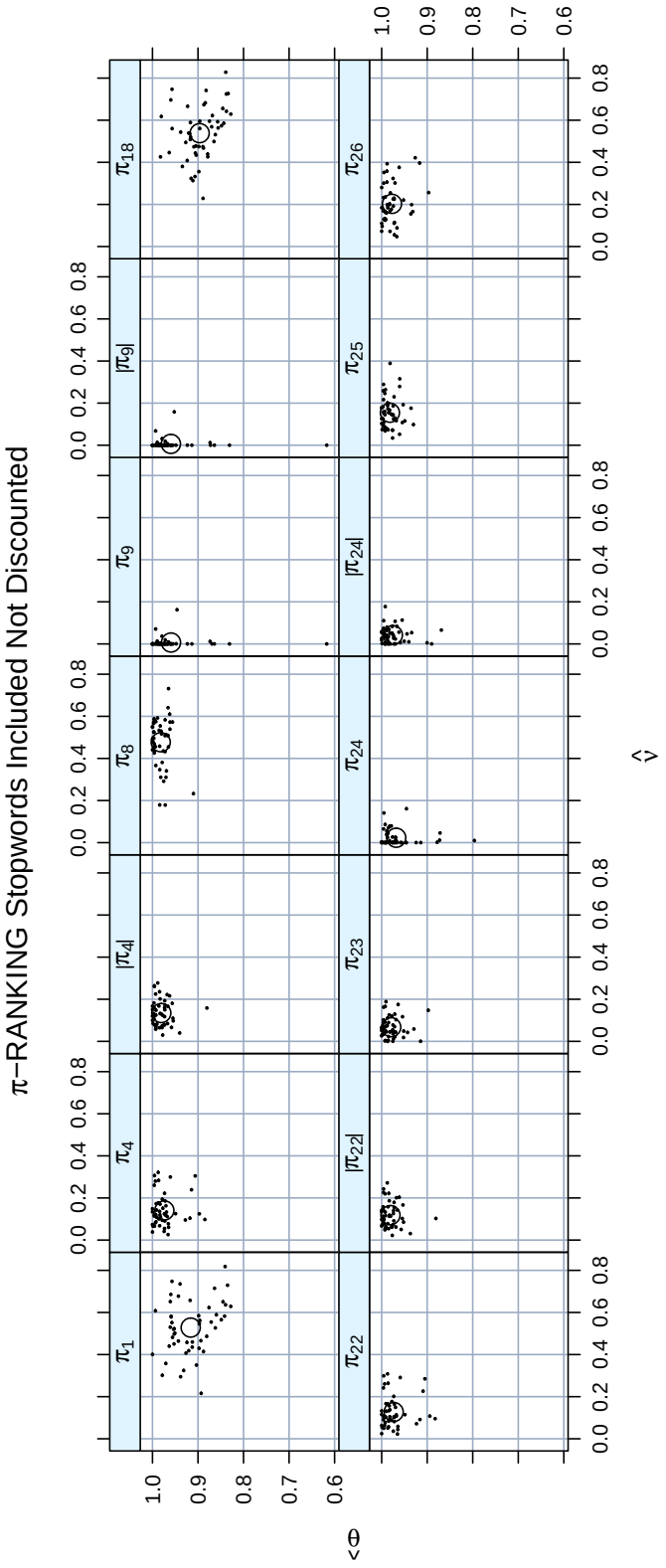

π−RANKING Stopwords Included Not Discounted
θ̂
V̂
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26

Figure P.1

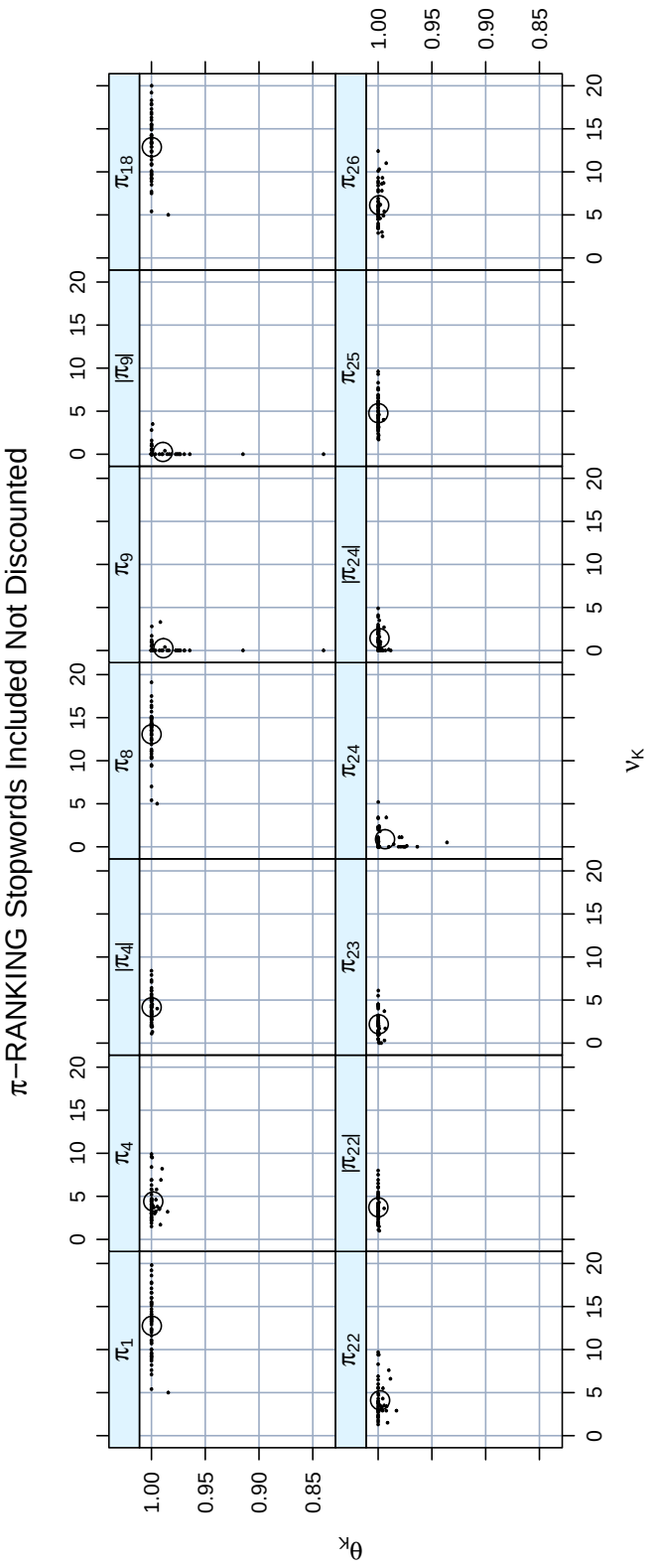

π−RANKING Stopwords Included Not Discounted
θκ
νκ
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26
1.00
0.95
0.90
0.85
0
5
10
15
20

Figure P.2

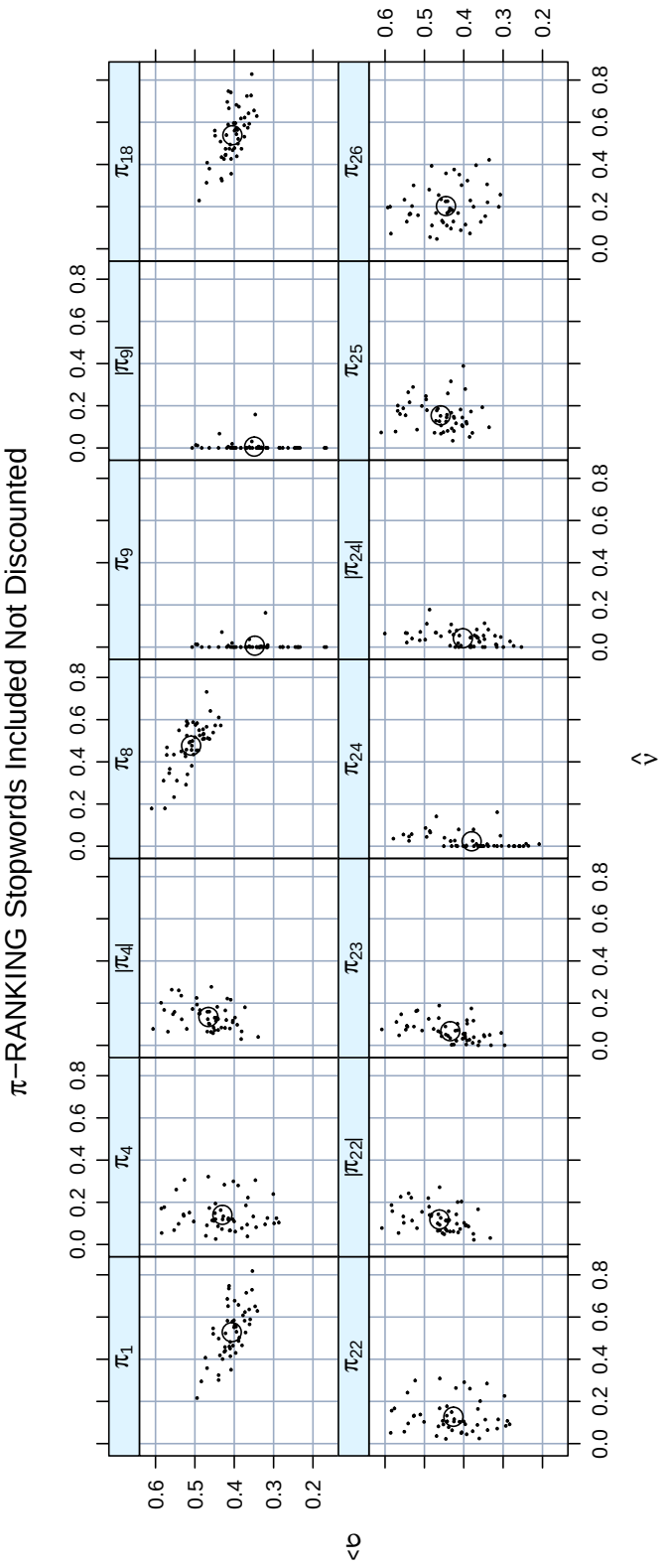

π−RANKING Stopwords Included Not Discounted
$\hat{\sigma}$
$\hat{v}$
π_1
π_4
$|\pi_4|$
π_8
π_9
$|\pi_9|$
π_{18}
π_{22}
$|\pi_{22}|$
π_{23}
π_{24}
$|\pi_{24}|$
π_{25}
π_{26}

Figure P.3

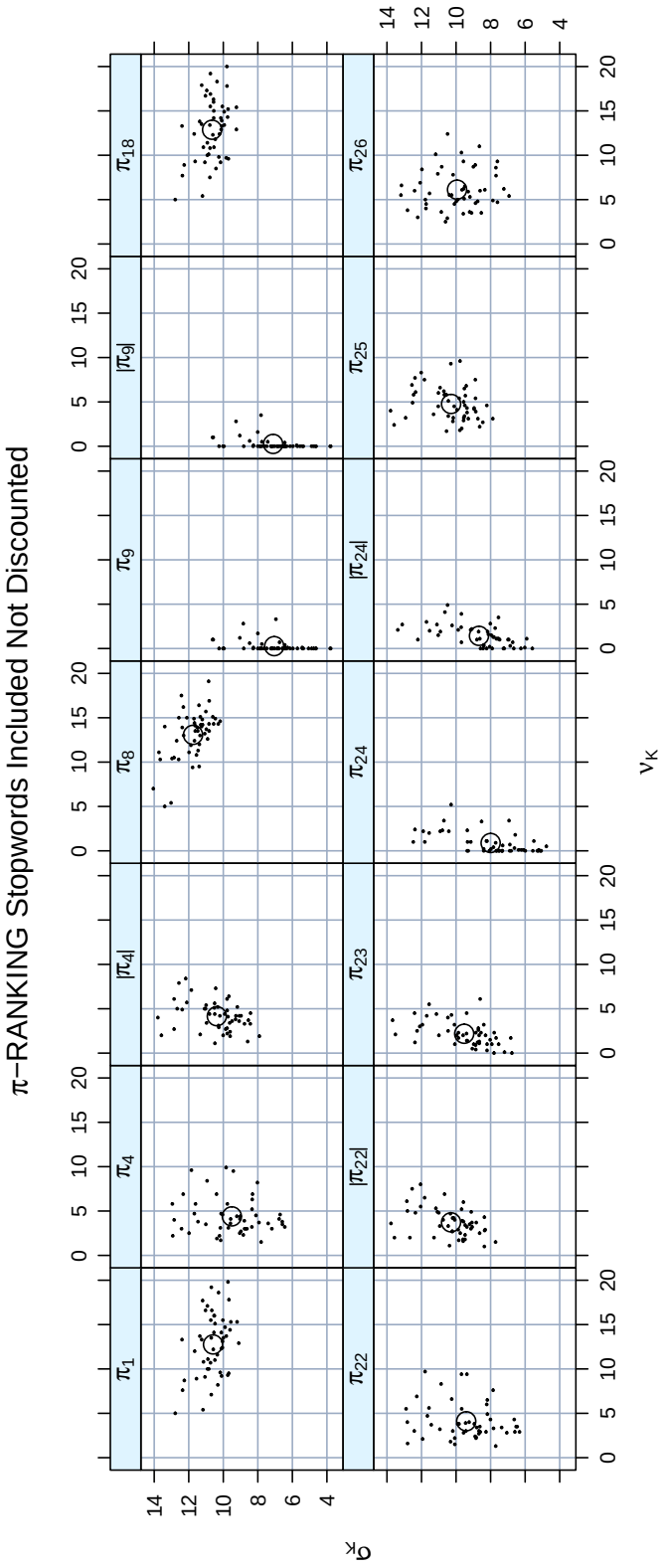

π−RANKING Stopwords Included Not Discounted
σK
νK
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26

Figure P.4

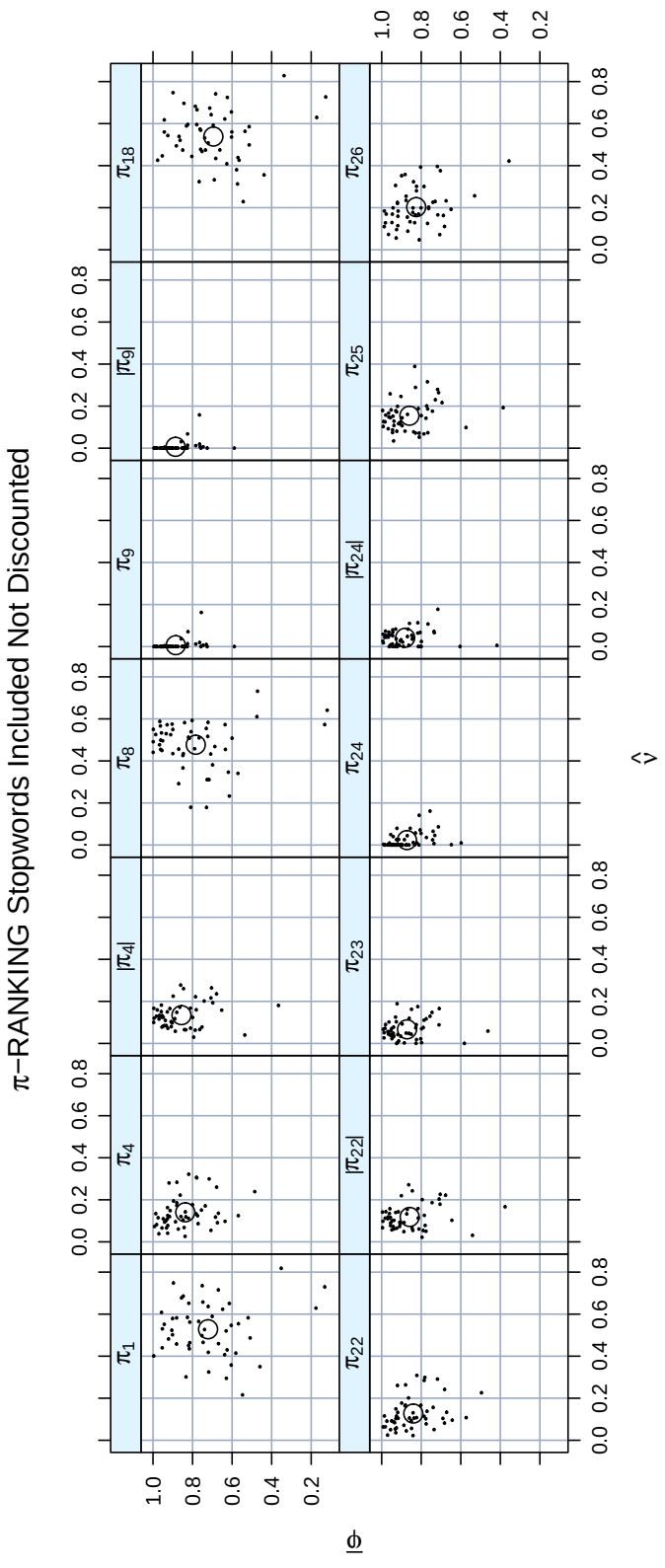

π−RANKING Stopwords Included Not Discounted
$\underline{\Phi}$
$\hat{V}$
π_1
π_4
$|\pi_4|$
π_8
π_9
$|\pi_9|$
π_{18}
π_{22}
$|\pi_{22}|$
π_{23}
π_{24}
$|\pi_{24}|$
π_{25}
π_{26}

Figure P.5

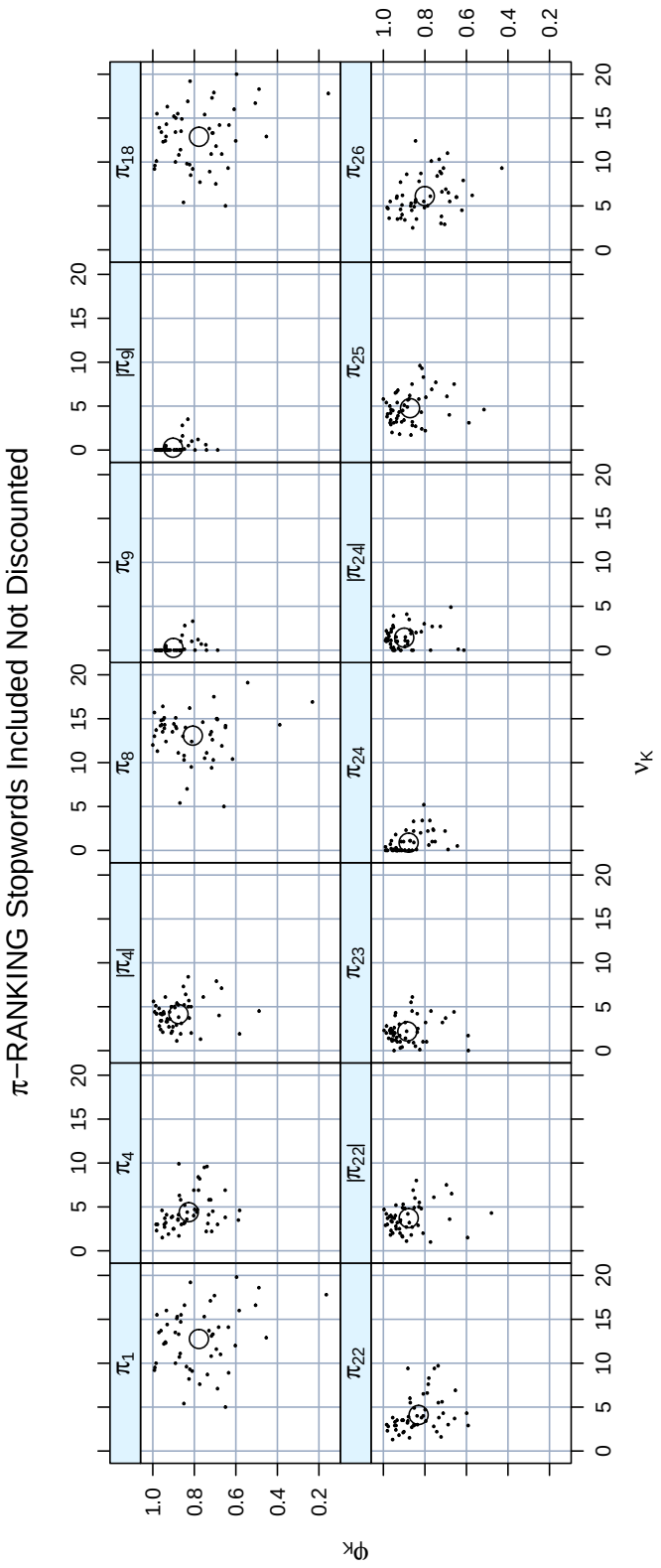

π−RANKING Stopwords Included Not Discounted
φK
νK
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26
0
5
10
15
20
0.2
0.4
0.6
0.8
1.0

