Eigenvalue-based Estimators for Optimal Dimensionality Reduction in Information Retrieval

by

Miles Efron

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2003

Approved by

Advisor: Professor Gregory B. Newby

Reader: Professor Michael L. Littman

Reader: Professor Robert M. Losee

Reader: Professor Gary Marchionini

Reader: Professor Paul Solomon

Abstract

MILES EFRON: Eigenvalue-based Estimators for Optimal Dimensionality Reduction in Information Retrieval

(Under the direction of Gregory B. Newby)

Latent Semantic Indexing (LSI) extends Salton's vector space model (VSM) of information retrieval, using dimensionality reduction to construct a statistical model of the relationships among the terms in a document collection. Though empirical studies have shown that such statistical models can improve retrieval over traditional key word-based approaches, dimensionality reduction raises an important question: if we are to reduce model dimensionality, how aggressively should we do so? Or conversely, what is the optimal value for k, the number of dimensions in an LSI system? In the unsupervised learning environment native to information retrieval, notions of model optimality and goodness of fit are difficult to define.

This dissertation pursues the viability of five statistical methods for estimating the optimal dimensionality of LSI systems. Though the five pursued methods entail different theoretical assumptions, they are all predicated on an analysis of the eigenvalues that arise naturally during LSI. This thesis contends that LSI's relation to principal component analysis makes an analysis of the eigenvalues the natural vehicle for dimensionality estimation.

To judge the utility of eigenvalue-based estimators for dimensionality estimation under LSI, two groups of experiments were performed. Both rounds of experimentation tested the accuracy of five dimensionality estimation criteria: eigenvalue-one, parallel analysis, percent-of-variance, Bartlett's test of isotrapy, and a novel method dubbed amended parallel analysis. During the first experiments, these criteria were applied to six standard information retrieval test collections and evaluated with regard to their accuracy. The second round of experiments applied each dimensionality estimator to simulated data.

The first round of experiments suggested that the family of estimation techniques based on the notion that dimensionality reduction is merited to the extent that the indexing features violate the VSM's assumption of statistical independence were especially accurate. Also during this round, amended parallel analysis yielded statistically significant improvements over traditional parallel analysis. The data simulations supported these findings. Applying parallel analysis and amended parallel analysis to simulated data of known intrinsic dimensionality yielded categorically superior estimates to all other tested estimation techniques.

Acknowledgement

While a full account of the generous help I recieved during completion of this dissertation is impossible, special thanks are due to several organizations and people. My advisor, Greg Newby provided consistent and insightful direction throughout the completion of this project. Likewise, the members of my dissertation committee offered compelling advice at every stage of my research. My sincerest thanks go to the entire committee for their efforts.

Additionally, this research was aided in formal and informal ways by `ibiblio.org`, whose director, Paul Jones offered thoughtful feedback and unflagging support in more ways than he knows. It was my privilege to share the company of `ibiblio`'s keen staff during my tenure at UNC. Without their influence and encouragement, this research would have suffered greatly.

Of course my family gets my most heartfelt thanks. In particular, I wish to thank Jessica Kilfoil, who inspired this work, not to mention its author. And finally, I acknowledge the direction of my father, Brad Efron, whose skill as a statistician I already knew, but whose talent as a teacher I discovered largely during the last two years. This dissertation is dedicated to him.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

One of the foundational approaches to information retrieval (IR), Salton's vector space model (VSM) implies a geometrical theory of information seeking (cf. Section2.1). Salton imagines documents as vectors in a high-dimensional space, with inter-document similarity measured by the corresponding vector cosine (cf. **[131]**). For Salton, similarity is thus a function of the orientation of documents in term space. Documents that are *about* similar topics lie near each other in the vector space. Under Salton's model, information retrieval is a matter of navigating this space; searchers attempt to locate regions of the vector space that hold documents relevant to their information needs.

An open question in the IR literature is, what should form the basis of the vector space that houses documents in an IR system? In Salton's formulation, documents are typically represented in the p-space spanned by a corpus' p indexing terms. However, elaborations on Salton's model, notably the generalized vector space model (GVSM) **[150]**, have suggested that alternatives to this space may be desirable. Due to the non-orthogonality of natural language terms, proponents of the GVSM argue that a transformation of the observed term space may improve retrieval.

Out of this conviction, latent semantic indexing (LSI) **[32]** derives a basis for a corpus' vector space by means of an orthogonal projection of its p-dimensional document vectors onto a k-dimensional subspace, where $k \ll p$ (cf. Section2.2). Proponents of LSI argue that this dimensionality reduction affords a robust representation of term-document associations, collocating similar objects by eliminating overspecification error among the observed data.

This dissertation is concerned with the parameterization of k, the number of dimensions retained during LSI. My analysis is aimed at discovering an effective means for selecting

k, and at considering how this selection reflects on the theoretical underpinnings of LSI. Throughout this study, I pursue the question:

> how effectively can an analysis of the eigenvalues derived during LSI be used to estimate the optimal representational dimensionality for IR?

Estimating a corpus' optimal dimensionality via eigenvalue analysis is attractive for a number of reasons. First, eigenvalues arise naturally during the generation of the derived LSI space. LSI projects terms and documents onto an orthogonal subspace of the term-document matrix $\mathbf{A}$ by means of the singular value decomposition (SVD, cf. Section2.2). This matrix factorization calculates the so-called singular values of $\mathbf{A}$, which are the positive square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$. Thus by virtue of performing the SVD, researchers can easily determine the eigenvalues of $\mathbf{A}'\mathbf{A}$.

Second, a number of eigenvalue-based dimensionality estimators have already been proposed in the statistical literature. The method of principal component analysis (PCA) is very similar to LSI (cf. Section2.2). Like LSI, PCA is based on the eigenvalue-eigenvector decomposition of an input matrix. Researchers in multivariate statistics have developed an advanced literature on the selection of k during PCA, and almost all of their theorizing focuses on methods of analyzing the distribution of eigenvalues. This study examines the suitability of five of these estimators for parameterizing LSI models.

Finally, basing dimensionality reduction on eigenvalue analysis implies a theoretical justification for truncating representational axes during LSI. As I show in Section 3.3 the magnitudes of a matrix's eigenvalues relates to the degree of redundancy among its variables. Following from this fact, I propose a novel method of dimensionality estimation, amended parallel analysis (APA). Dimensionality estimation by APA proceeds by analyzing the observed eigenvalues' departure from the eigenvalues that we would expect to see if the indexing terms were independent. The method counsels us to discard those dimensions whose eigenvalues are significantly smaller than the eigenvalues expected under the condition of term independence. By promoting APA, I suggest that dimensionality reduction during LSI is warranted insofar as the data violate the assumption of term orthogonality that is inherent in the VSM's conflation of similarity and vector orientation.

The remainder of this section describes the problem of dimensionality estimation in general terms. I offer several important definitions that limit the scope of this study, and give an example of the ramifications of selecting a proper dimensionality for IR systems. In Chapter 2 I describe my area of interest in more depth by reviewing related literature and pursuing a few extended examples based on my explication of earlier work. Chapter 3 describes the experiments that I undertook in efforts to address my research question. After this methodological discussion, I turn to an analysis of the experiments' resultant data. Chapter 4 describes the data generated by analysis of six standard IR test collections. To supplement this data analysis, in Chapter 5, I turn to simulated data sets, analyzing the dimensionality estimation problem in simplified environments where the right answer is known at the outset. Finally, I offer a synthesis of my findings in Chapter 6. This synthesis compares the relative merits of each dimensionality estimation technique. It also attempts to contextualize LSI's dimensionality reduction and the problem of optimizing LSI models in the larger domain of VSM-based retrieval.

1.1. Information Retrieval as a Geometrical Problem

Information retrieval systems attempt to discover documents in a corpus that are relevant to a searcher's stated information need. I discuss IR systems in depth in Section 2.1. Instead of belaboring the details of IR system operation, this section elaborates a geometrical interpretation of the IR task in anticipation of later discussions of dimensionality reduction. In particular, I introduce the notion of *information space* (cf. [**110**]), making efforts to elucidate some of the fundamental assumptions that inform a geometric theory of information retrieval.

Much of the difficulty of the IR task stems from the fact that three emphatically abstract notions inform its most basic mandates. For instance the subject matter of documents (e.g. books or articles), what Hutchins terms *aboutness* [**74**], is not typically open to direct observation, as is a person's height or the temperature of a chemical solution. Without a firm grasp on aboutness the *relevance* of a document to a query is hard to quantify. Furthermore, relevance between documents and queries is closely tied to a third abstraction native to the

IR problem—*similarity*. We intuit that documents that are relevant to a query are in some way similar to it. And relevant documents are similar to each other. *Aboutness*, *relevance*, and *similarity* are all crucial to IR, yet in most cases these variables are only latent in the data at hand. Thus information retrieval begins with the challenge of making inferences about these abstractions.

1.1.1. Information Spaces. Faced with an abstract problem, many IR systems resort to a geometrical metaphor to translate the task at hand into a readily computable form. The first step in this translation is adopting the notion of an *information space*. As Newby writes, an information space is the set of concepts and relations among them held by a computer system [**110**]. Though the philosophical status of concepts is the matter of ongoing debate in the field of cognitive science (cf. [**92, 123, 115**]), Peter Gärdenfors has recently advanced a geometrical theory of conceptualization [**56**]. According to Gärdenfors, concepts comprise variables that measure the properties of objects. An information space, then, may be understood as the set of variables observed by a system and the system's means of associating them. So **mass**, **volume**, and **velocity** might be concepts in an information space related to physical measurement. On the other hand **dogs**, and **cats** might be important concepts in the information space of an IR system related to household pets.

1.1.1.1. *The Definition of Dimension.* Crucial to the notion of an information space is the idea of dimension. Following Gärdenfors' argument, I define a dimension as a measurable direction in a given space. Thus dimensions provide the structure of the space, and as we shall see, define the topology that informs common notions of similarity and distance. As Gärdenfors writes, "dimensions form the framework used to assign *properties* to objects and to specify *relations* among them. The coordinates of a point within a conceptual space represent particular instances of each dimension..." [**56**].

Dimensions and concepts are thus closely related. They both comprise variables that describe objects contained in the space. The important point, however, is that a one-to-one correspondence between concepts and dimensions need not obtain. Thus, for instance, a particular dimension might be comprised of a linear combination of a number of concepts. For example, the dimension *size* might conflate ideas of **mass** and **volume**. On the other

Point	*Distance*
2	2.34
3	2.58
4	1.49
5	3.82
6	3.29
7	3.27
8	0.61
9	3.00
10	2.07

TABLE 1.1.1. Euclidean distance in an information space

hand, the dimension *motion* might simply include the concept **velocity**. Thus concepts and dimensions may be coequal, or they may be more loosely coupled.

The dimensions of an information space structure measurements of inter-object proximity within the space. Although specific distance measurements (e.g. Euclidean, Dice, city-block) vary in implementation, they are all based on the notion that objects are represented as points in the space spanned by the chosen dimensions. For example, consider the simulated information space depicted in Figure 1.1.1. These data are drawn from a multivariate normal distribution. Each observation from this distribution is scored on two variables: *Dimension 1* and *Dimension 2*. By sampling from this distribution and taking these variables as dimensions of a fictional information space, we derive the configuration represented in Figure 1.1.1. Points in the space represent objects, whose location on a given dimension corresponds to its observed value on that dimension's constituent variables. This allows us to discuss the distance between any two points i and j with respect to this space. Table 1.1.1 shows the Euclidean distance between point 1 and each other point in the space. Here point 8 is closest to point 1, while point 5 is farthest from it. The crucial idea is that any distance metric in an information space is necessarily taken with respect to the space's dimensions. That is, the distance between two points i and j is defined in terms of each point's location on all of the dimensions that define their mutual space.

1.1.1.2. *IR as a Spatial Problem.* Imagining that data reside in an information space is useful insofar as we admit some analogy between proximity and similarity. As I discuss in Section 2.1 Salton's VSM uses the vector cosine to measure document-document similarity.

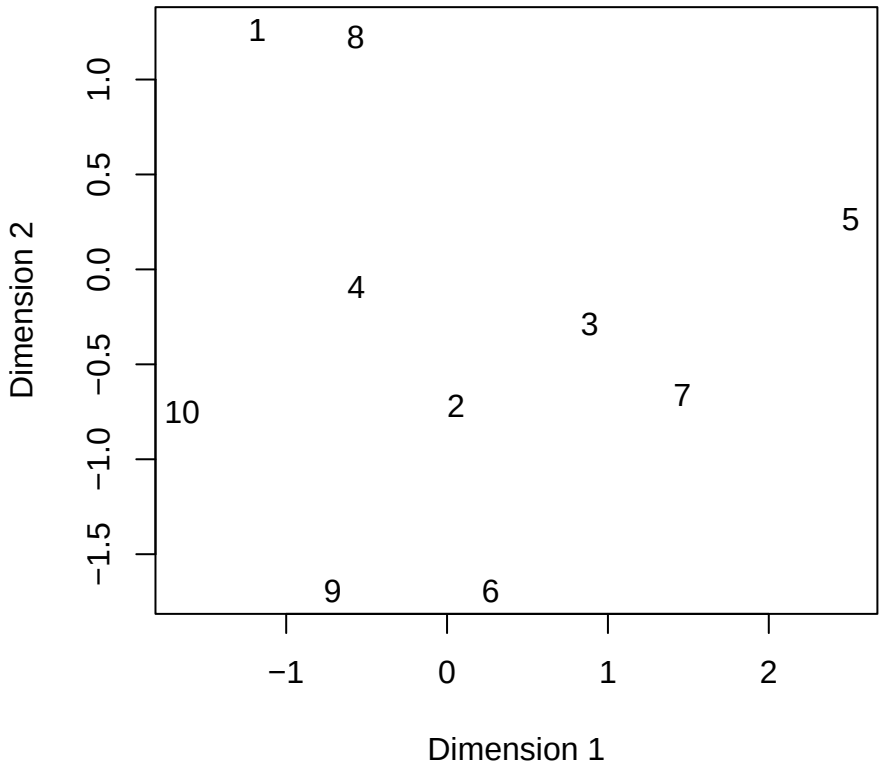

FIGURE 1.1.1. Representation of a simple information space

Geometrically based information retrieval takes as axiomatic the idea that objects that lie close together in information space are more similar than documents that are far apart. Thus the distance functions that are defined on the dimensions of an information space also act as similarity functions, suggesting that a given region of the space contains similarly informative objects.

As an example, consider Figure 1.1.2. These data represent a simulated collection of 20 documents about two topics, **dogs** and **cats.** For the sake of illustration, imagine that an omniscient indexer (whom we'll meet again in Section 2.2) has assigned to each of these documents two real-valued scores, one a variable called *dog-ness,* and another called *cat-ness.* These variables comprise the dimensions of the 2-dimensional information space shown in

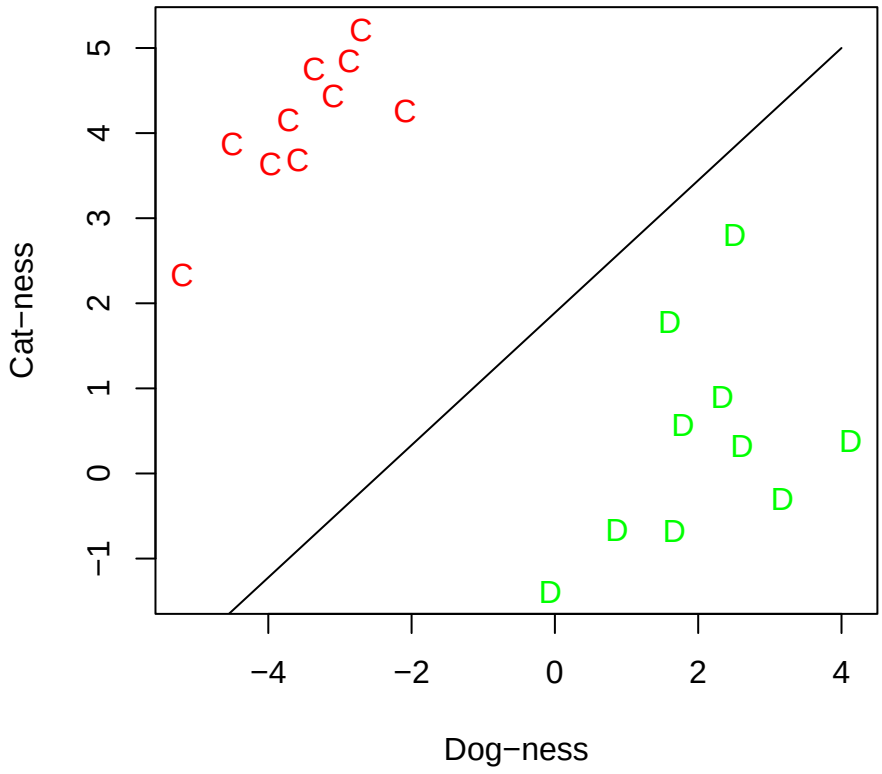

FIGURE 1.1.2. Documents in a 2D information space

the figure. Each document appears as a point plotted at its coordinates on the space's dimensions.

Assuming that documents that are near each other in this space are similar, information retrieval can be understood as a partitioning of this space. That is, a searcher might like to see documents about **dogs**, wishing not to be distracted by documents about **cats**. To accomplish this, an IR system attempts to isolate the region of information space relevant to **dogs**. The black line in Figure 1.1.2 represents such an attempt. According to some optimality criterion (this line was fitted by linear discriminant analysis) the system defines a hyperplane in the information space. All documents to one side of the hyperplane are classified as being about **dogs**, while those on the other side are assumed to be about **cats**.

This sort of 2-class classification problem exemplifies an extremely simple form of IR. For instance it ignores entirely the issue of document ranking. However, it is instructive insofar as it shows the utility of adopting a geometric approach to information retrieval. Constructing an information space from observed variables is useful because it manifests the abstractions discussed above into measurable features. Thus in information space, proximity constitutes a model of similarity. This model's approximation is motivated by assuming that *aboutness* is a function of the adopted dimensions. Finally, then, *relevance* is inferable by analyzing the distribution of similar objects within the space, as exemplified by defining the classification rule in Figure 1.1.2.

1.1.2. Improving Information Spaces. As exemplified by the successes of Salton's VSM (discussed in Section 2.1), modeling the IR problem geometrically offers an apt approach to retrieval. However, the VSM is not without its detractors. A common critique of the vector model cites its assumption of term independence. Under Salton's model, documents reside in the information space spanned by the system's indexing terms, and similarity is defined by the vector cosine. Thus if *car* and *automobile* are both present in the indexing vocabulary, systems based on the standard vector model will fail to retrieve documents indexed on *automobiles* for queries about *cars.*

To see why this is the case, consider the similarity function of the VSM[1]. Given an $n \times p$ document-term matrix $\mathbf{A}$ and a p-dimensional query vector $\mathbf{q}$, Equation 1.1.1 gives the VSM similarity function:

$$\mathbf{s} = \mathbf{q}\mathbf{A}' \tag{1.1.1}$$

where $\mathbf{s}$ is the n-vector of similarity scores. Under the standard VSM, dimensions of term space are assumed to be orthogonal; the model assumes that terms are statistically independent. Equation 1.1.1 may be re-written to emphasize its assumption of term independence:

$$\mathbf{s} = \mathbf{q}\mathbf{I}_p\mathbf{A}'. \tag{1.1.2}$$

[1]This discussion is based on the treatment of similarity models given in [**77**].

In this expression, the identity matrix articulates the assumed independence among the corpus' indexing variables. That is, no transformation is applied to $\mathbf{q}$ to account for correlation among the terms. In the case of *car* and *automobile* our intuition warns us that this model is not accurate. Rather these terms seem to have a great deal in common. If this is the case, perhaps this intuition should inform the system's similarity function.

In fact, Wong *et al.* argue in **[150]** that term co-occurrence information should inform the model of information space. Extending Salton's theory, Wong proposes the so-called generalized vector space model (GVSM):

$$\mathbf{s}_{GVSM} = \mathbf{qRA}' \tag{1.1.3}$$

where $\mathbf{R}$ is the $p \times p$ term correlation matrix calculated from $\mathbf{A}$. According to the GVSM, if *car* and *automobile* tend to co-occur in a corpus, an IR system ought to reflect their relationship. For Wong, the sample correlation matrix provides a model of the relationships that obtain among the corpus' indexing terms. Thus the GVSM attempts to improve Salton's model by altering the axes of information space to account for inter-term correlation. That is, by replacing the identity matrix of Equation 1.1.2 with the correlation matrix $\mathbf{R}$, the GVSM mitigates the error introduced to the VSM by assuming term independence.

1.1.3. LSI as a model of the Population Correlation Matrix. If the GVSM removes error from Salton's theory by accounting for the observed term correlations, LSI removes error from the GVSM by fashioning a model of the population correlation matrix based on the observed sample. Under LSI, we have the similarity function:

$$\mathbf{s}_{LSI} = \mathbf{qR}_k\mathbf{A}' \tag{1.1.4}$$

where $\mathbf{R}_k$ is the best rank-k approximation of $\mathbf{R}$, in the least-squares sense, and $k \leq rank(\mathbf{A})$. Section 2.2, treats the derivation of $\mathbf{R}_k$. For now, suffice it to say that the similarity model defined by Equation 1.1.4 supplements the traditional VSM with a linear model of the correlational structure among the columns (terms) of $\mathbf{A}$. Choosing an optimal value of k thus amounts to a problem of statistical model building. That is, the optimal

dimensionality of an LSI system will result in the best approximation of the population term-term correlation matrix.

LSI's improvement over Salton's VSM may be understood as deriving from two mechanisms. First, if $k = k_{max}$ then LSI converges on the GVSM. Thus LSI alters Salton's approach to IR by rotating the data onto independent axes. Instead of assuming that the terms of a collection are independent, LSI projects the documents onto an orthogonal information space. Second, LSI attempts to improve the GVSM model of term correlations by means of dimensionality reduction. When it employs a k-dimensional representation, LSI assumes that the sample correlation matrix $\mathbf{R}$ is similar to the population correlation matrix, modulo some perturbation. LSI's dimensionality reduction, may thus be understood as an effort to remove artifacts of this perturbation in efforts to derive the best model of the multivariate probability density function (PDF) that generated the data. In both mechanisms—its orthogonal projection and its dimensionality reduction—LSI is concerned with deriving the optimally oriented information space.

1.2. Dimensionality Reduction for Information Retrieval

Dimensionality reduction involves discarding putatively misleading dimensions from an information space. Because the features of the derived LSI space are orthogonal, they convey most of the variance of the observed (non-orthogonal) data using relatively few dimensions. By retaining only the k dimensions with the largest corresponding eigenvalues, LSI discards dimensions that capture small amounts of variation in the observed data. LSI's supporters argue that these dimensions contain mostly random noise, and that removing them from the representation effectively mitigates the problem of overfitting the model of the population correlation matrix.

As an example of dimensionality reduction, consider the earlier discussion of the dimensions of an information space describing physical systems. There I discussed conflating two features, *mass* and *volume*, onto a single new variable, *size*. This is the same operation performed by LSI. Each dimension of LSI's derived factor space is a linear combination of the input variables. If the first derived factor corresponds to a latent variable *size*, then,

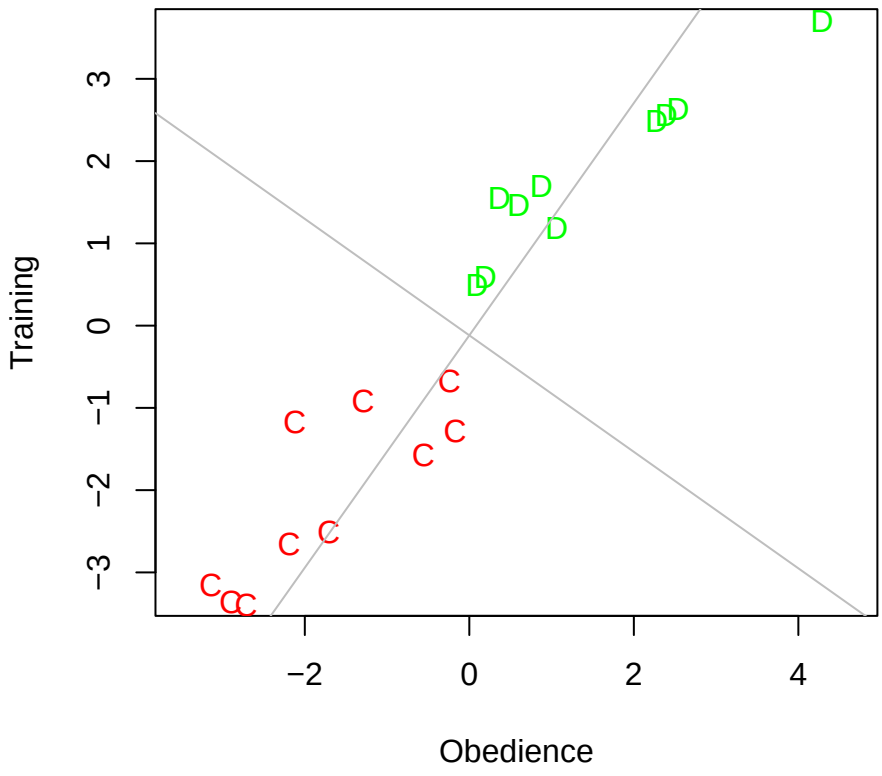

FIGURE 1.2.1. Simulated 2-class document collection in term space

the coefficients for *mass* and *volume* might be large, while the coefficient for *velocity* would be small. Having conflated *mass* and *volume* onto a single variable, we are left with an information space of reduced dimensionality. Whereas before, the example information space was spanned by the dimensions *mass*, *volume*, and *velocity*, a better model might use the two-dimensional representation spanned by the inferred dimensions *size,* and *motion* (on which *mass* and *volume* presumably score low, while *velocity* scores high).

To ground this discussion in an example related to IR, consider Figure 1.2.1. Here we see 20 documents in a simulated 2-dimensional term space. In this case, the dimensions of the space are the terms *obedience* and *training*. The documents in this imaginary corpus are of two classes, those about **dogs** (shown as D's) and those about **cats** (the C's). Each

document's location in the space is comprised of its (for now undefined) score on each dimension. As is evident from the figure, *obedience* and *training* are not independent; they appear to be positively correlated. In fact, the coefficient of correlation between these variables is $r = 0.96$. As I discuss in Section 2.1 such a correlation is worrisome insofar as the cosine similarity metric is predicated upon the orthogonality of a space's dimensions. The problem arises because the dimensions, *obedience* and *training,* are largely measuring the same thing. Thus applying a distance-based metric on the space spanned by these variables effectively counts the same information twice, giving skewed measurements of inter-object similarity.

If using both variables introduces error into the similarity model, perhaps we should only use one of the variables to represent our documents. The question then becomes, which variable to choose? The total variance of these data is 9.12, which is divided almost equally between the two dimensions; the variance of *obedience* is 4.1, or 45% of the total variance, while *training* captures the remaining 55%. Discarding either dimension of the space would thus incur a significant loss of measurement accuracy.

The grey lines in Figure 1.2.1 visualize the LSI approach to solving this problem. The lines represent new dimensions for the information space. They were generated by finding the two mutually orthogonal vectors in the space that capture maximal variance. Thus the first factor (the axis with a positive slope) is the vector that minimizes the squared distance between itself an each point in the space, while passing through the mean vector of the data. The second factor is the best least-squares fit that is orthogonal to the first factor[2]. The grey lines in Figure 1.2.1 comprise new dimensions for the information space. These dimensions were constructed by analyzing the covariance between the observed data, and rotating the axes to account for this correlation.

Figure 1.2.2 suggests why axis rotation leads to dimensionality reduction. The figure shows the 20 documents, projected onto the 1-space spanned by the first inferred dimension. As is evident from the orientation of documents in this space, this single dimension appears to capture the variation in the original data quite well. In fact, of the original 9.12 units

[2]I discuss the notion of least-squares fitting in Section 2.2.

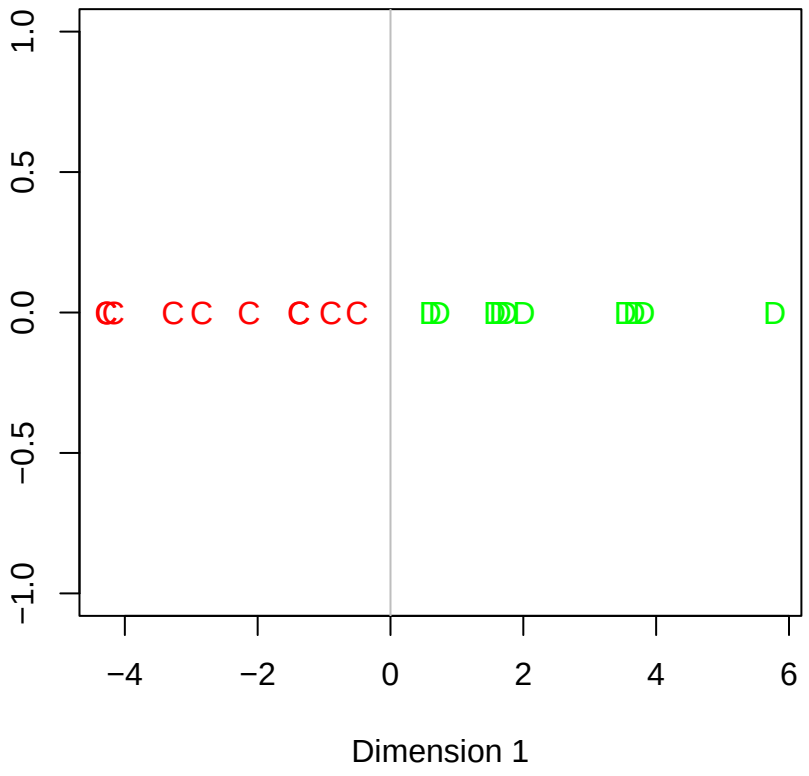

FIGURE 1.2.2. Simulated 2-class document collection in inferred 1-space

of variance, this factor describes 8.92 units, or 98% of the total variance. Thus the second inferred factor accounts for only 2% of the total variation. As stated above, we suspect that representing documents in the information space spanned by *obedience* and *training* introduces error into the VSM's distance measurements due to their statistical correlation. However, we worried about discarding either of the observed dimensions since they both captured a significant proportion of the variation among the data. By rotating the axes to derive two new dimensions, LSI makes the selection of an appropriate factor much easier. By rotating the data we obtain variables that describe 98% and 2% of the total variance, respectively.

Proponents of LSI argue that the 2% variance described by factor number 2 is negligible, that it probably describes sampling error among the data, as opposed to a systematic correlational pattern. In other words a model of term correlations that includes dimension number 2 would thus include random error, risking the problem of becoming overfitted to the observed data. Under LSI we would truncate the representation, projecting documents onto the 1-space defined by the first inferred dimension. The vertical grey line in Figure 1.2.2 shows a classification hyperplane in this 1-space; documents to the left of the line are assumed to be about **cats** while documents to its right are assumed to be about **dogs**. Thus dimensionality reduction has not made the classification task qualitatively harder. And if dimensionality reduction is merited, removing the second (weakly descriptive) dimension from the similarity function derives a superior model of the term correlations that obtain in the population. Proponents of LSI argue that this reduced-rank model will thus generalize to new data (i.e. new queries) more accurately than the model based on the full-rank sample correlation matrix.

1.3. Eigenvalues and Dimensionality Reduction

Closely related to techniques such as principal component analysis [**81**] and multidimensional scaling [**29**], LSI is based on the singular value decomposition of an input matrix, which I discuss in Section 2.2. Given an $n \times p$ matrix $\mathbf{A}$ of rank r let the singular value decomposition of $\mathbf{A}$ be given in Equation 1.3.1:

$$\mathbf{A} = \mathbf{T}\mathbf{\Sigma}\mathbf{D}' \tag{1.3.1}$$

where $\mathbf{T}$ is an $n \times r$ orthogonal matrix, $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix, and $\mathbf{D}$ is an $r \times r$ orthogonal matrix. Matrices $\mathbf{T}$ and $\mathbf{D}$ contain the left and right singular vectors of $\mathbf{A}$, respectively, and the main diagonal of $\mathbf{\Sigma}$ contains the singular values, which are the positive square roots of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ (which I shall call the co-occurrence matrices of $\mathbf{A}$). As shown in [**67**] and [**116**] the singular vectors of $\mathbf{A}$ define the rotated axes shown in the example above. Likewise, the eigenvalues of the co-occurrence matrices are the amount of variance described by each dimension in the rotated space. Because each dimension of the

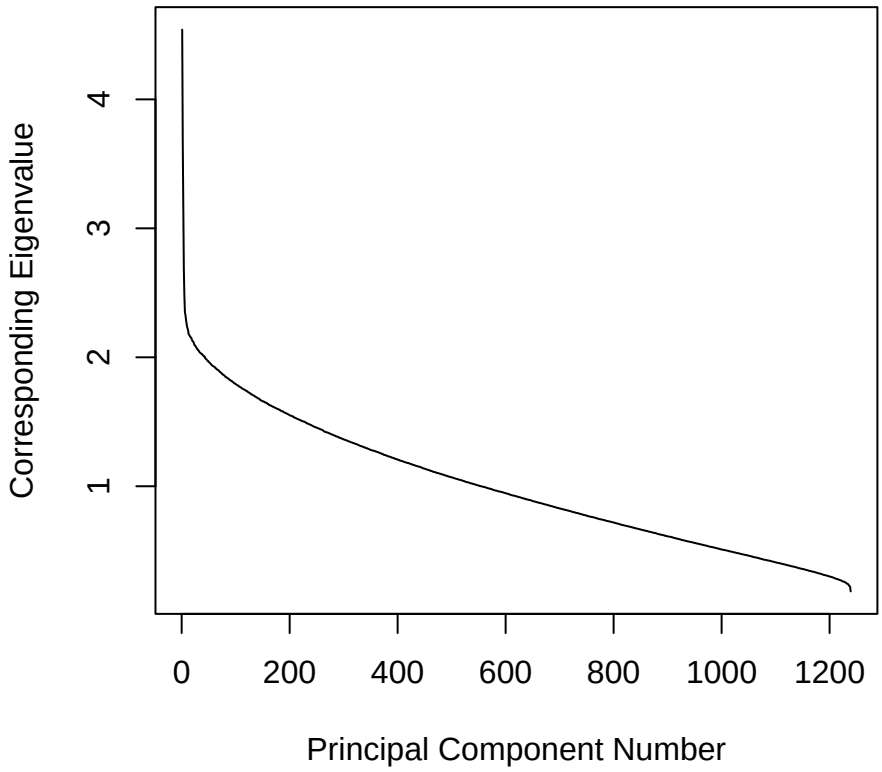

FIGURE 1.3.1. Scree plot of Cystic Fibrosis data

inferred space maximizes the available variance, σ_i the diagonal elements of $\mathbf{\Sigma}$, decrease in magnitude as i goes from 1 to r. Thus $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. As demonstrated in [**37**] and [**105**] the eigenvalues of matrices in IR applications tend to follow a power-law distribution. Thus the magnitude of an eigenvalue is related inversely and exponentially to its rank. This trend is visible in Figure 1.3.1, which is based on the eigenvalue decomposition of the Cystic Fibrosis database, a collection of 1239 medical documents (cf. Section2.4). Figure 1.3.1 is a so-called scree plot. It graphs the magnitude of the k^{th} eigenvalue versus its rank. The eigenvalues decrease in size quickly as we move from left to right along the graph. Because the eigenvalues describe the amount of variance captured by the corresponding dimensions, LSI's advocates suggest that one may improve an information space's model of

term correlations by ignoring axes with small eigenvalues. This hypothesis is contentious, but borne out by LSI's improvements in performance over traditional vector-based models.

By removing dimensions with small corresponding eigenvalues, Deerwester *et al.* report significant improvement over the VSM on several standard data sets [**32**]. Likewise, Dumais has applied LSI to several problems in the Text Retrieval Conferences (TREC), with promising results [**38, 39, 40**]. Dumais reports a 31% advantage over keyword-based methods for the filtering task, and a 16% improvement for *ad hoc* retrieval (cf. [**10**]). Ding reports improvements in performance of 30% above traditional VSM-based systems on the *ad hoc* retrieval task [**36, 37**]. Landauer and Dumais apply a variant of LSI to a vocabulary learning problem. They find that an LSI system is able to learn new vocabulary with accuracy above 50% [**90**]. What is particularly interesting about the study by Landauer and Dumais is the relationship between their system's dimensionality and its performance. As I discuss in Section 2.3, Landauer and Dumais find that retaining approximately 300 dimensions yields the best accuracy for the vocabulary inference problem. While it is not surprising that such a system requires more than one or two dimensions, Landauer and Dumais find that when k becomes much larger than 300, performance declines. They cite this decline as evidence that the factors corresponding to small eigenvalues contain essentially random noise. Theoretical and empirical work by Ding [**36, 37**] and Story [**140**] corroborates this hypothesis. In all of these studies, research suggests that selecting a proper value for k is crucial for good LSI performance. Thus an optimal LSI model would include factors whose corresponding eigenvalues are large, discarding those eigenvalues that are small.

The purpose of the present study is to analyze statistical methods for defining what constitutes a "small" eigenvalue for IR applications. In most LSI applications, system designers choose k, the number of retained dimensions, in some *ad hoc* manner. For instance the seminal work of, Deerwester *et al.,* calls the selection of an appropriate dimensionality "crucial" for good retrieval under LSI. Moving from $k = 1$ to $k = 100$ yields a 30% improvement in Deerwester's system performance. Yet their criterion for this parameterization is informal, to say the least. "We have reason to avoid both very low and extremely high numbers of

dimensions," they write. "In between we are guided only by what appears to work best. What we mean by 'works best' is ... what will give the best retrieval effectiveness" [**32**].

This approach is common in applications of LSI. Unfortunately, in practice it is difficult to judge what does work best. In the case of Deerwester *et al.* or Landauer and Dumais, selection of k was performed by recourse to pre-classified data. That is, these experiments make use of training data and test data that have been classified in advance, thus allowing the researchers to judge a given parameterization retrospectively by observing its accuracy on the test data. This approach is unsatisfying in two respects. First, from a practical standpoint, most operational IR systems do not have ready access to the type of relevance judgements that inform the retrospective performance analysis used by Deerwester *et al.* Second, we desire a more theoretically sound motivation for dimensionality reduction.

I argue that the eigenvalues that arise during LSI are useful on both of these counts; they allow researchers to estimate the optimal dimensionality for a data set without recourse to pre-classified data. Second, as I discuss in Sections 2.3.3 and 3.3, the method that I propose—amended parallel analysis (APA)—implies a theoretical justification for LSI. APA operates by judging the deviation of each observed eigenvalue λ_i, from the corresponding eigenvalue expected if the data were independent. The method rejects all eigenvalues that are significantly smaller (in the statistical sense) than their counterparts under the null hypothesis of independence. As noted above, dimensionality reduction is useful when the observed dimensions are non-orthogonal. Thus I suggest that analyzing eigenvalues for their departure from the eigenvalues expected given independent data provides a good estimate of the best parameterization of k. One goal of this dissertation, then, involves testing this hypothesis.

1.4. Intrinsic Dimensionality and Optimal Representations for IR

The difficulty inherent in the proposed research lies in knowing what constitutes "the best" parameterization of k. Does some optimal value of k exist, as a function of the matrix $\mathbf{A}$? If so, how can one ascertain it? Or does optimal k perhaps depend on the task that the LSI system will ultimately perform? To address these issues, I offer several definitions

in this section. First, I introduce the notion of a data set's intrinsic dimensionality (also known as its effective dimensionality, cf. [**108**]). This notion is common in the literature of principal component analysis (cf. [**81, 79, 116**]) and multivariate statistical theory [**3, 107**]. Following Fukunaga [**54**] Bishop describes the notion of intrinsic dimensionality as follows:

> Suppose we are given a set of data in a d-dimensional space, and we apply principal component analysis and discover that the first d' eigenvalues have significantly larger values than the remaining $d - d'$ eigenvalues. This tells us that the data can be represented to a relatively high accuracy by projection onto the first d' eigenvectors. We therefore discover that the effective dimensionality of the data is less than the apparent dimensionality d, as a result of correlations within the data.... More generally, a data set in d dimensions is said to have an *intrinsic dimensionality* equal to d' if the data lies entirely within a d'-dimensional subspace. [**12**]

Departing slightly from this definition, I define intrinsic dimensionality as a function of the multivariate probability density function (PDF) responsible for the $n \times p$ matrix $\mathbf{A}$. The intrinsic dimensionality of $\mathbf{A}$ is thus the number of statistically uncorrelated variables in the probability density function of $\mathbf{A}$. Assuming that $\mathbf{A}$ is drawn from a multivariate normal distribution, this is equivalent to the number of independent variables in $\mathbf{A}$**.** Alternatively, we may understand the intrinsic dimensionality of $\mathbf{A}$ to be the number of non-zero eigenvalues in the population covariance matrix for the PDF that generated $\mathbf{A}$. Because the intrinsic dimensionality is defined on the PDF of $\mathbf{A}$, it is a parameter—I abbreviate it k_{opt}—which I hope to estimate by recourse to statistical analysis of $\mathbf{A}$. The methods of eigenvalue analysis proposed in the current study provide techniques for computing such statistics.

As an example of the idea of intrinsic dimensionality and how it pertains to eigenvalue analysis, I generated two 1000×5 data sets, $\mathbf{A}_5$ and $\mathbf{A}_3$. Each data matrix thus contained 1000 observations on 5 variables. Both matrices were drawn from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$. The difference between these matrices lay in the covariance matrices of the distributions that generated them. Dataset $\mathbf{A}_5$ was created with covariance matrix $\boldsymbol{\Sigma}_5 = \mathbf{I}_5$. Thus the 5 variables of the distribution are independent, modulo

sampling error. The intrinsic dimensionality of $\mathbf{A}_5$ is thus 5. On the other hand, matrix $\mathbf{A}_3$ was generated from a distribution with covariance matrix $\mathbf{\Sigma}_3$:

$$\mathbf{\Sigma}_3 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.4.1}$$

which has the first three variables positively correlated with each other, with the last two independent. I call this a distribution of intrinsic dimensionality 3, counting the 3 correlated variables as a single dimension. That $\mathbf{A}_3$ is three dimensional is clear by inspection of its eigenvalues:

$$\boldsymbol{\lambda}'_3 = \begin{pmatrix} 3 & 1 & 1 & 0 & 0 \end{pmatrix}. \tag{1.4.2}$$

Matrix $\mathbf{A}_3$ has 5 variables, but only 3 non-zero eigenvalues. Thus due to inter-variable correlation, its intrinsic dimensionality is three, two less than its observed dimensionality. Data generated from a PDF parameterized with covariance $\mathbf{A}_3$ will thus be defined to be essentially three-dimensional.

Scree plots of $\mathbf{A}_5$ and $\mathbf{A}_3$ appear in Figure 1.4.1. For the matrix with intrinsic dimensionality of 5, all 5 eigenvalues are approximately equal. As discussed in the section about APA, this is what we would expect, and it constitutes APA's null hypothesis—that all five eigenvalues are non-zero in the population. On the other hand, matrix $\mathbf{A}_3$, with an intrinsic dimensionality of 3 has strongly descending eigenvalues. Most notably, it has 3 eigenvalues that are non-zero. This example constitutes an extreme case of the influence of intrinsic dimensionality on eigenvalues; under less artificial circumstances, sampling error would probably raise the values of several of the eigenvalues in both matrices (choosing $n = 1000$ on data with low variance induced this behavior). The point, however, is that eigenvalues for dimensions that exceed the intrinsic dimensionality of a matrix will tend to be small.

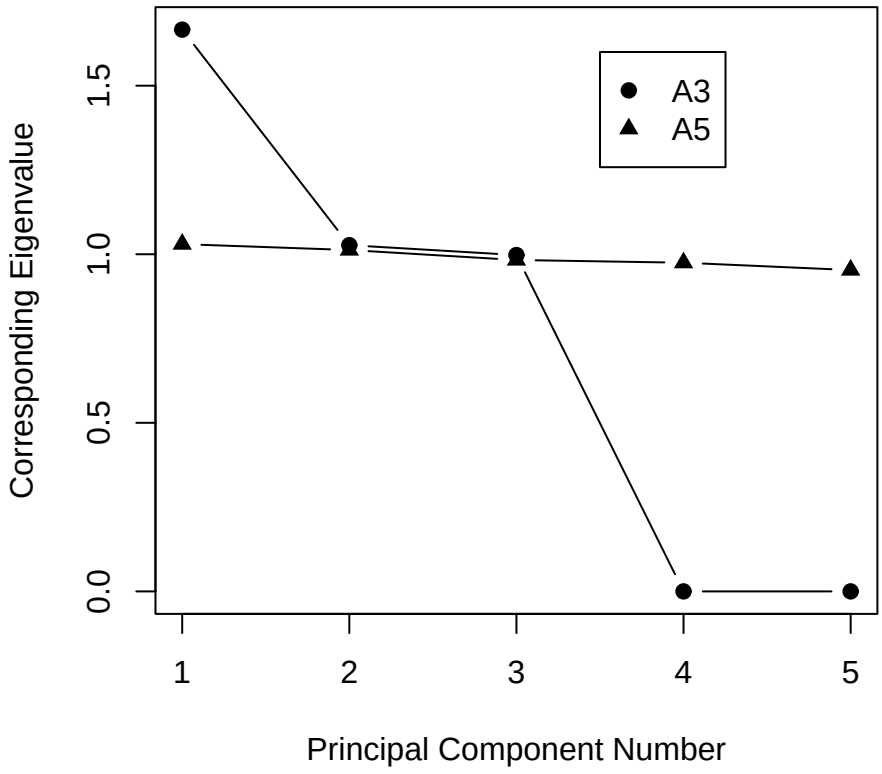

FIGURE 1.4.1. Scree plots of matrices with differing intrinsic dimensionalities.

In contrast to k_{opt}, which is defined on the PDF of $\mathbf{A}$, I also define the *observed optimal dimensionality* of $\mathbf{A}$ with respect to IR performance metric m. Under operational circumstances, we do not know the PDF that generated a given term-document matrix. But we can observe how well a certain estimation of k_{opt} fares when it is applied to a retrieval problem. As I discuss in Section 2.4, evaluating the performance of information retrieval systems is an unsolved problem. Due to the abstract imperatives of the task, judging the quality of a system's performance is non-trivial, and in many instances subjective. Nonetheless in Section 3.2 I discuss three performance metrics that informed my analysis during this study: average precision, the harmonic mean of precision and recall, and average search length. For a given corpus C, we may apply each of these metrics at all possible parameterizations

of k. By tracking the performance of a given metric m at each value of k and noting the dimensionality that optimizes the given performance measurement, we derive the observed optimal dimensionality of $\mathbf{A}$ with respect to m. A major element of this dissertation involves using retrospective performance analysis to judge the quality of various eigenvalue-based estimates of intrinsic dimensionality. That is, the experiments described in Chapters 3 and 4 use observed optimal dimensionality with respect to three evaluation metrics as a surrogate for knowledge of the true intrinsic dimensionality to enable comparison of the accuracy of five eigenvalue-based dimensionality estimators.

1.5. Conclusion

Automatic information retrieval systems face a difficult challenge: to offer a computational treatment of documents' *aboutness, similarity,* and *relevance*. While approaches to this task run a wide gamut, a large number of IR systems rely on geometric models such as those outlined in this chapter. Under these geometrically motivated models, retrieval is seen as a problem of navigation in information space. Thus similar documents are assumed to reside close together in the vector space spanned by the system's representational axes. Despite the success of the geometrical approach to IR, however, deriving the optimal model of information space remains an open problem.

Salton's foundational VSM sacrifices realism for the sake of simplicity when it assumes statistical independence among terms. In contrast, Wong's GVSM extends Salton's model, rotating the axes of its information space by including a model of term associations based on the sample correlation matrix. Finally, LSI extends the GVSM, attempting to improve the model by constructing a statistical model of the population correlation matrix via dimensionality reduction.

In a number of empirical studies LSI's statistical modeling approach has been shown to improve retrieval over traditional keyword-based techniques. However, dimensionality reduction saddles researchers with an important question. If we are going to reduce model dimensionality, how aggressively should we do so? Or conversely, what is k_{opt}, the optimal number of dimensions to retain under LSI? In the unsupervised learning environment of IR,

defining notions of model goodness of fit and optimality is non-trivial. Although *ad hoc* studies have shown good performance using $k \approx 100 \cdots 300$, a sound criterion for model selection during LSI has not been forthcoming. This dissertation attempts to remedy this omission. Specifically, I pursue the question of how best to estimate the intrinsic dimensionality of a corpus by recourse to a statistical analysis of its eigenvalues.

This study approaches the problem of dimensionality estimation for IR broadly, methodologically speaking. After a more thorough discussion of my research domain in Chapter 2, Chapter 3 outlines a series of experiments undertaken to compare the utility of five eigenvalue-based dimensionality estimators. Next, Chapter 4 reports the results of these experiments, comparing attempts to predict the observed optimal dimensionality of six test corpora, with respect to three performance metrics, via five eigenvalue-based statistical estimators. My goal in this analysis is to note conditions under which certain estimators succeed or fail in predicting the observed optimal dimensionality for retrieval. In Chapter 5 I supplement Chapter 4's empirical analysis by conducting similar experiments on simulated data whose intrinsic dimensionality is known beforehand. Finally, I conclude in Chapter 6 by synthesizing and interpreting my findings.

Throughout these experiments I give particular attention to interpreting the performance of my own proposed method, amended parallel analysis (cf. Sections 2.3.3 and 3.3). I suggest that due to APA's ability to account not only for the magnitude of eigenvalues, but also for their variability, the technique yields an accurate estimate of a corpus' effective dimensionality, and by extension, a superior estimate of its observed optimal dimensionality for retrieval. Such an analysis will hopefully prove useful for practitioners insofar as the selection of an appropriate value of k is a crucial step during LSI. However, I argue that this approach will also solidify the theoretical motivation for dimensionality reduction in IR.

CHAPTER 2

Literature Review

While the matter of dimensionality reduction for information retrieval has generated its own body of research, a full understanding of its intricacies demands reading across literatures. Of course information retrieval researchers do address LSI and its attendant intricacies; but scientists in machine learning have also made strides that shed light on empirically determined dimensionality estimation. Since the work of R. A. Fisher, statisticians have formulated theories of linear models, of which LSI is a special case [**50**]. Finally, mathematical work on the theory of eigensystems provides a foundation for any rigorous analysis of least-squares modeling, such as LSI. This chapter aims to contextualize my discussion by offering some background on the research areas germane to empirical dimensionality reduction.

The first of the following sections paves the way for my analysis of LSI by surveying the theory that underpins the vector space model of information retrieval. An extension of the generalized vector space model [**150**] (cf. Section 1.1.3), LSI owes its formulation and implementation not only to research within the IR community, but to a tradition of statistical and mathematical work in linear models and the theory of eigenvalues; Section 2.2 describes the mathematical and conceptual framework that informs LSI. Of particular importance to understanding and implementing LSI is the matter of dimensionality reduction: given a corpus with n documents and p terms, what is k_{opt}, the dimensionality of the optimal LSI model? A body of techniques for identifying a corpus' intrinsic dimensionality forms the subject of Section 2.3. In Section 2.4 I turn to the problem of evaluation in information retrieval systems. If a goal of IR research is to discover good models, it pays to be rigorous in our definitions—what does it mean to evaluate a model? What does it mean for one model

to improve on another? Section 2.4 surveys consensus and contention *vis a vis* evaluation of IR systems.

2.1. Intelligent IR via the Vector Space Model

Information retrieval systems search databases for documents that are relevant to a user's stated information need [**4, 118**]. Following Baeza-Yates, I adopt a "logical" definition of basic IR vocabulary, "[using] the term *document* to denote a single unit of information, typically text in a digital form, but [inclusive of] other media" [**4**, p. 141]. I consider queries to be analogous to documents, both mathematically and conceptually; that is, for our purposes, queries may be considered "pseudo-documents." In older IR systems, documents entailed just a few key words, titles, or abstracted surrogates of longer works [**22, 101**]. But thanks to improved computing resources and the growth of electronic corpora such as the World Wide Web, documents in many newer IR systems contain a full reproduction of electronic texts[1].

While the advent of richer document representations extended the possible merits of information retrieval systems, it also raised the challenges facing IR researchers. Early IR efforts, based on controlled vocabularies and well-structured document surrogates, resembled standard database systems. Retrieving documents from these systems entailed a simple lookup table—an inverted index—against which query terms could be compared using set-theoretic operations. This led to the traditional Boolean approach to IR. Commenting on the inadequacies of Boolean systems, W. S. Cooper wrote in 1988, "it should be possible to improve considerably upon the fundamental design features of most present day retrieval systems" [**26**]. Cooper suggests that modern IR—what I term "intelligent information retrieval"—should use *a priori* knowledge about language to improve upon the Boolean model.

As Cooper recommends, intelligent information retrieval systems borrow from machine learning, artificial intelligence, and linguistic research. The volume and complexity of research into intelligent IR prohibits a general treatment of the subject. Instead, anticipating

[1]For that matter, documents need not be textual at all. However, multimedia retrieval lies outside the scope of this study. For discussions of multimedia retrieval, see [**103**].

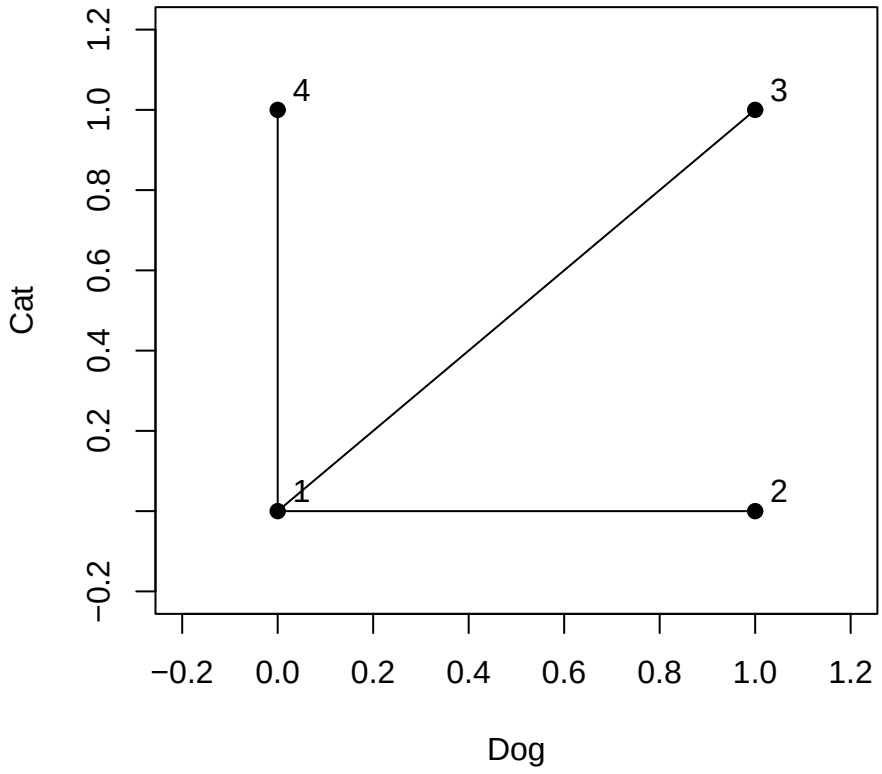

FIGURE 2.1.1. *Pets* data as vectors

our discussion of LSI, this section describes research into IR systems that build upon the vector space model of retrieval.

2.1.1. The Vector Space Model. Salton's vector space model (VSM) of IR imagines retrieval in linear algebraic terms **[125, 131, 129]**. Under Salton's model, each document represents a vector in a p-dimensional vector space, where p is the number of indexing terms used by the system. The location of the i^{th} document d_i along the j^{th} axis corresponds to the presence or absence of the j^{th} term in the i^{th} document. The simplest expression of the vector space model treats terms as binary data. Thus $d_{ij} = 1$ if the j^{th} term appears in the i^{th} document. Otherwise $d_{ij} = 0$.

Document	*Contents*
1	The Iguana: man's best friend
2	Guide to raising a dog
3	Raininig Cats and Dogs
4	A Cat-lover's Guide to Cats

TABLE 2.1.1. The *pets* data

Table 2.1.1 contains an imaginary, very small document collection, that I call the *pets* data. Figure 2.1.1 depicts the *pets* data as points in a vector space. In this example, four documents are represented by two terms, *cat* and *dog*. The vector space shown in Figure 2.1.1 is defined as the space spanned by the rows of matrix **A**:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{2.1.1}$$

Matrix **A** is known as a term-document matrix; the i^{th} column of **A** represents the i^{th} indexing term in document space. On the other hand, the j^{th} row represents document j as a vector in term-space. Document number 1 contains neither indexing term, and thus the system locates it at vector (0, 0). On the other hand document 3 contains both *cat* and *dog*, and thus becomes (1, 1). Under the vector space model, similarity between two documents i and j is defined as the inner product between the i^{th} and j^{th} document vectors:

$$sim(i,\, j) = \mathbf{i} \cdot \mathbf{j} = \sum_{m=1}^{t} i_m \cdot j_m. \tag{2.1.2}$$

If the document vectors have been normalized to unit length, then Equation 2.1.2 becomes the vector cosine:

$$sim(i,\, j) = cos_{ij} = \frac{\mathbf{i} \cdot \mathbf{j}}{\|\mathbf{i}\|\, \|\mathbf{j}\|}. \tag{2.1.3}$$

Thus $sim(1,\, 4) = 0$, while $sim(3,\, 4) = 0.71$. This notion of similarity is intuitively appealing insofar as it conflates ideas of geometrical proximity and lexicographic content; documents that share indexing terms are considered close together in term space. Documents 1 and 4

Document	*Sim(q,di)*
3	1
4	0.71
2	0.71
1	0

TABLE 2.1.2. Query-document similarity

share no terms, whereas documents 3 and 4 share one term. Thus documents 3 and 4 are closer together than documents 1 and 4 under Salton's model.