Figure P.6

Appendix Q

π-RANKING Results Stopwords Included Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
π_0	Train	0.13	0.23	0.01	0.27	0.04	0.88	0.06	0.10
π_0	Test	0.11	0.21	0.01	0.24	0.04	0.88	0.04	0.07
π_0	$\Delta\%$	-11.16	-10.01	-11.87	-11.11	0.00	0.00	-36.03	-33.34
π_1	Train	0.86	0.99	0.23	5.54	0.53	12.76	0.74	0.86
π_1	Test	0.86	0.99	0.22	5.37	0.53	12.76	0.73	0.84
π_1	$\Delta\%$	-0.29	-0.11	-3.00	-3.16	0.00	0.00	-2.14	-2.28
π_4	Train	0.97	0.99	0.36	7.49	0.14	4.40	0.85	0.87
π_4	Test	0.96	0.99	0.34	6.98	0.14	4.40	0.83	0.85
π_4	$\Delta\%$	-0.53	-0.29	-5.49	-6.79	0.00	0.00	-2.64	-2.31
$\lvert\pi_4\rvert$	Train	0.98	1.00	0.40	8.55	0.13	4.16	0.86	0.89
$\lvert\pi_4\rvert$	Test	0.98	1.00	0.39	8.26	0.13	4.16	0.84	0.87
$\lvert\pi_4\rvert$	$\Delta\%$	-0.11	-0.02	-3.03	-3.36	0.00	0.00	-2.18	-2.23
π_8	Train	0.97	1.00	0.30	6.13	0.48	13.07	0.80	0.86
π_8	Test	0.97	1.00	0.29	5.97	0.48	13.07	0.78	0.84
π_8	$\Delta\%$	-0.11	-0.04	-1.86	-2.70	0.00	0.00	-2.06	-2.04
π_9	Train	0.96	0.99	0.35	6.95	0.01	0.27	0.88	0.90
π_9	Test	0.95	0.98	0.32	6.36	0.01	0.27	0.86	0.87
π_9	$\Delta\%$	-0.63	-0.59	-6.70	-8.44	0.00	0.00	-2.85	-3.36
$\lvert\pi_9\rvert$	Train	0.96	0.99	0.35	7.01	0.01	0.27	0.89	0.90
$\lvert\pi_9\rvert$	Test	0.95	0.98	0.32	6.43	0.01	0.27	0.86	0.87
$\lvert\pi_9\rvert$	$\Delta\%$	-0.63	-0.59	-6.66	-8.35	0.00	0.00	-2.84	-3.24
π_{18}	Train	0.84	0.99	0.22	5.54	0.54	12.88	0.72	0.86
π_{18}	Test	0.84	0.99	0.22	5.37	0.54	12.88	0.70	0.84
π_{18}	$\Delta\%$	-0.24	-0.04	-2.81	-2.98	0.00	0.00	-2.09	-2.25
π_{22}	Train	0.97	0.99	0.37	7.53	0.13	4.11	0.86	0.87
π_{22}	Test	0.96	0.99	0.34	7.01	0.13	4.11	0.83	0.85
π_{22}	$\Delta\%$	-0.54	-0.30	-5.58	-6.83	0.00	0.00	-2.60	-2.12
$\lvert\pi_{22}\rvert$	Train	0.98	1.00	0.41	8.63	0.12	3.72	0.86	0.89
$\lvert\pi_{22}\rvert$	Test	0.98	1.00	0.39	8.35	0.12	3.72	0.84	0.87
$\lvert\pi_{22}\rvert$	$\Delta\%$	-0.13	-0.03	-3.10	-3.34	0.00	0.00	-2.17	-1.89
π_{23}	Train	0.98	1.00	0.40	8.61	0.07	2.18	0.87	0.89
π_{23}	Test	0.98	1.00	0.39	8.24	0.07	2.18	0.85	0.87
π_{23}	$\Delta\%$	-0.19	-0.03	-4.10	-4.31	0.00	0.00	-2.23	-2.13
π_{24}	Train	0.97	0.99	0.37	7.57	0.02	0.88	0.87	0.89

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
π_{24}	Test	0.96	0.99	0.35	7.00	0.02	0.88	0.85	0.87
π_{24}	$\Delta\%$	-0.41	-0.26	-6.33	-7.52	0.00	0.00	-2.54	-2.19
$\lvert\pi_{24}\rvert$	Train	0.97	1.00	0.39	8.14	0.04	1.43	0.88	0.90
$\lvert\pi_{24}\rvert$	Test	0.97	1.00	0.37	7.80	0.04	1.43	0.86	0.88
$\lvert\pi_{24}\rvert$	$\Delta\%$	-0.13	-0.04	-3.98	-4.14	0.00	0.00	-2.03	-1.76
π_{25}	Train	0.98	1.00	0.39	8.29	0.15	4.77	0.86	0.88
π_{25}	Test	0.98	1.00	0.38	7.94	0.15	4.77	0.84	0.86
π_{25}	$\Delta\%$	-0.19	-0.01	-4.06	-4.23	0.00	0.00	-2.37	-2.32
π_{26}	Train	0.97	0.99	0.35	7.24	0.20	6.12	0.85	0.86
π_{26}	Test	0.96	0.99	0.33	6.77	0.20	6.12	0.83	0.83
π_{26}	$\Delta\%$	-0.53	-0.46	-5.75	-6.52	0.00	0.00	-2.79	-3.64

Table Q.1: π-RANKING Stopwords Included Discounted

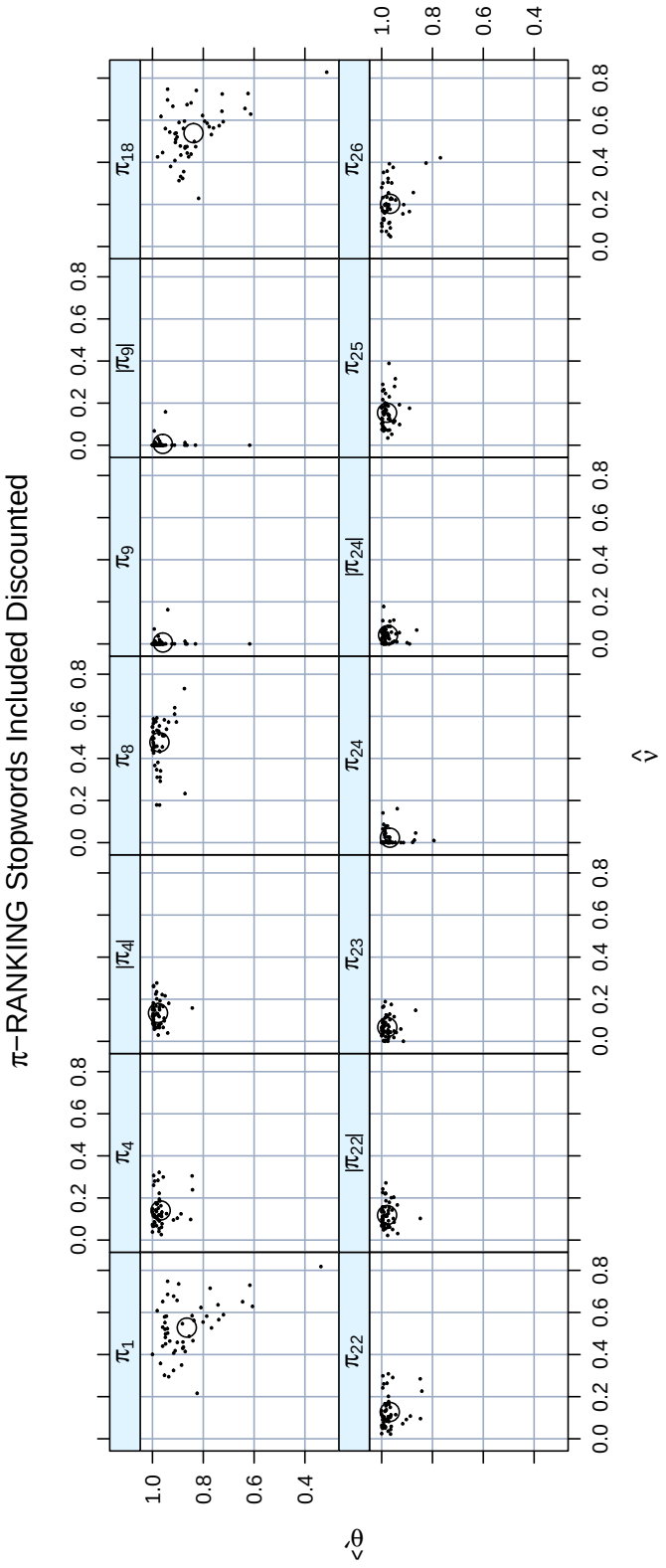

π−RANKING Stopwords Included Discounted
$\hat{\theta}$
$\hat{V}$
π_1
π_4
$|\pi_4|$
π_8
π_9
$|\pi_9|$
π_{18}
π_{22}
$|\pi_{22}|$
π_{23}
π_{24}
$|\pi_{24}|$
π_{25}
π_{26}

Figure Q.1

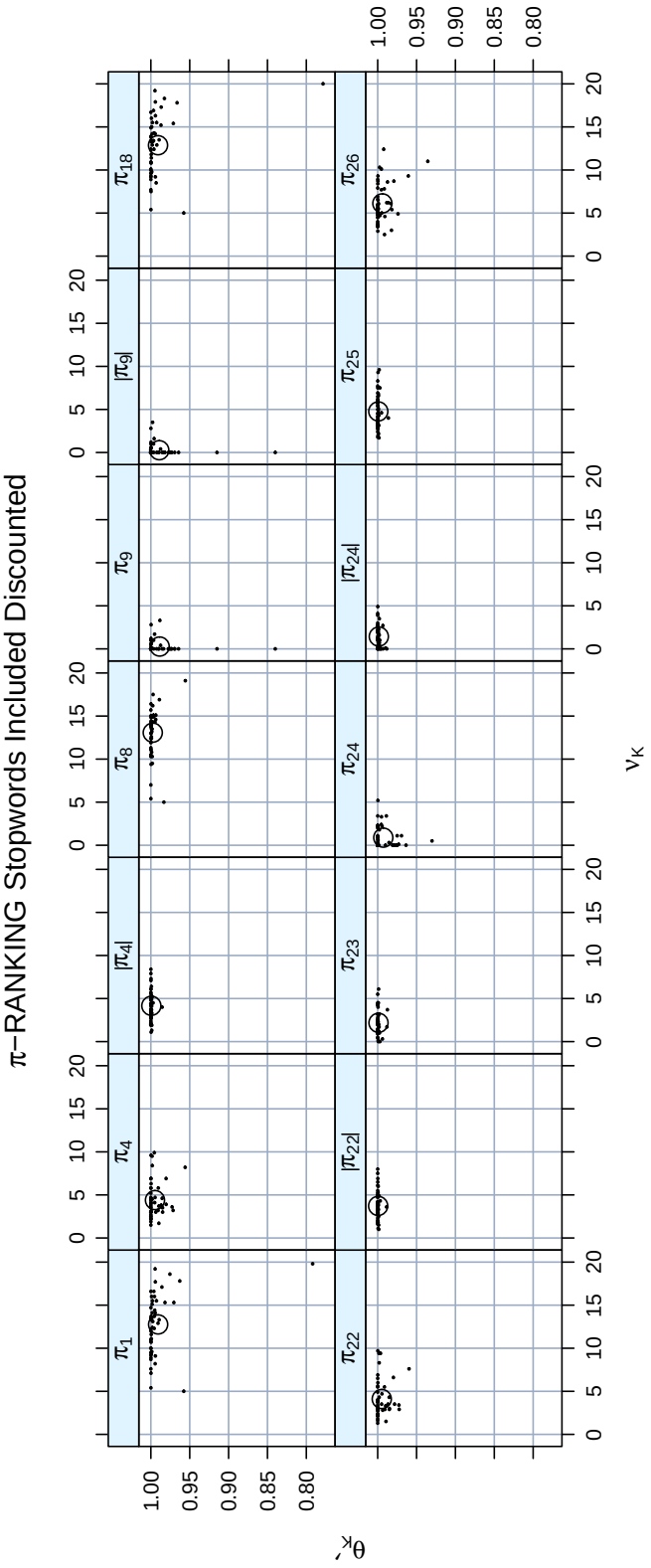

π−RANKING Stopwords Included Discounted
θκ'
νκ
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26

Figure Q.2

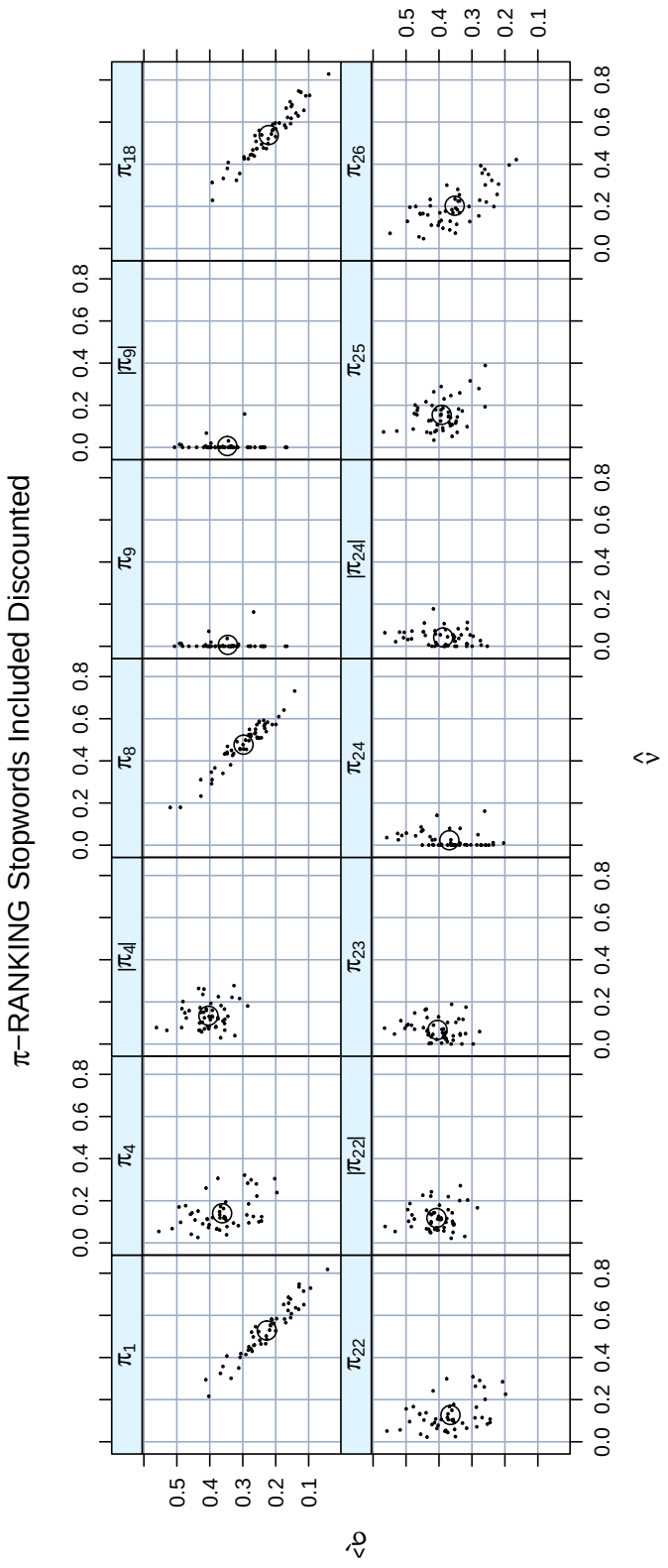

π−RANKING Stopwords Included Discounted
$\hat{\sigma}^2$
$\hat{v}$
π_1
π_4
$|\pi_4|$
π_8
π_9
$|\pi_9|$
π_{18}
π_{22}
$|\pi_{22}|$
π_{23}
π_{24}
$|\pi_{24}|$
π_{25}
π_{26}

Figure Q.3

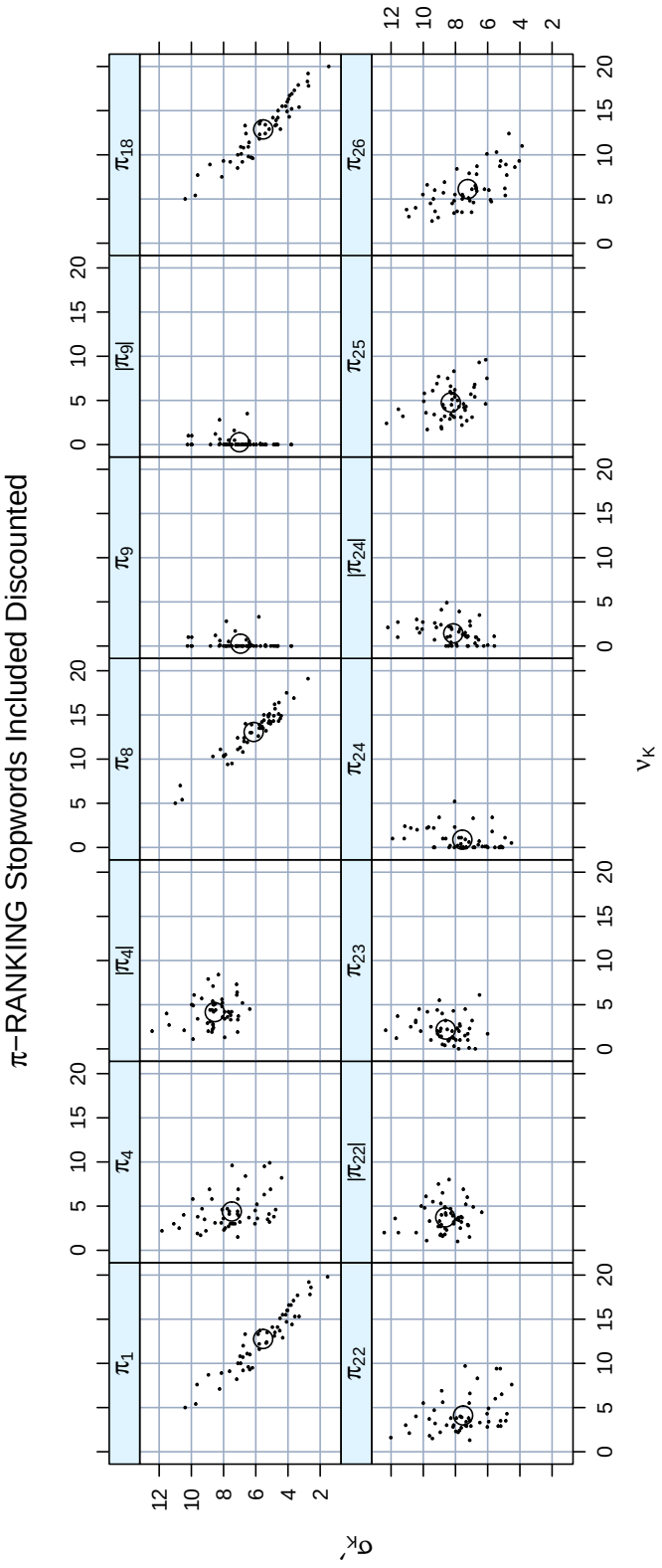

π−RANKING Stopwords Included Discounted
σκ
νκ
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26