Given some document d_i, we may rank all other documents $d_{i'}$ in a collection by their pairwise similarity with d_i via Equation 2.1.3. It is often useful in IR settings to obtain a list of all documents, ordered by their similarity to some d_i. This is particularly common in *ad hoc* retrieval situations, where the system is to be presented with a user's query. Here, the query is represented as a "pseudo-document," q_i. After translating a query into q_i, vector space retrieval entails calculating $sim(q_i, d_i)$ for all d_i. The system then presents items to the user, ranked by their similarity to q_i.

Returning to the *pets* example, imagine a query: *titles about 'dogs' or titles about 'cats and dogs.'* Projecting this query into the vector space spanned by $\mathbf{A}$ yields query vector $\mathbf{q}$:

$$\mathbf{q} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Ranking each document against $\mathbf{q}$ using Equation 2.1.3 yields the query-document similarity table shown in Table 2.1.2. A system based on the vector model would present documents to the author of $\mathbf{q}$ in this order.

Variations of Salton's model have performed well experimentally [**130**], and have become a mainstay of IR research. The vector space model is especially important in the historical development of academic IR due to its formalization of the notion of inter-document similarity. Unlike Boolean approaches (based on set-theoretic ideas, cf. [**118**, chs. 3, 5]), Salton's vector space method implies that similarity is analogous to geometric proximity; inter-document similarity is a matter of degree, and $sim(d_i, d_{i'}) = sim(d_{i'}, d_i)$ for all $i \neq i'$.

Document	$sim(q, d_i)$
3	0.95
2	0.9
4	0.45
1	0

TABLE 2.1.3. Query-document similarity

This implication lends vector space IR intuitive and mathematical appeal (discussed below), but it also bears unfavorable baggage. By defining document-document similarity as the vector cosine, Salton suggests that similarity is linear on the collection's indexing features. That is, vector space IR assumes that indexing terms are statistically independent. This assumption is patently false **[102, 111, 26, 27]**, although it is unclear exactly how grievously the assumption of term independence degrades the performance of IR systems **[96]**.

2.1.2. Elaborating on Linearity: Term Weighting and Query Expansion as Primitive Data Reduction. In the previous example, we represented the *pets* data using the binary matrix $\mathbf{A}$. However, a more realistic model defines $\mathbf{A}$ as real-valued, where a_{ij} is, for example, an integer that records the number of times that the i^{th} term occurs in the j^{th} document. Such a scenario locates the pets documents in $\Re^2$ via matrix $\mathbf{A}_r$:

$$\mathbf{A}_r = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix}.$$

The vector space described by $\mathbf{A}_r$ is shown in Figure 2.1.2, and query q is shown in vector form as:

$$\mathbf{q}_r = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Figure 2.1.3 shows the relevance table under this model. The user has requested *titles about 'dogs' or titles about 'cats and dogs'*. By storing term-document count information,

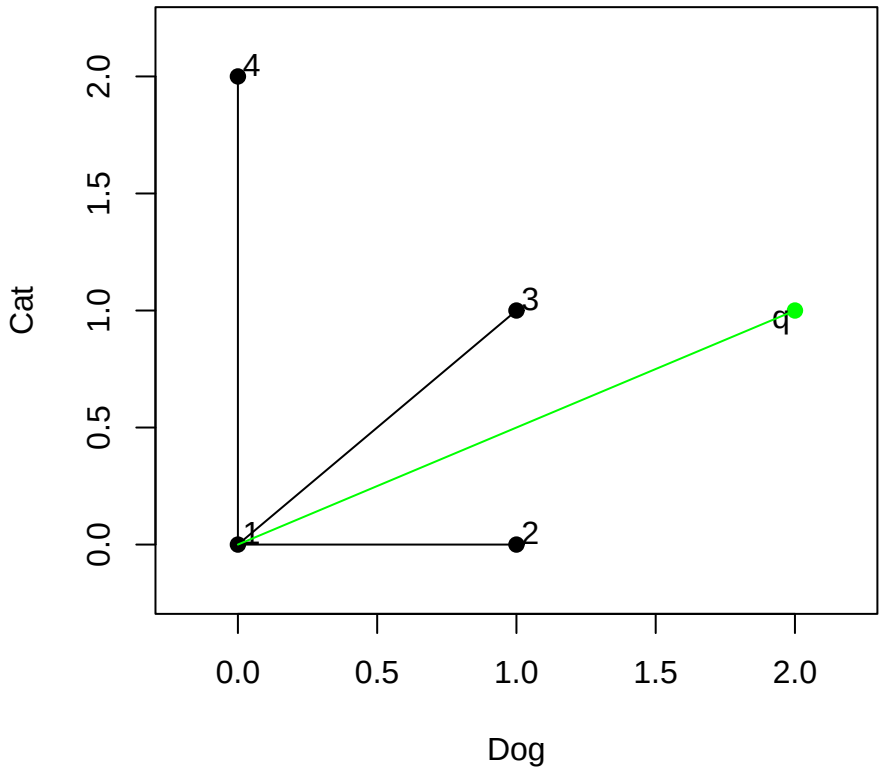

FIGURE 2.1.2. *Pets* data as integer-valued vectors

the new model gives a novel ranking for the same data, effectively weighting *dogs* in the user's query. Whether this ranking is better than the original is difficult to say. Instead, the important point is that by altering the mechanics of our document representation, we have imposed two different semantic models on the same data. In fact the vector space model thus refers to a general family of retrieval models, whose notions of similarity are analogous but distinct.

While term-document counts provide more information than binary data, in practice IR systems tend to use more elaborate methods of term weighting. Salton argues that "distinctions must be introduced between individual terms, based on their presumed value

as document descriptors" [**126**]. Viable term weights tend to be based on several factors; Salton identifies two in particular:

- *term frequency:* how many times the term appears in the document
- *inverse document frequency:* how often the term appears in the database

By taking these factors into account, Salton argues, IR systems may develop weights whose interactions provide suitably realistic models of natural language semantics.

Even the earliest work in IR recognized the importance of term frequency (tf) data for text analysis. In the mid-1950s, H. P. Luhn argued that the most important terms in a document were those that occur with middling frequency [**99, 100**]. Extremely common terms such as *the, in,* and *it* are over-represented in almost all English corpora; their presence or absence thus conveys very little information about document "aboutness." On the other hand, many terms in a corpus will occur rarely, once or perhaps twice. These so-called *hapax legomena* provide too little information for useful text processing. Instead, Luhn suggests that those terms that occur with mid-range frequency should be weighted when computing inter-document similarity. Thus Salton argues that any term weighting model should account for term frequency.

In [**82, 83**] Karen Sparck Jones introduces the notion of *inverse document frequency* (idf). According to Sparck Jones it is not sufficient to consider a term's global frequency (tf) when estimating its usefulness for discrimination. She argues that some analysis of a term's distribution across documents should supplement tf analysis. This consideration stems from the possibility that a term could be quite common, but present in only a small subset of a corpus' documents. A purely tf-based model would demote such a term due to its common appearance, although its localized distribution suggests that it could serve as a useful marker for a subclass of document. Thus "the idf factor," writes Salton, "varies inversely with the number of documents n to which a term is assigned in a collection of N documents. A typical idf factor may be computed as $\log N/n$" [**126**].

Having identified two components of a term weighting model, Salton offers the general idea of *term discrimination* as a weighting scheme for IR [**132**]. A term with high discrimination value is "able to distinguish certain individual documents from the remainder of

the collection. This implies that the best terms should have high term frequencies but low overall collection frequencies" **[126]**. As an estimate of a term's discrimination value, Salton recommends the product of its *tf* and *idf* scores ($tf \times idf$). Although the use of $tf \times idf$ term weighting continues to draw criticism due to its lack of theoretical foundation **[28, 14]** it also sees widespread use in IR research **[13, 78, 114]**.

In its simplest, binary articulation, the vector space model imagines that all terms are equally important, and that their mere presence or absence—as opposed to the frequency of their repetition—determines the conceptual content of a document. The term discrimination model attempts to remedy this oversimplification. By accounting for the number of times a term appears in a document (tf), the term discrimination model implies that terms that appear often in a document convey more of the document's content than do terms that appear just once. And by analyzing each term's distribution across documents (idf), the model accounts for a feature's semantic range, suggesting that those terms that are heavily used in a small group of documents will be strong discriminators for retrieval purposes.

We may understand the development of term weighting schemes as a primitive effort to improve an IR system's native language model via data reduction. Without term weighting, an IR system has little choice but to represent each document in p-space, where p is the total number of terms in the collection. However, many words in a corpus are only marginally useful for IR purposes. Thus stop-lists **[4, 129, 125]** are useful for removing high-frequency terms. Likewise, the use of stemming **[113]** can reduce p by mapping variants of a stem down to a single form. In both cases, researchers hope to eliminate document features that introduce noise into the document ranking process. Term weighting schemes provide a statistical means of effecting the same result. "Is the determination of the terms weights," asks Salton, "capable of distinguishing the important terms from those less crucial for content identification?" **[126]**. Assuming that our weighting model is up to the task, we may derive the k most important features in a collection by ranking the terms by idf weight and keeping the top k.

To demonstrate the utility of removing variables from a document representation consider the following simulated data set. This imaginary corpus contains 300 documents, each

Class	*Mean Vector*	*Cov. Matrix*
1	$[1, 0]$	$\mathbf{I}_2$
2	$[1, 3]$	$\mathbf{I}_2$
3	$[1, -3]$	$\mathbf{I}_2$

TABLE 2.1.4. Distributions for feature selection simulation

of which belongs to one of three semantic classes—say, documents about *dogs*, *cats,* and *mice*. There are 100 documents in each class. Each document in this example is represented as a 2-vector, where one variable is intended to be useful for discriminating between classes, and the other variable is the same across classes—i.e. it constitutes a noise variable. Documents were drawn from three multivariate normal distributions, described in Table 2.1.4. Obviously variable number 2 provides better evidence for classifying these documents than does variable 1, as is evident in Figures 2.1.3 and 2.1.4, which show the simulated data in alternate 1-spaces.

Figure 2.1.3 shows the 300 simulated documents (with each class in a distinct character) represented only by variable 1, the noise variable. Defining a classification rule for these data would be almost impossible. On the other hand, Figure 2.1.4 represents each document by its score on variable 2, a rendering that isolates the classes quite well. That variable 2 constitutes a better indexing variable (for this classification task) than variable 1 is also expressible numerically. I subjected this simulated data set to the k-means clustering algorithm twice. The first time I clustered the data using only variable 2. This clustering yielded a 94.3% classification accuracy. During the second clustering, both variables were used. Adding variable 1 to the data representation yielded almost no improvement in classification; this round of clustering classified 94.7% of the data correctly. Representing the data using only variable 1 would reduce the classification process to sheer guesswork. For the purposes of document classification, then, we would do well to set more stock in each observation's score on variable 2 than on its variable 1 score. This is the end result of a feature weighting model. But for that matter, we might ignore variable 1 altogether, which corresponds to dropping those variables with low discrimination values.

The point of this somewhat gratuitous example is to suggest that all variables are not created equal. The binary vector space model expressed in Figure 2.1.1 assumes that each

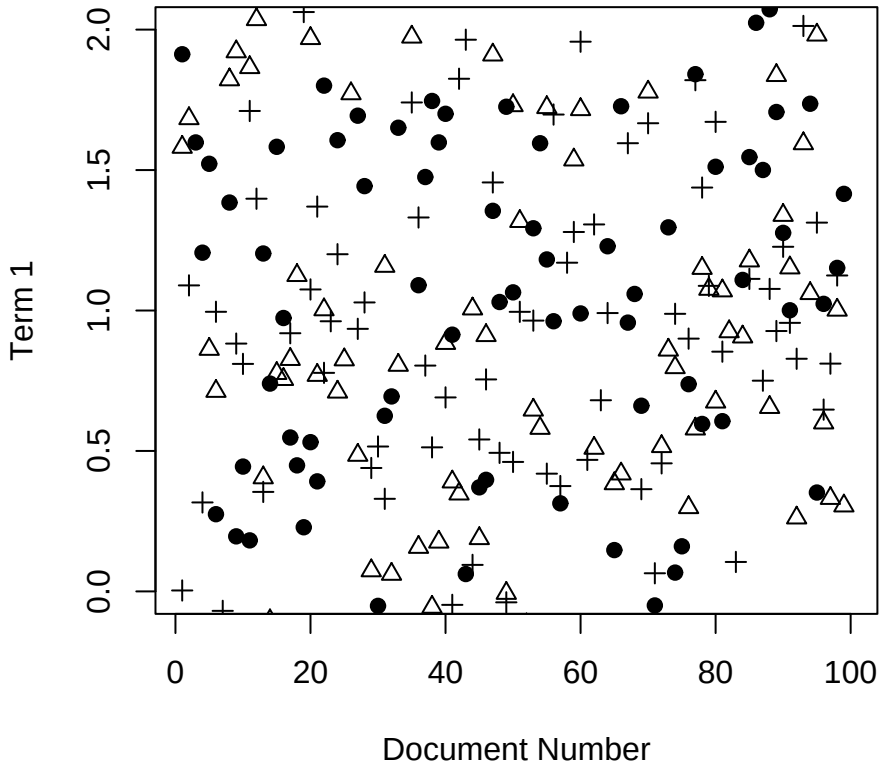

FIGURE 2.1.3. Simulated data in noisy 1-space

variable is as important as its neighbor. On the other hand, the term weighting models described in this section admit a more nuanced semantics. Not only does it matter how many times a term appears in a document, for the term discrimination model, it matters how many documents contain the term. Such a model allows an IR system to restrict its analysis to those terms deemed promising discriminators. By reducing its document model from a vector in p-space to a vector in k-space, where $k < p$ (whether via outright variable exclusion, or by judicious feature weighting), an IR system stands to mitigate the effect of so-called "function words," freeing resources for analysis of "content words," words likely to convey topical information about a document's semantics.

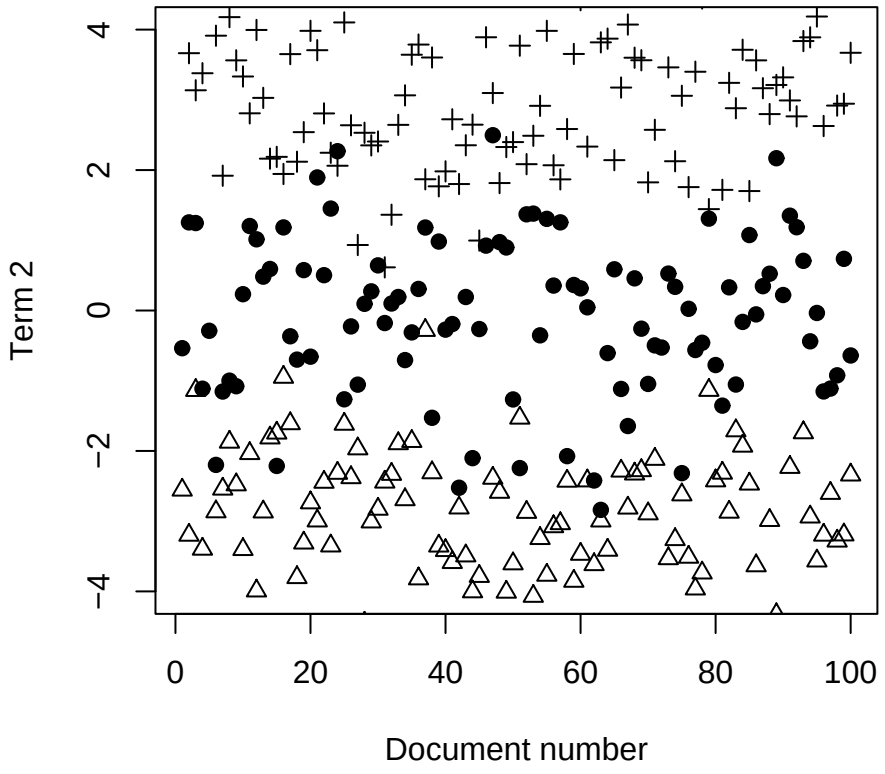

FIGURE 2.1.4. Simulated data in 1-space

2.1.3. Thesauri and Query Expansion. Singling out good indexing terms is a prerequisite for robust indexing. Without this step, a system may be led astray by purely functional (as opposed to topical) words. However, the methods described so far do nothing to address the assumption of linearity inherent in the vector space model. Joining Cooper's earlier criticism [**27**], Salton suggests that in addition to term discrimination models IR systems should account for statistical dependencies among terms.

> In most of the early [IR] experiments, *single terms* alone were used for content representation, often consisting of words extracted from the texts of documents and from natural language query formulations....Ultimately, however, sets of single terms cannot provide complete identifications of

> document content. For this reason, many enhancements in content analysis and text indexing procedures have been proposed over the years in an effort to generate complex text representations. [**126**]

Salton identifies four means of extending the vector space model to include term-term correlation information:

(1) Generating sets of related terms by observing co-occurrence data from online corpora (cf. [**94, 117, 21**]).

(2) Deriving n-gram features (cf. [**84**]). i.e. Identifying common n-word phrases and considering them indexing features akin to individual words.

(3) Use of online thesauri (cf. [**2, 84, 53, 49**]).

(4) Construction of knowledge bases or other encodements of logical relations among indexing terms (cf. [**30, 31**]).

These techniques address the vector model's assumption of term independence by articulating relationships between terms at a global (i.e. corpus-wide) level. Thus a thesaurus-based system might observe, thanks to its *a priori* encoded knowledge of term relationships, that *car* and *automobile* tend to co-occur. When such a system encounters a document or query d_i that contains *automobile* but not *car,* the system might supplement its representation to obtain $d_{i'}$, a representation identical to d_i, except that the vector entry for *car* would show a positive value (instead of the observed 0 value). In the final analysis, such a system still resorts to a linear similarity function. However, under the models discussed here, some notion of non-linearity is imposed through the application of *a priori* knowledge about term-term correlations.

Closely related to such methods of enhancing document representation is the family of techniques known as query expansion [**4**, ch. 5]. These methods address shortcomings in the VSM by local analysis (i.e. interrogating a query-specific neighborhood of documents). Under the vector space model, variants of the Rocchio approach to relevance feedback [**121**] have been particularly successful [**127**]. In thesaurus-aided retrieval, documents are augmented by recourse to previously observed term-term relationships. Relevance feedback takes another approach. Here the goal is to fashion an optimal query, q_{opt} by analyzing

Symbol	*Meaning*
D_r	set of relevant documents among retrieved documents
D_n	set of non-relevant documents among retrieved documens
C_r	set of all relevant documents
$\|D_r\|,\|D_n\|,\|C_r\|$	number of elements in each set of documents
α, β, γ	constant parameters

TABLE 2.1.5. Notation for Rocchio Relevance Feedback

the content of the set of documents, C_r, that are retrieved by a given query q. Table 2.1.5 is adapted from [**4**] and [**121**]; it describes notation used in Equation 2.1.4. The Rocchio method begins by imagining that we have knowledge of the relevance value *vis a vis* query q for every document in our corpus. Assuming that this is the case, Rocchio proves Equation 2.1.4:

$$\mathbf{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \mathbf{d}_j \in C_r} \mathbf{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \mathbf{d}_j ! \in C_r} \mathbf{d}_j. \tag{2.1.4}$$

The optimal query is thus a weighted sum of relevant and non-relevant document vectors, where the weights depend on the size of C_r in relation to the size of the collection. In practice, however, we do not have access to the requisite sets of relevant and non-relevant documents. In practice, the final query vector under Rocchio expansion is thus generated by Equation 2.1.5.

$$\mathbf{q}_m = \alpha \mathbf{q} + \frac{\beta}{|D_r|} \sum_{\forall \mathbf{d}_j \in D_r} \mathbf{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \mathbf{d}_j \in D_n} \mathbf{d}_j \tag{2.1.5}$$

Whether derived via Equation 2.1.5 or by another method (cf. [**75**]) the goal of relevance feedback is to infer relationships between terms. By utilizing what amounts to weighted averages of the feature vectors associated with relevant and non-relevant documents, these methods estimate the degree to which each indexing term participates in the topic addressed by the query q. The idea here is that q is over-specified. That is, q contains non-zero entries only on those terms that the user has chosen to name explicitly. But there may be many terms that are related to q implicitly; an optimal query, under the Rocchio model, should model this implicit evidence. Thus q refers to an intangible conceptual topic, but it provides incomplete evidence about this reference; the query vector is parsimonious in its topical

representation. Thus relevance feedback constructs an idealized solution vector $\mathbf{q}_m$, which is the best linear approximation of $\mathbf{q}_{opt}$ insofar as $\mathbf{q}_m$ maximizes the similarity between itself and the centroid of the set of relevant documents while maximizing its distance from the centroid of the set of non-relevant documents.

Query expansion and relevance feedback methods attempt to remedy the linearity inherent in the vector space model of IR. Just as term weighting models permit important feature selection operations, these techniques improve the VSM language model. The goal in all of the schemes addressed here is to improve the orientation of the vector space spanned by the features of a document-term matrix $\mathbf{A}$. Because natural language terms exhibit non-linear relationships, researchers use query expansion and relevance feedback to augment the VSM. These augmentations are intended to improve the information space spanned by the columns (terms) of $\mathbf{A}$ by modeling its implicit correlational structure. In this section I have discussed improvements of a fairly primitive nature. The following section treats more advanced means of improving an information space.

2.2. Latent Semantic Indexing

It has been argued extensively that term-based information discovery places undue cognitive burden upon end-users, whose interest tends to lie in abstract concepts rather than in specific words [**55, 110**]. Arguing that modern IR should account for linguistic and psychological developments in the cognitive sciences, Newby calls for "computerized representations of data sets (e.g. document collections) that are consistent with human perception of the data sets" [**110**]. Toward this goal, Newby identifies two useful notions, which I adopt in this discussion. An *information space* is the set of concepts and relations among them held by a computer system. Information spaces are comprised of words, documents, and the relations among them. In an information space, documents are represented in *term space,* while terms are represented as vectors in *document space.* On the other hand, a *cognitive space* is the set of concepts and relations among them held by a human. Although it is difficult to identify the fundamental elements of cognitive spaces, Newby finds a high degree of similarity among psychometric analyses of individual habits of linguistic association. This

finding is in keeping with a body of psychological research that has advanced the notion of conceptualization as a statistical association of language and experience [**122, 124, 149**].

Latent semantic indexing begins with the assumption that modeling the term correlations in IR reduces the cognitive burden on searchers. While a full treatment of the psychometric validity of LSI is beyond the scope of this dissertation (cf. [**90, 89, 91, 51, 56**]), the LSI approach to retrieval assumes that accounting for the correlational structure of a document-term matrix $\mathbf{A}$ improves the representation afforded by the vector space model. Latent semantic indexing thus takes an empirical approach to addressing the gap between information spaces and cognitive spaces. As discussed in Sections 1.1.2 and 1.1.3 LSI constructs a statistical model of the population term correlation matrix by the method of least-squares. Proponents of dimensionality reduction argue that such a model improves the representation of the VSM by mitigating the error introduced by the assumption of term independence.

In efforts to improve the representation of terms and documents, latent semantic indexing [**32, 10, 73**] extends the VSM by means of statistical modeling. Closely related to principal component analysis, multidimensional scaling, and factor analysis, LSI derives a low-rank approximation, $\widehat{\mathbf{A}}_k$, of the term-document matrix, $\mathbf{A}$, where $\widehat{\mathbf{A}}_k$ provides the best rank-k fit of $\mathbf{A}$, in the least-squares sense. By projecting a system's information space onto the low-rank $\widehat{\mathbf{A}}_k$, LSI achieves two putative benefits over the standard vector model: a rotation onto independent axes, and dimensionality reduction. Describing LSI, Deerwester *et al.* argue the merits of these transformations:

> A fundamental deficiency of current information retrieval methods is that the words searchers use often are not the same as those by which the information they seek has been indexed. There are actually two sides to the issue; we will call them broadly *synonymy* and *polysemy.* [**32**]

Synonymy intrudes on retrieval when a searcher uses different terms in a query than an author or indexer used in a relevant document. Thus elementary retrieval systems would fail to deliver documents about *automobiles* when presented with a query about *cars.* On the other hand, retrieval suffers due to polysemy because natural language terms tend to

have several meanings, or senses. The term *car* can imply quite different topics in different contexts; for example, documents about *railroad cars* are of little interest to searchers hoping to buy a new automobile.

Although its goals are more complex than this brief outline suggests, the aim of LSI is in essence to negotiate such vagaries of natural language. The remainder of the current section describes the method of LSI, emphasizing first its conceptual basis, then moving on to a treatment of the mathematics behind the derivation of an appropriate $\widehat{\mathbf{A}}_k$. Because many earlier studies [**32, 10, 90**] have explained LSI clearly and thoroughly, I take a more specialized approach. Anticipating a later treatment of optimal dimensionality reduction, my goal throughout this discussion is to articulate the details of LSI in the context of statistical model selection.

2.2.1. Rationale behind LSI—Improving the Vector Space Model via Statistical Modeling. To ground their introduction to LSI, Deerwester *et al.* describe their method programatically:

> The proposed approach tries to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a statistical problem. We assume there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. [**32**]

The goal of LSI is thus to construct a statistical model. With this model we may estimate the degree to which a document d_i (or a term t_j) participates in a given topic, or factor f_k. If the model is well-formed, the set of derived factors F will, Deerwester *et al.* argue, approximate the set of concepts present in a human user's cognitive space. From a statistical standpoint, we may understand this argument in terms of the representation of the term-term correlation matrix. That is, the optimal LSI model yields the best representation of the relationship between terms that obtains in the probability density function that generated the data. An IR system may use such a model for retrieval purposes by allowing it to inform its model of inter-item similarity. Under traditional vector space models, documents reside in term space. LSI replaces this arrangement by locating each document in the inferred

factor space. That is, each document d_i is represented by its vector of scores obtained from the model. Likewise, each term t_i resides at the vector defined by its fitted parameters under each of the models.

The literature on statistical modeling is vast, and a full exploration of it is beyond the scope of this study. Instead, I present a brief discussion of the major results from the literature, emphasizing their applicability to information retrieval under LSI.

A statistical model approximates the dynamics of a variable, stochastic system. According to Neter *et al.* **[109]**, statistical models contain two components:

- *A Functional Element.* The model expresses the relations among system variables as a mathematical function.
- *A Stochastic Element.* We assume that the behavior of the system is non-deterministic, but rather that its dynamics is in part governed by a set of probability distributions.

A mathematical model describes a system deterministically. For example, we may construct a model to calculate a firm's monthly income based on the number of products sold that month. Such a model defines two variable types:

- *Dependent Variables.* We predict the value of a dependent variable based on given knowledge of other variables in the system. In this case, the value of monthly income depends on monthly sales. As such, it acts as a dependent, or *response* variable.
- *Independent Variables.* Independent variables, also called *predictors,* provide the information by which we predict the value of a dependent variable. Here, monthly sales is an independent variable.

To predict y, monthly income, based on monthly sales, x, we might define the function $y = f_1(x) = price * x$, where *price* is a constant. Setting *price* =\$2, for example, we can predict y via $f_1(x)$, as shown in Figure 2.2.1. The points in Figure 2.2.1 represent the data points for this deterministic system for $x = 1, 2, ..., 50$, while the solid line represents the corresponding predictions under $f_1(x)$. Due to the non-random behavior of this system, the data observations and the model predictions are identical.