Figure Q.4

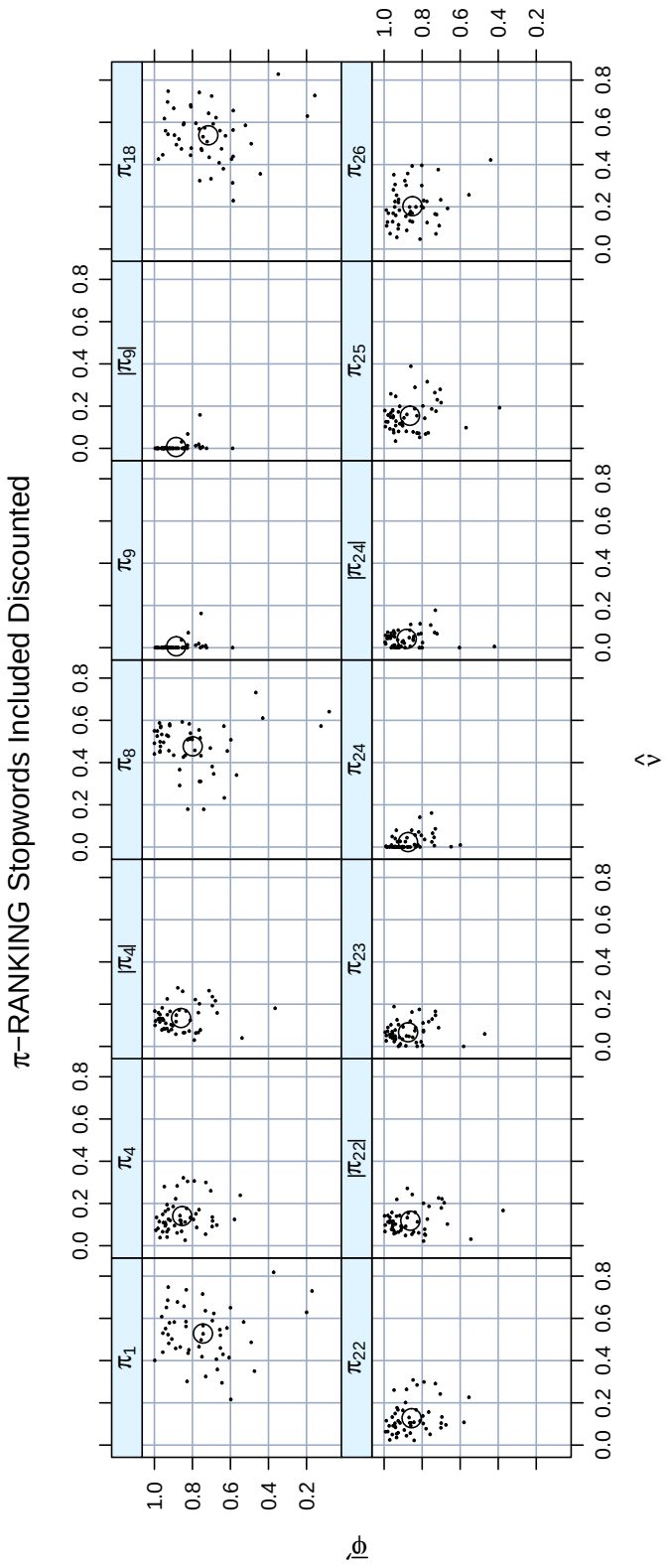

π−RANKING Stopwords Included Discounted
$\hat{\phi}$
$\hat{v}$
π_1
π_4
$|\pi_4|$
π_8
π_9
$|\pi_9|$
π_{18}
π_{22}
$|\pi_{22}|$
π_{23}
π_{24}
$|\pi_{24}|$
π_{25}
π_{26}

Figure Q.5

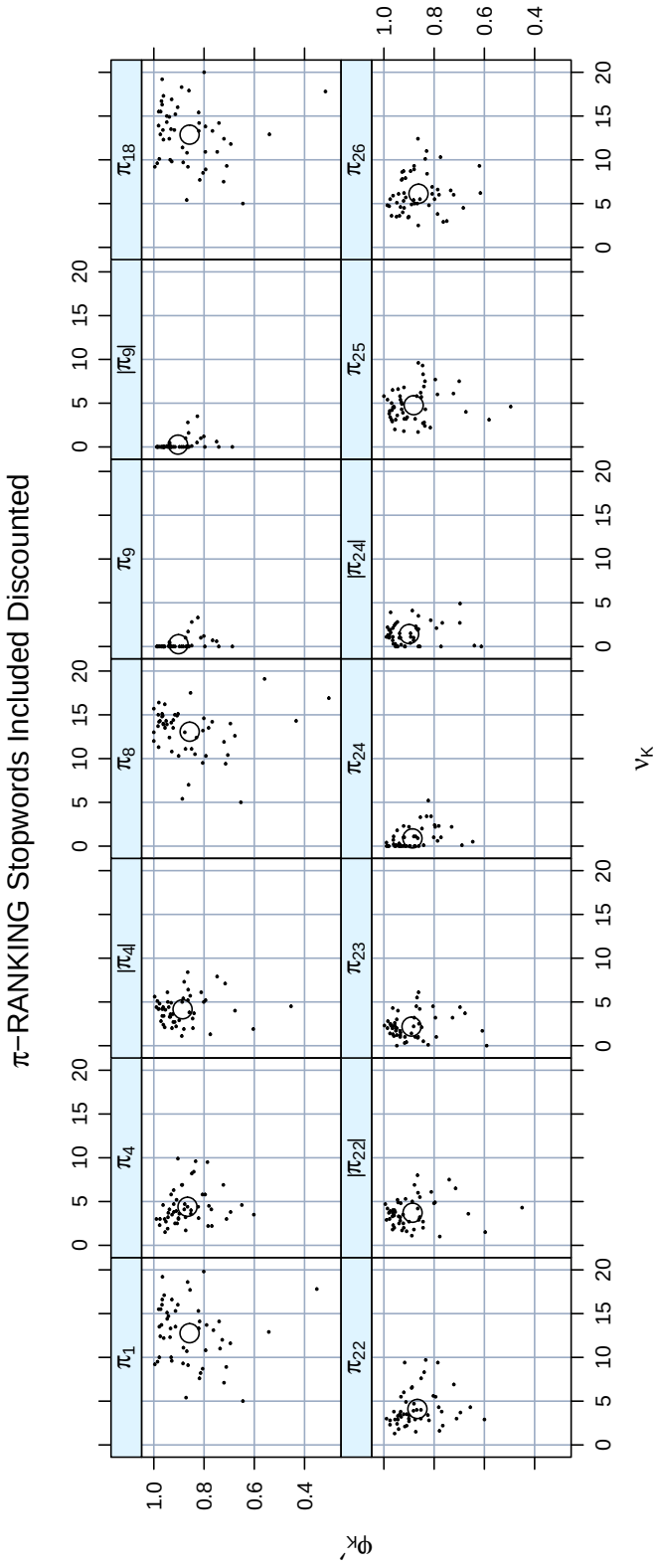

π−RANKING Stopwords Included Discounted
φk'
vk
π1
π4
|π4|
π8
π9
|π9|
π18
π22
|π22|
π23
π24
|π24|
π25
π26

Figure Q.6

Appendix R

h_k-GREEDY Results Stopwords Included Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\phi_{\alpha=\epsilon}$	Train	0.99	1.00	0.47	11.32	0.43	12.00	0.83	0.83
$\phi_{\alpha=\epsilon}$	Test	0.98	1.00	0.46	11.00	0.43	12.00	0.80	0.80
$\phi_{\alpha=\epsilon}$	Δ%	-0.40	-0.00	-2.79	-2.82	0.00	0.00	-3.37	-3.14
$\phi_{\alpha=0.25}$	Train	0.99	1.00	0.52	11.67	0.48	12.75	0.82	0.83
$\phi_{\alpha=0.25}$	Test	0.99	1.00	0.50	11.39	0.48	12.75	0.80	0.80
$\phi_{\alpha=0.25}$	Δ%	-0.21	0.00	-2.23	-2.40	0.00	0.00	-2.85	-3.09
$\phi_{\alpha=0.5}$	Train	0.99	1.00	0.53	12.00	0.48	12.97	0.81	0.83
$\phi_{\alpha=0.5}$	Test	0.99	1.00	0.52	11.73	0.48	12.97	0.79	0.80
$\phi_{\alpha=0.5}$	Δ%	-0.19	0.00	-2.12	-2.23	0.00	0.00	-2.81	-3.29
$\phi_{\alpha=0.75}$	Train	0.99	1.00	0.53	12.25	0.49	13.49	0.79	0.80
$\phi_{\alpha=0.75}$	Test	0.98	1.00	0.52	11.97	0.49	13.49	0.77	0.78
$\phi_{\alpha=0.75}$	Δ%	-0.15	0.00	-2.16	-2.29	0.00	0.00	-2.67	-2.91
$\phi_{\alpha=0.99}$	Train	0.99	1.00	0.54	12.43	0.50	14.38	0.78	0.75
$\phi_{\alpha=0.99}$	Test	0.98	1.00	0.53	12.14	0.50	14.38	0.76	0.72
$\phi_{\alpha=0.99}$	Δ%	-0.13	0.00	-2.33	-2.26	0.00	0.00	-2.64	-3.13
$\phi^{\Delta}_{\alpha=\epsilon}$	Train	0.86	0.97	0.19	4.28	0.01	0.35	0.88	0.94
$\phi^{\Delta}_{\alpha=\epsilon}$	Test	0.82	0.94	0.17	3.68	0.01	0.35	0.80	0.86
$\phi^{\Delta}_{\alpha=\epsilon}$	Δ%	-5.58	-2.75	-12.10	-14.11	0.00	0.00	-9.23	-8.31
$\phi^{\Delta}_{\alpha=0.25}$	Train	0.98	1.00	0.31	6.10	0.02	0.60	0.91	0.94
$\phi^{\Delta}_{\alpha=0.25}$	Test	0.97	0.99	0.29	5.44	0.02	0.60	0.87	0.88
$\phi^{\Delta}_{\alpha=0.25}$	Δ%	-1.22	-0.61	-8.17	-10.76	0.00	0.00	-4.57	-6.47
$\phi^{\Delta}_{\alpha=0.5}$	Train	0.98	1.00	0.35	7.05	0.03	0.85	0.90	0.92
$\phi^{\Delta}_{\alpha=0.5}$	Test	0.97	1.00	0.32	6.41	0.03	0.85	0.86	0.88
$\phi^{\Delta}_{\alpha=0.5}$	Δ%	-0.97	-0.33	-7.37	-9.09	0.00	0.00	-3.64	-4.88
$\phi^{\Delta}_{\alpha=0.75}$	Train	0.98	1.00	0.39	8.13	0.06	1.69	0.88	0.89
$\phi^{\Delta}_{\alpha=0.75}$	Test	0.97	1.00	0.36	7.43	0.06	1.69	0.85	0.86
$\phi^{\Delta}_{\alpha=0.75}$	Δ%	-0.64	-0.19	-7.12	-8.61	0.00	0.00	-3.10	-3.94
$\phi^{\Delta}_{\alpha=0.99}$	Train	0.98	1.00	0.44	9.62	0.14	4.56	0.84	0.82
$\phi^{\Delta}_{\alpha=0.99}$	Test	0.97	1.00	0.41	8.94	0.14	4.56	0.81	0.79
$\phi^{\Delta}_{\alpha=0.99}$	Δ%	-0.33	-0.08	-6.15	-7.03	0.00	0.00	-3.02	-3.17
θ	Train	0.99	1.00	0.51	12.13	0.46	13.70	0.81	0.77
θ	Test	0.98	1.00	0.49	11.84	0.46	13.70	0.79	0.75
θ	Δ%	-0.39	-0.00	-2.56	-2.40	0.00	0.00	-3.10	-2.93
θ^{Δ}	Train	0.82	0.95	0.22	4.89	0.02	0.60	0.85	0.92

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
θ^{Δ}	Test	0.78	0.92	0.20	4.26	0.02	0.60	0.77	0.83
θ^{Δ}	Δ%	-4.35	-3.21	-11.37	-12.98	0.00	0.00	-8.76	-9.12
ρ	Train	0.99	1.00	0.53	12.26	0.49	14.05	0.78	0.76
ρ	Test	0.98	1.00	0.52	11.97	0.49	14.05	0.76	0.74
ρ	Δ%	-0.13	0.00	-2.39	-2.31	0.00	0.00	-2.68	-2.72
ρ^{Δ}	Train	0.98	1.00	0.41	9.05	0.13	4.16	0.84	0.83
ρ^{Δ}	Test	0.97	1.00	0.39	8.39	0.13	4.16	0.82	0.80
ρ^{Δ}	Δ%	-0.40	-0.18	-6.35	-7.28	0.00	0.00	-3.20	-3.34

Table R.1: h_k-GREEDY Stopwords Included Not Discounted

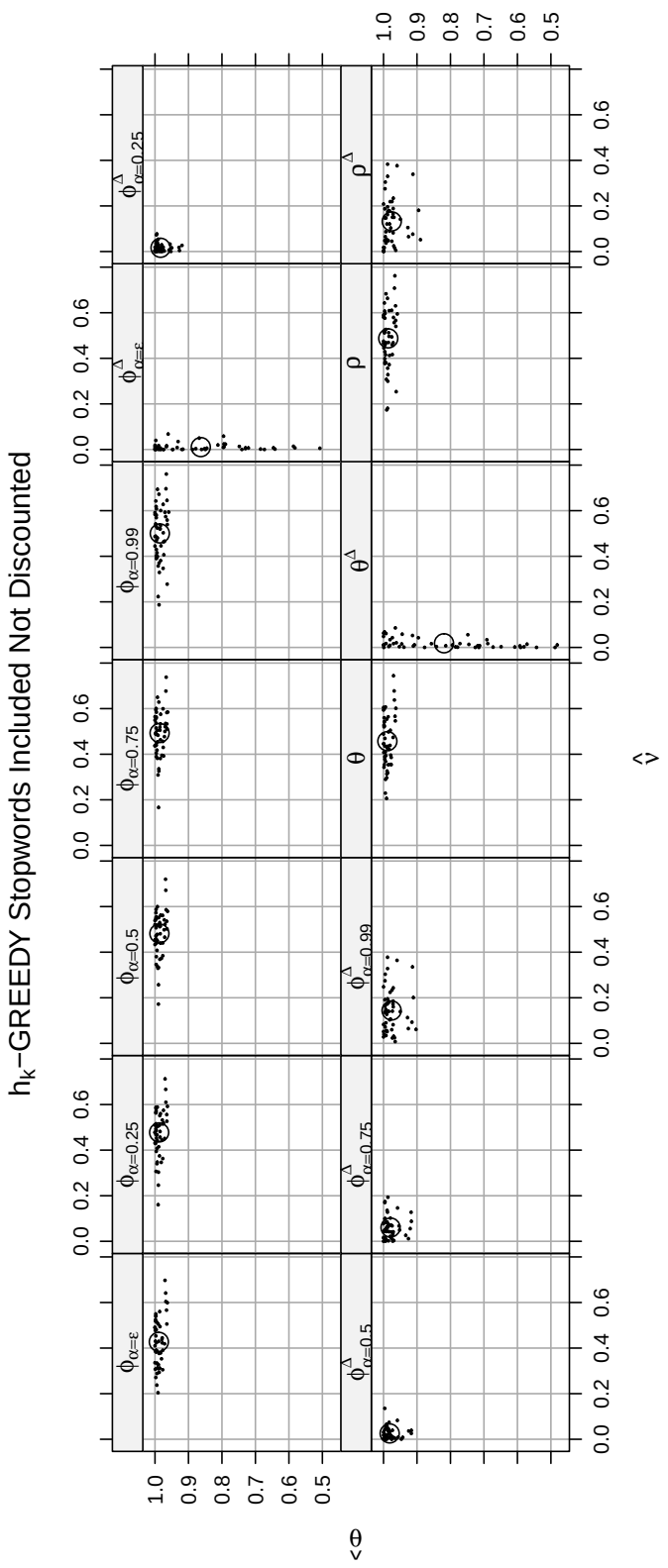


Figure R.1

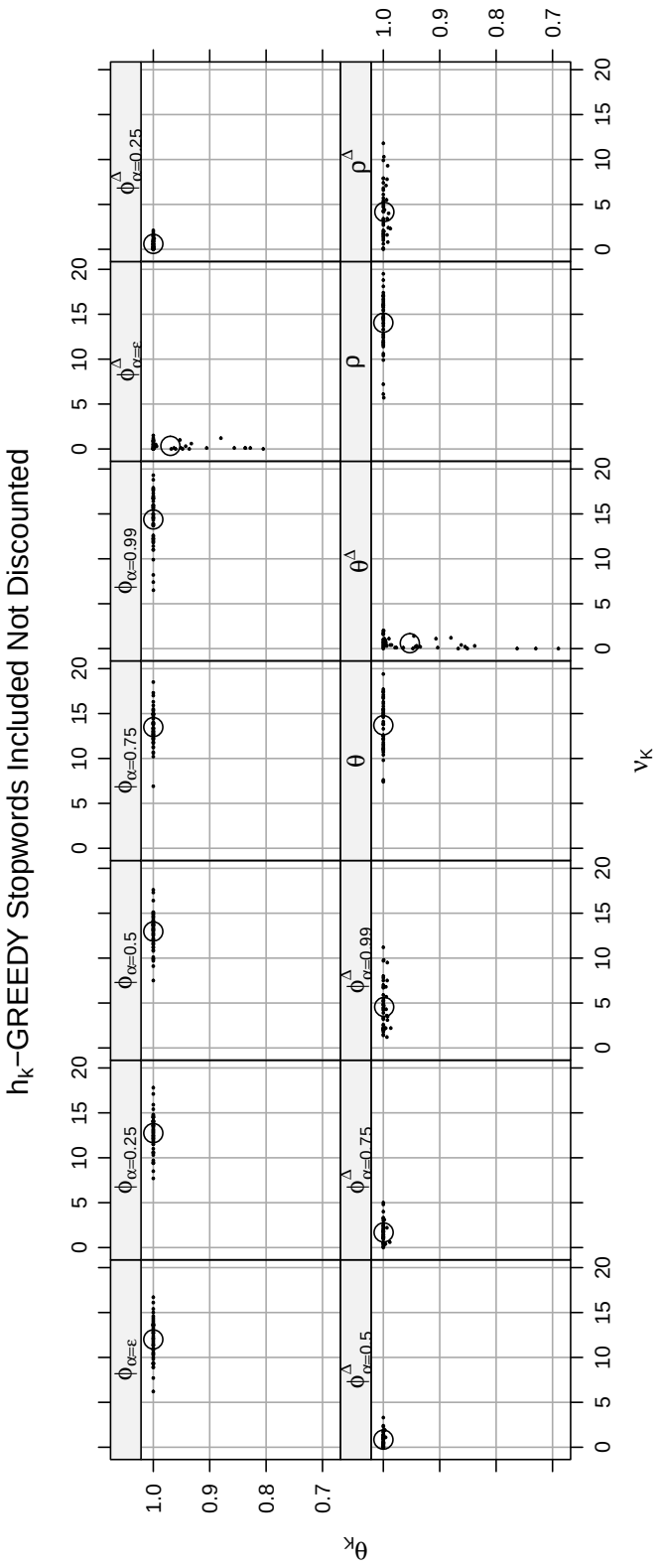

hk−GREEDY Stopwords Included Not Discounted
θκ
νκ
ϕα=ε
ϕα=0.25
ϕα=0.5
ϕα=0.75
ϕα=0.99
ϕα=F
ϕΔα=0.25
θ
θΔ
ρ
ρΔ
ϕΔα=0.5
ϕΔα=0.75
ϕΔα=0.99