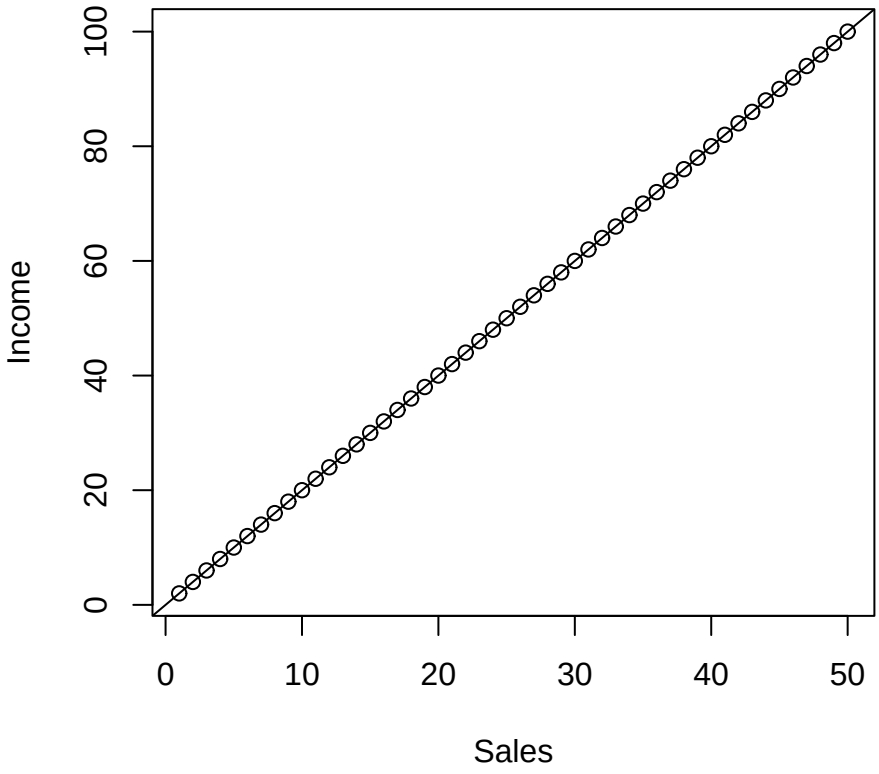

FIGURE 2.2.1. A Mathematical Model of Company Income

Unlike mathematical models, statistical models include some fuzziness. They describe what Bhattacharyya and Johnson term "semi-deterministic" systems [**11**]. Suppose that again we wish to predict income based on monthly sales. However, our company has now moved online, conducting all transactions on Ebay. Because we operate by auction now, *price* is no longer a constant factor. Instead, it becomes a random variable, subject to some distribution. For the sake of argument, assume that *price* follows a Gaussian distribution; thus $price \propto N(\overline{price}, \sigma^2)$, where $\overline{price}$ is the population mean of *price* and $var(price) = \sigma^2$. The behavior of such a system is shown in Figure 2.2.2 for a set of simulated data. Under such circumstances, our deterministic model no longer provides a perfect fit to the data; the data points do not lie upon a straight line. While we could derive an extremely elaborate function

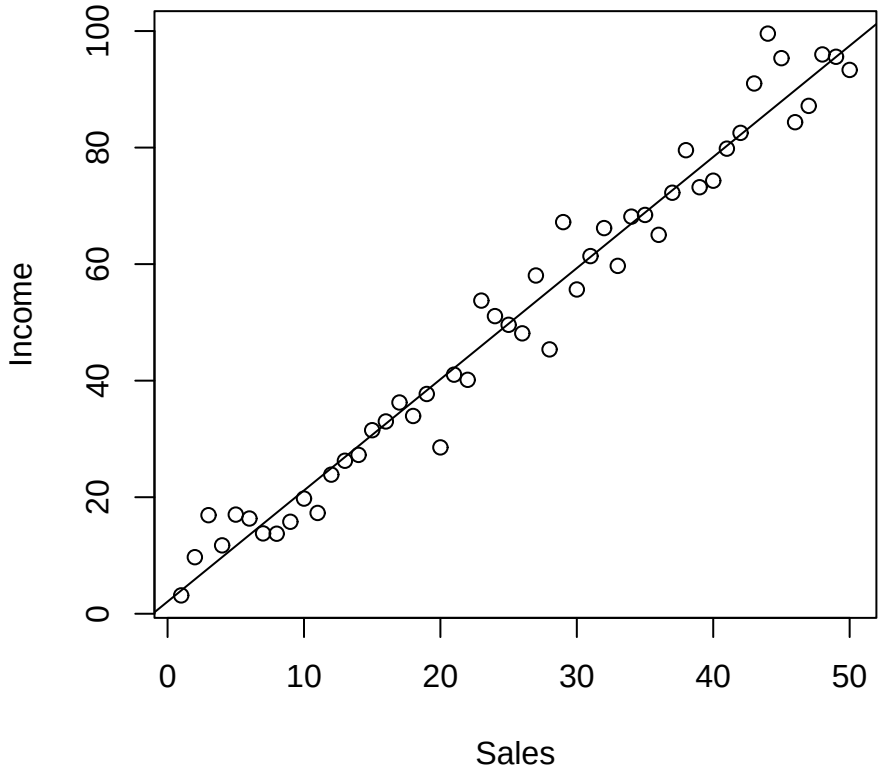

FIGURE 2.2.2. Statistical Model of Company Income

to describe these data as a deterministic system, we suspect that a simpler explanation is more appropriate. Instead of a deterministic model we use statistical methods to derive a useful *approximation* of the data.

We might model the system using the function $y = f_2(x) = \overline{price} * x + \epsilon$, where ϵ is a random variable called the error-term. We thus assume that a functional relationship—a systematic correlation among variables—drives the behavior of the system, but that some random noise obscures direct observation of the function. Moreover, we assume that the system's randomness is structured insofar as it is governed by a probability distribution. The goal of statistical modeling is thus to estimate the functional portion of the system. The straight line visible in Figure 2.2.2 approximates the trend visible in the stochastic data,

assuming that their underlying relationship is still governed by $f_2(x)$, which is to say that $\overline{\epsilon} = 0$.

The process of building a statistical model is empirical. We begin by observing data and using this observation to draw conclusions about them. The process of observation includes 3 main steps [**79, 116, 109**]:

- We must choose the family of functional relations likely to describe the system's behavior. Although a large variety of functions have been explored [**104, 19, 67, 145**], the family of linear functions are widely used, due to their mathematical tractability and descriptive power [**109, 116**].
- The researcher must identify the probability distribution that governs the variability of the system.
- A method of parameterizing the model function must be chosen.

The method of linear regression is one of the most commonly applied modeling procedures, and is closely related to LSI. Using linear regression we predict a dependent variable y based on a linear combination of the p predictor variables, $x_1, x_2, ...x_p$. I shall return to the p-variate regression problem momentarily. But I first motivate our discussion by describing the mathematics of simple linear regression, where we recognize only a single predictor, x. A similar discussion of the relation between retrieval and linear regression is available in [**140**].

Simple linear regression formalizes the notion of correlation between variables [**109, 11**]. The model is expressed by Equation 2.2.1:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.2.1}$$

where y_i is the i^{th} response, β_0 and β_1 are fitted parameters, x_i is the i^{th} observation, and ϵ_i is the i^{th} error term. The linear regression model makes several assumptions about the error terms, which appear in table 2.2.1. Thus y_i is a random variable, with $E(y_i) = \beta_0 + \beta_1 x_i$ and $var(y_i) = \sigma^2$ and $cov(y_i, y_j) = 0$ for all $i \neq j$.

Fitting the regression model entails suitable parameterization—assigning to the regression coefficients β_0 and β_1 values that lead to a "good" approximation of the observed

$E(\epsilon_i) = 0$ for $i = 1, 2, ..., n$
$Var(\epsilon_i) = \sigma^2$ for $i = 1, 2, ..., n$
$Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

TABLE 2.2.1. Assumptions on Regression Error Terms

data. To parameterize the model, linear regression employs the method of least-squares **[109, 116, 79, 52**, ch. 9]. Thus we choose those regression coefficients that minimize the squared error between the observed data, and the predictions at each observation x_i. For each x_i, we define a fitted value for the response, $\widehat{y}_i = \beta_0 + \beta_1 x_i$. Let Equation 2.2.2 define the sum of squared errors (SSE) under a given parameterization of our model:

$$SSE = \sum (y_i - \widehat{y}_i)^2. \tag{2.2.2}$$

To find the least-squares estimate of the regression parameters, we choose those values that give $min(SSE)$. By definition, the model parameters that minimize SSE, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (i.e. those that minimize the so-called *residual deviance* of the model), are the optimal estimates in the least-squares sense. Minimizing the SSE is feasible using numeric or analytical methods, and interested readers are refered to **[50, 67, 79, 109]** for a complete discussion of this matter.

The method of least squares is attractive because, under the previously stated assumptions about the regression model, the fitted values that it derives comprise the best linear unbiased estimates of the true regression coefficients. In other words, $E(\widehat{\beta}_0) = \beta_0$, $E(\widehat{\beta}_1) = \beta_1$, and $\sigma^2(\widehat{\beta}_0)$ and $\sigma^2(\widehat{\beta}_1)$ are the lowest variances of all possible linear estimators.

In many cases, however, a system contains not one predictor variable, but, $p > 1$ variables. This will be the case I describe the relation between regression and LSI. The simple regression model is easily generalized to the p-variate case. To describe the multiple regression procedure, we use the following matrix notation. Let $\mathbf{X}$ be the $n \times p$ matrix of data observations. Let $\mathbf{y}$ be the n-vector of responses on $\mathbf{X}$, such that the i^{th} element of $\mathbf{y}$ is the response observed for the i^{th} row of $\mathbf{X}$, the p-vector $\mathbf{x}_i'$. Our regression parameters form the p-vector $\boldsymbol{\beta}$, and the error terms are defined by the n-dimensional vector $\boldsymbol{\epsilon}$. Given this

notation, the regression model may be written as Equation 2.2.3:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.2.3}$$

This model is approximated empirically by selecting optimal regression coefficients, yielding Equation 2.2.4:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{2.2.4}$$

As shown in **[79, 109, 116]** the least-squares estimate of the p regression coefficients is derivable via Equation 2.2.5:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.2.5}$$

Matrix $\mathbf{X}'\mathbf{X}$ describes the covariance among the data, and $\mathbf{X}'\mathbf{y}$ is used to obtain the covariance between the dependent and independent variables[2].

Having developed a statistical model it is often of interest to quantify its descriptive power. Although a variety of measures aid in this regard **[16]** perhaps the simplest is the coefficient of multiple determination, or R^2, defined by Equation 2.2.6:

$$R^2 = \frac{\widehat{\boldsymbol{\beta}}\mathbf{x}'\mathbf{y} - n\overline{y}^2}{\mathbf{y}'\mathbf{y} - n\overline{y}^2}. \tag{2.2.6}$$

Equation 2.2.6 gives the proportion of the variance observed in the data that is captured by the model. The numerator measures the squared deviation of the model's fitted values from the sample mean $\overline{y}$, while the denominator measures the squared deviation of the observed y's from $\overline{y}$. The statistic R^2 thus measures the percent of the sample variance described by a given model.

A statistical model provides an approximation of a system. Linear regression uses the matrix of predictors, $\mathbf{X}$, to approximate the vector of responses, $\mathbf{y}$. The R^2 statistic measures the fidelity of this approximation. Linear models thus allow us to generalize from observation. This is precisely the goal of LSI, where we assume that authors' word choices

[2]This Equation assumes that matrix $\mathbf{X}$ is non-singular, a condition that holds, provided that $n > p+1$ and that $\mathbf{X}$ is of full rank.

dog	*canine*	*bark*	*paw*	*cat*	**dog**	**cat**	**misc.**
1	1	1	1	0	1	0	0
1	1	1	0	0	0.7	0.2	0.1
1	0	0	0	0	0.6	0.3	0.1
0	0	0	0	1	0.3	0.5	0.2
0	0	0	0	0	0.3	0.7	0
0	1	0	0	1	0.1	0.7	0.2
0	0	0	1	1	0	1	0
0	0	1	0	0	0	0.2	0.8

TABLE 2.2.2. A pre-classified document collection

are partially random and partially governed by a latent semantic structure native to the strictures of language. LSI employs a linear model to approximate this structure. At the risk of oversimplification, we may characterize LSI as a process analogous to multiple linear regression. As I shall show in another section, this analogy is more than casually apt[3]. At this point, however, I demonstrate only the conceptual relation between these two methods, using a simplified, but hopefully illustrative example.

Let **A** be the term-document matrix defined by the first five columns of Table 2.2.2. For the sake of illustration, imagine that documents in this collection may only be "about" three topics—**dog**, **cat**, and **miscellanea**; this is a small semantic universe. Stretching the fiction a bit further, imagine that an omniscient indexer has classified each document d_i in **A**, assigning to it three scores: D_i, C_i, and M_i, which describe the degree to which d_i is about each topic, **dog**, **cat**, and **miscellanea**, respectively. These scores are shown in the three rightmost columns of table 2.2.2.

Under the standard vector space model, we represent documents in the term space of matrix **A**. However the scenario at hand gives us extra information to work with: the document classifications. To improve the representation of this information space, we may use this information to construct a statistical model of the relation between terms and topics. Instead of representing documents in term space, we construct three models and project documents into the 3-space that they define. To predict a document d_j's score on

[3]Sections 2.2.2 and 2.2.3 demonstrate that LSI is, like linear regression, a linear statistical model of the input data.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
dog	1.57	1.39	1.30	0.09	0.37	0.37	0.28	0.19
cat	0.30	-0..01	0.57	1.10	1.14	1.25	1.41	0.40
misc.	-0.08	0.79	-0.10	-0.95	0.41	0.21	0.08	2.05

TABLE 2.2.3. Fitted values for linear topic model

$topic_k$, we construct the linear regression model shown in Equation 2.2.7:

$$topic_{kj} = \beta_0 + \beta_1 dog_j + \beta_2 canine_j + \beta_3 bark_j + \beta_4 paw_j + \beta_5 cat_j + \epsilon_i. \tag{2.2.7}$$

Thus regression coefficient β_{ki} describes the relation between the i^{th} term and the k^{th} topic. A large positive β_{ki} implies that term i is highly correlated with topic k, while a negative score implies that the i^{th} term rarely appears in documents that are about topic k.

The fitted values derived by applying the method of least-squares to our example data appear in table 2.2.3. Let $\mathbf{A}_3$ be the 3×8 topic-document matrix defined by table 2.2.3. Each column vector $\mathbf{d}_i$ of $\mathbf{A}_3$ represents a document in our new model space. This modeling procedure amounts to a projection of our 5-dimensional vector space onto a newly devised 3-space that we may represent visually, as in Figure 2.2.3, where the documents appear in the 2-space defined by the **dog** and **cat** models. Figure 2.2.3 shows a nicely segmented space, with documents mostly about **dogs** close together, yet distanced from documents about **cats**. The one document that is mostly about **miscellanea** (d_8) appears in relative isolation.

To project a new document or query, q, into this model space, we simply generate a predicted value for each topic model based on the query's term vector $\mathbf{q}$. Figure 2.2.4 shows the location in 2-space of a vector for a query containing the term *canine*. Note this query's proximity to the other documents that are mostly about **dogs**.

Projecting documents into model space amounts to changing the representational axes of the system. This projection removes error from the VSM representation by accounting for the correlational structure of the input data. Thanks to the omniscient indexer we know the true dimensionality of this information space; there are only three topics in this abbreviated semantic universe. Under the standard vector space model, we assume a one-to-one relationship between observed indexing features and topics of potential interest (hence the

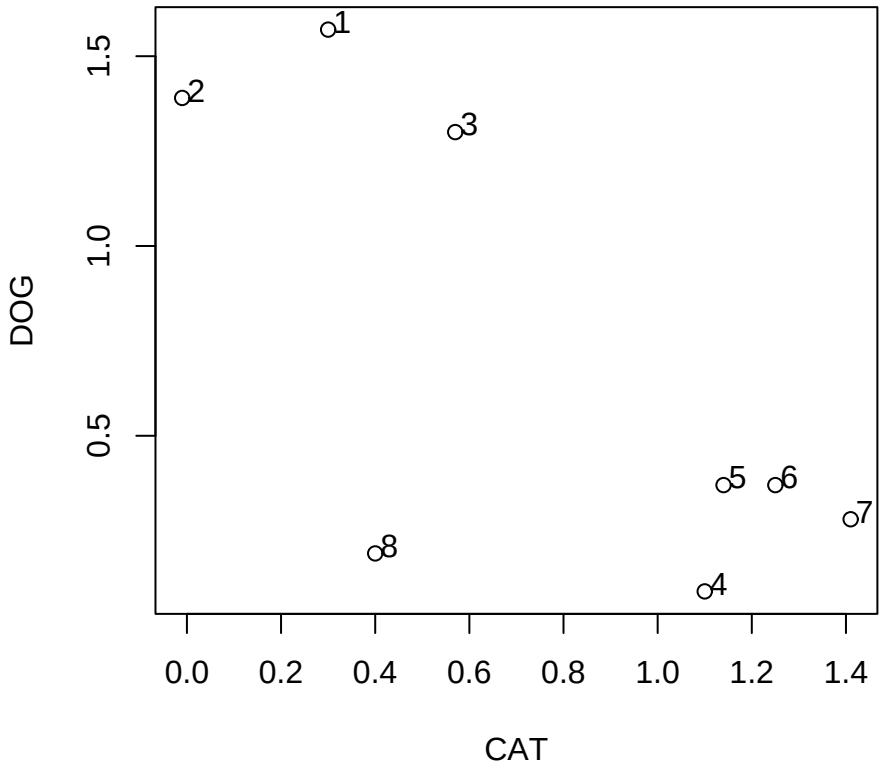

FIGURE 2.2.3. Documents in derived 2-D topic space

need to choose good indexing features). However, a casual inspection of our data reveals redundancy among the observed terms. For instance, *dog, canine,* and *bark* all refer to aspects of **dog**-ness, while *paws* appear on both dogs and cats. Our one-to-one correspondence of word to topic seems wrong. Thus constructing a new space based on a statistical model of feature correlations is appealing

As a final example of the utility of linear regression to augment the vector space approach to IR, imagine that our previously infallible indexer admits that he might have been wrong in his assignment of topics to documents. In particular, he has second-guessed himself and is no longer sure that the **miscellaneous** topic should exist in this little universe. A more hard-nosed indexer, he admits, would only recognize the **dog** and **cat** topics. **Miscellanea**

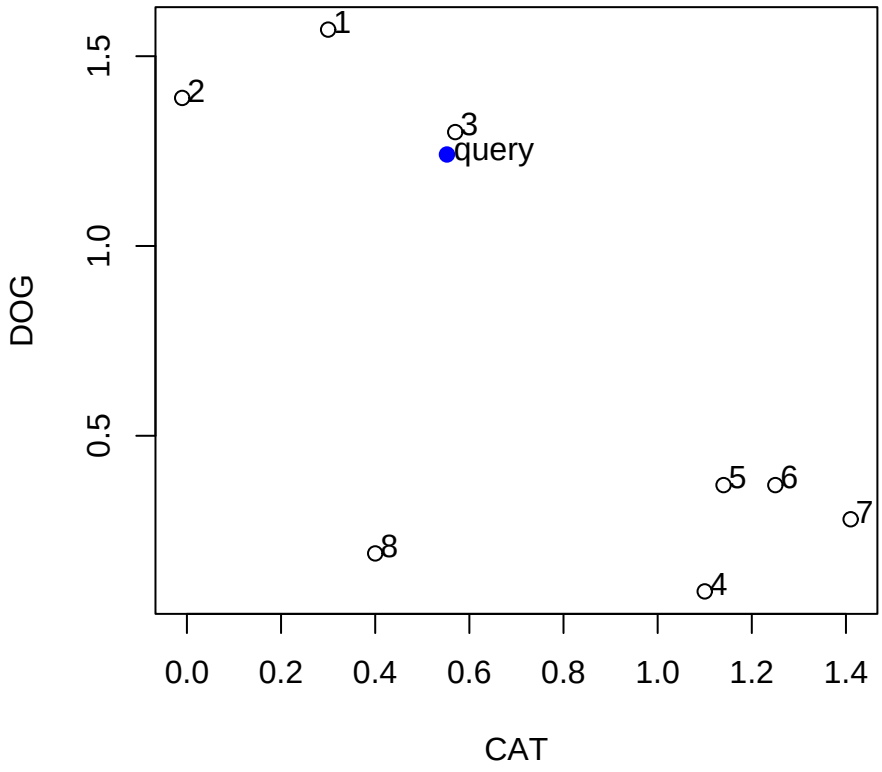

FIGURE 2.2.4. A query in 2-D topic space

was just a catch-all category he invented because document 8 didn't seem to fit anywhere else.

Not eager to be fooled twice, we want to decide for ourselves how viable the **miscellaneous** topic really is. Table 2.2.4 shows one means of making such a decision. The table lists the adjusted R^2 for each of our three models. This metric shows the proportion of variance in the training data that is described by a given model, also taking into account the number of degrees of freedom used by the model. A high adjusted R^2 implies that a given model is likely to be useful for descriptive purposes. The models derived for the **dog** and **cat** topics score nearly identical adjusted R^2 coefficients, while the fit of the **miscellaneous** model is approximately half of the other two. This suggests that we are not gaining much

Topic	*Adjusted* R^2
Dog	0.43
Cat	0.42
Misc.	0.23

TABLE 2.2.4. Adjusted R^2 for linear topic models

descriptive power from our **miscellaneous** model, and that its role in inter-document similarity calculations is suspect. Using linear models for document representation allows us to exploit such variance-based diagnostic measures. By projecting our 5-dimensional document representation onto a 2- or 3-dimensional topic space, we effect a dimensionality reduction in our system, thus omitting error introduced by treating redundant vectors as though they were orthogonal. Using statistical model selection techniques gives us a rationale for rejecting dimensions. Given a faulty indexer (or no indexer at all, as we shall see), we retain those models whose descriptive power accounts for a "large" portion of the observed variance among our data (I leave the term "large" undefined for now).

2.2.2. Principal Component Analysis and Matrix Approximation. In the previous section I discussed how linear models can be used to improve the similarity model of a VSM-based IR system. However, that discussion assumed the availability of an omniscient indexer. In particular, the method of linear regression depends on the presence of dependent variables in order to define its notion of goodness of fit. Thus in the previous example, we discovered our model parameters by minimizing the squared vertical distance between the vector of observed document scores on a given topic t_k and their predicted values $\widehat{t_k}$ under a given linear combination of term weights m_i. Thus we needed t_k to serve as a dependent variable in order for the model to be defined. In *ad hoc* IR applications, however, we usually lack authoritative topical classifications of documents.

The method of principal components analysis (PCA) allows us to extend the notion of linear modeling to data that lack dependent variables. This is precisely how LSI operates, and thus I offer the following extension of my earlier example to initiate of our discussion of the mathematics of LSI.

	dog	*canine*	*bark*	*paw*	*cat*
dog	0.63	0.29	0.29	0.10	-0.38
canine	0.29	0.63	0.29	0.10	-0.04
bark	0.29	0.29	0.63	0.10	-0.38
paw	0.10	0.10	0.10	0.75	0.10
cat	-0.38	-0.04	-0.38	0.10	0.63

TABLE 2.2.5. Sample covariance matrix

Group	*Terms (in order of importance)*
1	*dog, canine, bark*
2	*cat*
3	*paw*

TABLE 2.2.6. Putative term associations

I begin discussion of PCA informally, returning to the term-document matrix **A** defined by the first five columns of table 2.2.2. For the present discussion I assume that the omniscient indexer has been fired, and thus we no longer have access to the previously discussed topical categories; for all we know, each of the five terms corresponds to a unique topic. Given this state of affairs, it is tempting to revert to the standard vector space model. However, we can do better than this.

Table 2.2.5 shows the covariance matrix for this document collection. By scanning the first column of the table, we see that the terms *canine* and *bark* are positively correlated with *dog,* which is in turn negatively correlated with *cat.* It seems as if *dog, canine,* and *bark* have something in common. Moreover, the bloc comprised by these terms seems quite distinct from *cat.* Finally, the term *paw* gives us little information about its habits of association; it is equally likely to appear with *dog* or *cat.*

This cursory analysis suggests that there are roughly three patterns of term co-occurrence in our data. These patterns are summarized in Table 2.2.6. If we are willing to admit that these groupings are due to something more meaningful than chance, we are on our way to defining an *ad hoc* version of our previous model of topic space. Instead of the response variables **dog**, **cat**, and **miscellaneous**, we now have simply "factors" 1, 2, and 3. As initial guesses at the weights of such a model of topic space we could import the scores from our covariance matrix, thus predicting a given document d_i's score on factor 1 via Equation

Document	*Factor 1 Score*
1	1.85
2	1.85
3	0.96
4	-0.58
5	0.00
6	-0.14
7	-0.58
8	0.44

TABLE 2.2.7. Documents in an *ad hoc* 1-space

2.2.8.

$$factor_{i1} = 0.63(dog) + 0.29(canine) + 0.29(bark) - 0.38(cat) \tag{2.2.8}$$

Table 2.2.7 shows our eight documents projected into the one-dimensional topic space defined by Equation 2.2.8. Equation 2.2.8 gives a linear combination of the original variables. Due to its reliance on the correlational structure of the data, we suspect that this linear combination is in a sense a natural axis of the data. In other words, by analyzing the corpus' term-term covariance we have derived a new variable, a factor defined by Equation 2.2.8. Despite its highly provisional nature, this one-dimensional representation collocates the "dog" documents (documents 1, 2, and 3), while keeping them at some remove from the "cat" documents (documents 4, 5, 6, and 7). Finally, the "miscellaneous" document 8, appears in between the other two sets.

By analyzing the term-term covariance in the observed data, we have discovered several putative axes along which we may represent the documents. Using this *ad hoc* method we discovered linear combinations of terms that comprised a new set of variables that represents the variance among our data more concisely than the observed variables (terms) did. Principal component analysis uses the method of least squares to effect the same result in a well-defined fashion:

> In principal component analysis, we seek to maximize the variance of a linear combination of the variables....In regression, we have linear combinations of the independent variables that best predict the dependent

> variable(s)....Principal components, on the other hand, are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent....In seeking a linear combination with maximal variance, we are essentially searching for a dimension along which the observations are maximally separated or spread out. **[116]**

Applying principal component analysis generates a matrix $\mathbf{D}$ that defines a rotation of $\mathbf{A}$ onto independent axes.

Deriving the principal components of the $n \times p$ matrix $\mathbf{A}$, where $rank(\mathbf{A}) = p$, is an example of the eigenvalue-eigenvector problem **[79, 141]**. Given a square matrix $\mathbf{M}$ of rank r, it can be proved **[141, 148]** that r scalars comprising the vector $\boldsymbol{\lambda}$ and and r non-zero n-vectors that comprise the matrix $\mathbf{v}$ exist that satisfy Equation 2.2.9.

$$\mathbf{M}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v} \tag{2.2.9}$$

The elements of $\boldsymbol{\lambda}$ are the eigenvalues of $\mathbf{M}$, and the r columns of matrix $\mathbf{v}$ contain its eigenvectors.

Finding $\mathbf{z}_k$, the k^{th} principal component of the $n \times p$ matrix $\mathbf{A}$ (assuming that the data have been centered around their means) involves an orthogonal projection via equation 2.2.10.

$$\mathbf{z}_k = \mathbf{D}\mathbf{a}_k \tag{2.2.10}$$

where $\mathbf{a}_k$ is the k^{th} column of $\mathbf{A}$ and $\mathbf{D}$ is an orthogonal matrix, such that $\mathbf{D}'\mathbf{D} = \mathbf{I}_n$ and $\mathbf{d}_i \cdot \mathbf{d}_j = 0$ for all $i \neq j$ **[116]**[4]. We thus seek the orthogonal matrix $\mathbf{D}$ such that $\mathbf{z} = \mathbf{D}\mathbf{A}$ yields the matrix $\mathbf{z}$ whose columns contain the p uncorrelated variables such that each variable p_i captures maximal variance among all variables orthogonal to each p_j, $i \neq j$. In other words, we seek a transformation of $\mathbf{A}$ that yields $\mathbf{z}$ with a diagonal covariance matrix $\mathbf{S}_z$.

[4]For a discussion of orthogonal projections and the properties of orthogonal matrices, see **[141]**.

We diagonalize the covariance matrix $\mathbf{S}$ by use of the spectral decomposition (cf. [**116**, ch. 2] and [**141**]). If $\mathbf{S}$ is the $p \times p$ covariance matrix of $\mathbf{A}$ and $\mathbf{z} = \mathbf{DA}$, then $\mathbf{S}_z = \mathbf{DSD'}$ is diagonal, as shown in Equation 2.2.11.

$$\mathbf{S}_z = \mathbf{DSD'} = \begin{pmatrix} s_{z1}^2 & 0 & \dots & 0 \\ 0 & s_{z2}^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_{z3}^2 \end{pmatrix}. \tag{2.2.11}$$

By definition of the spectral decomposition $\mathbf{D}$ contains the normalized eigenvectors of $\mathbf{S}$ and $\mathbf{S}_z$ is diagonal, containing the eigenvalues of $\mathbf{S}$. Thus the eigenvalues of the covariance matrix provide the desired orthogonal projection matrix $\mathbf{D}$. The principal components are derived by projecting $\mathbf{A}$ into the orthogonal space spanned by its eigenvectors: $z_1 = \mathbf{d}_1'\mathbf{A}$, $z_2 = \mathbf{d}_2'\mathbf{A}$, ..., $z_p = \mathbf{d}_p'\mathbf{A}$.