Figure R.2

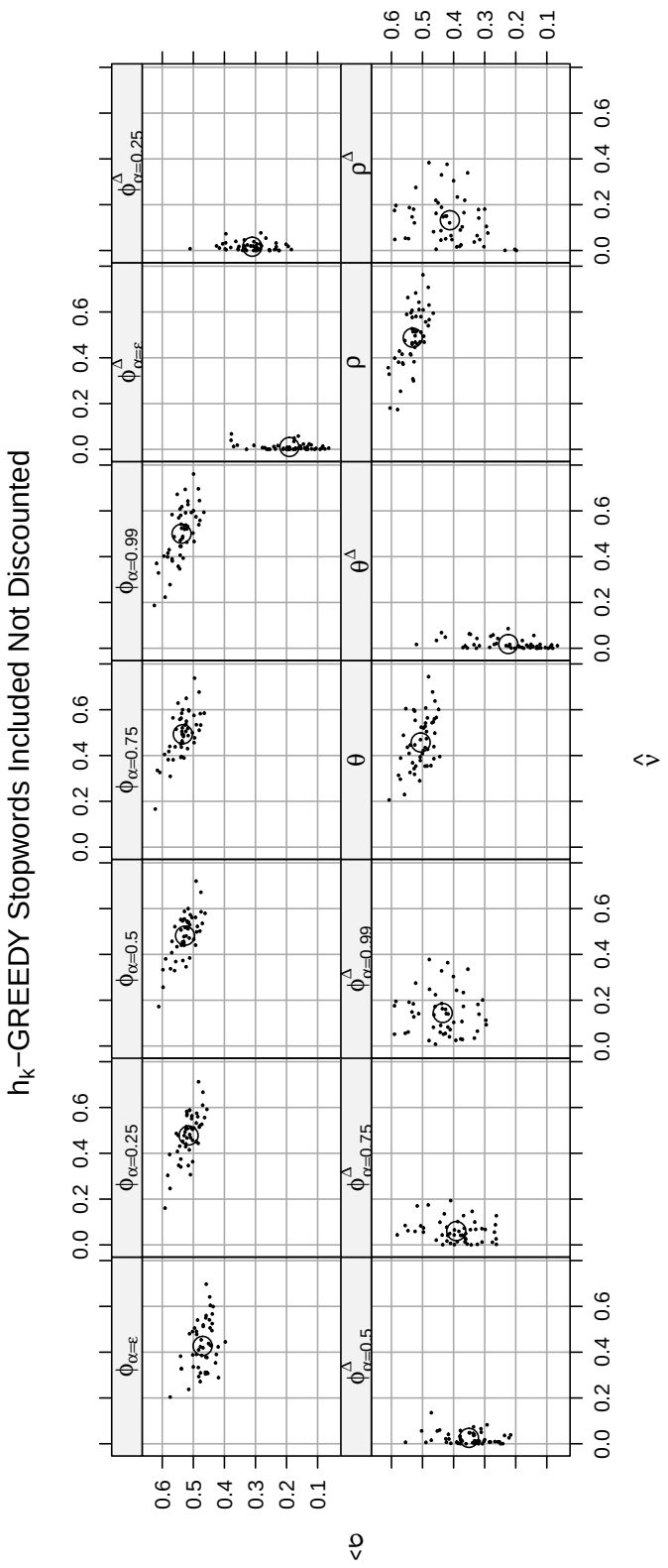

h_k−GREEDY Stopwords Included Not Discounted
σ̂
ν̂
ϕ_{α=ε}
ϕ_{α=0.25}
ϕ_{α=0.5}
ϕ_{α=0.75}
ϕ_{α=0.99}
ϕ^Δ_{α=ε}
ϕ^Δ_{α=0.25}
ϕ^Δ_{α=0.5}
ϕ^Δ_{α=0.75}
ϕ^Δ_{α=0.99}
ρ
θ
ρ^Δ
θ^Δ

Figure R.3

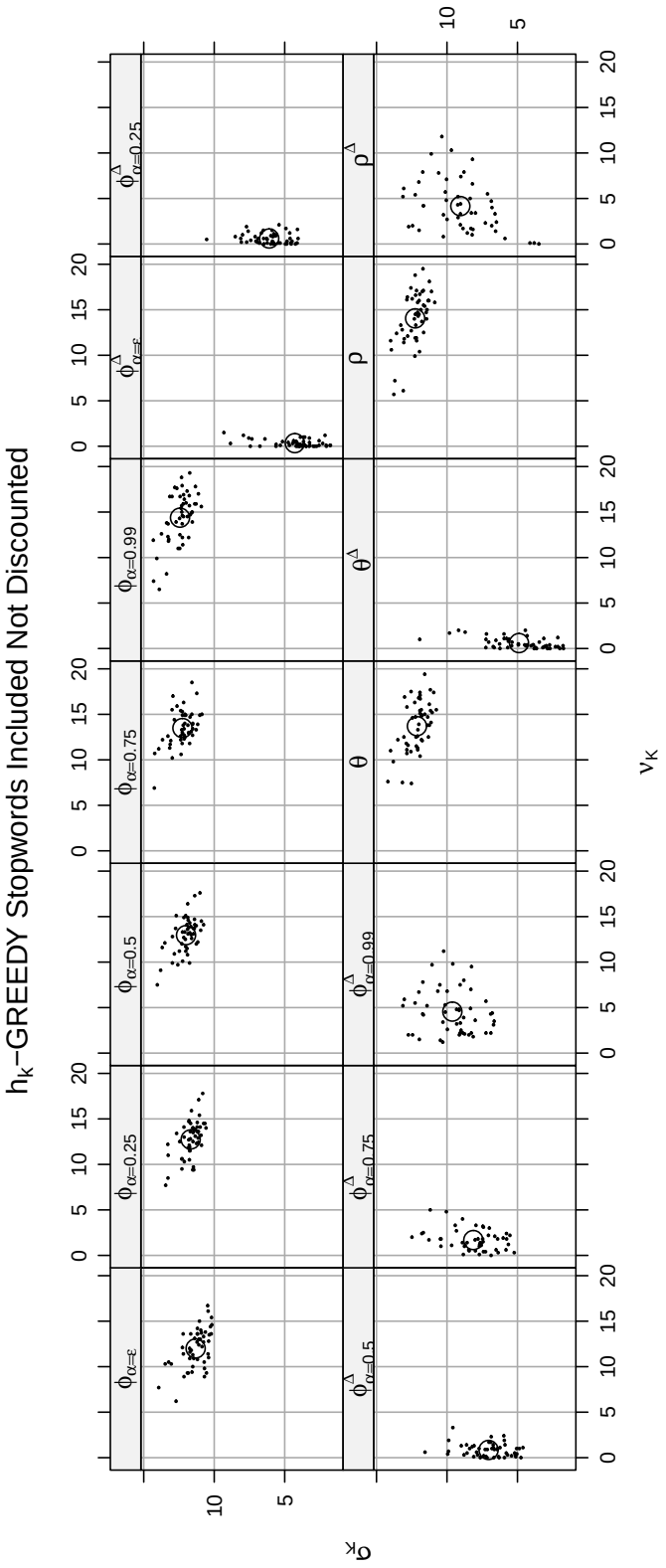

h_k−GREEDY Stopwords Included Not Discounted
σ_K
ν_K
φ_α=ε
φ_α=0.25
φ_α=0.5
φ_α=0.75
φ_α=0.99
φ^Δ_α=ε
φ^Δ_α=0.25
φ^Δ_α=0.5
φ^Δ_α=0.75
φ^Δ_α=0.99
θ
θ^Δ
ρ
ρ^Δ

Figure R.4

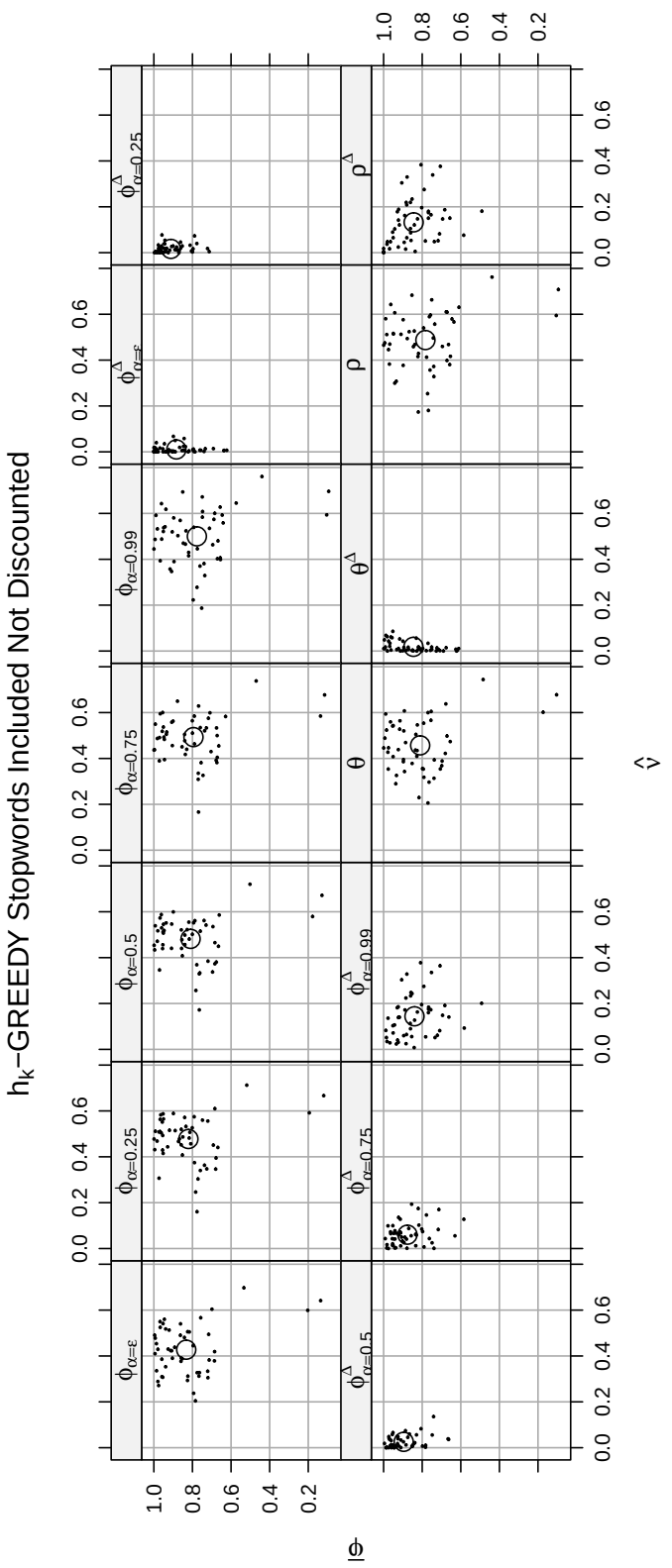

h_k−GREEDY Stopwords Included Not Discounted
Φ
V̂
ϕα=ε
ϕα=0.25
ϕα=0.5
ϕα=0.75
ϕα=0.99
ϕα=ε
ϕα=0.25
ϕΔα=0.5
ϕΔα=0.75
ϕΔα=0.99
θ
θΔ
ρ
ρΔ

Figure R.5

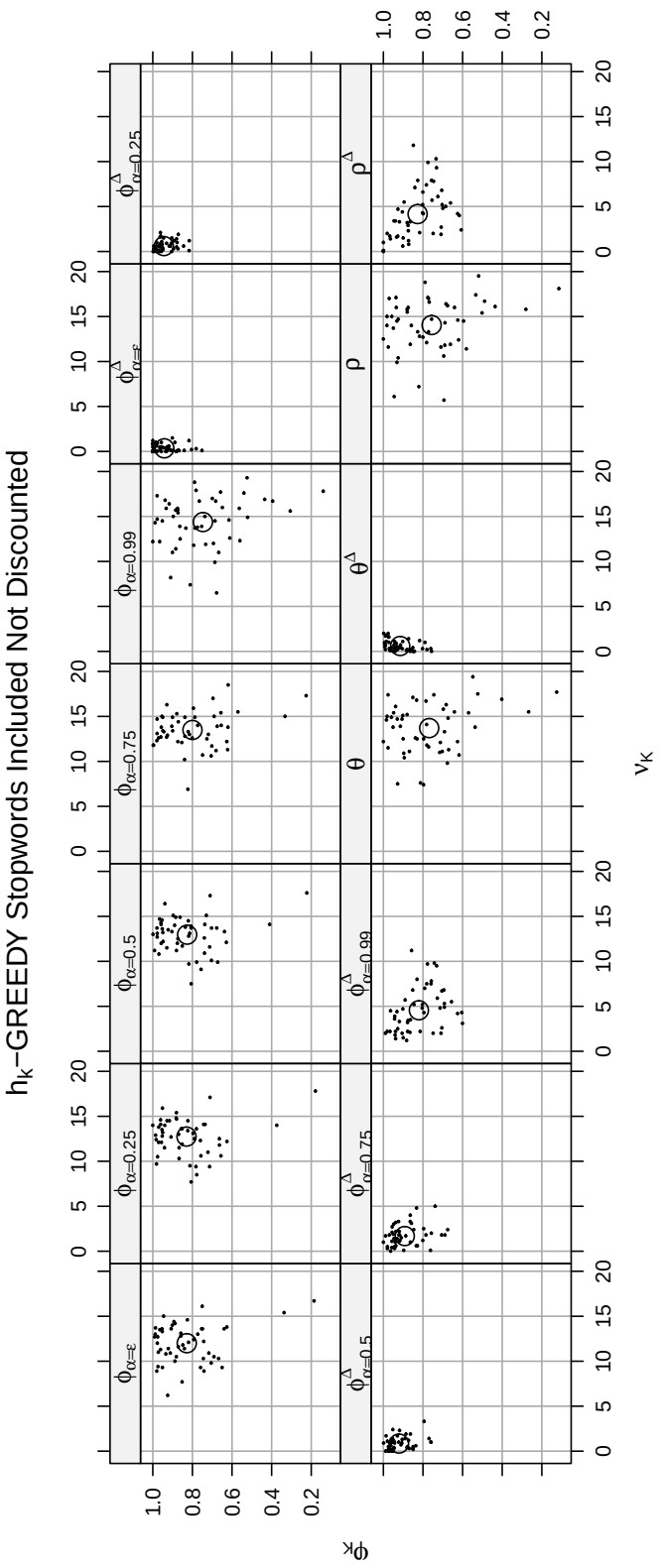

h_k−GREEDY Stopwords Included Not Discounted
φ_K
ν_K
φ^Δ_α=ε
φ_α=0.25
φ_α=0.5
φ_α=0.75
φ_α=0.99
φ^Δ_α=ε
φ^Δ_α=0.25
φ^Δ_α=0.5
φ^Δ_α=0.75
φ^Δ_α=0.99
θ
θ^Δ
ρ
ρ^Δ

Figure R.6

Appendix S

h_k-GREEDY Results Stopwords Included Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
$\phi_{\alpha=\epsilon}$	Train	0.98	1.00	0.28	5.99	0.43	12.00	0.84	0.88
$\phi_{\alpha=\epsilon}$	Test	0.97	1.00	0.27	5.80	0.43	12.00	0.82	0.85
$\phi_{\alpha=\epsilon}$	Δ%	-0.76	-0.10	-3.16	-3.20	0.00	0.00	-3.04	-2.63
$\phi_{\alpha=0.25}$	Train	0.98	1.00	0.29	5.94	0.48	12.75	0.83	0.87
$\phi_{\alpha=0.25}$	Test	0.97	1.00	0.28	5.77	0.48	12.75	0.81	0.85
$\phi_{\alpha=0.25}$	Δ%	-0.38	-0.06	-2.48	-2.92	0.00	0.00	-2.43	-2.36
$\phi_{\alpha=0.5}$	Train	0.98	1.00	0.29	6.05	0.48	12.97	0.83	0.87
$\phi_{\alpha=0.5}$	Test	0.97	1.00	0.29	5.89	0.48	12.97	0.81	0.85
$\phi_{\alpha=0.5}$	Δ%	-0.31	-0.06	-2.33	-2.60	0.00	0.00	-2.34	-2.13
$\phi_{\alpha=0.75}$	Train	0.98	1.00	0.29	5.95	0.49	13.49	0.82	0.86
$\phi_{\alpha=0.75}$	Test	0.97	1.00	0.29	5.78	0.49	13.49	0.80	0.85
$\phi_{\alpha=0.75}$	Δ%	-0.23	-0.04	-2.37	-2.84	0.00	0.00	-2.05	-1.70
$\phi_{\alpha=0.99}$	Train	0.98	1.00	0.29	5.71	0.50	14.38	0.81	0.85
$\phi_{\alpha=0.99}$	Test	0.97	1.00	0.29	5.56	0.50	14.38	0.80	0.83
$\phi_{\alpha=0.99}$	Δ%	-0.19	-0.05	-2.54	-2.66	0.00	0.00	-2.17	-2.89
$\phi^{\Delta}_{\alpha=\epsilon}$	Train	0.86	0.97	0.19	4.20	0.01	0.35	0.88	0.94
$\phi^{\Delta}_{\alpha=\epsilon}$	Test	0.81	0.94	0.17	3.60	0.01	0.35	0.80	0.86
$\phi^{\Delta}_{\alpha=\epsilon}$	Δ%	-5.73	-2.98	-12.17	-14.21	0.00	0.00	-9.19	-8.17
$\phi^{\Delta}_{\alpha=0.25}$	Train	0.98	1.00	0.31	5.99	0.02	0.60	0.91	0.94
$\phi^{\Delta}_{\alpha=0.25}$	Test	0.97	0.99	0.28	5.35	0.02	0.60	0.87	0.88
$\phi^{\Delta}_{\alpha=0.25}$	Δ%	-1.22	-0.62	-8.05	-10.60	0.00	0.00	-4.54	-6.44
$\phi^{\Delta}_{\alpha=0.5}$	Train	0.98	1.00	0.34	6.81	0.03	0.85	0.90	0.92
$\phi^{\Delta}_{\alpha=0.5}$	Test	0.97	1.00	0.32	6.21	0.03	0.85	0.87	0.88
$\phi^{\Delta}_{\alpha=0.5}$	Δ%	-0.97	-0.36	-7.16	-8.86	0.00	0.00	-3.50	-4.54
$\phi^{\Delta}_{\alpha=0.75}$	Train	0.98	1.00	0.36	7.41	0.06	1.69	0.88	0.90
$\phi^{\Delta}_{\alpha=0.75}$	Test	0.97	1.00	0.34	6.79	0.06	1.69	0.86	0.87
$\phi^{\Delta}_{\alpha=0.75}$	Δ%	-0.78	-0.33	-6.84	-8.40	0.00	0.00	-2.88	-3.60
$\phi^{\Delta}_{\alpha=0.99}$	Train	0.97	1.00	0.37	7.52	0.14	4.56	0.86	0.87
$\phi^{\Delta}_{\alpha=0.99}$	Test	0.96	0.99	0.34	6.98	0.14	4.56	0.83	0.84
$\phi^{\Delta}_{\alpha=0.99}$	Δ%	-0.56	-0.34	-5.93	-7.11	0.00	0.00	-2.72	-2.90
θ	Train	0.98	1.00	0.29	5.77	0.46	13.70	0.83	0.86
θ	Test	0.97	1.00	0.28	5.60	0.46	13.70	0.81	0.83
θ	Δ%	-0.64	-0.09	-3.00	-2.97	0.00	0.00	-2.86	-3.23
θ^{Δ}	Train	0.82	0.95	0.22	4.68	0.02	0.60	0.85	0.91