According to Rencher, "since the [PCA] rotation lines up with the natural extensions of the swarm of points, z_1... has the largest (sample) variance and z_p ... has the smallest sample variance" [**116**]. Thus by retaining only the first $k < p$ principal components we achieve the best rank-k approximation of the covariance matrix, in the least squares sense. Let $\widehat{\mathbf{S}}_k = \mathbf{D}_k\mathbf{\Sigma}_k\mathbf{D}_k'$, where $\mathbf{D}_k$ contains the first k eigenvectors of $\mathbf{S}$, and $\mathbf{\Sigma}_k$ is diagonal, containing the first k eigenvalues arranged in descending order of magnitude. Given this definition $\widehat{\mathbf{S}}_k$ is the k-dimensional least-squares approximation of the sample covariance matrix.

Such an approximation is useful insofar as it addresses the difference between the population and sample. "If the variables [of a p-dimensional matrix] are highly correlated," writes Rencher, "the essential dimensionality is much smaller than p; that is, the first few eigenvalues will be large...[thus they will account for a large portion of the sample variance]. On the other hand, if the correlations among the variables are small, the dimensionality is close to p and the eigenvalues will be nearly equal. In this case, no useful reduction in dimension is achieved, because the principal components essentially duplicate the variables" [**116**]. This point is reinforced in [**79, 3, 35**]. Dimensionality reduction is merited insofar

Document	*Score on first PC*
1	-1.89
2	-1.63
3	-0.18
4	1.37
5	0.65
6	0.75
7	1.11
8	-0.18

TABLE 2.2.8. Example documents along their first principal component

PC	*Cumulative Percent of Variance*
1	0.458
2	0.714
3	0.870
4	0.973
5	1.000

TABLE 2.2.9. Variance of example principal components

as our sample contains random error. Although we have observed p dimensions, the multivariate PDF that generated the data has $k \leq p$ dimensions. Thus by projecting our data onto the first k eigenvectors, we are implicitly stating that the last $p - k$ dimensions are the product of sampling error. If this is true then the reduced rank model $\widehat{\mathbf{S}}_k$ yields a model of the population covariance matrix that is superior to the sample covariance matrix $\mathbf{S}$.

Table 2.2.8 shows the scores of the documents from our earlier example (cf .table 2.2.2) on their first principal component. If we compare table 2.2.8 with our *ad hoc* linear combination shown in table 2.2.8 we see a high degree of correspondence. Like our *ad hoc* model, the first principal component groups documents about **dogs** together, while isolating them from documents about **cats**, as is evident in Figure 2.2.5, which represents the documents from Table 2.2.2 along their first two principal components. And as we can see from Table 2.2.9, the first two or three principal components account for the lion's share of the variance in the data. A two-dimensional representation of these data accounts for over 70% of the variance observed on five variables. By adding a third principal component, our model captures 87% of the variation.

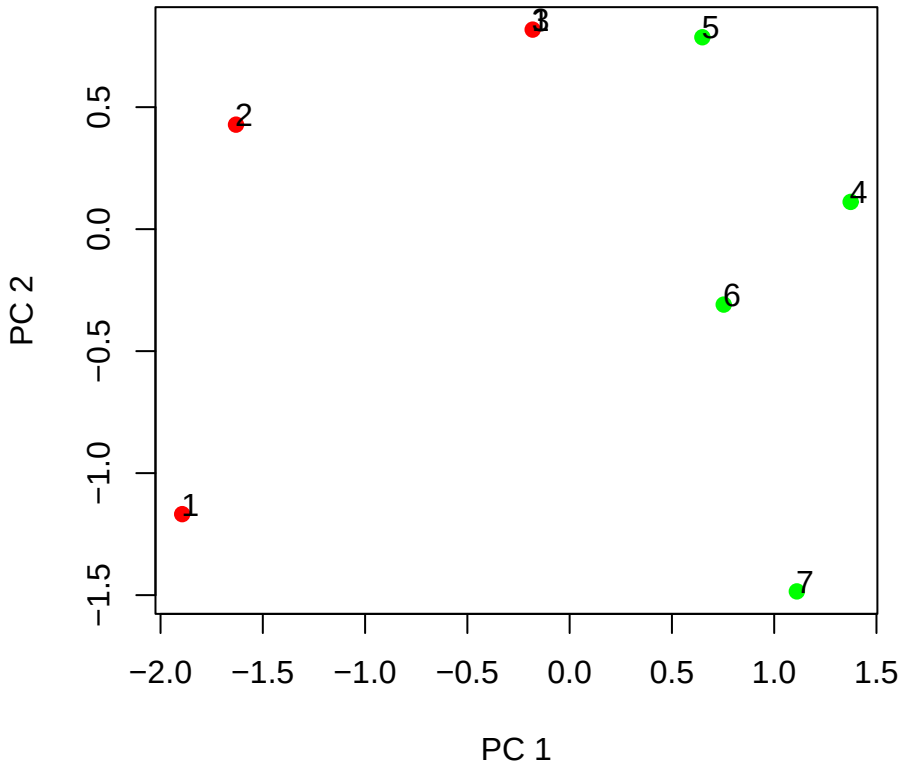

FIGURE 2.2.5. Example data in principal component space

The motivation behind PCA's dimensionality reduction stems from the belief that the observed covariance matrix contains artifacts of sampling error. In our ongoing example we observed five variables (terms). However, the covariance among these terms suggests that the PDF that generated the data contains $k < 5$ non-zero eigenvalues. Given unlimited data from this distribution, we expect that one or more eigenvalues would converge on zero. Due to inter-term correlation, then, representing these data in 5-space constitutes error in the vector space model. Projecting the data onto the first $k < 5$ principal components thus implies a model of the population covariance matrix that is based on, but not identical to the sample covariance matrix. Insofar as the intrinsic dimensionality of these data is less than their observed rank, this reduced model thus constitutes a representational improvement.

2.2.3. The Singular Value Decomposition. Whereas principal component analysis uses the data's correlational structure to approximate a square matrix, latent semantic indexing derives a low-rank approximation of the $n \times p$ matrix $\mathbf{A}$. To affect its dimensionality reduction, LSI uses the singular value decomposition (SVD), a standard least-squares matrix factorization method from linear algebra [**58, 52, 141, 10**].

Let $\mathbf{A}$ be an $n \times p$ matrix of rank r. The singular value decomposition of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{T\Sigma D'} \tag{2.2.12}$$

where $\mathbf{T}$ and $\mathbf{D}$ are orthogonal matrices. $\mathbf{T}$ is $n \times r$, with columns t_i containing the *left singular vectors* of $\mathbf{A}$. $\mathbf{D}$ is an $r \times r$ matrix with columns d_i called the *right singular vectors* of $\mathbf{A}$. Finally matrix $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix, with diagonal elements $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r \geq 0$ called the *singular values* [**32, 10, 67, 102**].

It is worth noting the similarity between PCA and SVD. In the previous section we derived the principal components of the square matrix $\mathbf{S} = \mathbf{A'A}$, the term-term co-occurrence matrix in the case of vector space IR. We could just as well have defined it in terms of $\mathbf{S} = \mathbf{AA'}$, the document similarity matrix. The singular value decomposition yields both solutions simultaneously [**67**, p. 487]. Assuming centered, normalized data, the columns of $\mathbf{T\Sigma}$ are the principal components of $\mathbf{A'A}$, while $\mathbf{D\Sigma}$ gives the principal components of $\mathbf{AA'}$. Not surprisingly, then, the singular values of $\mathbf{A}$ are the positive square roots of the eigenvalues of $\mathbf{A'A}$ and $\mathbf{AA'}$.

In the case of IR, matrix $\mathbf{A}$ is a term-document matrix. Thus SVD yields a number of desirable properties. The factorization derives the r axes of the data that capture maximum variance while maintaining mutual orthogonality. Thus each singular vector comprises an artificial variable, or factor; like PCA, SVD derives—by the method of least-squares—an orthogonal factor space. The left singular vectors serve as a projection matrix for the rows of matrix $\mathbf{A}$, while the right singular vectors project the documents onto the subspace. In the standard vector space model, documents reside in term-space, while terms reside in document-space. Using the SVD-derived projection matrices, LSI represents both terms

and documents in the same factor space. This permits us to define not only document-document similarity, but also term-document similarity as the inner product between two vectors [**32**]. It also enables us to project new documents (or queries) into the derived factor space by means of a simple matrix multiplication [**10**]. Finally, by virtue of its relation to the eigensystem of $\mathbf{A}$, the matrix of singular values $\mathbf{\Sigma}$ acts as an indication of the amount of variance described by each factor k in the derived factor space [**79**]. As I discuss below, this property of the SVD is useful when selecting singular vectors to retain during dimensionality reduction.

According to Deerwester *et al.,* because most entries of $\mathbf{\Sigma}$ are very small (cf. [**73, 77**]), we may omit them and their corresponding singular vectors from further analysis, "leading to an approximate model that contains many fewer dimensions. In this reduced model all the term-term, document-document and term-document similarities are now approximated by values on this smaller number of dimensions" [**32**, p. 395]. As in the case of PCA, we thus use SVD to derive a least-squares approximation of $\mathbf{A}$:

$$\widehat{\mathbf{A}}_k = \mathbf{T}_k \mathbf{\Sigma}_k \mathbf{D}'_k \tag{2.2.13}$$

where $\mathbf{T}_k$ contains the first k columns of $\mathbf{T}$, $\mathbf{\Sigma}_k$ contains the first k rows and columns of $\mathbf{\Sigma}$, and $\mathbf{D}_k$ has the first k columns of $\mathbf{D}$. Under this reduced-rank model, we define document-document similarity as before in Equation 2.1.2. The similarity between two document vectors $\widehat{\mathbf{d}}_i$ and $\widehat{\mathbf{d}}_j$ is simply the inner product between the i^{th} and j^{th} rows of $\widehat{\mathbf{D}}_k = \mathbf{D}_k \mathbf{\Sigma}_k^2$.

A new document, or pseudo-document such as an *ad hoc* query $\mathbf{q}$, may be projected into the factor space via Equation 2.2.14[**10**]:

$$\widehat{\mathbf{q}}_k = \mathbf{q}' \mathbf{T}_k \mathbf{\Sigma}_k. \tag{2.2.14}$$

Having defined the projection of a query into factor space, we may then calculate the similarity between the query and each document in the corpus by applying Equation 2.2.14 to find $sim(\widehat{\mathbf{q}}_k, \widehat{\mathbf{d}}_{ik})$, where $\widehat{\mathbf{d}}_{ik}$ is the i^{th} row of $\widehat{\mathbf{D}}_k$.

Assuming that $k \ll r$, $\widehat{\mathbf{A}}_k$ will provide a loose approximation of the observed relationships between terms and documents. Proponents of LSI argue that by virtue of the randomness inherent in natural language, such an approximation yields a more robust similarity model than is native to the standard vector space approach to IR, where similarity judgements are prone to error due to the overdetermined nature of a given corpus [**10, 32, 90**]. The matrix approximation used in LSI affords a least-squares model of the linguistic system that generated the term-document matrix $\mathbf{A}$. Such a model augments standard vector space retrieval to include an analysis of the correlational structure of the input data. In this way it is an extension of Wong's generalized vector space model [**77, 150**] (cf. 1.1.2). Insofar as this model is well built and applied to appropriate data (the subject of our next section), such an approach improves similarity judgements by supplementing the VSM with the best model of the population term correlation matrix, in the least-squares sense.

Before turning to a discussion of building the LSI model, I conclude this introduction to linear modeling for IR with a return to our ongoing numerical example. Let matrix $\mathbf{A}$ be the 5×8 term-document matrix defined by taking the transpose of the matrix analyzed in our previous example regarding PCA. The singular value decomposition of this matrix is given in Figure 2.2.6. Figure 2.2.7 shows the projection of these terms and documents into the 2-space spanned by the first singular vectors of the SVD. As in the case for PCA, the SVD-derived factor space arranges the documents in a way amenable to semantic discrimination. We can easily construct a linear discriminator to distinguish **dog** documents from **cat** documents. For instance, I applied the k-means clustering algorithm to the 2-dimensional matrix derived by retaining the first two singular triplets of $\mathbf{A}$. The k-means procedure grouped together documents 1, 2, and 3 (those documents that seem mostly to be about *dogs*). It also isolated documents 4, 5, 6, and 7 (our putative *cat* documents). Admittedly the k-means approach does include document 8, our dark horse document, with the **dog** set. However, this is not surprising, as document 3 and document 8 map onto the same location in 2-space. That this occurs is understandable if we consider that document vector 3 and document vector 8 each have a non-zero entry for only a single term. Although each vector's non-zero entry is for a different term, they are each for terms that occur only in **dog** documents. Thus, the

$$\mathbf{\Sigma} = \begin{pmatrix} 4.35 & 0 & 0 & 0 & 0 \\ 0 & 2.88 & 0 & 0 & 0 \\ 0 & 0 & 1.99 & 0 & 0 \\ 0 & 0 & 0 & 1.53 & 0 \\ 0 & 0 & 0 & 0 & 1.19 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} -0.50 & 0.32 & -0.01 & -.71 & -0.38 \\ -0.53 & 0.00 & 0.50 & 0.00 & 0.68 \\ -0.50 & 0.32 & -0.01 & -0.71 & -0.38 \\ -0.42 & -0.38 & -0.79 & 0.00 & 0.26 \\ -0.20 & -0.81 & 0.36 & 0.00 & -0.42 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} -0.72 & 0.09 & -0.37 & 0.00 & 0.30 \\ -0.54 & 0.34 & 0.37 & 0.00 & -0.10 \\ -0.18 & 0.17 & -0.01 & 0.71 & -0.49 \\ -0.07 & -0.43 & 0.27 & 0.00 & -0.53 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.26 & -0.43 & 0.66 & 0.00 & 0.34 \\ -0.25 & -0.67 & -0.47 & 0.00 & -0.13 \\ -0.18 & 0.17 & -0.01 & -0.71 & -0.49 \end{pmatrix}$$

FIGURE 2.2.6. SVD of example term-document matrix

SVD model appears to have derived something akin to **dog**-ness as one of its initial factors; documents 3 and 8 both contain the same amount and quality of evidence for this factor. If we—perhaps too easily—interpret the second factor as **cat**-ness, document's 3 and 8 again map to the same location.

Several things are worth noting about this example. The low-dimensional model seems to do a good job in isolating classes of terms and documents that have similar correlational patterns. Thus the model infers that *bark* provides about the same information as does *dog*; queries about *barking* would map to the semantic region devoted to **dog**-ness, without needing to contain the word *dog*. A similar, but weaker pattern emerges between *dog* and *canine*. The ability of LSI to identify such topical clusters of terms and documents is among its strong suits.

However, the example also points out a problem inherent in dimensionality reduction methods. In our derived 2-space, documents 3 and 8 map to the same location, despite the fact that they contain distinct term vectors in the original data. By analyzing matrix **D** in Figure 2.2.6 we can see the cause of this problem. The vectors for documents

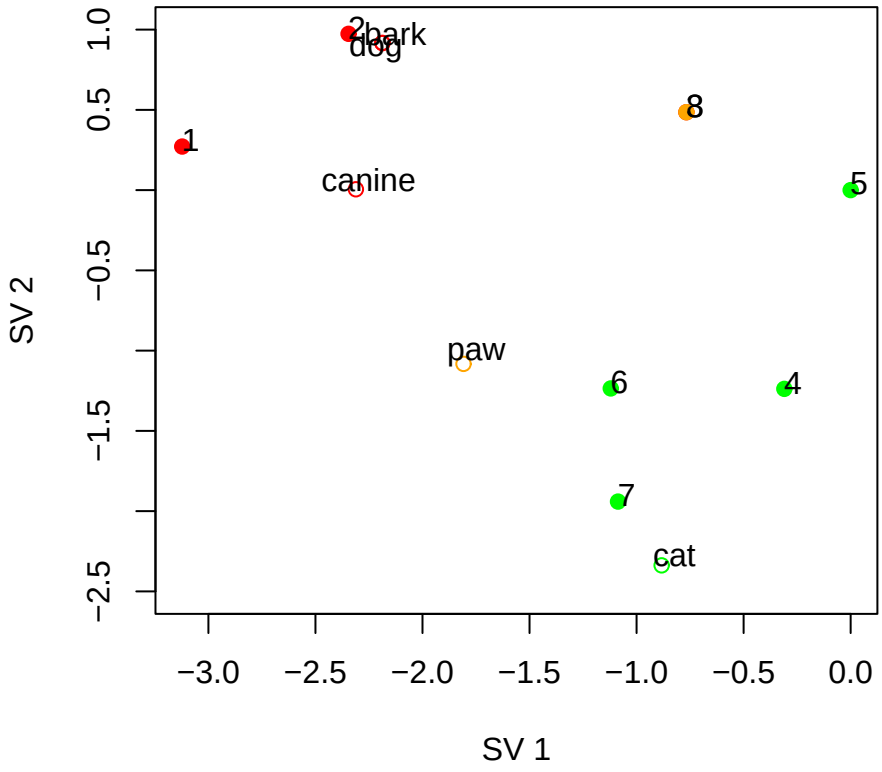

FIGURE 2.2.7. Terms and documents in SVD-derived 2-space

3 and 8 are indeed identical in matrix **D**, except for their scores on the fourth singular vector. By truncating our representation to a two-dimensional model, we have thus elided the information that allows documents 3 and 8 to remain distinct. By omitting the third, fourth, and fifth, singular vectors, we assumed that we were removing erroneous data from our model. By and large this may be true, as the 2-dimensional representation otherwise seems to capture the correlational structure of the data well. However, we have, in a sense, allowed the model to infer too much. By limiting our document representation to the first two singular vectors, we have thrown useful information out of the system. As Deerwester *et al.* note, "the representation of a conceptual space for any large document collection will require more than a handful of underlying independent 'concepts' ... [thus] the amount of

dimension reduction, the choice of k, is crucial to [LSI]." Choosing a dimensionality that manifests the correlational structure of the population from which a data sample is drawn is an open problem in the LSI literature.

2.3. Discovering the Optimal Dimensionality

The goal of LSI is to improve retrieval by projecting the term-document matrix $\mathbf{A}$ of rank r onto an orthogonal subspace of rank k where $k \ll r$. According to LSI's proponents, the resultant matrix approximation, $\widehat{\mathbf{A}}_k$, represents the semantics of the corpus more faithfully than does the putatively overspecified observed matrix. However, k, the dimensionality of the subspace is a poorly understood free parameter in applications of LSI. Arguments in favor of LSI typically suggest that dimensionality reduction removes "noise" from the representation of term-document relationships. However, the source and characteristics of this noise are difficult to identify. And unlike in classic regression analysis, even the distribution of such error is obscure [**73, 37, 102**]. Without a clear sense of what the singular values and singular vectors represent, discarding the last $r - k$ of them seems risky at best.

This section treats the details of dimensionality reduction under LSI. Earlier I articulated the motivation behind dimensionality reduction. Here I discuss a more detailed problem: the choice of k—how this choice impacts LSI performance and techniques for identifying a suitable dimensionality for a given corpus. First I review what is at stake in the selection of k, showing that the choice of an appropriate dimensionality bears heavily on LSI performance. I then turn to a discussion of several efforts to put the singular value truncation on firm theoretical ground. Finally I discuss several methods of dimensionality estimation that arise in the literature of principal component analysis. As mentioned earlier, PCA is closely related to LSI, and I argue that its attendant techniques for selecting an appropriate representational dimensionality should be of interest to researchers in IR.

2.3.1. Optimal k—Selecting an Appropriate Dimensionality for LSI. Since Deerwester *et al.* proposed the idea of LSI, researchers have noted that properly parameterizing k, the representational dimensionality of an LSI system, is a vital part of so-called latent

semantic analysis. Deerwester and his co-authors call this parameter "crucial" to successful application of LSI [**32**], noting a 30% improvement in average precision as they changed k from 1 to 100 on the Medline dataset[5]. It is not surprising that setting $k = 2$ deprives a model of important descriptive power to perform robust retrieval. And in light of our earlier discussion about statistical models, it is not surprising that a relatively low number of factors, $k = 100$, yields good performance.

What is not obvious is the fact that setting too high a value for k may also lead to decreased retrieval quality. Describing their analysis of SVD applications for language learning, Landauer and Dumais write, "using too many factors [for LSI representations] also resulted in very poor performance" [**90**]. They note that setting $k = 1$ leads to accuracy slightly below 16% on a synonym learning test. In the region of $k \approx 300$, Landauer and Dumais report accuracy above 50%. However, as they increase k, letting it approach the full dimensionality of their corpus, accuracy dips back to the 15% level. Landauer and Dumais test the validity of this "strong nonmonotonic relation between [the] number of LSA dimensions and [the] accuracy of simulation," by recourse to a statistical hypothesis test, noting a p-value below 0.0002. Their work formalizes a pattern that emerges often applied LSI research: for many collections there exists a region of optimal dimensionality less than the rank of the data. Setting k below this region deprives the system of important descriptive power, while setting a value of k that is too high appears to overfit the model, causing it to learn spurious term-document relations, and thus impeding its predictive ability.

Other researchers have discovered the same phenomenon. For instance, Ding notes that adding factors to an LSI model quickly improves performance until a certain threshold is reached. After this region of optimality, performance decreases steadily as one adds more singular vectors [**36, 37**]. Likewise, Manning and Schütze mention a region of optimality with regard to parameterizing k in LSI systems. These observations suggest that observing the performance of an LSI system at various parameterizations of k gives us information about the intrinsic dimensionality of a corpus. As I discuss in Section 3.2, I denote the value of k that optimizes observed LSI performance with respect to a given metric *observed*

[5] I discuss precision and other IR performance metrics below.

optimal dimensionality, or *observed* k_{opt}. Throughout the data analysis of Chapter 4 I rely on observed k_{opt} to estimate the intrinsic dimensionality of a corpus.

In most early LSI work, researchers were content to approximate k_{opt} by recourse to *ad hoc* methods. Deerwester *et al.* note the presence of a region of optimal dimensionality, which they aim to approximate by choosing k wisely. "We have reason to avoid both very low and extremely high numbers of dimensions," they write. "In between we are guided only by what appears to work best. What we mean by 'works best' is ... what will give the best retrieval effectiveness" [**32**]. Likewise, Landauer and Dumais note that identifying k_{opt} for a given corpus is a highly complex matter:

> How much improvement results from optimal dimensionality choice depends on empirical issues, the distribution of inter-word distances, the frequency and composition of their context in natural discourse, the detailed structure of distances among words estimated with varying precision, and so forth. [**90**] (cf. [**51, 91**])

Arguing that k_{opt} is an intractably complex parameter, Landauer and Dumais suggest that it is discernible mostly in its effect. Thus they discover a region of optimal dimensionality by finding a value of k that results in good system performance. This amounts to finding the intrinsic dimensionality by using training and test data. While this is common to much machine learning research (cf. [**106**]) its use for LSI parameterization is somewhat unsatisfactory. Most problematic is the method's reliance on query-specific evaluation methods (see discussion of IR evaluation below). That is, the *ad hoc* methods cited here depend on pre-classified test data to define a well-constructed model. While retrospective evaluation is common to IR (and will form part of my analysis), we desire a notion of model goodness-of-fit that is applicable to the unsupervised learning environment of LSI.

Nonetheless, parameterizing k by retrospective, Cranfield-style analysis is the norm for applied LSI problems (cf. [**45, 59**]). In IR, this has led to a fairly standard range for estimates of k_{opt}. For collections on the order of several thousand documents, a dimensionality reduction of approximately 95% is common. As Manning and Schütze note, "values of k that are frequently chosen are 100 and 150" [**102**]. Deerwester *et al.* mention a similar

range [**32, 41**]. The problem of estimating k_{opt} for larger applications yields less consensus. Large corpora appear to demand a larger k-value than modest data sets do. But the rate of optimal k's increase appears to be sub-linear on the rank of the term-document matrix. Describing her application of LSI to the routing and *ad hoc* retrieval problems in several meetings of TREC (the Text Retrieval Conference), Dumais needs more representational detail than 100 factors can afford [**39, 40, 38**]. To represent a 742,331 document by 104,533 term matrix, Dumais derives a smaller, more tractable matrix by document sampling. Analyzing these sampled matrices by SVD, Dumais chooses values of k ranging from 200 to 300 [**39**]. Dumais' results suggest that while larger corpora demand more factors, this increase is sub-linear. Whereas small collections might perform well under $k \approx 5\%$ of the number of documents, representing a large corpus may only require $k \approx .005\%$ of the number of documents. However, in [**77**] Jiang and Littman find no evidence for $k_{opt} < 800$ on the TREC AP 1990 data. Their results call into question the generality of Dumais' findings.

2.3.2. The Theoretical Basis for Dimensionality Truncation. Although some consensus about likely values of k_{opt} has emerged in the LSI literature, the matter of dimensionality selection remains an open and important problem. As Hofmann notes in the context of fitting LSI models, "[deriving] conditions under which generalization on unseen data can be guaranteed is actually *the* fundamental problem of statistical learning theory" [**68**, p. 51]. Theorizing the relation between LSI and Bayesian regression, Roger Story suggests that "there is a certain amount of 'art' in LSI procedures which selectively round [singular values] to zero ..." [**140**, p. 329]. Likewise Ding calls dimensionality estimation a central and unsolved question in LSI research [**36, 37**].

Particularly vexing is the fact that dimensionality reduction for IR lacks a rigorous theoretical basis. Several authors have tried to put LSI's dimensionality truncation on firmer theoretical ground. Common to most of these attempts is the notion of statistical likelihood [**11, 72, 104**]. Fitting a model by the method of maximum likelihood involves choosing those parameter values that define the model most likely to have generated the observed data. The approaches to selecting k_{opt} that I discuss in this section operate by the method of maximum likelihood.

In **[70]** Hofmann articulates a common critique of LSI's theoretical underpinning:

> While SVD by itself is a well-understood and principled method, its application to count data in LSA remains somewhat *ad hoc*. From a statistical point of view, the utilization of a L_2-norm approximation principle is reminiscent of a Gaussian noise assumption which is hard to justify in the context of count variables.

Hofmann's criticism stems from the normality assumption native to least-squares methods. As Manning and Schütze note **[102]** the link between the normal distribution and least-squares is evident in the definition of the Gaussian density function:

$$n(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \tag{2.3.1}$$

where μ is the mean and σ is the variance. Due to the numerator in the final term of the equation, the smaller the squared deviation from the mean, the larger the probability of observing $X = x$. The method of least-squares minimizes a model's squared error, $(x-\mu)^2$. Thus the least-squares solution is identical to the maximum likelihood solution, if the data are normally distributed.

However, term-document associations are notoriously non-normal in their distribution. A well-known result from computational linguistics holds that term count data tend to follow a Zipf-like distribution **[102, 85, 151, 152, 105, 37]**. The Zipf distribution is a so-called power law model, which suggests that the rank and frequency of terms in a corpus will be inversely and exponentially related. Thus many terms occur once or twice, while only a few terms occur often. Figures 2.3.1 and 2.3.2 exemplify Zipfian term distributions.

These figures show the relation between term rank and frequency in the Cystic Fibrosis database, a corpus of 1239 documents and roughly 45,000 terms (after removing stop-words). The x-axis of Figure 2.3.1 shows the log of all observed word counts in the collection. The sharp decline in frequency as we increase the rank of a term is the hallmark of the Zipf distribution. As a member of the power-law family of distributions, Zipfian data is linearly expressible on a log-log scale. Thus Figure 2.3.2 plots the log-rank versus log-frequency. The solid line in the figure depicts the least-squares fit of these data. Although the model is

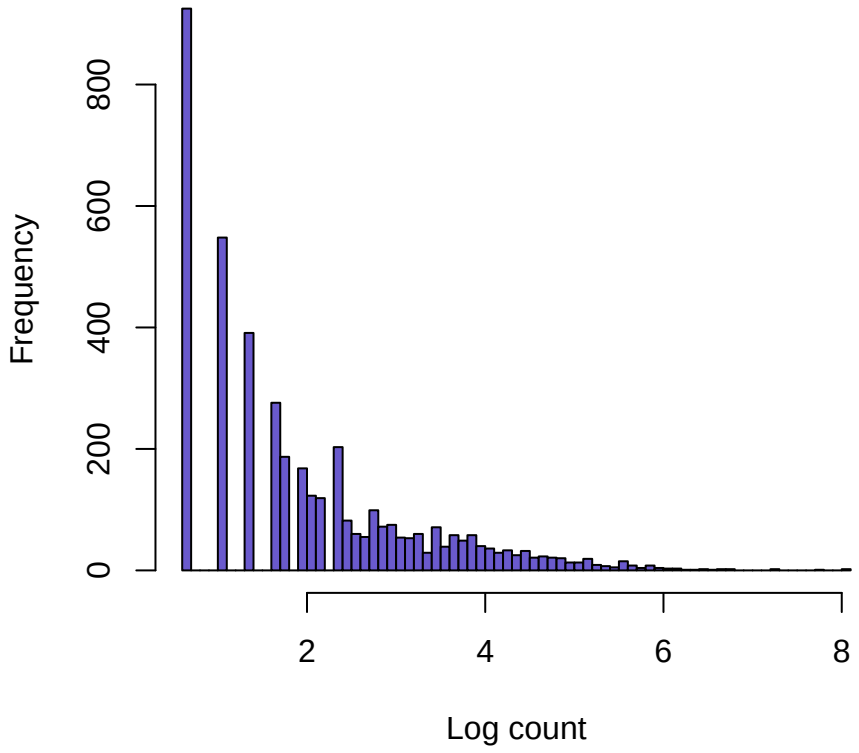

FIGURE 2.3.1. Word frequency for CF Data

somewhat skewed due to the preponderance of data at the left end of the plot, the model still accounts for a large portion of the overall variance, scoring $R^2 = 0.78$. For larger data sets, the linear fit on a log-log scale is apt to be more pronounced **[152]**. Nonetheless, although 78% fit is fairly weak in the world of power-law distributions, it does suggest that these data are far from Gaussian. Without Gaussian data, the application of a least-squares fitting method is indeed, as Hofmann notes, an *ad hoc* proposition[6] .

To motivate dimensionality reduction for IR more properly, Hofmann proposes so-called probabilistic LSA (PLSA) **[68, 69, 70]**. PLSA uses a mixture model known as the *aspect*

[6]Instead of modeling term-document data via the Gaussian distribution, Bookstein and Swanson suggest the *two-Poisson* model **[14]**. Other approaches use the negative binomial distribution to model term count data **[20]**.

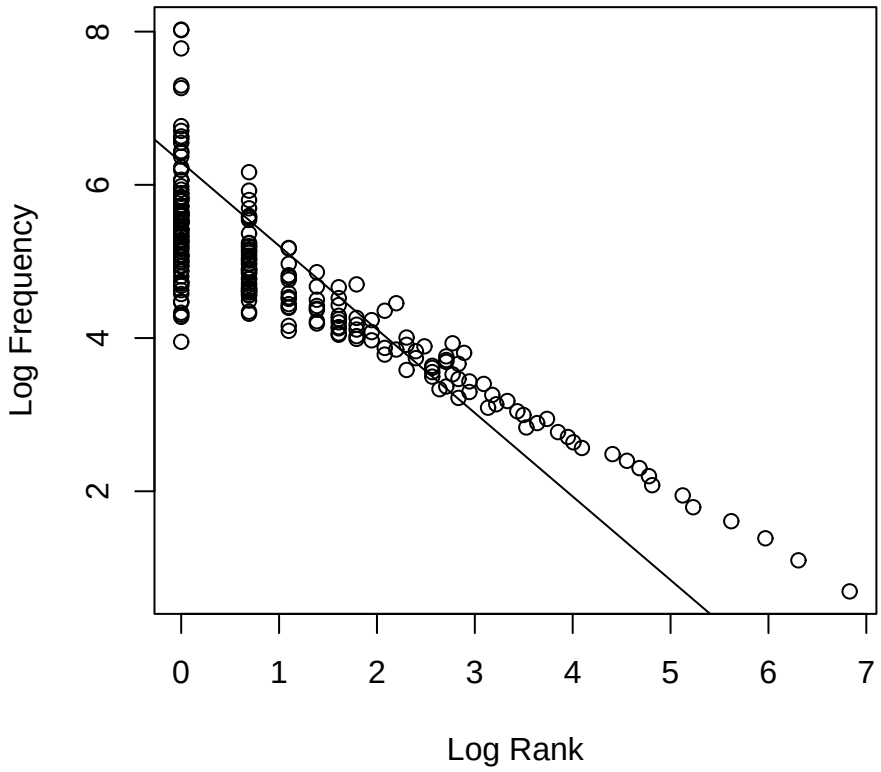

FIGURE 2.3.2. Power law distribution for CF Data terms

model, "a latent-variable model for co-occurrence data" that posits the influence of k latent class variables, c_1, c_2, ..., c_k on the generation of the term-document matrix $\mathbf{A}$. Under the aspect model, each term-document observation is assumed to be generated by a probability distribution, conditioned on the likelihood of choosing a particular word w_i given the influence of latent class variable c_j. Hofmann uses a version of the EM (expectation maximization) algorithm [**33**] to parameterize the latent class variables C. Under Hofmann's model, the k factors derived by LSI "are seen to correspond to the mixture components of the aspect model." As such, "the mixing proportions in PLSA substitute for the singular values of the SVD in LSA" [**70**, p. 184]. Thus stronger latent classes exert more influence

over the generative document model than do weaker classes. This suggests that by omitting the weakest $r - k$ classes, the model rejects the weakest correlational patterns in the observed data. An interesting corollary of this approach: Hofmann finds the best retrieval performance by using a linear combination of models, each fitted with a different k-value **[68, 69, 70]**.

Hofmann's approach is reminiscent of a more general class of probabilistic explanations of LSI **[36, 37, 68, 69, 112]**. These models imagine the influence of k latent variables over the distribution of terms and documents in a corpus. Such probabilistic models help us understand LSI more rigorously. Informed by this work, we may consider each document d_i to have been generated by a linear combination of latent document class distributions. Under LSI these distributions are estimated by the singular vectors derived under SVD. Thus during LSI we approximate each latent class c_k via the method of least-squares, whereas Hoffman's approach derives the k latent classes by the Kullback-Leibler projection **[68, 69, 70]**. The precise means of projection—orthogonal versus an objective projection—need not concern us here. Instead, the important point is that probabilistic models of LSI suggest that the role of the singular values of $\mathbf{A}$ is to weight the influence of each latent class variable on the likelihood of seeing a given document. Thus small singular values for a latent class c_j suggest that low likelihood that the j^{th} class is responsible for much variation in the data.

Like Hofmann, Ding uses the method of maximum likelihood to justify dimensionality reduction for retrieval. Unlike Hofmann, however, Ding proposes a theoretical model for LSI itself, omitting mention of any alternative method. Although LSI may violate certain assumptions inherent in the least-squares model, it's good performance and its mathematical simplicity (as a least-squares method) argue for its merits, claims Ding. To motivate dimensionality reduction under LSI, Ding proposes a "dual probabilistic model" **[36, 37]**, finding that LSI is the optimal solution of the model.

As in Hofmann's model, Ding begins with the assumption that the term-document matrix bears the influence of k latent class variables, or "characteristic document vectors,"

$c_1 ... c_k$ (collectively called C_k). For Ding, each document in $\mathbf{A}$, a_i, is drawn from a probability distribution such that:

$$Pr(a_i|c_1 ... c_k) = e^{(a_i \cdot c_1)^2} ... e^{(a_i \cdot c_k)^2} / Z(C_k) \tag{2.3.2}$$

where $Z(C_k)$ is a normalization constant. Thus the probability of seeing document a_i is proportional to its similarity to $c_1 ... c_k$. This leads to Equation 2.3.3 which shows the likelihood for a k-dimensional generative model for documents under LSI:

$$\ell(U_k) = \lambda_1 + \cdots + \lambda_k - n \log Z(U_k) \tag{2.3.3}$$

where λ_j is the j^{th} eigenvalue of $\mathbf{A'A}$. The generative model for term vectors is defined analogously to this, but instead of depending on the term-term similarity matrix $\mathbf{A'A}$, the term's model is based on $\mathbf{AA'}$, the document similarity matrix. Omitting discussion of $n \log Z(U_k)$ here (Ding excludes it from analysis on the basis of its slowly changing nature), we note that both term and document probability models have the same maximum log-likelihood:

$$\ell_k = \sigma_k^2 + \cdots + \sigma_k^2 \tag{2.3.4}$$

where σ_k is the kth singular value of $\mathbf{A}$. Thus the solution given by SVD gives the maximum likelihood solution under Ding's formulation. Despite the data's putative lack of normality, then, Ding shows that LSI's least-squares projection is in fact the model generated via maximum likelihood.

A useful consequence of Ding's model is that we acquire a precise definition of the contribution of each singular vector to the overall representation. As Ding writes, "the contribution (or the statistical significance) of each LSI dimension is approximately the square of its singular value" [**37**, p. 11]. By analyzing the eigenvalues of $\mathbf{A'A}$ or $\mathbf{AA'}$, we gain insight into the statistical significance of each LSI factor. This provides a more formal meaning to the *ad hoc* argument used earlier, that small singular values were insignificant because their associated singular vectors described only a little variance. Now we may understand the singular value truncation in LSI in terms of model likelihood. Due to Equation 2.3.4 adding

weak singular vectors increases the model likelihood only a small amount, while using up an additional degree of freedom (corresponding to the normalization constant of Equation 2.3.2). Because of this relation, adding small eigenvalues actually reduces the overall likelihood of the model. Ding thus defines k_{opt} as the value for k that maximizes Equation 2.3.4.

Applying his model, Ding finds encouraging results. Perhaps most importantly, he discovers strong evidence of the existence of an "intrinsic semantic subspace" in several test collections. Ding finds fairly close correspondence between his theoretical predictions of k_{opt} and those discovered by more traditional *ad hoc* methods [**36, 37**]. However, in several cases, his model appears to overestimate the observationally derived k_{opt}. Nonetheless, his model predicts the same "nonmonotonic relation" between dimensionality and performance observed by Landauer and Dumais.

Ding's model is satisfying in many respects. Its argument that the model likelihood attributable to a singular vector is proportional to its corresponding eigenvalue is especially interesting. This means that each factor's statistical significance is guided by a quadratic relation to the magnitude of its corresponding singular value. Thus small singular values correspond to *very small* eigenvalues, and as such, to negligible improvements in model likelihood. Ding's work thus puts dimensionality reduction by SVD in a stronger position, theoretically speaking.

However, Ding's model does not solve the problem of selecting k for an LSI system. Nor is his analysis clear on the characteristics of a collection $\mathbf{A}$ that bear on k_{opt}. Of particular concern is the generative model given in Equation 2.3.2. This model still depends implicitly on the assumption of normality. The likelihood of a given document a_i is proportional to its similarity to the latent characteristic vectors $c_1 ... c_k$. Ding's similarity metric here is the dot product, the same measure that informs the standard vector model. Thus the identity of the maximum likelihood solution and the least-squares-derived eigenvalues is not surprising insofar as the characteristic document vectors (i.e. the latent class variables) are assumed to be those that are closest to the most documents. In other words, the eigenvectors are the maximum likelihood solution precisely because the principal of least-squares defines

the principal of likelihood under the model. Thus, although Ding's approach gives the most satisfying theory for LSI's dimensionality reduction, it does not solve the problem completely.

2.3.3. Selecting an Optimal Semantic Subspace–Methods from Multivariate Statistics. Although Ding's model does not answer every question about choosing optimal k, it gives us a strong apparatus for undertaking an analysis of this choice. In particular, let us begin with Ding's observation that the eigenvalues of the similarity matrices that arise during SVD describe the model likelihood attributable to a given factor f_k. Following the mainstream of literature from multivariate statistics I suggest that the best means of selecting factors for inclusion in a model of reduced dimensionality is by recourse to analysis of their associated eigenvalues; large eigenvalues correspond to factors that exert strong influence over the generative model of terms and documents.

Rencher argues that the amount of dimensionality reduction warranted by a particular corpus is proportional to the degree of correlation among its variables. "If the variables are highly correlated," he writes, "the essential dimensionality is much smaller than p [the matrix rank]; that is, the first few eigenvalues will be large .... On the other hand, if the correlations among the variables are all small, the dimensionality is close to p and the eigenvalues will be nearly equal. In this case, no useful reduction in dimension is achieved, because the principal components essentially duplicate the variables" [**116**]. For Rencher and others (cf. [**3, 35, 79**]) the key to choosing the severity of dimensionality reduction lies in an analysis of inter-variable correlation among the data. Although a number of methods could enable this analysis, the most frequently employed techniques make use of the data's eigenvalues.

Figure 2.3.3 shows the eigenvalues of the covariance matrix for data gathered in a study of the physiology of athletes. The data are from [**116**]. Each of these 60 observations measures 6 variables: head width, head circumference, front-to-back depth of skull, ear-to-crown height, and jaw width. The x-axis of the plot is simply the rank of a given eigenvalue, $\lambda_k = 1...6$. The y-axis shows the magnitude of the k^{th} eigenvalue. Figure 2.3.3 is known as a scree plot, a name attributed to Cattell [**17**]. Although the magnitude and number of eigenvalues vary according to the rank of data matrices and the amount of variance they describe, their general shape is highly characteristic [**105, 37**]. Figure 2.3.4 shows a more realistic scree plot

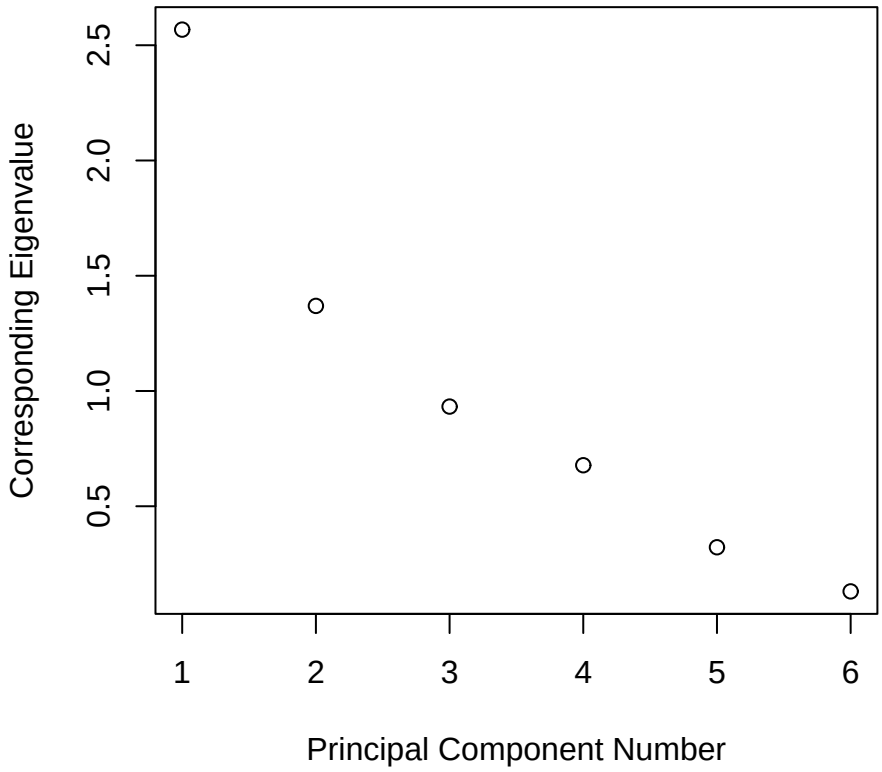

FIGURE 2.3.3. Scree plot for athletic physiology data

for IR applications. This example shows the eigenvalues for the document similarity matrix from the CF database. As in these examples, the first few eigenvalues of a data set usually capture the lion's share of system variance. Quickly, then, the size of eigenvalues decreases, usually reaching a near-horizontal plateau. Assuming that large eigenvalues improve the accuracy of statistical approximations, the bulk of statistical analysis of eigenvalues for dimensionality estimation operates by trying to discover the "elbow" in a scree plot—that point where "large" eigenvalues give way to "small" ones [**18**]. The literature of principal component analysis offers several candidate criteria. Given the similarity between LSI and PCA articulated above in Section 2.2.2, the remainder of this section details several such criteria.

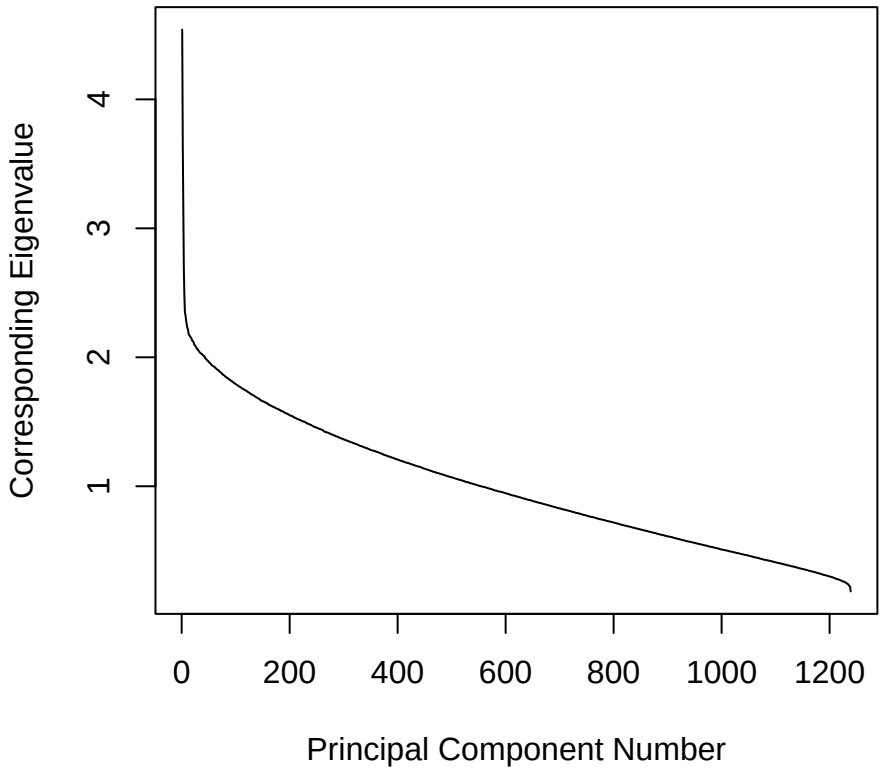

FIGURE 2.3.4. Scree plot for CF data

Perhaps the simplest (and also most popular) method of identifying significant principal components is the so-called "eigenvalue-one criterion", also known as the Kaiser-Guttman rule [**61**]. Under this technique we retain all factors whose corresponding eigenvalues are greater than the average of all the eigenvalues. The technique's name stems from its application to principal component analysis on correlation matrices; in such a situation, the mean eigenvalue, $\overline{\lambda} = 1$. Thus retaining all eigenvalues greater than the average implies retaining correlation matrix eigenvalues greater than 1.

To understand the motivation behind eigenvalue-one, consider a common result from linear algebra:

$$trace(\mathbf{S}) = \sum_{i=1}^{n} \lambda_i \tag{2.3.5}$$

where $trace(\mathbf{S})$ is the sum of the diagonal elements of the square symmetric matrix $\mathbf{S}$ [**141, 148, 116**]. Thus if $\mathbf{S}$ is the covariance matrix of $\mathbf{A}$, the average among $\mathbf{S}$'s eigenvalues is the average variance among the variables of $\mathbf{A}$. Retaining all eigenvectors whose corresponding eigenvalues are greater than $\overline{\lambda}$ entails keeping those factors (i.e. those artificial variables) that describe more variance than the average observed variable in $\mathbf{A}$.

Why should the average eigenvalue constitute this stopping point for principal component inclusion? Arguing from a psychometric standpoint, Dickman [**34**] defends the eigenvalue-one approach. Insofar as principal components comprise "fundamental" dimensions, Dickman argues that it is unreasonable to retain any factors whose variance is less than the unity accorded to an observed variable in the standard score space (cf. [**71**]). According to Kaiser, "for a principal component to have positive KR-20 internal consistency, it is necessary and sufficient that the associated eigenvalue be greater than one" [**86**]. Horn argues in [**71**] that unity thus entails the upper bound for psychometric interpretability of principal components.

But eigenvalue-one is especially useful due to its statistical motivation. Describing the Kaiser-Guttman rule, Jolliffe writes:

> The idea behind the rule is that if all elements of [$\mathbf{A}$] are independent, then the principal components are the same as the original variables and all have unit variances in the case of a correlation matrix. ... If the data set contains groups of variables having large within-group correlations, but small between group correlations, then there is one PC associated with each group whose variance is > 1, whereas any other PCs associated with the group have variances < 1. Thus the rule will generally retain one, and only one, PC associated with each such group of variables....[**81**]

If the columns of $\mathbf{A}$ are orthogonal, then all eigenvalues are equal, and Kaiser-Guttman advocates a model of full dimensionality. Likewise, if all columns of $\mathbf{A}$ are linearly dependent, then $rank(\mathbf{A}) = 1$, and Kaiser-Guttman delivers a 1-dimensional model. These cases—orthogonality and complete linear dependence of variables—display the extrema of eigenvalue-one behavior. Between these extremes lie cases of middling inter-variate correlation, under which eigenvalue-one delivers models of middling complexity. The crucial point, however, is that by using eigenvalue-one, we assume that dimensionality reduction is merited because the variables of $\mathbf{A}$ are correlated; the severity of an optimal dimensionality truncation is proportional to the degree of inter-variable correlation. Thus, under eigenvalue-one, we consider dimensionality reduction to be a form of error correction for the standard VSM. Insofar as Salton's model assumes mutual orthogonality among the terms, it incurs some amount of error when applied to non-orthogonal data. When we use eigenvalue-one, then, we assume that the difference between k_{opt} and p (the number of terms) is proportional to the amount of error incurred by the VSM's assumption of independence.

The eigenvalue-one criterion is laudable for its rigor and its simplicity[7]. Moreover, its demonstrated accuracy has led to widespread deployment of Guttman's approach. However, eigenvalue-one evinces a glaring defect. To make his analysis more tractable, Guttman elides the distinction between samples and populations in his exposition. "In this paper," he writes, "we do not treat the problem of ordinary sampling error....We assume throughout that population parameters are used, and not sample statistics" [**61**]. However, in common practice we work with samples, not parameters. Problems in applications of eigenvalue-one arise because Guttman's procedure does not recognize the distinction between the observed correlation matrix $\mathbf{R}$ and the population correlation matrix $\mathbf{P}$.

In 1965 Horn proposed an adaptation of the eigenvalue-one criterion, suited for application to sample correlation matrices [**71**] (cf. [**35, 142**]). Horn's method, called parallel analysis, is a resampling procedure (cf. [**43, 44, 67**]) which is closely related to my own method (proposed in Section 3.3). To perform parallel analysis on the principal components

[7]Guttman's approach is based on a rigorous optimization of the common factor analysis problem. Because factor analysis entails a different model than LSI and PCA, I omit a full treatment of Guttman's results, instead referring readers to the canonical literature found in [**61, 62, 86, 93, 8**].

of $\mathbf{R}$, the correlation matrix of $\mathbf{A}$, we generate many, say B, $n \times p$ data sets $\mathbf{A}_0^*$ from a multivariate normal distribution with the mean vector of $\mathbf{A}$ and $\mathbf{I}_p$ for a covariance matrix. In other words, the variables of each $\mathbf{A}_0^*$ are uncorrelated, modulo sampling error. For each $\mathbf{A}_0^*$ we calculate the principal components with corresponding "null" eigenvalues $\lambda_{01}^*, \lambda_{02}^*, \cdots \lambda_{0p}^*$. Since the variables of $\mathbf{A}_0^*$ are uncorrelated, $E(\lambda_{0k}^*) = 1$. But due to sampling error, the first $p/2$ eigenvalues will be greater than one, while the remainder will be less than one. The analysis proceeds by averaging the eigenvalues $\lambda_{01}^*, \lambda_{02}^*, \cdots \lambda_{0p}^*$ across all B samples, to derive $\widehat{\boldsymbol{\lambda}}_0^*$, a vector of eigenvalues generated from p independent variates. To complete the analysis the scree plot of $\widehat{\boldsymbol{\lambda}}_0^*$ is superimposed on the plot generated by the observed data. Horn suggests that k_{opt} corresponds to the last eigenvalue before the two superimposed scree plots cross one another.

Although I discuss the motivation behind parallel analysis in more depth in Section 3.3, it is worth stressing the fundamental similarity between parallel analysis and the eigenvalue-one criterion. As noted in [**142**], because the columns of $\mathbf{A}_0^*$ are independent, the expected value of a given null eigenvalue λ_{0k}^* is 1. That is, because the population correlation matrix is $\mathbf{I}_p$, the population null eigenvalues are all 1 (since the eigenvalues of a diagonal matrix are the elements of the main diagonal, cf [**141**]). Due to sampling error,however, the observed correlation matrix $\mathbf{R}$ will evince some opportunistic correlation, leading to $p/2$ eigenvalues greater than 1, and $p/2$ less than 1. If $\mathbf{A}$ were infinitely large—i.e. if we had unlimited data—then by the law of large numbers $\mathbf{R}$ converges on the population correlation matrix, and the null eigenvalues converge on unity. Under the condition of infinite data, parallel analysis thus converges on the eigenvalue-one criterion. We may understand parallel analysis as in improvement upon eigenvalue-one insofar as parallel analysis accounts for the fact that $n < \infty$.

Although eigenvalue-one and parallel analysis diverge in their definition of the null case, they rely on the same rationale. They both imply that LSI's dimensionality reduction entails a removal of error from the VSM. The source of this error is the VSM's assumption that the terms (columns) of $\mathbf{A}$ are orthogonal. That is, both criteria assume that dimensionality reduction is merited to the extent that the data depart from independence. For

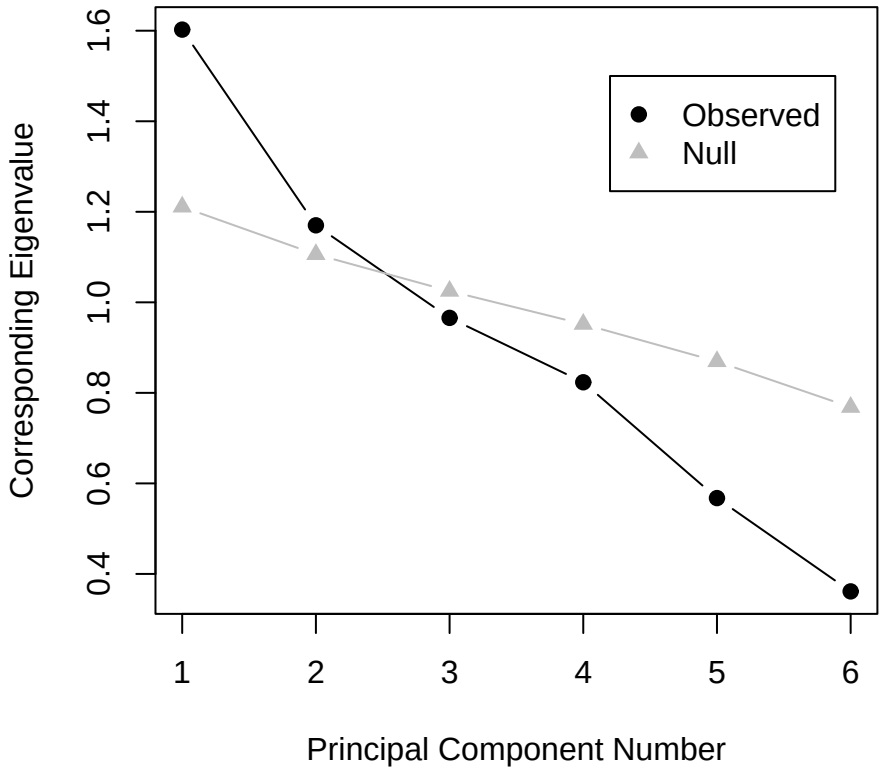

FIGURE 2.3.5. Parallel analysis on athletic data

each technique, k_{opt} is defined by the number of observed eigenvalues that do not fall below the respective null condition. The methods differ with respect to how they define the null case. While eigenvalue-one treats the observed covariance matrix as a population parameter, parallel analysis accounts for the fact that we have access only to a sample. As I argue in Section 3.3, my own method of amended parallel analysis improves upon both of these methods by defining another, more realistic null condition for implementation of the error-correction rationale.

Figure 2.3.5 shows the result of performing parallel analysis on the athletic physiology data. The simulation shown here used $B = 100$ resampling rounds. In other words, $\boldsymbol{\lambda}_0^*$ was

calculated 100 times to settle on $\widehat{\boldsymbol{\lambda}}_0^*$, whose values appear as triangles in Figure 2.3.5, alongside the observed eigenvalues, which appear as circles. This approach predicts an intrinsic dimensionality of $k_{opt} = 2$. It is worth noting that using lower values for B produced similar results (even $B = 2$ yielded the same k_{opt} prediction). How many resampling iterations are required is an open question in the statistical literature. However, setting $B = 100$ is a standard approach [**43, 44**]. I return to the number of bootstrap samples required during parallel analysis in Section 4.2.2.1.