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
θ^{Δ}	Test	0.78	0.92	0.19	4.06	0.02	0.60	0.77	0.83
θ^{Δ}	Δ%	-4.36	-3.20	-11.40	-13.13	0.00	0.00	-8.57	-9.09
ρ	Train	0.98	1.00	0.29	5.73	0.49	14.05	0.82	0.85
ρ	Test	0.97	1.00	0.29	5.57	0.49	14.05	0.80	0.83
ρ	Δ%	-0.21	-0.06	-2.63	-2.77	0.00	0.00	-2.24	-2.65
ρ^{Δ}	Train	0.97	0.99	0.35	7.13	0.13	4.16	0.86	0.87
ρ^{Δ}	Test	0.96	0.99	0.33	6.61	0.13	4.16	0.84	0.84
ρ^{Δ}	Δ%	-0.62	-0.46	-6.22	-7.34	0.00	0.00	-2.93	-3.04

Table S.1: h_k-GREEDY Stopwords Included Discounted

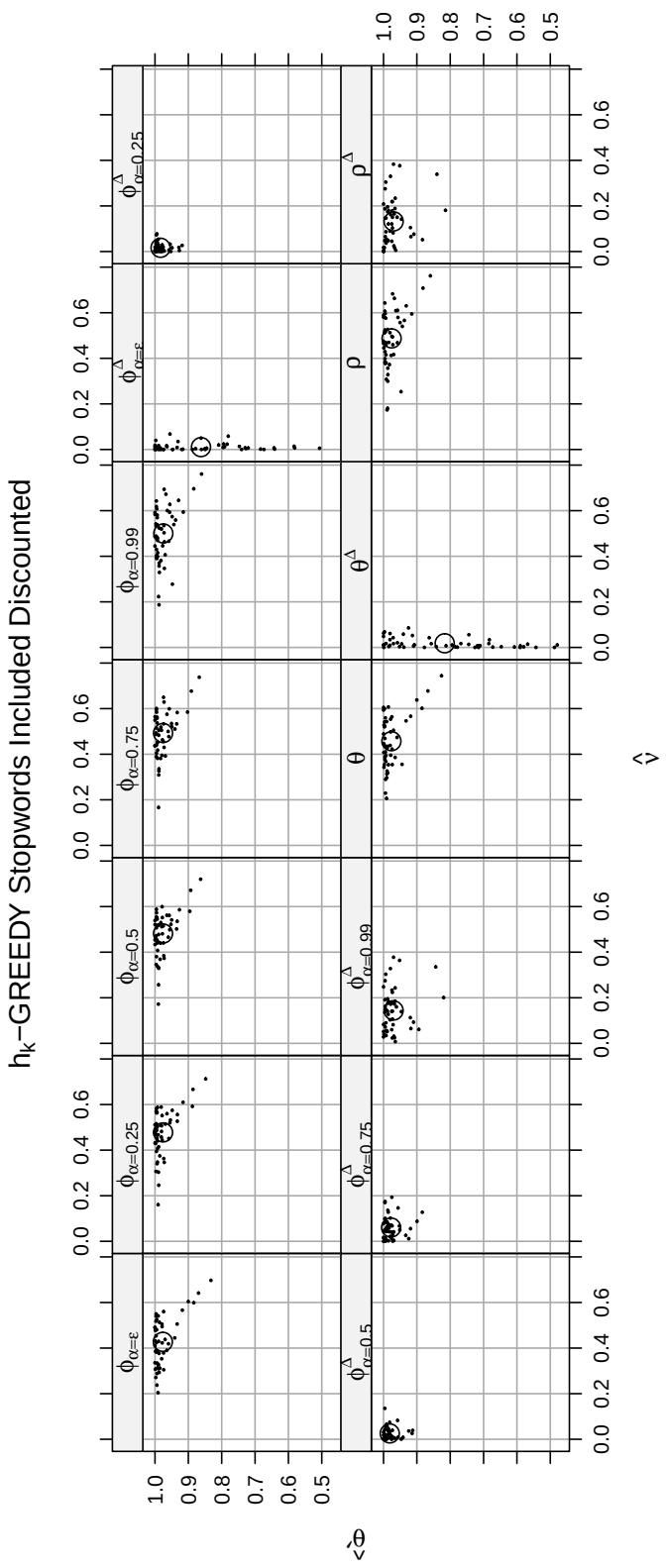


Figure S.1

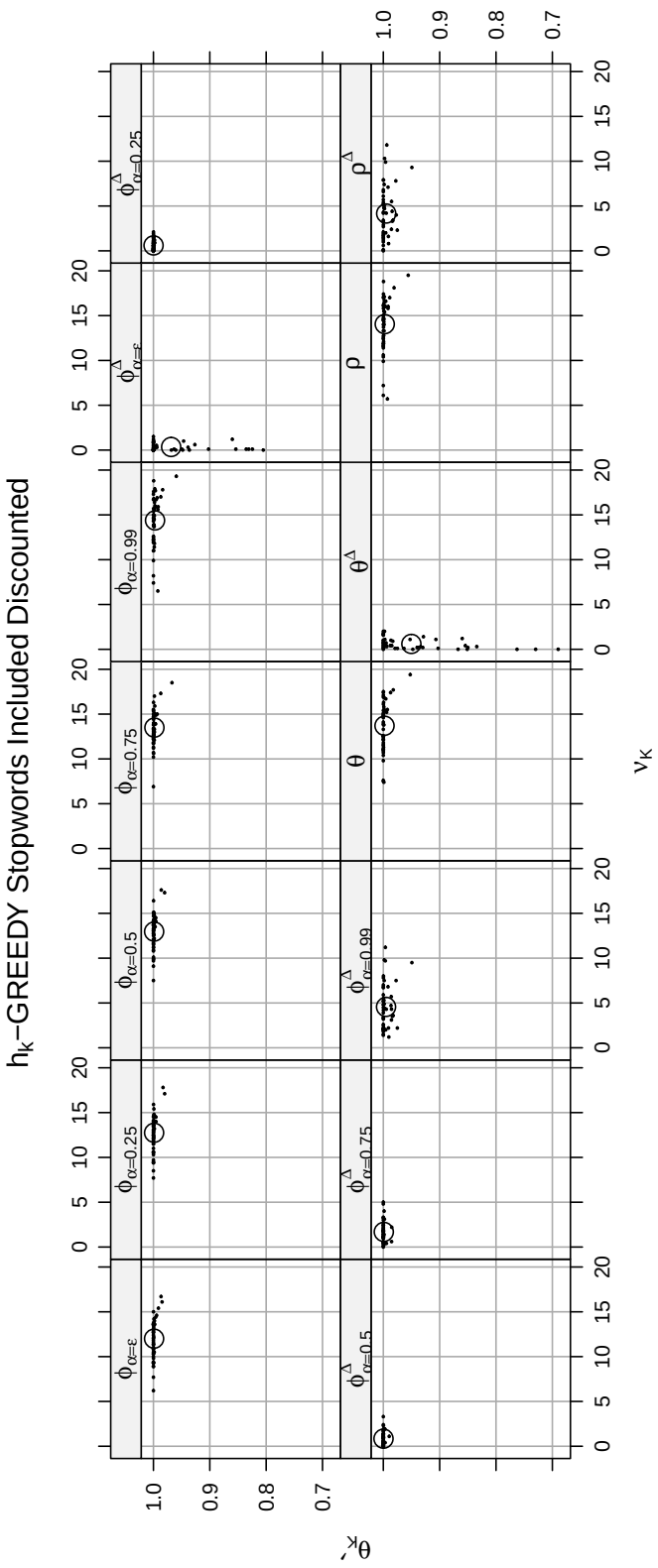

h_k–GREEDY Stopwords Included Discounted
θ_k'
v_k
ϕ α=ε
ϕ α=0.25
ϕ α=0.5
ϕ α=0.75
ϕ α=0.99
ϕ α=F
ϕ α=0.25
ϕ α=0.5
ϕ α=0.75
ϕ α=0.99
θ
θ
ρ
ρ

Figure S.2

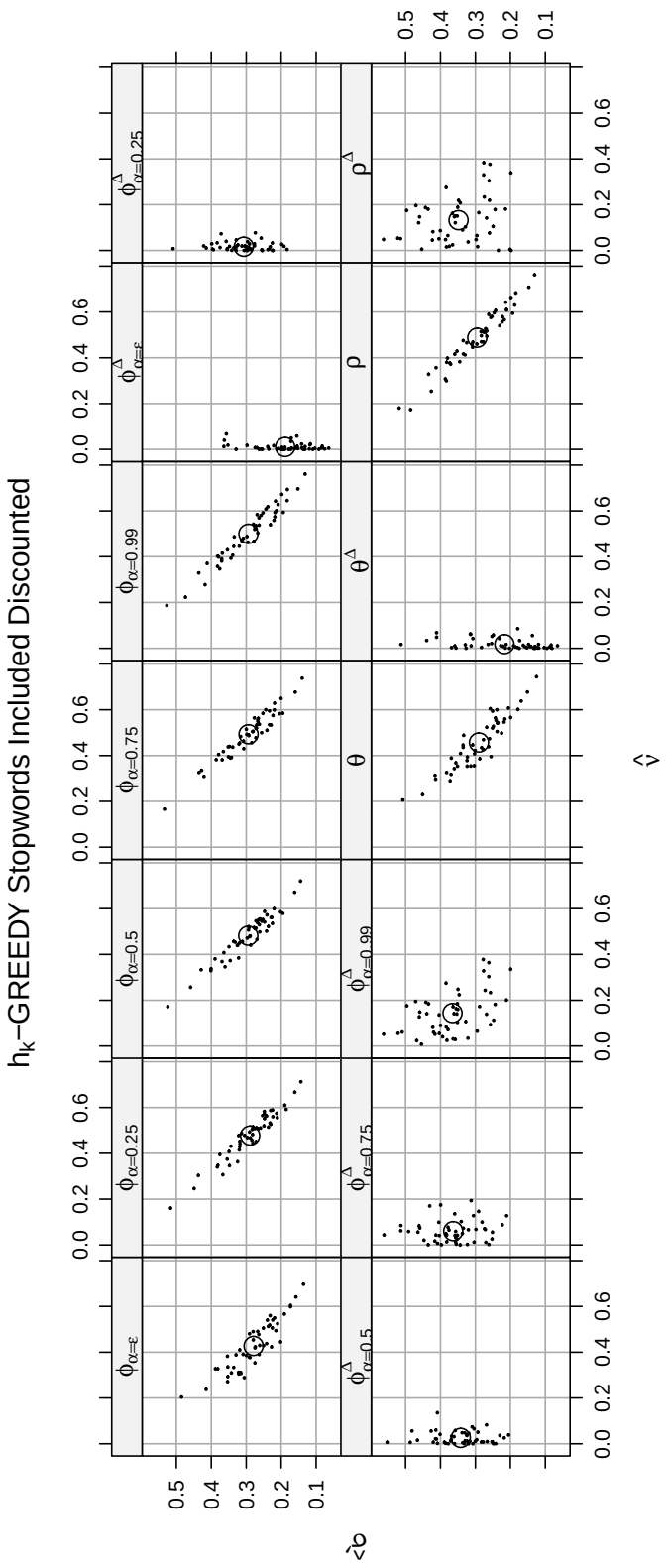

h_k−GREEDY Stopwords Included Discounted
σ²
v̂
φ_α=ε
φ_α=0.25
φ_α=0.5
φ_α=0.75
φ_α=0.99
φ^Δ_α=ε
φ^Δ_α=0.25
φ^Δ_α=0.5
φ^Δ_α=0.75
φ^Δ_α=0.99
θ
θ^Δ
ρ
ρ^Δ
0.0 0.2 0.4 0.6
0.1 0.2 0.3 0.4 0.5

Figure S.3

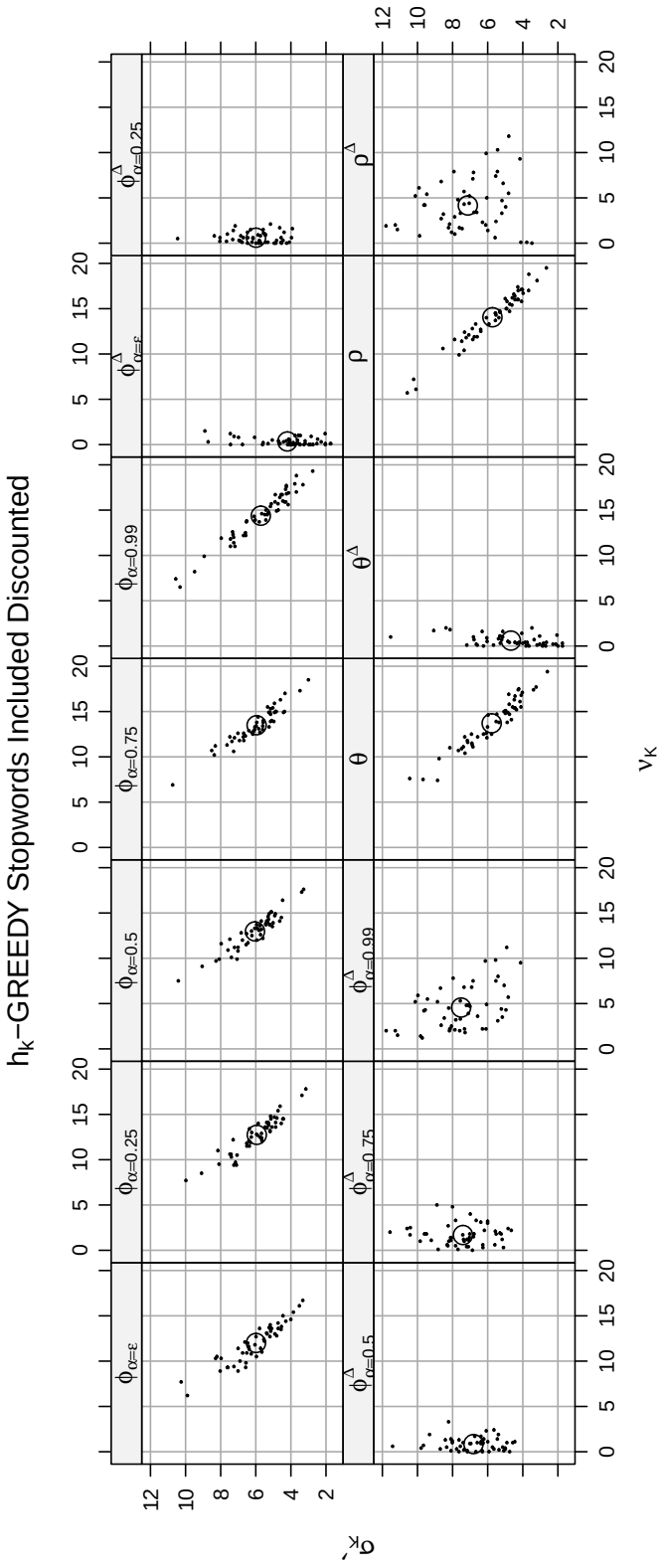

h_k–GREEDY Stopwords Included Discounted
σ_k'
ν_K
φ α=ε
φ α=0.25
φ α=0.5
φ α=0.75
φ α=0.99
φ α=1
φ Δ α=0.25
φ Δ α=0.5
φ Δ α=0.75
φ Δ α=0.99
θ
θ Δ
ρ
ρ Δ

Figure S.4

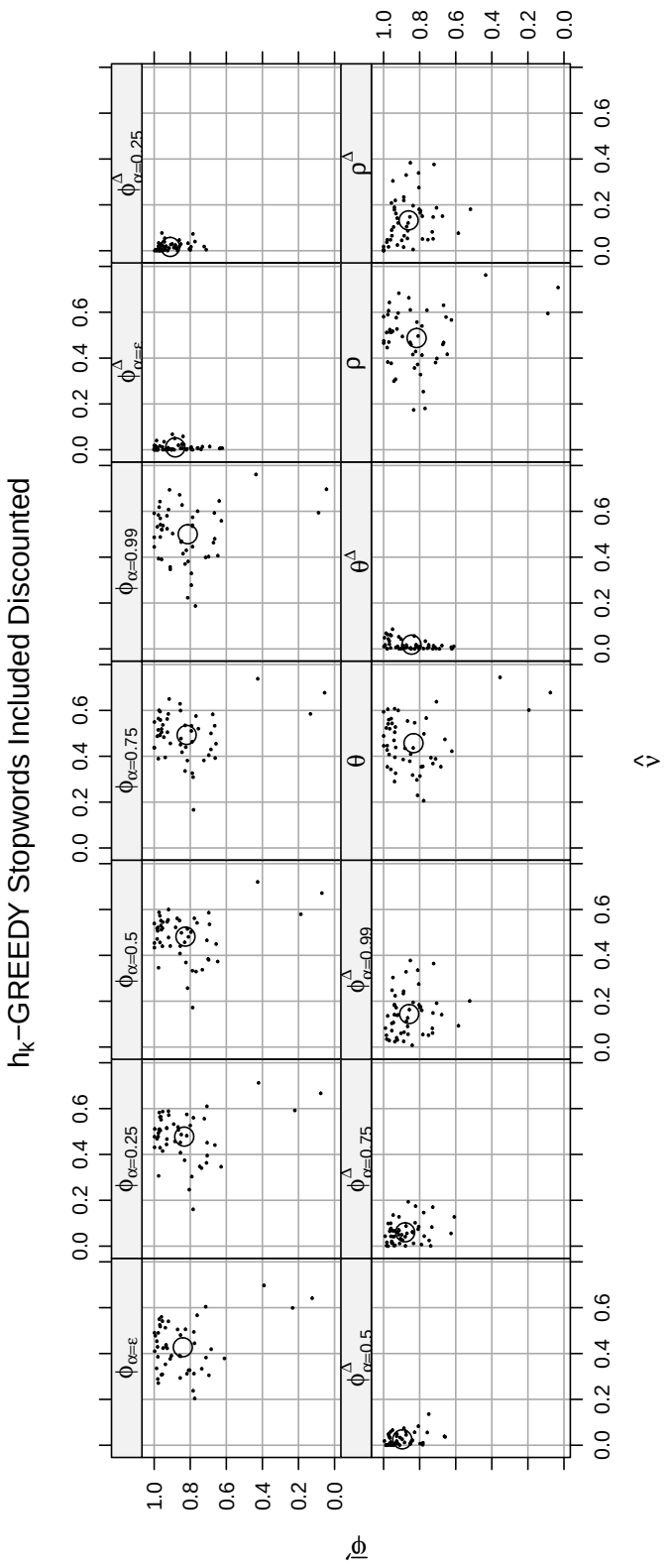

h_k−GREEDY Stopwords Included Discounted
ϕ_{α=ε}
ϕ_{α=0.25}
ϕ_{α=0.5}
ϕ_{α=0.75}
ϕ_{α=0.99}
ϕ^Δ_{α=ε}
ϕ^Δ_{α=0.25}
ϕ^Δ_{α=0.5}
ϕ^Δ_{α=0.75}
ϕ^Δ_{α=0.99}
θ
θ^Δ
ρ
ρ^Δ
0.0 0.2 0.4 0.6
0.0 0.2 0.4 0.6 0.8 1.0
V̂
ϕ̂