To gauge the accuracy of eigenvalue-one and parallel analysis, I shall compare their estimates to estimation techniques with different rationales. One such method is what Dillon and Goldstein term "the percentage-of-variance criterion" [**35**]. With this method, the researcher chooses a cut-off point, m, the proportion of observed variance that the final model should describe; the researcher retains the fewest eigenvalues sufficient to account for $m\%$ of the variation among the original data. This is easy to implement since we can calculate the percent of variance captured by the first k eigenvalues via Equation 2.3.6:

$$m_k = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}. \tag{2.3.6}$$

A common value of m is 90% or 95% [**116, 76**]. For LSI model selection, however, initial experiments suggest that a 95% of variance criterion is too liberal. As Jolliffe notes, complex data sets with numerous variables are probably amenable to more aggressive dimensionality reduction, with $m \approx 85\%$ a good value [**81**]. However, this equivocation raises an important criticism of the percent-of-variance approach to dimensionality estimation. The choice of m for such a criterion is inherently *ad hoc*. As Jackson argues in [**76**], the percent-of-variance approach to model selection is unhelpful insofar as it tells us nothing about *why* dimensionality reduction is called for in an application. In other words, a model selected by the percent-of-variance criterion is difficult to interpret.

The use of the eigenvalue-one and percent-of-variance criteria is widespread in applied statistics, often comprising the default method of dimensionality estimation in statistical software packages. Figure 2.3.6 shows the application of each method to the athletic physiology data. Figure 2.3.7 applies these methods to the eigenvalues from the *CF* database.

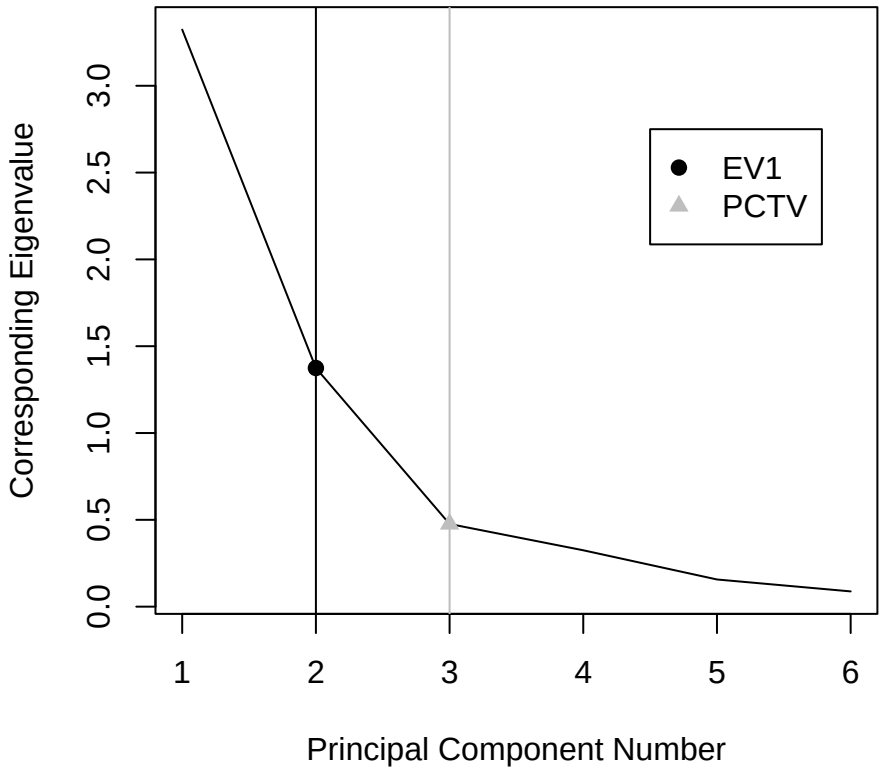

FIGURE 2.3.6. Eigenvalue-one and percent-of-variance criteria for athletic data

In both figures, the eigenvalue-one estimation of k_{opt} appears as a dot, with a vertical line showing the location of the associated cut-off. The percentage-of-variance criterion, with $m = 85\%$, appears as a grey triangle. While both methods appear to work quite well on the athletic data (identifying fair approximations of where the "elbow" in the scree plot appears), they seem to over-estimate the dimensionality of the data for the much larger *CF* dataset. Without more rigorous evaluation it is premature to conjecture any particular *true* dimensionality of the *CF* data. Yet the elbow near $k \approx 150$ seems pronounced, and both methods fail to identify it accurately. As we shall see, such over-estimation of effective dimensionality is common to many IR applications of eigenvalue-based estimators.

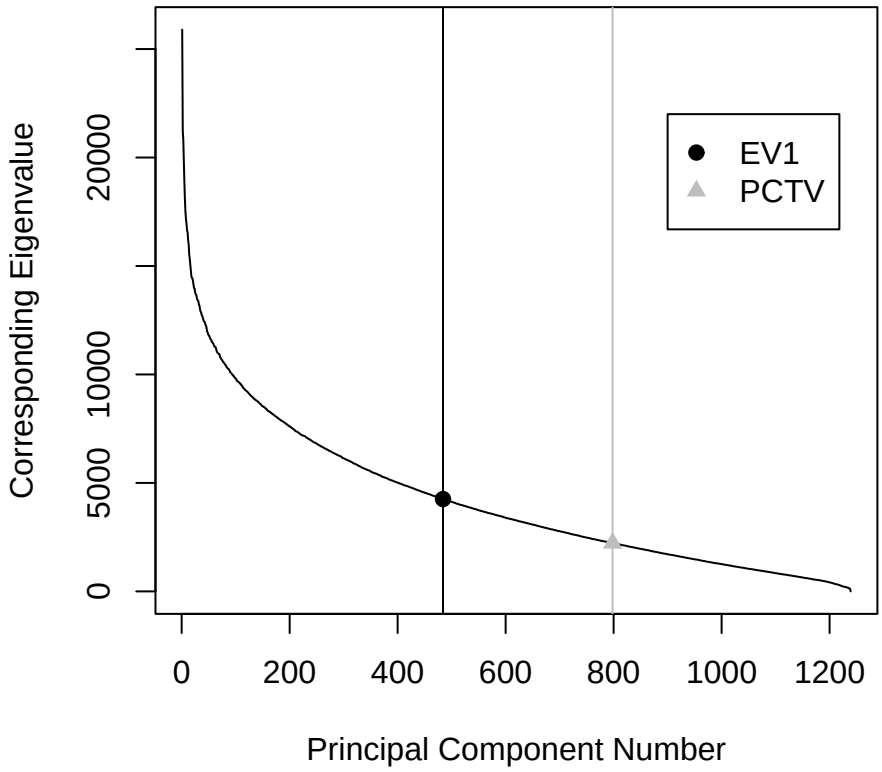

FIGURE 2.3.7. Eigenvalue-one and percent-of-variance criteria for CF data

In addition to the percent of variance approach, we may compare eigenvalue-one and parallel analysis to a technique based on parametric hypothesis testing. The idea here is to see whether the $(p-k)$ smallest eigenvalues are equal. Thus we test the null hypothesis $H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p$. As Krzanowski notes, "if H_0 is true, then there exists no one preferred direction in the subspace spanned by the last $(p-k)$ eigenvectors, so there is no reason to choose any one eigenvector in preference to any of the others. Thus we should reduce dimensionality to k dimensions, or not reduce dimensionality at all" [**87**, pp. 257-8]. To put it another way, if H_0 is true, then the last $(p-k)$ eigenvalues comprise a "shelf" in the scree plot, suggesting that the desired elbow in eigenvalue magnitude has already been reached by the time we reach the k^{th} factor.

Although a number of hypothesis tests for eigenvalue equality exist, the most common is known at Bartlett's test of isotropy **[5, 6, 7]**. The analysis begins by calculating the average of the last $(p-k)$ eigenvalues:

$$\overline{\lambda}_k = \sum_{i=k+1}^{p} \frac{\lambda_i}{(p-k)} \tag{2.3.7}$$

where λ_i is the i^{th} eigenvalue of $\mathbf{S}$. Under Bartlett's method we wish to decide whether any of the last $(p-k)$ eigenvalues are significantly different from $\overline{\lambda}$. To test this, we calculate the test statistic:

$$u = (n - \frac{2p+11}{6})(k \ln \overline{\lambda}_k - \sum_{i=k+1}^{p} \ln \lambda_i) \tag{2.3.8}$$

where n is the number of data observations. Bartlett argues that for traditional PCA applications, the statistic u is approximately χ^2-distributed. We thus reject H_0 if $u \geq \chi^2_{\alpha,\nu}$, where $\nu = \frac{1}{2}(p-k-1)(p-k+2)$. To find k_{opt} by Bartlett's isotropy, we test $H_{02} : \lambda_{p-1} = \lambda_p$. If H_{02} yields adequate confidence, we test $H_{03} : \lambda_{p-2} = \lambda_{p-1} = \lambda_p$, and so on until we find H_0 where we are no longer sufficiently confident that the last $(p-k)$ eigenvalues are equal, assuming that our stopping point corresponds to k_{opt} (cf. **[3, 79]**).

While the method of Bartlett's hypothesis test is more satisfying theoretically than the percentage-of-variance criteria, the technique is prone to stark over-estimation of a data's intrinsic dimensionality **[116]**. As in our comparison of eigenvalue-one and percent-of-variance, Bartlett's method appears to work well on the fairly simple athletic data set, as shown in Figure 2.3.8. However, Bartlett's performance on the *CF* data is nearly useless. Even setting a high confidence level, $\alpha = 0.001$, yields almost no dimensionality truncation. As is evident from Figure 2.3.9 applying this test to the *CF* data yields an estimation of 1237 (out of a possible 1239) for the intrinsic dimensionality. Because of its tendency to eschew even highly merited parsimonious representations, Bartlett's method is rarely used in practice **[60, 142]**.

As Ding shows, the eigenvalues of the similarity matrices of $\mathbf{A}$, i.e. $\mathbf{A'A}$ and $\mathbf{AA'}$, provide a natural vehicle for estimating the intrinsic dimensionality of a dataset. From linear algebra we know that the k^{th} eigenvalue is the amount of variance described by the

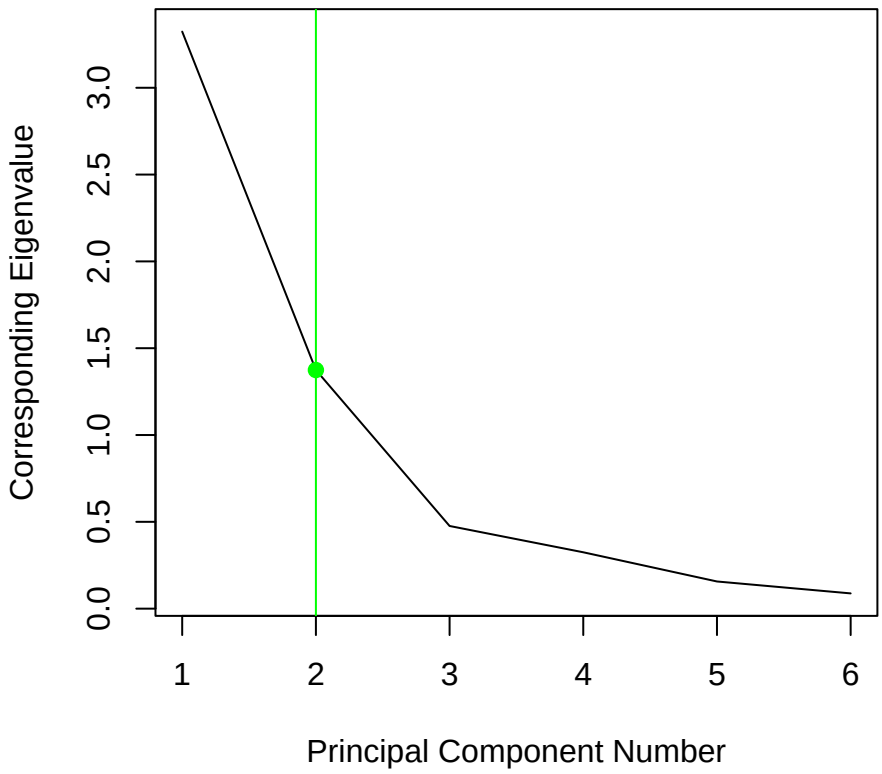

FIGURE 2.3.8. Bartlett's test of isotropy applied to athletic data

k^{th} principal component. Ding relates this result specifically to LSI, showing that the k^{th} eigenvalue gives the amount of model likelihood attributable to the k^{th} LSI factor. Because these eigenvalues comprise the squares of the singular values of $\mathbf{A}$, a small λ_k suggests that the k^{th} factor is *very* weakly (the relationship is quadratic) associated with the generative model of the documents and terms of $\mathbf{A}$. The methods discussed here attempt to define where "large" eigenvalues give way to "small" ones.

Analyses of eigenvalue significance have a long history in the multivariate statistical literature. The eigenvalue-one and parallel analysis criteria approach dimensionality reduction form a principled and well-articulated theory. Moreover, they have evinced good results in previous studies. But many other promising dimensionality estimators exist as well, with

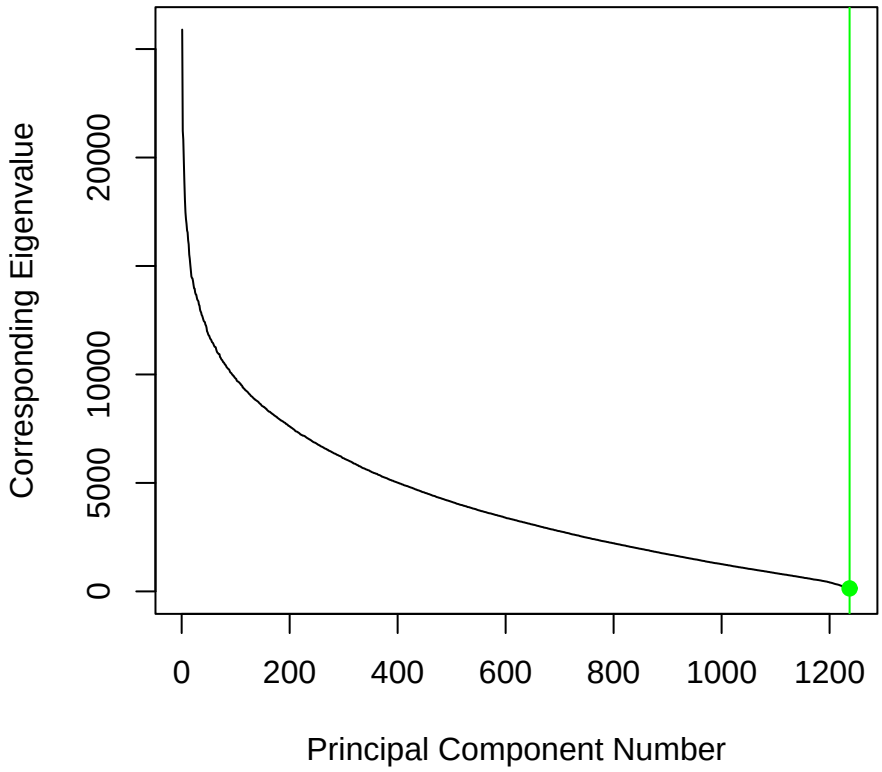

FIGURE 2.3.9. Bartlett's test of isotropy applied to the CF data

their own motivations and assumptions. Implementations of the percent-of-variance criterion suffer some problems due to their heuristic nature. But as this discussion has shown, they have performed well enough to see widespread deployment in standard software packages. On the other hand Bartlett's test of isotropy provides a basis for dimensionality truncation, basing the selection of k on a traditional, parametric hypothesis test. But Bartlett's approach appears to scale poorly to large data sets, suggesting that the χ^2 distribution of its test statistic does not hold in all cases.

To prepare our comparison of these estimation techniques, the remainder of this chapter turns to a description of methods for evaluating IR systems. The goal of this dissertation is to discover how various methods of estimating k_{opt} bear on LSI performance. To accomplish

this, my analysis will rely on the Cranfield style of IR performance evaluation. I wish to discover which values of k lead to good performance under LSI. To define "good performance," my analysis will make use of the most standard methods of system evaluation in the field of experimental IR.

2.4. IR Evaluation

Most performance evaluation in retrieval involves analyzing the quality of a system's output. For instance, it is often the case that we have two models or systems M_1 and M_2 whose performance we wish to compare. Perhaps M_2 makes use of a novel indexing method, whose value we would like to measure against the performance of a baseline system M_1. In the present study M_1 and M_2 might be LSI systems parameterized with different k-values. The goal of IR evaluation is to analyze the performance of a system with an eye toward quantifying its merits in an operational setting. As William Cooper writes, "an ideal evaluation methodology must somehow measure the ultimate worth of a retrieval system to its users in terms of an appropriate unit of utility" [**25**]. If we meet this expectation, measuring performance will give the researcher an apparatus with which to compare models and to argue for the benefits or disadvantages of a particular model.

As Van Rijsbergen notes, an important assumption underpins the mainstream of IR evaluation:

> It is a general assumption in the field of IR that should a retrieval strategy fare well under a large number of experimental conditions then it is likely to perform well in an operational situation. [**118**, ch. 7]

Nowhere is this assumption more evident than in the so-called Cranfield model of IR evaluation. Named for a series of experiments performed by Cleverdon during the 1960's [**23, 22**], Cranfield has become the mainstay of performance evaluation in IR. Not only was the Cranfield model vital to the historical development of IR methods (cf. [**4, 129**]), it continues to guide research in the field by its incorporation into TREC, the text retrieval conferences [**64, 146**]. Due to its canonical status in current IR (David Ellis goes so far as to called Cranfield the "archetypal model" of IR evaluation [**46**]), in this section I describe the assumptions

and methods associated with Cleverdon's measurements, with the goal of adapting them to the problem of choosing k_{opt} in an LSI system.

The starting point of Cranfield-style evaluation is the idea of *relevance* [**66**]. In gauging IR performance, Cleverdon aims to measure a system's ability to deliver relevant information—to discriminate between relevant and non-relevant documents. Thus good performance implies that a system is able to discern and act favorably upon a user's stated information need. For Cleverdon's studies, however, relevance is a simplified ideal: "it has to be assumed in [IR research] that we are considering idealized conditions, and do not have to take into account losses due to human error" [**23**]. Instead of any intuitive, subjective notion of relevance, Cleverdon implicitly posits a binary, objective relevance function : for a given query q_i and a given corpus D comprised of n documents d_j, $j = 1...n$, there exists a function $R(q_i, d_j)$ such that $R(q_i, d_j) = 1$ if document j is relevant to q_i, and $R(q_i, d_j) = 0$ otherwise.

It is important to note that the problem of relevance continues to receive attention in the IR literature (cf. [**139, 133, 66**]). Obviously the assumption of binary, objective relevance between queries and documents is problematic. As described by Schamber *et al.,* relevance is in fact a highly subjective construct [**135, 134**]. Contextual details of a search bear heavily on whether a searcher finds a given document relevant to his or her information needs. Likewise, a user's idea of what constitutes relevant information is liable to change over time, as he or she encounters new data [**144**].

Noting the disjunction between psychological relevance and Cleverdon's operationalized objective variety, Harter argues that Cranfield-type evaluation is prone to egregious measurement error [**65**]. However, empirical research has demonstrated that despite its shortcomings, evaluation based on objective, binary relevance does yield useful information for IR research. As noted by Salton and Lesk [**128**] and more recently by Voorhees [**146**], objective relevance judgements provide strong information about the *comparative* benefits of one IR system over another. That is, testing an individual model's ability to perform on a single data set based on objective relevance yields little in the way of useful measurement. However, in comparing between two or more systems, metrics based on objective relevance appear to provide good evidence about the relative merits of each model *vis a vis* the other

Test Collection	*Abbreviation*	*Subject Matter*
Communications of the ACM	CACM	Computer Science
Cystic Fibrosis	CF	Cystic Fibrosis (medicine)
Cystic Fibrosis (full text version)	CF_FULL	Cystic Fibrosis (medicine)
Institute of Scientific Information	CISI	Information Science
Cranfield	CRAN	Aeronautics
Medline	MED	Medicine

TABLE 2.4.1. IR Test Collections

models. Because my aim in the present study lies in comparing the performance of LSI systems parameterized with different dimensionalities, such an analysis is highly desirable.

Capitalizing on the assumption that results from experimental data translate to similar results in operational settings, evaluation in the Cranfield tradition makes use of so-called test collections [**4**]. Table 2.4.1 lists six test collections that informed this study. These collections provide the data necessary to deploy an IR system on realistic data, while also giving the researcher information with which to measure the effectiveness of his or her model. Thus test collections usually include three components:

- A corpus of documents
- A set of queries
- A set of relevance judgements

For instance, the Cystic Fibrosis (CF) test collection [**138**] includes a corpus of 1239 documents. The corpus contains all documents from the National Library of Medicine's Medline database that are indexed with the subject heading 'cystic fibrosis.' The CF collection also includes 100 queries, statements of information need generated by subject experts in medicine. The relevance judgements contain a list of all documents judged relevant to each of the 100 queries by a panel of expert reviewers. Statistics for the test collections used in this study appear in table 2.4.2. By the standards of today's web-based IR research paradigm, these collections are small. For instance, the largest of these collections, CACM uses 2.2 megabytes of storage space. On the other hand, the data used for TREC-6 occupied over a gigabyte of disk [**4**]. The matter of performing LSI on large corpora is a research area in its own right [**9, 73**]. While studying k_{opt} in TREC-sized data sets would be desirable, doing so in an intensive, systematic fashion is computationally impractical at the current

Collection	*Num. Docs*	*Num. Terms*	*Num. Queries*
CACM	3200	4867	64
CRAN	1400	4612	225
CF	1239	5116	100
CF_FULL	379	9549	100
MED	1033	5831	30
CISI	1460	5615	112

TABLE 2.4.2. Statistics from IR test collections

time. My interest lies in comparing estimates of intrinsic dimensionality in a wide range of settings, and thus demands the use of several corpora. Despite their small size, the collections shown in table 2.4.2 were useful for this analysis by virtue of their diversity. With regard to corpus size, topical domain, and document representation, these collections span a large area. Moreover, they have become standard in the IR literature [**4**]. Finally, several of these collections have informed the most significant theoretical studies of dimensionality reduction in IR [**70, 37**]. I discuss my rationale for choosing these test collections in more depth in Section 3.1.

Although the method of gathering relevance judgements varies from collection to collection [**146**], the result is the same: test collections give the researcher information about the documents that a system *should* retrieve when presented with a given query. Armed with putatively exhaustive relevance judgements, IR researchers typically employ two measures to evaluate their systems:

- *Precision:* the proportion of relevant to non-relevant documents in the set of retrieved documents
- *Recall:* the proportion of relevant documents in the retrieved set to the total number of relevant documents.

To define precision and recall formally I adopt the notation shown in table 2.4.3. Given a query q and an exhaustive list of relevance judgements, we define the following variables: Following Van Rijsbergen [**118**] I define precision and recall for a collection with n documents by the contingency table of Table 2.4.4. This allows the following definitions:

$$precision = \frac{|Rel \cap Ret|}{|Ret|} \tag{2.4.1}$$

Symbol	*Definition*
Rel	the set of relevant documents
$\overline{Rel}$	the set of non-relevant documents
Ret	the set of retrieved documents
$\overline{Ret}$	the set of non-retrieved documents

TABLE 2.4.3. Notation for IR evaluation metrics

	Relevant	*Non-Relevant*	
Retrieved	$Rel \cap Ret$	$\overline{Rel} \cap Ret$	Ret
Non-retrieved	$Rel \cap \overline{Ret}$	$\overline{Rel} \cap \overline{Ret}$	$\overline{Ret}$
	Rel	$\overline{Rel}$	n

TABLE 2.4.4. Precision/recall contingency table

Model	*Ranking*
M_1	*N R R R N N N N R N R N N N N R N R N R N N R N N R N N N N*
M_2	*R R R N R N N N R N R N N N N R N R N N N N R N N R N N N N*

TABLE 2.4.5. Two fictional document rankings

$$recall = \frac{|Rel \cap Ret|}{|Rel|} \tag{2.4.2}$$

Under these definitions, precision and recall are mutually dependent measures. Because most IR systems rank documents against queries, we may measure precision at a given recall level. Likewise, we measure recall for a given level of precision. Thus we might define $precision_{0.5}$ to be the ratio of relevant documents to the total number of documents retrieved when 50% of the relevant documents for query q have been retrieved.

Not surprisingly, precision and recall exhibit a negative correlation. It is trivial to achieve 100% recall by retrieving every document in a collection. However, doing so without recourse to a more sophisticated ranking strategy is liable to force a user to view many non-relevant documents before finding the relevant ones. Because of the inverse relationship between precision and recall, many studies describe performance by reporting the observed precision at a variety of recall levels. For example, consider the fictional document rankings of Table 2.4.5. This simulated data set contains 30 documents. Of those 30, 10 have been judged relevant to a given query. Table 2.4.5 lists the ordered output of two retrieval systems M_1 and M_2, where R represents a relevant document and N represents a non-relevant document.

Given these data, we would calculate $precision_{0.1}$ for each model by calculating the precision score at that point in the ranking where ten percent (i.e. 1) of the relevant documents has been retrieved. For model M_1 we calculate precision on the ranking $M_1(0.1) = N\,R$. Thus $precision_{0.1}(M_1) = 1/2$. On the other hand calculating precision for the ranking while $M_1(0.1) = R$ yields $precision_{0.1}(M_1) = 1/1$. At the 10% recall level, then, M_2 yields better precision than M_1.

A common goal of performance evaluation is to discover which model, M_1 or M_2, is better overall. This is a complicated construct to measure. But a typical approach to such measurement involves graphical inspection (and associated statistical analysis [**129**]) of a precision/recall plot. Figure 2.4.1 shows a graphical representation of the data from Table 2.4.5. This plot shows the precision obtained by each model at five levels of recall (0.1, 0.25, 0.5, 0.75, and 1.0). By virtue of its higher precision scores model M_2 seems better than M_1 at low levels of recall. However as we observe precision at higher levels of recalls, the margin of improvement afforded by M_2 becomes negligible, disappearing altogether at $precision_{1.0} = 0.4$.

In the previous example, I plotted precision versus recall for a given query q_i. Of course standard test collections come with a variety of queries q_j, $j = 1...n$. Thus when evaluating a given model M, we usually create a precision/recall curve in which each point is the average precision at recall level r across each of the n queries. Thus we plot the average precision at each recall level r by:

$$\overline{prec}_r(M) = \sum_{i=1}^{n} \frac{prec_{r,i}}{n} \tag{2.4.3}$$

where $prec_{r,i}$ is the observed precision at recall r for the i^{th} query.

Although precision/recall graphs are ubiquitous in the IR literature, the use of precision and recall is often criticized. A common criticism of the recall/precision metric involves the need for interpolation in its use. As noted in [**4**, p. 78] the most common means of reporting precision is at 11 standard recall levels, $r_j, j \in \{0.0, 0.1, 0.2, ..., 1.0\}$. However, depending on the number of documents that are relevant to a query q_i, distinct precision values may not be defined at each r_j. Returning to our earlier example, consider a new query, $q_{i'}$ for

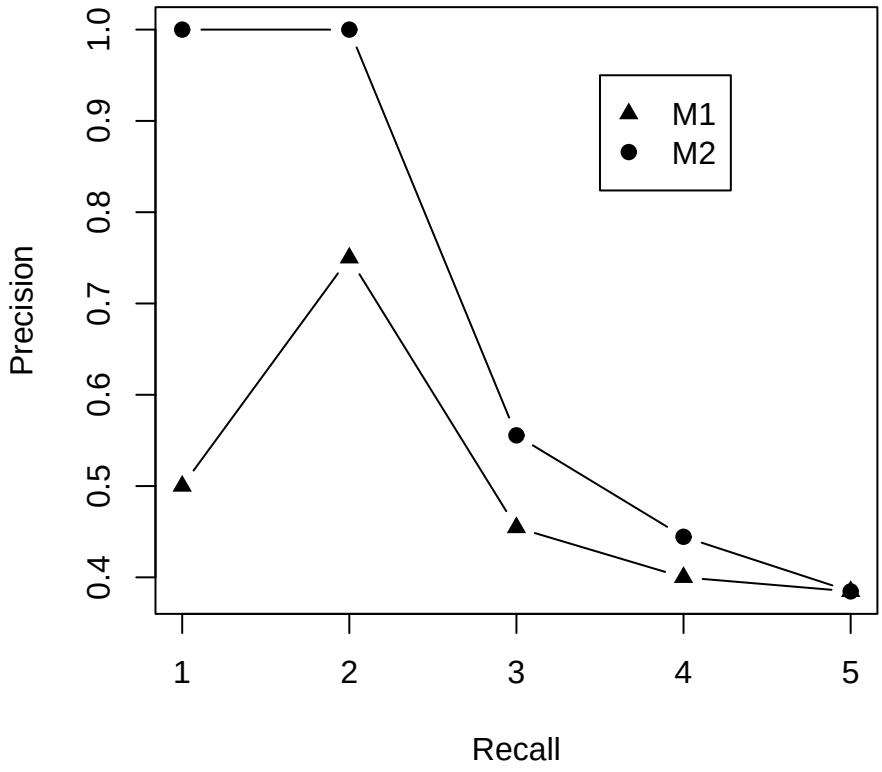

FIGURE 2.4.1. Fictional precision/recall graph

Model	*Ranking*
M_1	*R N N R R N N N N N N N N N N N N N N N N N N N N N*
M_2	*R R N N N N R N N N N N N N N N N N N N N N N N N N*

TABLE 2.4.6. Two fictional document rankings

which only three documents are relevant. This situation might yield the rankings shown in Table 2.4.6. Considering the ranking of M_1, after retrieving the first relevant document, we have $precision = 1.0$, with $recall \approx 0.33$. Because so few documents are relevant to $q_{i'}$, recall at 10%, for instance, is undefined. If we want to measure precision at the 11 standard recall intervals, we need to derive interpolated approximations of the returned set at each level of recall. Following Baeza-Yates and Ribiero-Neto [**4**], I define interpolated precision

for a given level of recall:

$$prec_{r_j} = max(prec_{r_j \leq r \leq r_{j+1}}) \tag{2.4.4}$$

such that the interpolated precision at the j^{th} recall level is the maximum known precision at any recall level between the j^{th} and $j^{th}+1$ levels. In our example, then, the first known precision level for M_1 is $prec_{0.33} = 1$. By Equation 2.4.4 $prec_{0.0}(M_1) = \cdots = prec_{0.3}(M1) = 1$. The need for such interpolation renders the use of precision/recall curves somewhat problematic, as artifacts from the interpolation process can color the evaluation process.