Figure S.5

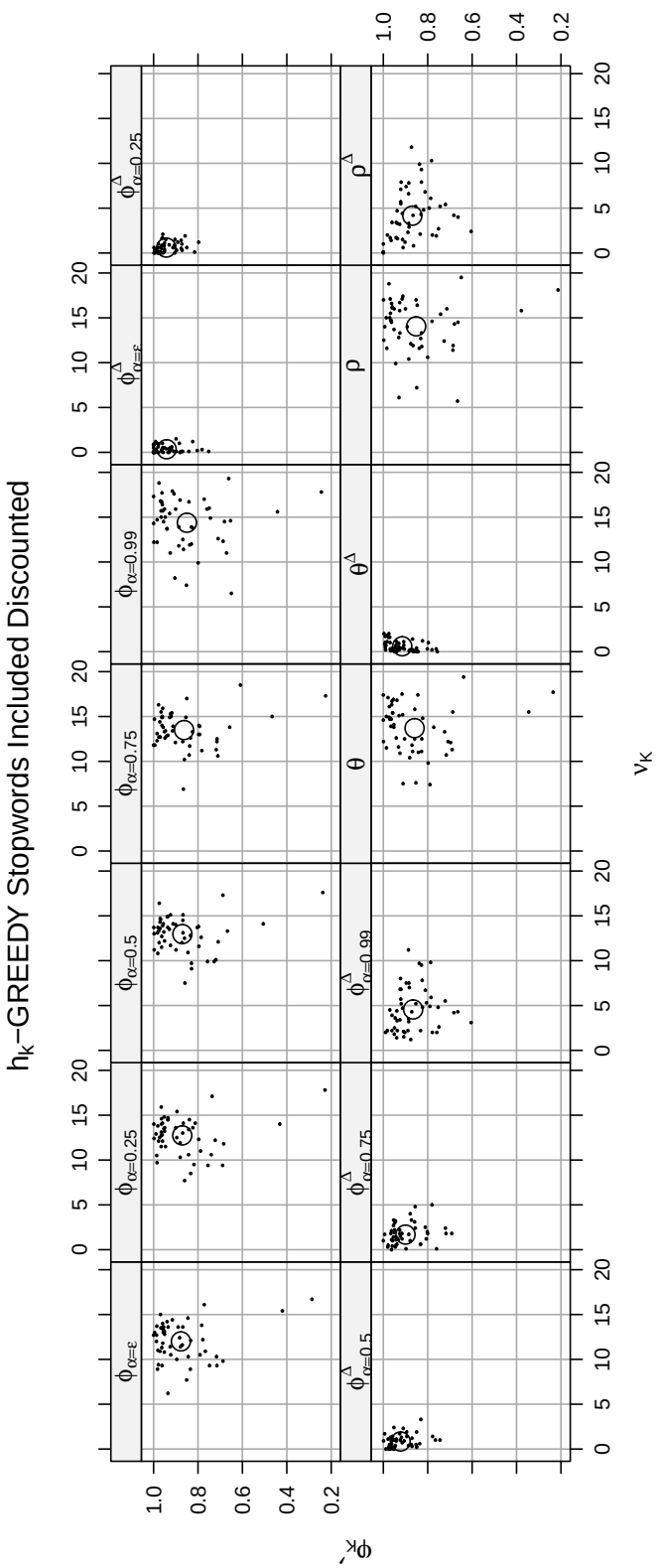

h_k–GREEDY Stopwords Included Discounted
φ_k^
ν_k
φ_{α=ε}
φ_{α=0.25}
φ_{α=0.5}
φ_{α=0.75}
φ_{α=0.99}
φ^Δ_{α=ε}
φ^Δ_{α=0.25}
φ^Δ_{α=0.5}
φ^Δ_{α=0.75}
φ^Δ_{α=0.99}
θ
θ^Δ
ρ
ρ^Δ

Figure S.6

Appendix T

Ensemble Results Stopwords Included Not Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
ℓ_1	Train	0.97	1.00	0.43	9.49	0.13	4.15	0.84	0.83
ℓ_1	Test	0.97	1.00	0.40	8.80	0.13	4.15	0.82	0.80
ℓ_1	Δ%	-0.43	-0.14	-6.28	-7.27	0.00	0.00	-3.16	-3.40
ℓ_2	Train	0.96	1.00	0.42	9.28	0.11	3.70	0.85	0.83
ℓ_2	Test	0.96	1.00	0.39	8.59	0.11	3.70	0.82	0.80
ℓ_2	Δ%	-0.63	-0.09	-6.46	-7.38	0.00	0.00	-3.06	-3.02
ℓ_3	Train	0.97	1.00	0.43	9.49	0.13	4.15	0.84	0.83
ℓ_3	Test	0.97	1.00	0.40	8.80	0.13	4.15	0.82	0.80
ℓ_3	Δ%	-0.42	-0.14	-6.27	-7.24	0.00	0.00	-3.16	-3.42
ℓ_4	Train	0.97	0.99	0.38	8.01	0.02	0.92	0.88	0.88
ℓ_4	Test	0.96	0.99	0.35	7.39	0.02	0.92	0.85	0.86
ℓ_4	Δ%	-0.55	-0.32	-6.44	-7.70	0.00	0.00	-2.60	-2.72
ℓ_5	Train	0.95	0.99	0.31	6.58	0.05	1.40	0.85	0.85
ℓ_5	Test	0.94	0.98	0.28	5.92	0.05	1.40	0.82	0.82
ℓ_5	Δ%	-1.18	-0.68	-9.15	-10.12	0.00	0.00	-3.81	-4.01
α_1	Train	0.97	1.00	0.39	8.27	0.02	0.97	0.87	0.87
α_1	Test	0.97	0.99	0.36	7.58	0.02	0.97	0.84	0.85
α_1	Δ%	-0.49	-0.12	-6.67	-8.32	0.00	0.00	-2.68	-2.45
α_2	Train	0.97	0.99	0.37	7.53	0.01	0.38	0.87	0.89
α_2	Test	0.96	0.99	0.35	6.91	0.01	0.38	0.85	0.86
α_2	Δ%	-0.50	-0.25	-6.48	-8.24	0.00	0.00	-2.76	-3.23
α_3	Train	0.97	0.99	0.37	7.70	0.01	0.34	0.87	0.89
α_3	Test	0.96	0.99	0.35	7.04	0.01	0.34	0.85	0.86
α_3	Δ%	-0.49	-0.03	-6.68	-8.49	0.00	0.00	-2.60	-2.54
α_4	Train	0.96	0.99	0.36	7.45	0.01	0.59	0.88	0.90
α_4	Test	0.96	0.99	0.33	6.82	0.01	0.59	0.86	0.87
α_4	Δ%	-0.60	-0.16	-6.75	-8.40	0.00	0.00	-2.68	-3.24
α_5	Train	0.96	0.99	0.35	7.23	0.01	0.32	0.89	0.91
α_5	Test	0.96	0.99	0.33	6.61	0.01	0.32	0.87	0.87
α_5	Δ%	-0.61	-0.52	-6.91	-8.54	0.00	0.00	-2.68	-3.52

Table T.1: μ-ENSEMBLE Stopwords Included Not Discounted

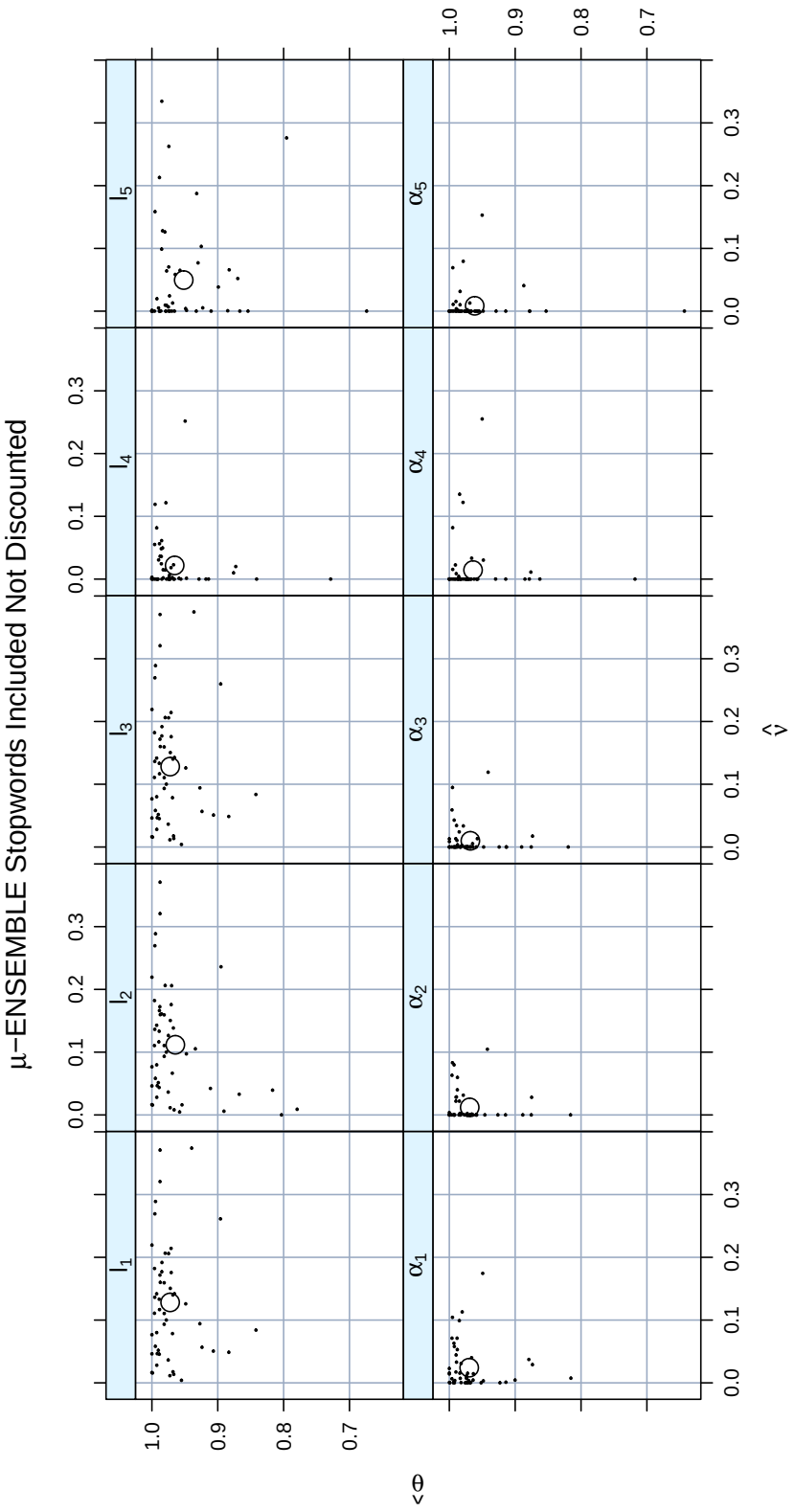

μ−ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
$\hat{\theta}$
$\hat{v}$

Figure T.1

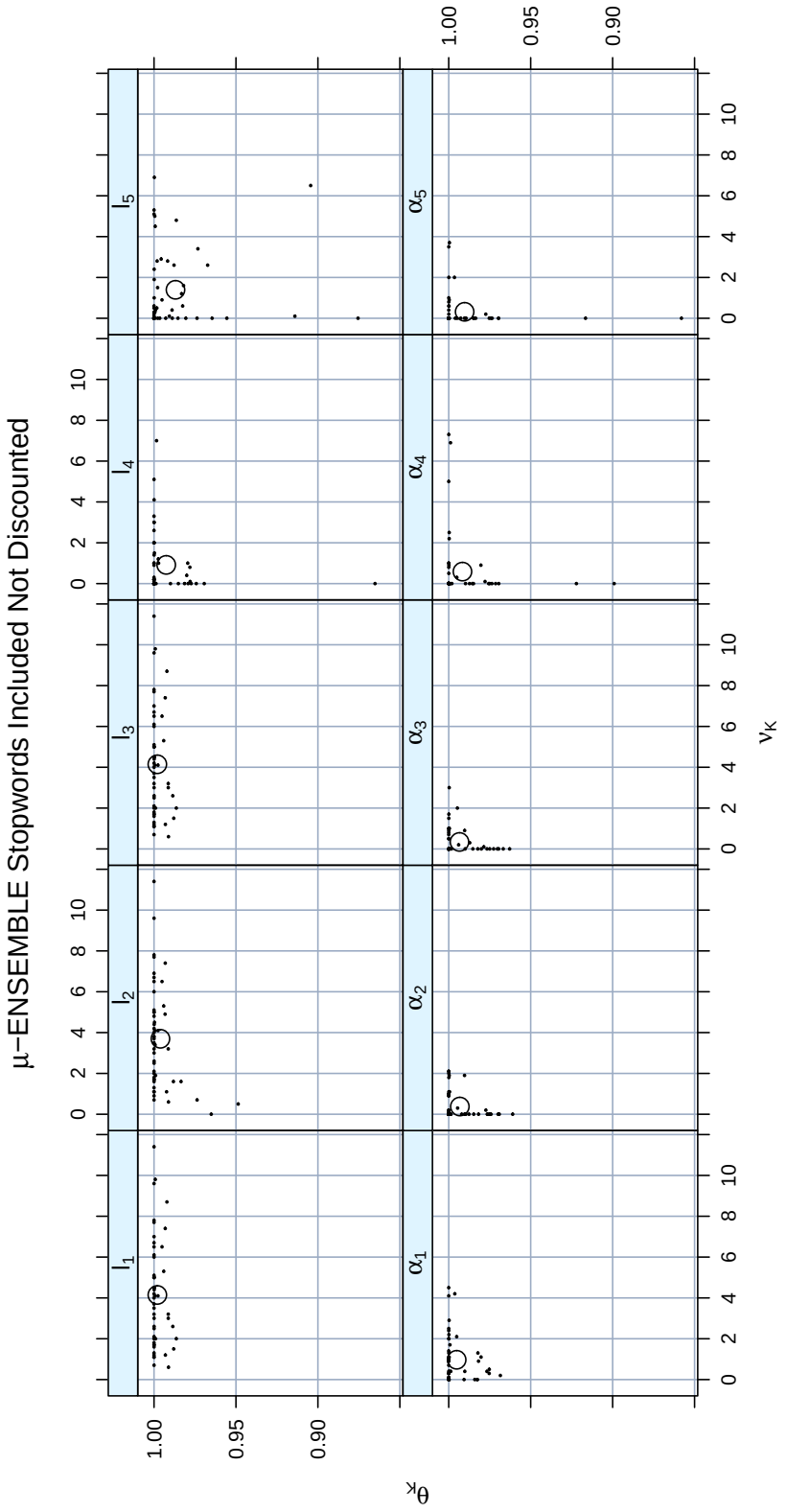

μ–ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
θ_K
ν_K
1.00
0.95
0.90
0
2
4
6
8
10

Figure T.2

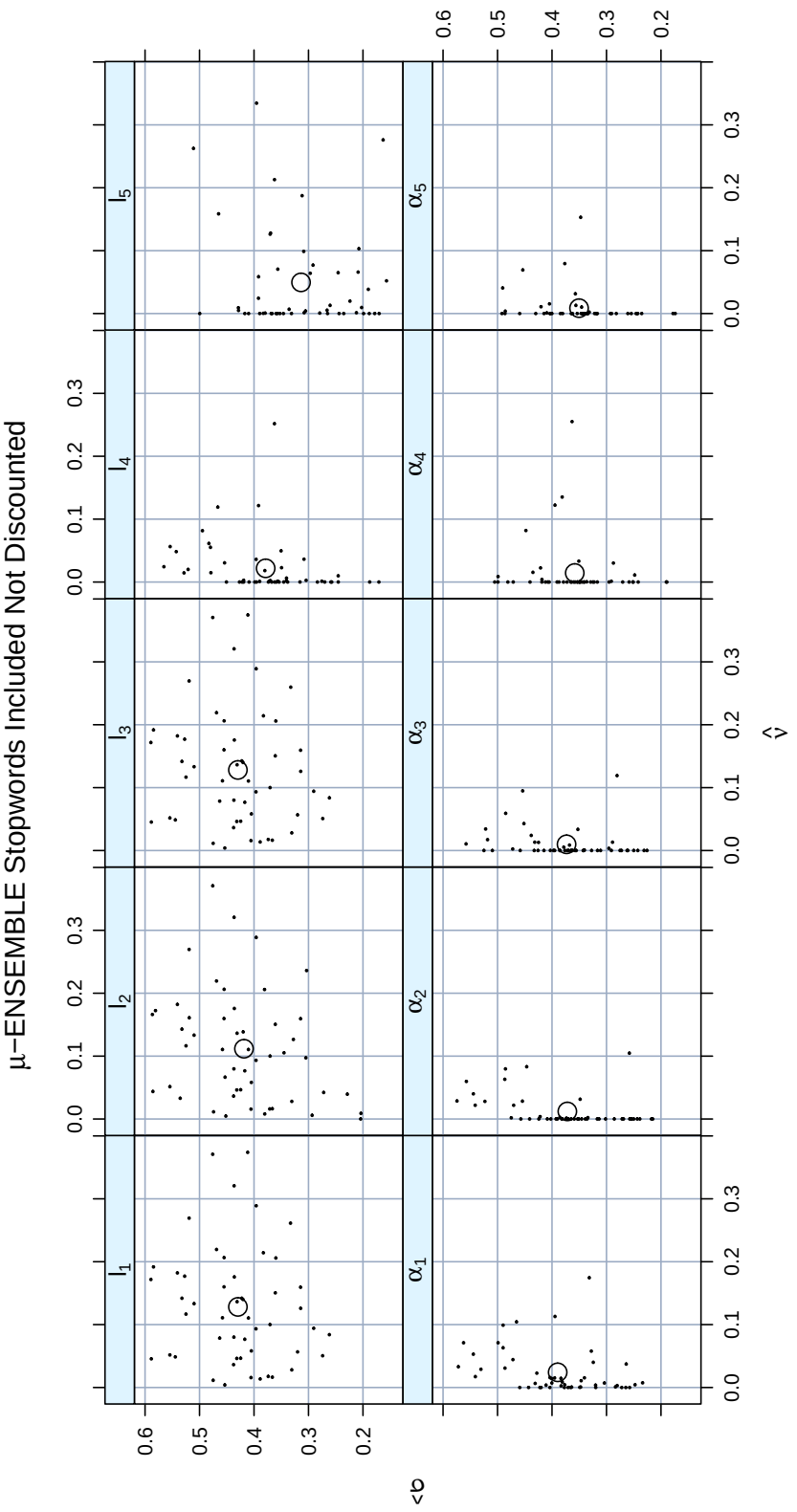

μ−ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
$\hat{\alpha}$
$\hat{V}$

Figure T.3

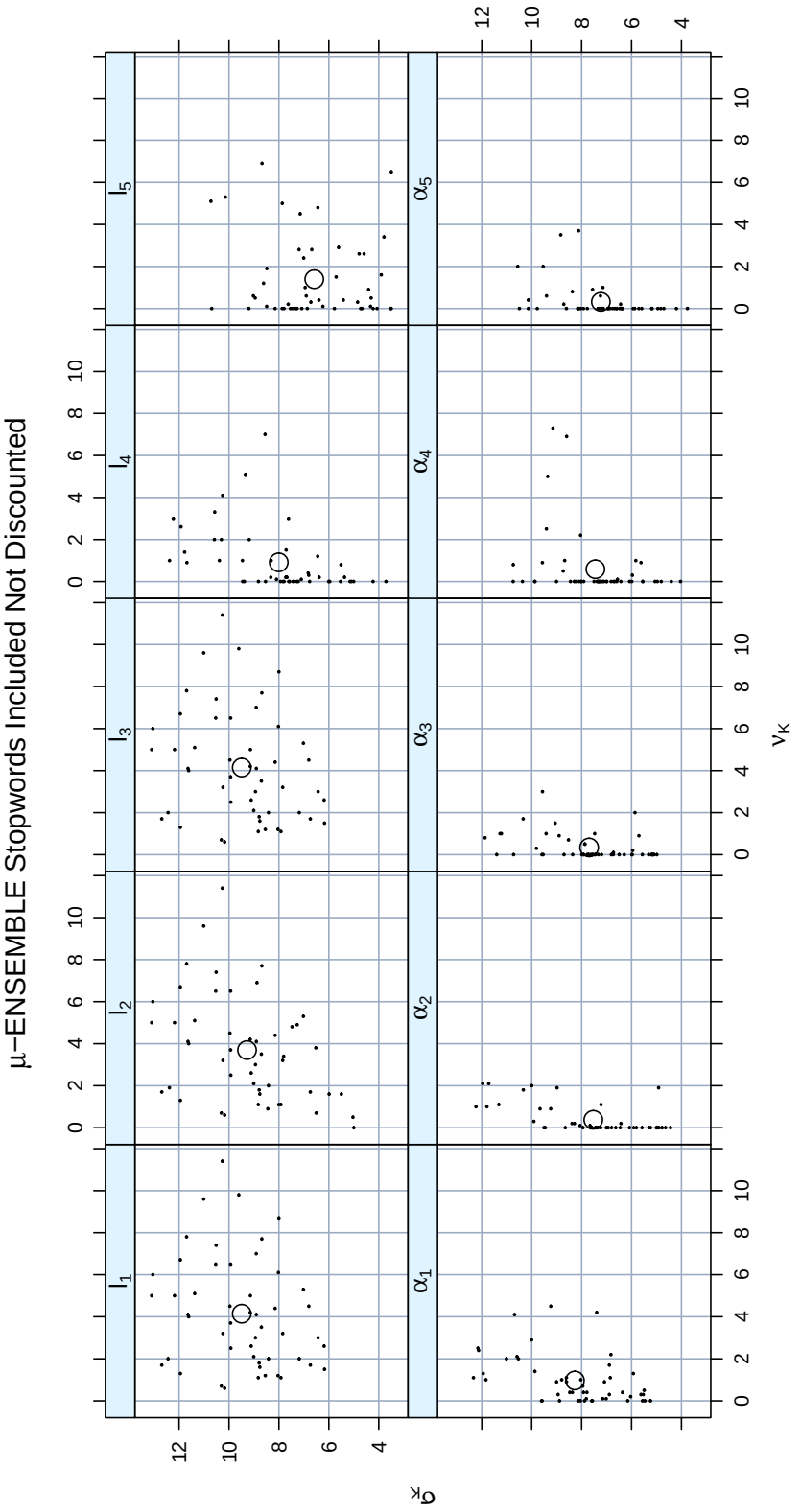

μ−ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
σ_K
ν_K

Figure T.4

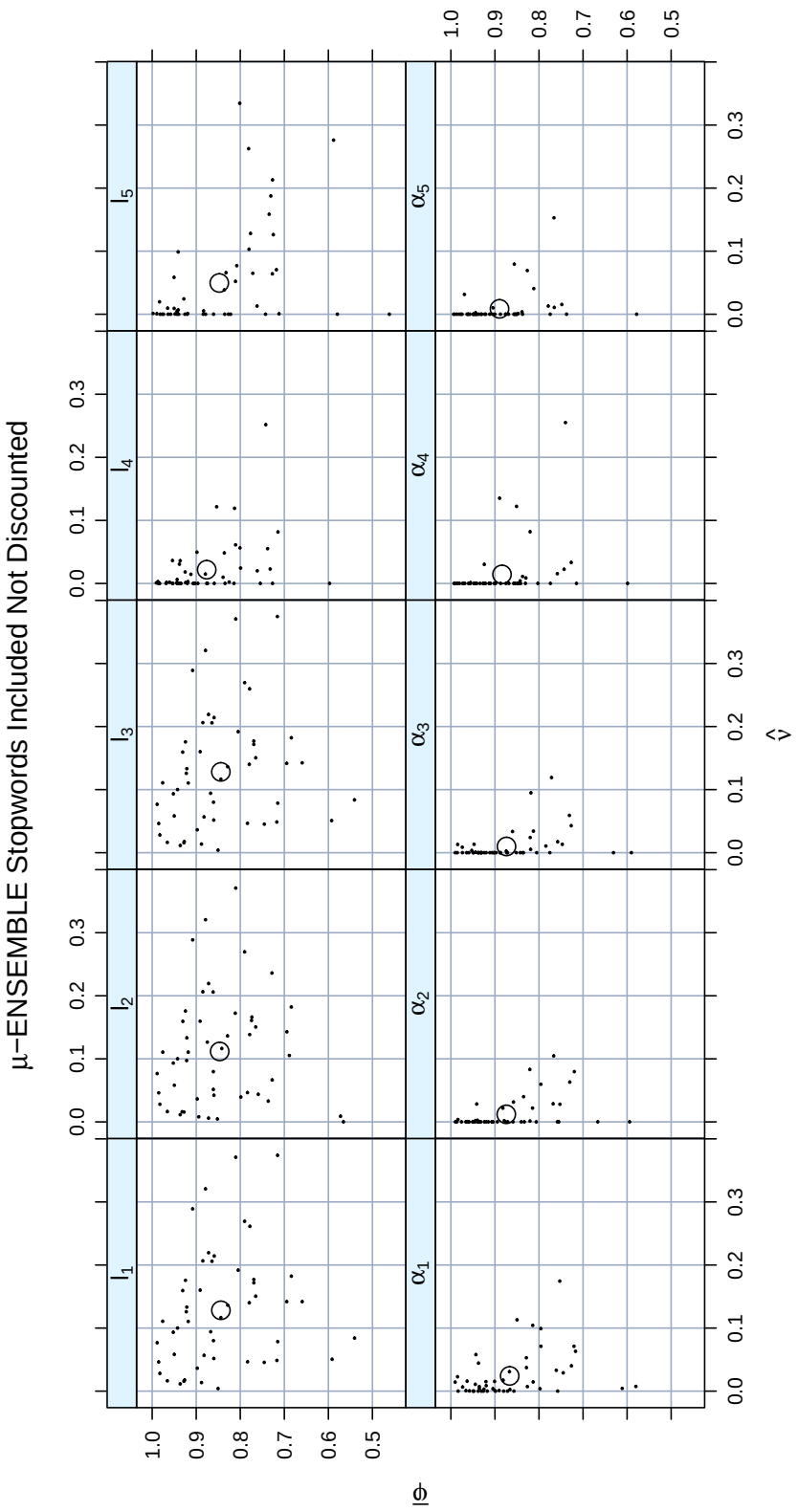

μ−ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
$\hat{\phi}$
$\hat{\nu}$

Figure T.5

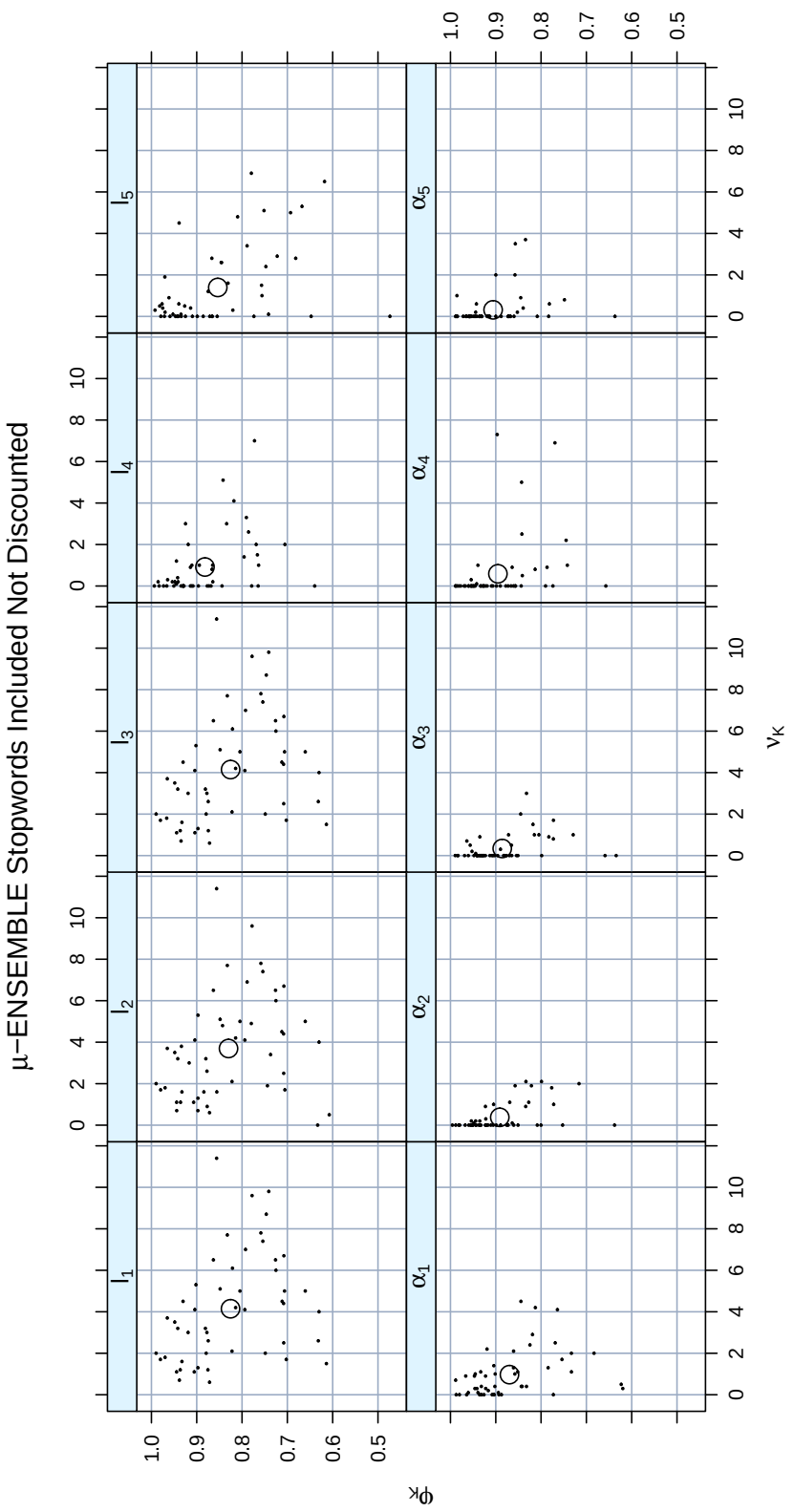

μ–ENSEMBLE Stopwords Included Not Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
ϕ_K
ν_K

Figure T.6

Appendix U

Ensemble Results Stopwords Included Discounted

Parameter	Type	$\widehat{\theta}$	θ_K	$\widehat{\sigma}$	σ_K	$\widehat{\nu}$	ν_K	$\bar{\varphi}$	φ_K
ℓ_1	Train	0.97	1.00	0.37	7.58	0.13	4.15	0.86	0.87
ℓ_1	Test	0.96	0.99	0.35	7.03	0.13	4.15	0.84	0.84
ℓ_1	$\Delta\%$	-0.60	-0.33	-6.07	-7.27	0.00	0.00	-2.83	-2.88
ℓ_2	Train	0.96	0.99	0.37	7.59	0.11	3.70	0.86	0.87
ℓ_2	Test	0.95	0.99	0.34	7.03	0.11	3.70	0.84	0.84
ℓ_2	$\Delta\%$	-0.68	-0.16	-6.26	-7.37	0.00	0.00	-2.77	-2.54
ℓ_3	Train	0.97	1.00	0.37	7.58	0.13	4.15	0.86	0.87
ℓ_3	Test	0.96	0.99	0.35	7.03	0.13	4.15	0.84	0.84
ℓ_3	$\Delta\%$	-0.59	-0.33	-6.06	-7.24	0.00	0.00	-2.83	-2.87
ℓ_4	Train	0.97	0.99	0.37	7.56	0.02	0.92	0.88	0.89
ℓ_4	Test	0.96	0.99	0.34	6.97	0.02	0.92	0.86	0.87
ℓ_4	$\Delta\%$	-0.55	-0.32	-6.40	-7.79	0.00	0.00	-2.55	-2.55
ℓ_5	Train	0.95	0.98	0.29	6.06	0.05	1.40	0.85	0.86
ℓ_5	Test	0.94	0.98	0.27	5.45	0.05	1.40	0.82	0.83
ℓ_5	$\Delta\%$	-1.32	-0.84	-9.17	-10.17	0.00	0.00	-3.80	-4.11
α_1	Train	0.97	0.99	0.38	7.84	0.02	0.97	0.87	0.88
α_1	Test	0.96	0.99	0.35	7.20	0.02	0.97	0.85	0.86
α_1	$\Delta\%$	-0.50	-0.12	-6.53	-8.07	0.00	0.00	-2.60	-2.38
α_2	Train	0.97	0.99	0.37	7.35	0.01	0.38	0.88	0.89
α_2	Test	0.96	0.99	0.34	6.74	0.01	0.38	0.85	0.87
α_2	$\Delta\%$	-0.50	-0.26	-6.44	-8.26	0.00	0.00	-2.70	-3.08
α_3	Train	0.97	0.99	0.37	7.57	0.01	0.34	0.87	0.89
α_3	Test	0.96	0.99	0.35	6.94	0.01	0.34	0.85	0.86
α_3	$\Delta\%$	-0.49	-0.03	-6.62	-8.38	0.00	0.00	-2.59	-2.48
α_4	Train	0.96	0.99	0.35	7.20	0.01	0.59	0.88	0.90
α_4	Test	0.96	0.99	0.33	6.58	0.01	0.59	0.86	0.87
α_4	$\Delta\%$	-0.60	-0.15	-6.75	-8.53	0.00	0.00	-2.64	-3.16
α_5	Train	0.96	0.99	0.35	7.10	0.01	0.32	0.89	0.91
α_5	Test	0.96	0.99	0.32	6.49	0.01	0.32	0.87	0.87
α_5	$\Delta\%$	-0.61	-0.52	-6.88	-8.56	0.00	0.00	-2.66	-3.53

Table U.1: μ-ENSEMBLE Stopwords Included Discounted

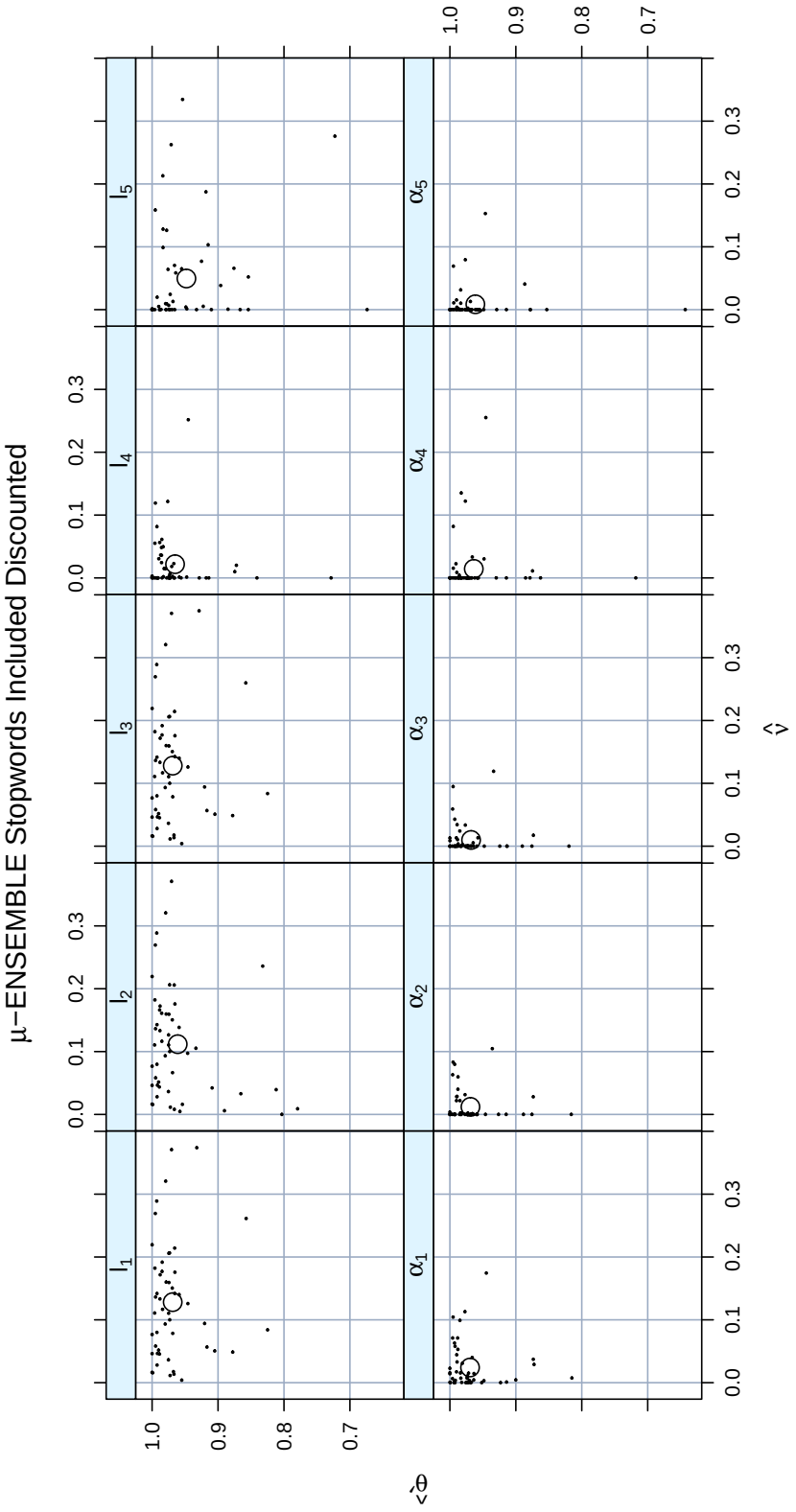

μ−ENSEMBLE Stopwords Included Discounted
$\hat{\theta}$
$\hat{\nu}$
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
0.0
0.1
0.2
0.3
0.7
0.8
0.9
1.0

Figure U.1

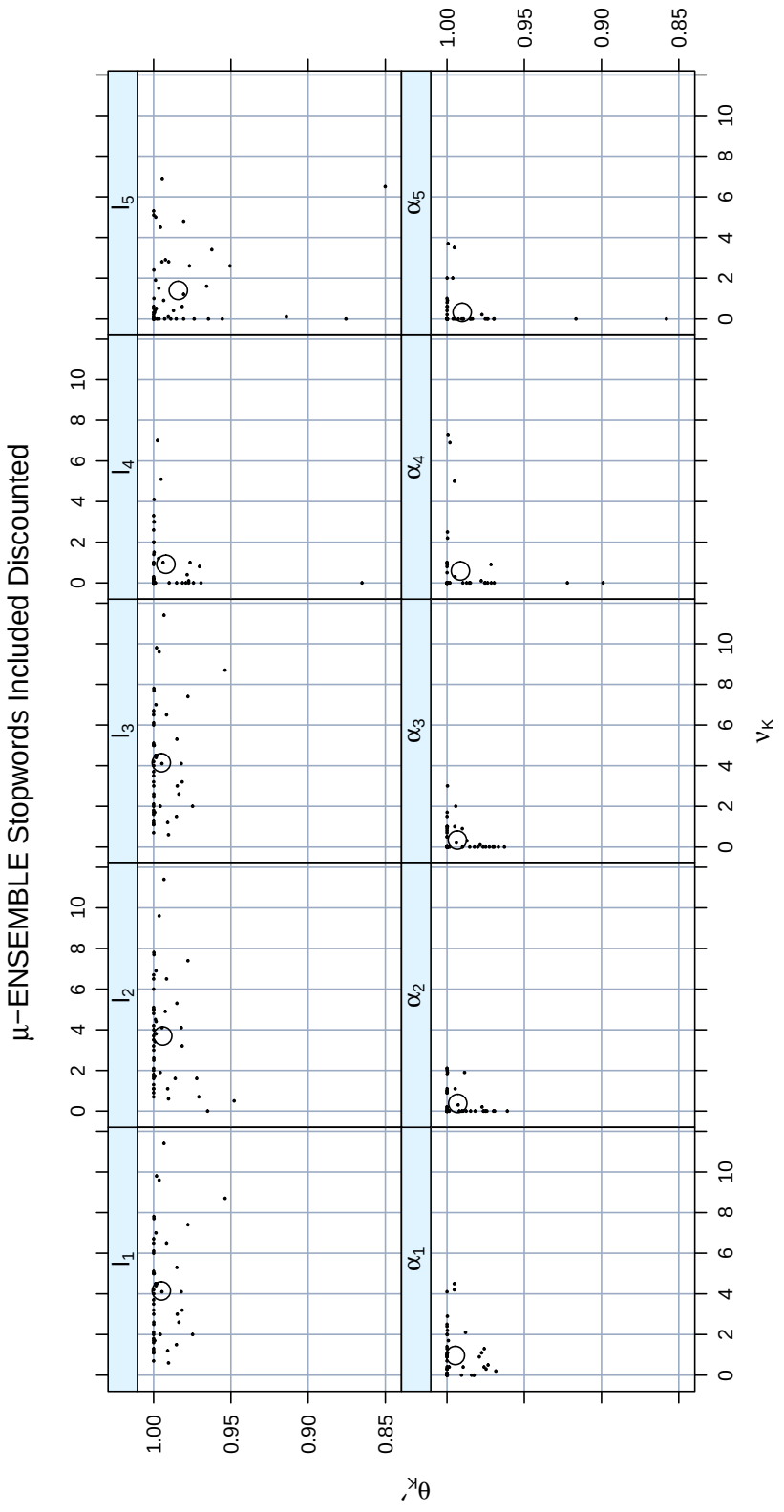

μ–ENSEMBLE Stopwords Included Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
$\theta_{k'}$
ν_K
1.00
0.95
0.90
0.85
0
2
4
6
8
10

Figure U.2

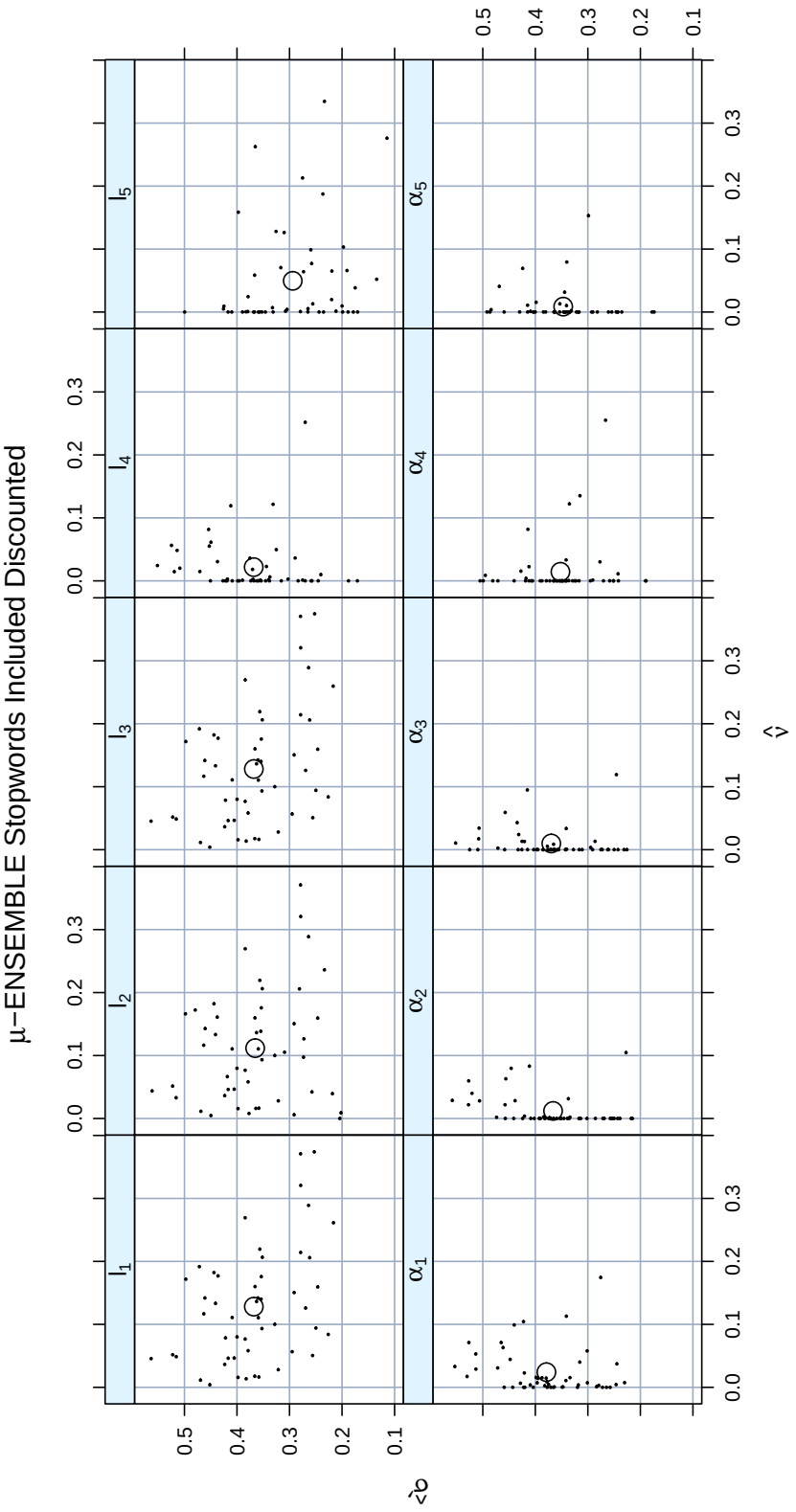

μ−ENSEMBLE Stopwords Included Discounted
$\hat{\alpha}$
$\hat{\nu}$
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
0.0
0.1
0.2
0.3
0.4
0.5

Figure U.3

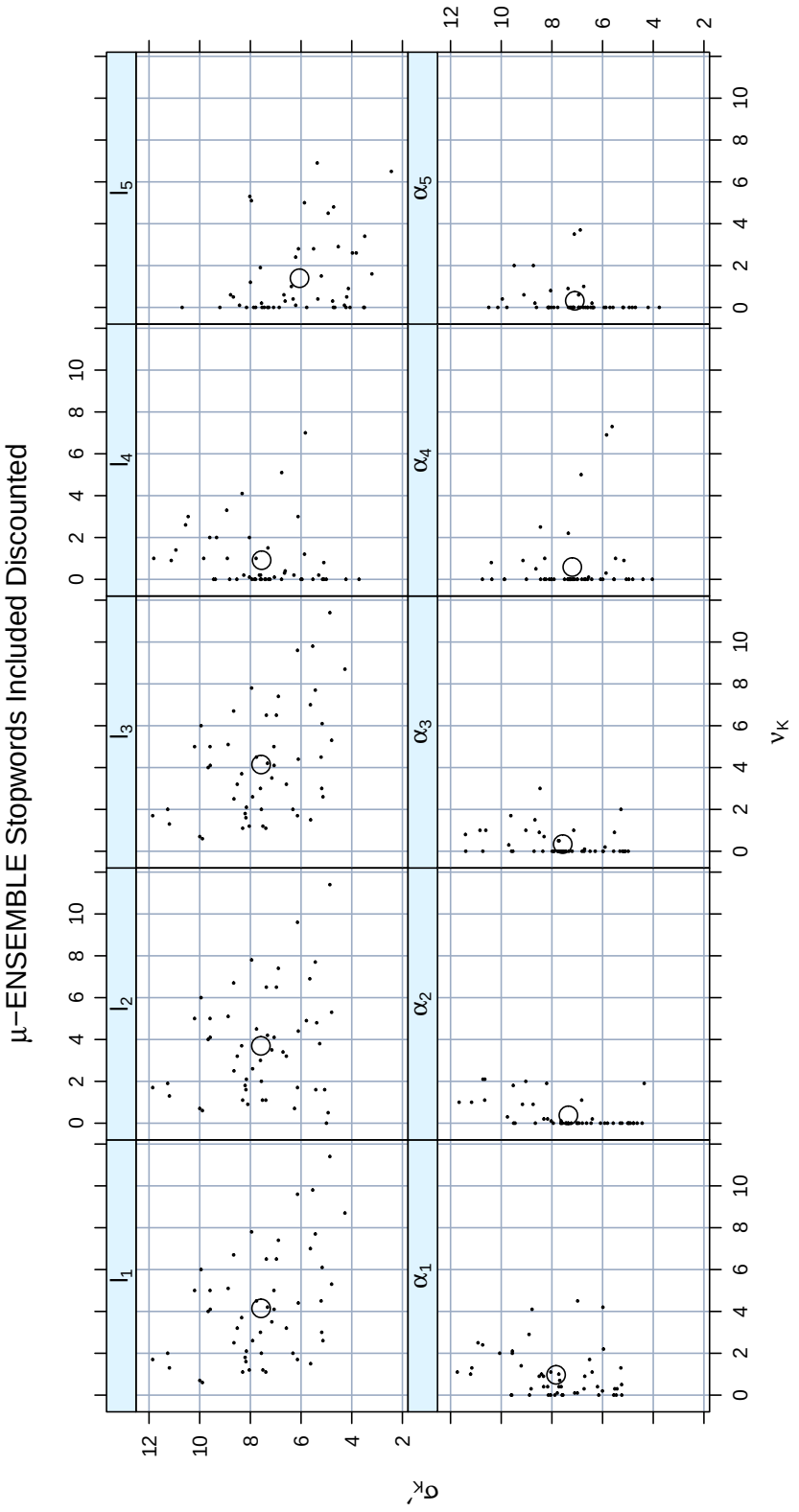

μ−ENSEMBLE Stopwords Included Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
σ_K
v_K

Figure U.4

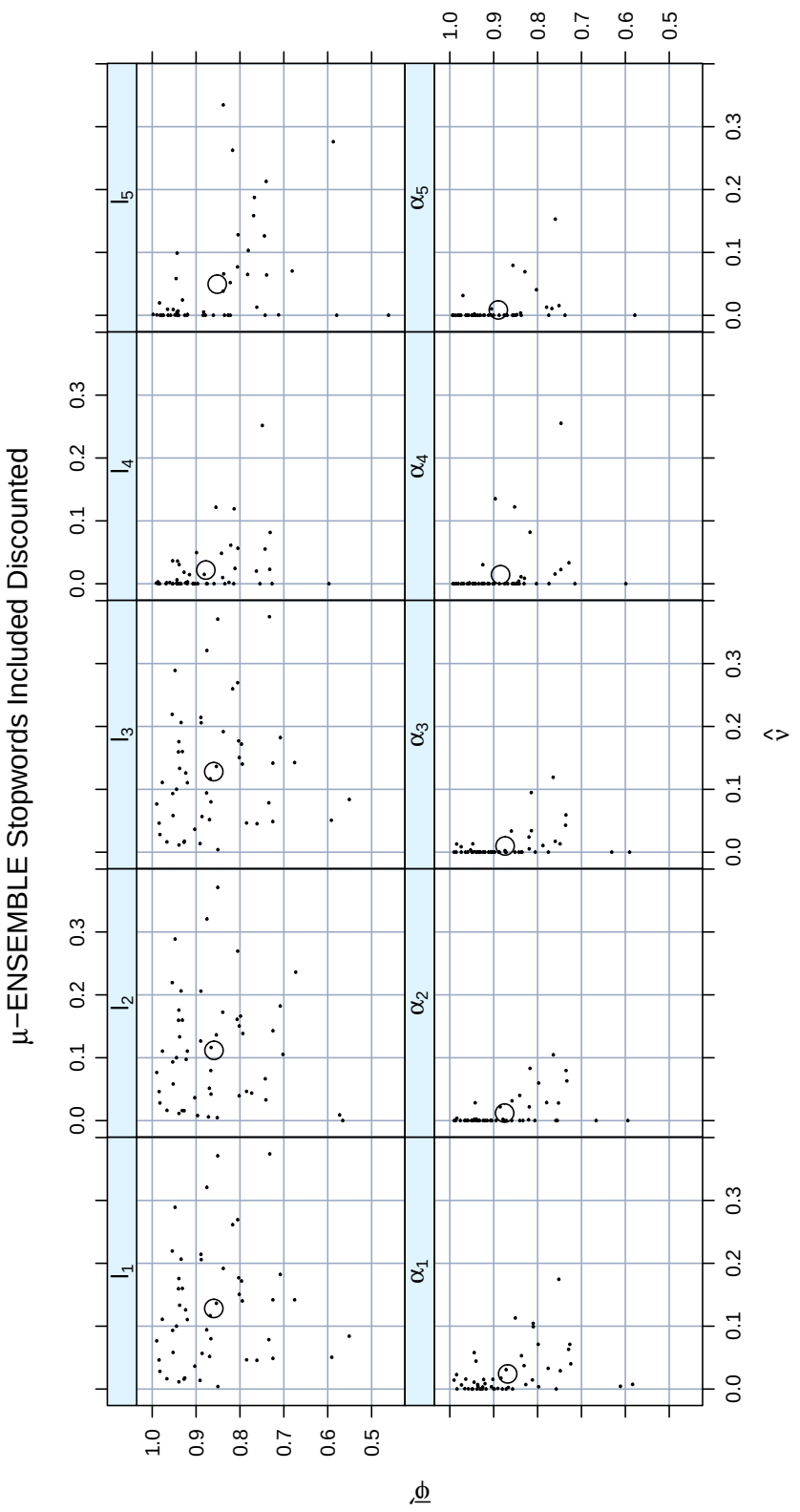

μ−ENSEMBLE Stopwords Included Discounted
$\hat{\phi}$
$\hat{\nu}$
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5

Figure U.5

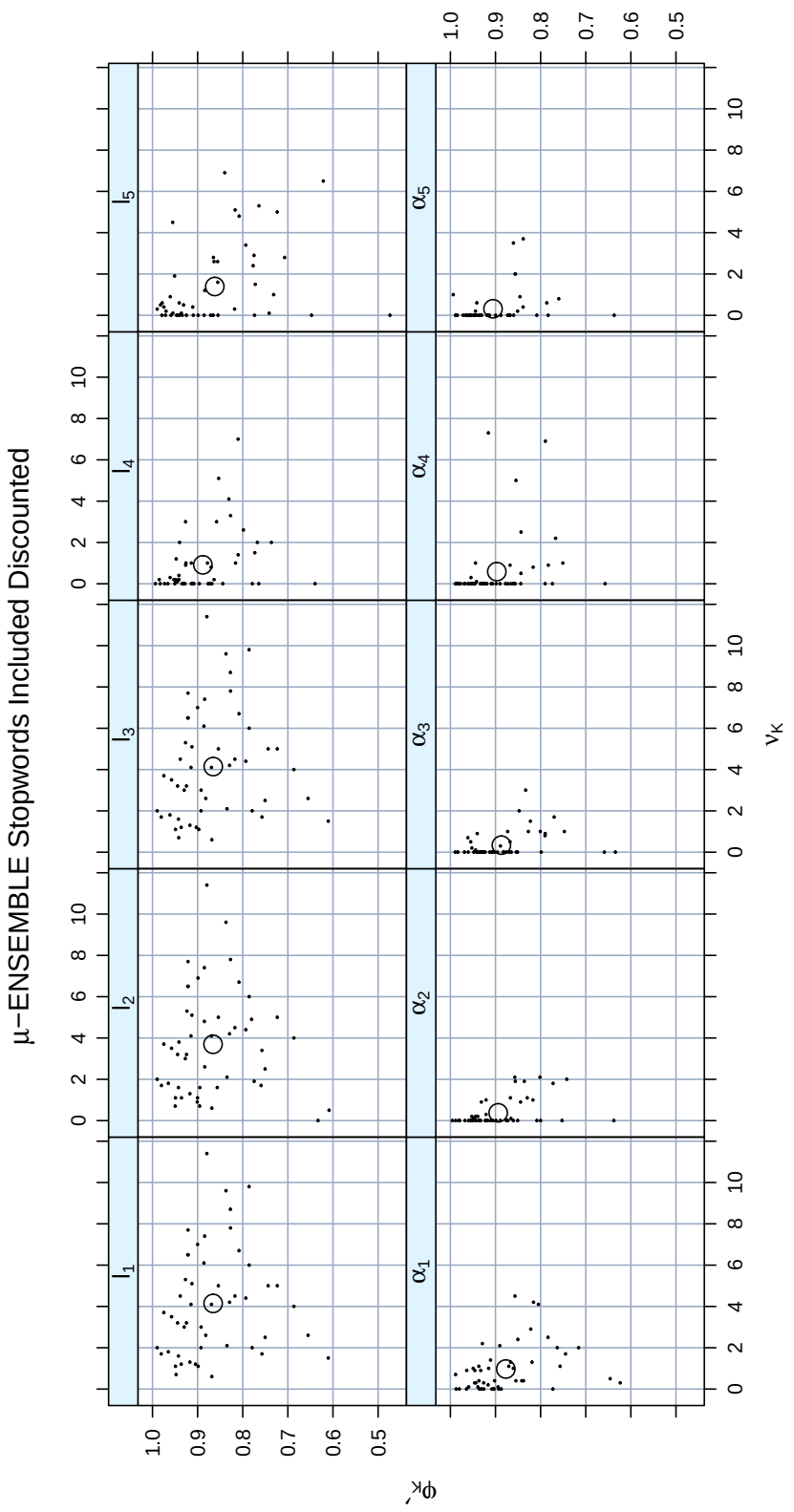

μ–ENSEMBLE Stopwords Included Discounted
l_1
l_2
l_3
l_4
l_5
α_1
α_2
α_3
α_4
α_5
ϕ_K
ν_K

Figure U.6

Vita

David J. Neu

EDUCATION

2012	Ph. D. in Operations Research, Rutgers University
1990	M.S. in Computer Science, Stevens Institute of Technology
1986	B.E. in Computer Science, Stevens Institute of Technology
1982	Graduated from Columbia High School, Maplewood, N.J.

EXPERIENCE

06/1994 – Present UTRS, Inc.

09/1997 – 5/2001 Rutgers University

11/1994 - 10/1995 Stevens Institute of Technology

07/1990 - 10/1997 Hilton Systems, Inc./SysTeam, Inc.

06/1986 - 07/1990 Teledyne Brown Engineering, Inc.