Another criticism of precision/recall is the interdependent relationship between its two measures. According to Salton, many critics dismiss the idea of contingency-table based metrics due to the confusing nature of their interrelated elements **[129, 97]**. Instead of precision and recall, critics argue that an ideal effectiveness measure would express retrieval effectiveness as a single number. Moreover, that number should be independent of any specific cutoff point in the retrieval process **[143, 15, 119, 120]**.

Perhaps the simplest single-valued performance metric in IR involves simply reporting precision at a given (and ostensibly important) level of recall. For instance, we might compare two systems by calculating $\overline{prec}_{0.5}$ for each and noting the difference between the resultant values. In some sense this measure offers an "average" picture of the recall/precision ratio for a given model. But a more sophisticated approach (one used much more frequently) is to calculate average precision across several levels of recall:

$$avPrec = \frac{\sum_{i=1}^{r} \overline{prec}_i}{r} \tag{2.4.5}$$

where $\overline{prec}_i$ is the across-query-averaged precision at recall level i, and r is the number of recall levels observed. As Losee argues, reporting $avPrec$ tends to provide a less biased account of retrieval performance than simply relying on $\overline{prec}_i$, insofar as $\overline{prec}_i$ amounts to taking a heavily weighted average of precision scores at various recall levels, where such a weighting may not be warranted **[98]**.

However, other single-term performance measures exist in the IR literature. Closely related to the precision/recall measure is the metric known as F, the harmonic mean of

precision and recall. As discussed in [4] the harmonic mean of precision and recall for the j^{th} document in the ranked list of n documents is given by Equation 2.4.6

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \tag{2.4.6}$$

where $r(j)$ is the recall measured at the j^{th} document in the ranking and $P(j)$ is precision at the j^{th} document. Until a relevant document is retrieved, $F = 0$. And if all documents up to point j are relevant, then $F(j) = 1$. As discussed in [4] and [**137**], F approaches 1 when most of the ranked documents are relevant. Thus reporting $max(F(j))$ for $j = 1 \cdots n$ is common insofar as it offers a good compromise between precision and recall. This measure—the maximum value of F found under a given model—is known as optimal F.

Another important single-number performance metric is the Average Search Length (ASL), which describes the expected position of a relevant document in the ranked output from a given retrieval system [**97, 98**][8]. In the example shown in Table 2.4.5 we see two document rankings. We compute the ASL for each by summing the position of each relevant document in each ranking and dividing by the number of relevant documents. Thus $ASL(M_1) = 13.2$, while $ASL(M_2) = 11.4$. On average, then, M_2 moves relevant documents closer to the front of the ranked list than M_1 can. Thus because its ASL is *lower,* we consider M_2 an improvement on M_1, with respect to ASL. Because the units of ASL are documents Losee argues that the measure is easily interpreted, "the ASL is measured in units of 'documents'; knowing that the average position of a relevant document is 23 or 500 or 2 million documents into the ranked list of documents conveys useful information to the searcher" [**98**]. In addition to its ready interpretability, ASL has been shown to correlate well with other metrics, thus giving a good, albeit highly digested, picture of system performance in a single measure [**97, 98**].

Performance evaluation in IR is clearly prone to certain subjectivities. It's reliance on the idea of relevance problematizes the notion of *good* retrieval. However, despite the obviously oversimplified idea of relevance that enables Cranfield-style evaluation, a good deal

[8]As Losee mentions ASL is related to Cooper's expected search length measure [**24**].

of empirical work [**128, 146**] suggests that the errors introduced by such an operationalization of relevance are not too egregious. I argue that measures such as precision/recall and ASL provide strong evidence for the comparative evaluation of IR systems. In other words, achieving 80% precision at 20% recall is meaningless without greater contextualization. However, if two systems that vary only with respect to a single parameter consistently score analogously on a variety of measures taken on a variety of test collections, I argue that this evidence implies a real difference between the methods. Such differences suggest that the variable in question will continue to bear on performance in an analogous fashion even in a non-experimental, fully operational setting.

2.5. Conclusion

This chapter has contextualized the problem of estimating the optimal dimensionality for an LSI system by pursuing a survey of the relevant literatures. The discussion began with an overview of the vector space model of IR. Section 2.1.1 gave a basic introduction to the VSM. In Sections 2.1.2 through 2.1.3 I argued that extensions of the VSM such as query expansion and relevance feedback comprise early efforts to mitigate some of the oversimplifications in the most basic articulation of Salton's model. Most notably, these extensions address the assumption of term independence in the VSM. This discussion paved the way for my analysis of principal component analysis and LSI in Section 2.2. Statistical modeling techniques allow LSI to derive a representation of the population term correlation matrix, a model that stands to improve retrieval over the standard VSM by describing the dynamics of inter-term relationships in a given corpus. However, model building under LSI is a poorly understood problem. Unlike traditional linear regression, for example, the unsupervised learning environment of LSI lacks a native measure of model goodness of fit. Thus in Section 2.3 I detailed several approaches to estimating k_{opt}, the intrinsic dimensionality of a corpus. Finally, in Section 2.4 I offered an overview of the Cranfield model of IR performance evaluation. The research reported in this dissertation is concerned with identifying optimal LSI models. My analysis, described in Chapter 3, relies heavily on the techniques developed to enable Cranfield-style performance evaluation.

Dimensionality reduction under LSI is a vital topic in IR. As an extension to the VSM, LSI marks an advanced point in the evolution of geometrically based retrieval models. It has been shown to improve keyword-based retrieval in many instances. But in order to function properly, the LSI model must be well constructed. In particular, k, the dimensionality of an LSI system is crucial to the effective use of dimensionality reduction for retrieval. Too few dimensions robs the model of important descriptive power, while models of too many dimensions risk becoming overfitted to the sample. Thus parameterizing k is a classic example of the bias-variance trade off in statistical modeling. Each method of eigenvalue analysis described in Section 2.3.3 offers a rationale for selecting a given value of k. In the remaining chapters of this study, I turn to an experimental comparison of these estimation techniques, with an eye towards describing their suitability for practical IR applications and their respective theoretical implications.

CHAPTER 3

Methods

As described in Chapter 1, the current study addresses the question:

> how effectively can an analysis of the eigenvalues derived during LSI be used to estimate the optimal representational dimensionality for IR?

To assess the suitability of such metrics I analyzed a variety of data by a variety of measures. Because the best means of measuring IR performance is open to debate, I approached this analysis broadly, arguing that if a method performs well on a variety of data according to a variety of criteria it is apt to perform well in operational settings, too. Likewise if a method performs well under certain conditions but not others, I suggest that such disparity merits further reflection.

My analysis began by estimating k_{opt} for the six data sets described in Section 2.4. To generate these estimates I used the four eigenvalue-based statistics of Section 2.3. In addition, in this chapter I describe a novel eigenvalue-based dimensionality estimator, amended parallel analysis (discussed in Section 3.3 below). Having derived five estimates of k_{opt} for each collection, I then used Cranfield-style methodology described in Section 2.4 to evaluate performance at a broad range of k values. My goal, then, was to note how well the eigenvalue-based dimensionality estimations correlated with the best observed LSI performance.

Under the rubric of my general research question are three related questions, which this study also aimed to address:

(1) Do performance measures based on Cranfield-style evaluation agree on the optimal dimensionality for performing LSI on each data set?

(2) If an intrinsic dimensionality seems evident from the Cranfield-style performance analysis, which eigenvalue-based estimators best predict it?

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
number docs	3204	1239	392	1460	1398	1033
number terms	4867	5116	9549	5615	4612	5831
median term freq.	17	17	26	15	28	12
max term freq.	41	49	126	46	67	42
variance term freq.	30.707	41.128	197.468	28.817	68.704	28.256
median doc length	12	41	536	34	43	52
max doc length	966	3307	6268	1942	1595	834
median docs/term	13	12	13	12	18	9
variance docs/term	12.925	11.269	12.459	12.3	17.142	8.967

TABLE 3.1.1. Summary statistics of IR test collections

(3) In what ways does the suitability of a given measure depend on the statistical characteristics of the data set to which it is applied?

This section describes in detail the methods I used to address these research questions. In Section 2.4 I mentioned the test collections that informed this analysis. In Section 3.1 I motivate the selection of these data in greater detail, discussing the appealing characteristics about each corpus. In Section 3.2 I describe the IR performance metrics that are relevant to this study, discussing my rationale for choosing this particular battery of measures. In Section 3.3 the discussion turns to a full description of my own proposed dimensionality estimation method, amended parallel analysis (APA). Finally, Section 3.5 treats the computational tools that enabled my experiments.

3.1. IR Test Collections

Because optimal k depends on the correlational structure of the input matrix, I hypothesized that different corpora will demand different representational dimensionalities. My aim was to evaluate eigenvalue-based dimensionality estimators for operational settings (not for a particular corpus), so I undertook experiments on a variety of corpora. Table 2.4.2 describes the test collections that were used to conduct this study. To motivate my selection of these collections Tables 3.1.1 and 3.1.2 display more information about each data set.

The first two rows of Table 3.1.1 show the number of documents and the number of indexing terms in each test collection. The number of indexing terms is the final number of features used to represent each document after removing stop-words and eliminating words

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
number of queries	64	100	83	76	225	30
median rel docs/query	12	6	3	30.5	7	22.5
variance rel docs/query	154.531	212.126	36.056	1308.373	29.031	75.821

TABLE 3.1.2. Query-related statistics for IR test collections

that occur only once in the corpus. No stemming was used to derive these counts. By scanning the first rows of the table 3.1.1 it is clear that these collections are quite different with regard to the number of constituent documents and the size of their vocabulary. If the intrinsic dimensionality of a data set bears some relation to these characteristics, such a spread is desirable.

The third through fifth rows of Table 3.1.1 describe term-related aspects of each data set. For instance, row three, *median term frequency,* shows the number of times that a term with middling frequency appears in each database. Likewise *variance of term frequency* gives the variance of term frequencies across all terms in a given collection. Finally, *maximum term frequency* is the number of occurrences of the most common term in the collection. Again, the selected data sets evince a wide variety of values on these criteria. For instance, although *CRAN* contains a similar number of documents to *CF*, its term variance is much higher than *CF*'s. On the other hand, *CACM* contains many more documents than *CRAN*, yet its term variance is relatively low.

Rows six and seven of Table 3.1.1 relate information about each collection's document characteristics. Thus *median document length* is based on a count of the number of distinct terms per document. The variance of distinct word counts per document is given in *variance of document length.*

The last two rows of Table 3.1.1 pertain to the relationship between terms and documents. Row eight, *median documents per term* describes, on average, how many documents a given term appears in. Thus for *CACM*, the average (i.e. median) term occurs at least once in 13 documents, 0.4% of the total number of documents. On the other hand, *MEDLINE*'s median term occurs in only 9 documents, which is 0.8% of that collection's total document count. *CRAN*'s median-frequency term occurs in 18 documents, approximately 1.3% of its total document population.

In choosing my test collections I relied on the metrics of Table 3.1.1 as a rough approximation of several informal corpus characteristics. First, to maximize the diversity of my experimental parameters, I aimed for data sets that were of various sizes, both with regard to document count and vocabulary size. At the outset I suspected that the intrinsic dimensionality of a data set would be substantially less than the rank of the data set. But it was of interest to note how k_{opt} relates to these features of a data set, and how various dimensionality predictors fare on data of diverse rank. Second, I desired input data with distinct patterns of term distribution. I hoped that such distributions would reflect distinct domains of coverage across these databases. By selecting data sets with varying numbers of documents per term, I tried to ensure that a different relationship between terms, documents, and concepts obtained among the various collections.

While the criteria that bear on the intrinsic dimensionality of a data set are, I believe, query-independent, I also consulted the data shown in Table 3.1.2 when selecting test collections. This table describes the relationship between queries and documents in each data set. Because Cranfield-style evaluation is query-specific, the identification of an observed k_{opt} may be colored by the statistical relationship between queries and documents. As shown in Table 3.1.2, I have selected test collections that vary not only in the overall number of queries, but also in the median number of relevant documents per query and in the variance of this measure. As discussed in Section 4.1, this distribution came to the fore in the analysis of the *CISI* data, where the observed k_{opt} measurements showed signs of interpolation artifacts.

It is also worth noting that four of the selected test collections, *CACM*, *CISI*, *CRAN*, and *MED*, were also appealing because they have been used in Ding's theoretical work on dimensionality reduction for IR **[36, 37]**. Thus using these collections for my own analysis allows direct comparison with the results obtained under other studies of similar problems. Finally, in addition to these four collections, I included two versions of the Cystic Fibrosis database, *CF* and *CF_FULL*, because of their unique relationship. That is, in many respects, *CF* is quite similar statistically to other selected collections. However, I include it because it makes an interesting baseline with which to describe the intrinsic dimensionality

of *CF_FULL*, the only selected database that represents documents in their entirety, instead of simply reporting abstracts or titles. Thus it will be of particular interest to compare the intrinsic dimensionality of these two data sets that treat identical material, but that do so in differing formats.

3.2. Performance Measures

Eigenvalue-based methods of estimating a corpus' intrinsic dimensionality are especially attractive because they are query independent. While most *ad hoc* approaches to dimensionality estimation in LSI depend on analyzing system performance on a pre-classified test collection of documents, the methods described in Section 2.3 may be applied without recourse to pre-existing relevance judgements. For example, we may use the eigenvalue-one rule or parallel analysis to estimate the intrinsic dimensionality of any corpus, regardless of the presence or absence of relevance judgements.

Nonetheless, of special interest during my analysis was judging the quality of each dimensionality estimation technique using traditional, Cranfield-style evaluation. My goal was to discover IR models—i.e. dimensional parameterizations—that optimize retrieval performance. To align this analysis with the mainstream of IR research, I define observed optimal performance in terms of Cranfield-based metrics. In particular, I offer the following definition:

- *Optimal Observed Performance (observed $k_{opt}(c_r, p_i)$):* An IR system is performing at its observed optimum with respect to a corpus c_r and a given performance measure p_i if its current parameterization m_j yields the best value of p_i across all observed parameterizations, $m_j, j = 1 \cdots n$.

For the purposes of this study, I have based the identification of observed optimal performance on the performance measures listed in Table 3.2.1. Descriptions of these measures is given in Section 2.4. To discover the optimal observed performance of a given LSI model requires that we employ a single-measure yardstick for IR performance. This allows us to take such a measurement at each parameterization of k, and to define the *observed* k_{opt} as that value of k that gives the best measured performance. It is important to stress that

Measure	*Abbreviation*	*Description*
Average Precision	PR	average precision at 25%, 50%, and 75% recall
Optimal F	opt. F	max. observed harmonic mean of precision & recall
Av. Search Length	ASL	avg. location of a rel. doc. in ranked output

TABLE 3.2.1. Analyzed performance measures

	ASL	*F*	*Pr*
ASL	1.000	-0.906	-0.883
F	-0.906	1.000	0.997
Pr	-0.883	0.997	1.000

TABLE 3.2.2. Correlation between ASL, opt. F, and PR on *MEDLINE* data

observed k_{opt} is taken with respect to a given corpus and a given performance metric. In this sense it is distinct from the intrinsic dimensionality of the corpus (which I denote simply as k_{opt}), a parameter whose value is invariant across performance metrics.

I have selected the metrics shown in Table 3.2.1 for a number of reasons. First, average precision has become the *lingua franca* of IR research. As such, defining performance in terms of precision and recall is appealing insofar as it aligns my results with the majority of work in the field. However, there is little theoretical basis to the use of average precision. Thus I supplement my analysis with two other metrics (optimal F and ASL) in hopes of assessing the validity of my observations. In other words, it will be of interest to note whether observed optimal performance with regard to precision, ASL and optimal F comes at similar parameterizations of k. If the notion of observed optimal performance is valid—that is, if observed optimality relates to the intrinsic dimensionality of a corpus—I expect to see overlap among the three performance measures. Such a hypothesis is borne out by Figures 3.2.1, 3.2.2 and 3.2.3, which show performance according to each measure as k is increased (in increments of 15 dimensions) from 1 to k_{max} on the *MED* data.

As is evident in these plots, there seems to be a strong correspondence between the three measures. All three show poor performance for very low values of k. Adding the first hundred singular triplets improves performance according to all three measures, as well. Moreover, all three measures decline as k increases from approximately $k = 150$ to $k = k_{max}$. Table 3.2.2 shows the correlation matrix for these data. The strong monotonicity among ASL,

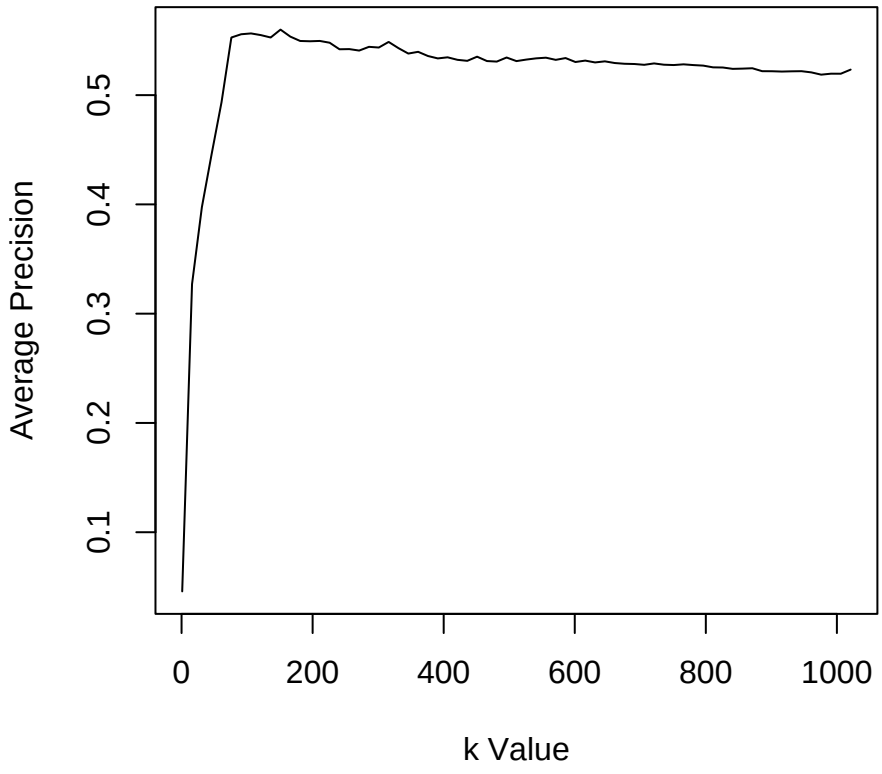

FIGURE 3.2.1. Pr for increasing k-values on Medline data

optimal F, and Pr suggests that all three metrics are measuring the the same phenomenon. Insofar as these measures agree on a range of optimality in the vicinity of $k \approx 150$ they suggest that the intrinsic dimensionality of the *MEDLINE* data is also in this range. Thus the bulk of my analysis involved comparing the five eigenvalue-based estimations of k_{opt} against this observation.

3.3. Amended Parallel Analysis

In this section I describe amended parallel analysis (APA), a novel method for estimating the intrinsic dimensionality of a data set. Like other eigenvalue-based criteria, APA

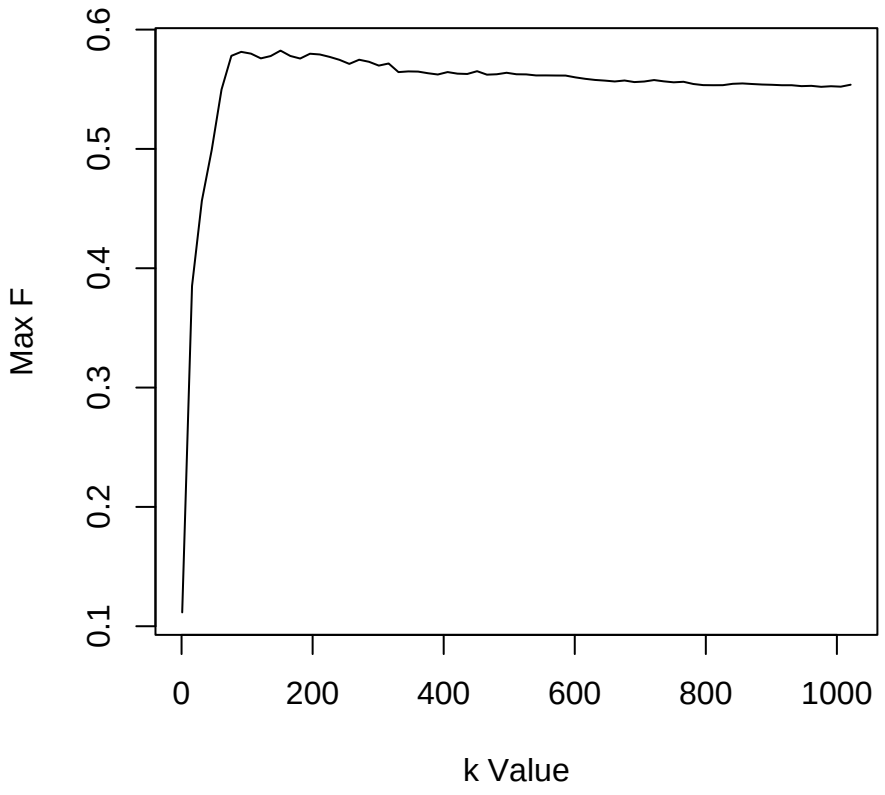

FIGURE 3.2.2. Optimal F for increasing k-values on *MEDLINE* data

attempts to discover which of the r eigenvalues derived from the $n \times p$ matrix $\mathbf{A}$ are "significant." Dimensionality reduction then proceeds by discarding singular vectors associated with insignificant eigenvalues. As the name suggests, APA is based on Horn's so-called parallel analysis which was introduced in Section 2.3. The technique involves estimating for each eigenvalue λ_k its departure from λ_{0k}, the k^{th} eigenvalue expected if the variables (columns) of $\mathbf{A}$ were statistically independent. The goal of APA is to reject those principal components (or singular vectors, in the case of LSI) whose corresponding eigenvalues are significantly less than the eigenvalues expected under the null case of independence. However, the proposed method differs from Horn's insofar as we derive confidence intervals for

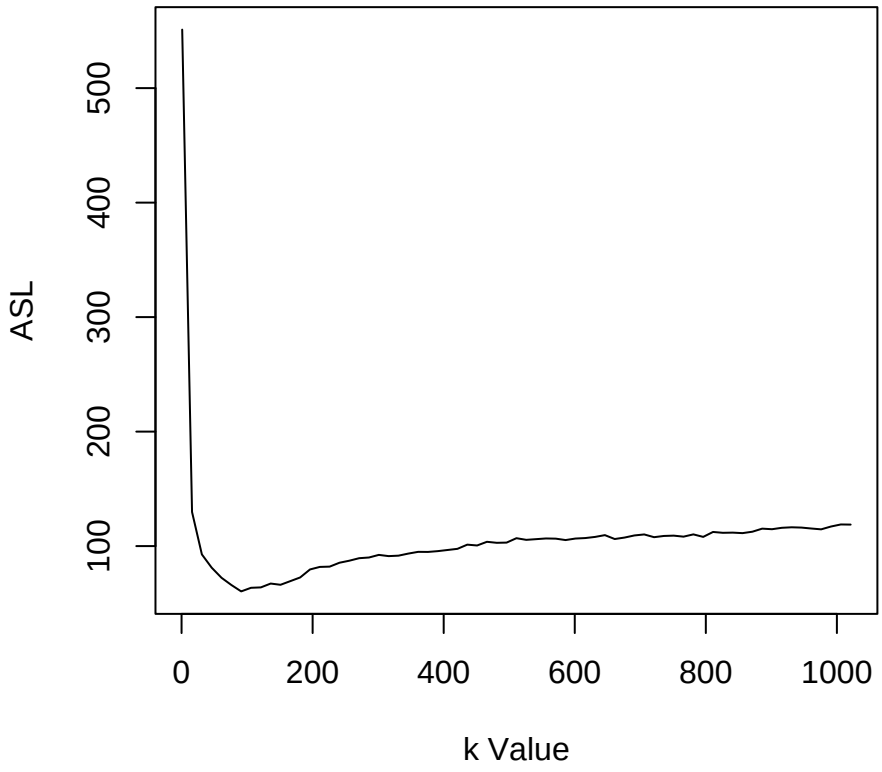

FIGURE 3.2.3. ASL for increasing k-values on *MEDLINE* data

λ_{0k}. Under APA a given eigenvalue λ_k is rejected if it lies below the lower bound of the $1 - \alpha\%$ confidence interval for λ_{0k}.

Defining the null case in terms of confidence intervals leads APA to offer equal or higher dimensionality estimates than those afforded by traditional parallel analysis. This is somewhat at odds with the mainstream of research on parallel analysis. As described in [**47**] and [**57**], a number of statisticians have noted the tendency of parallel analysis to overestimate the number of factors. These researchers have proposed methods of adapting parallel analysis to produce lower dimensionality estimates. However, I argue that the IR task—with its extremely large, sparse matrices—presents a dimensionality estimation problem that is

substantively different from the psychometric tasks studied in the mainstream PA literature. Initial research suggested that parallel analysis tends to underestimate the intrinsic dimensionality for IR data. Thus amended parallel analysis seeks to moderate PA's dimensionality truncation. As described in the following sections, this approach not only yields larger models (as I hypothesize IR requires), but also brings the apparatus of traditional hypothesis testing to the problem of dimensionality estimation.

3.3.1. The Method of APA. To estimate the intrinsic dimensionality of the $n \times p$ matrix $\mathbf{A}$ we begin by computing the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_r$ of the $p \times p$ covariance matrix $\mathbf{S}$ of rank r. As discussed in Section 2.3 each eigenvalue λ_k is the amount of variance captured by the k^{th} principal component of $\mathbf{S}$. To estimate the intrinsic dimensionality of $\mathbf{A}$ we wish to discriminate between "large" eigenvalues and "small" ones. Thus the problem involves deriving a suitable criterion for judging the significance of principal components based on the magnitude of their corresponding eigenvalues.

APA begins with the assumption, noted in **[3, 71, 79, 116]** that the utility of dimensionality reduction is predicated on the presence of correlation among the variables of $\mathbf{A}$. That is, if the p columns of $\mathbf{A}$ were independent there would be no room for representational improvement by projecting $\mathbf{A}$ onto a low dimensional subspace because the columns of $\mathbf{A}$ would, in such a case, already be its principal components. For example, consider a data set that possesses five independent variables, represented by the $n \times 5$ matrix $\mathbf{A}$, with covariance matrix $\mathbf{S} = \mathbf{I}_5$:

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

As an identity matrix $\mathbf{S}$ is orthogonal, so that $\mathbf{S}'\mathbf{S} = \mathbf{S}\mathbf{S}' = \mathbf{I}$. A standard result from linear algebra reminds us that the eigenvalues of a diagonal matrix are in fact the elements that appear on its main diagonal. Thus constructing a scree plot for the principal components

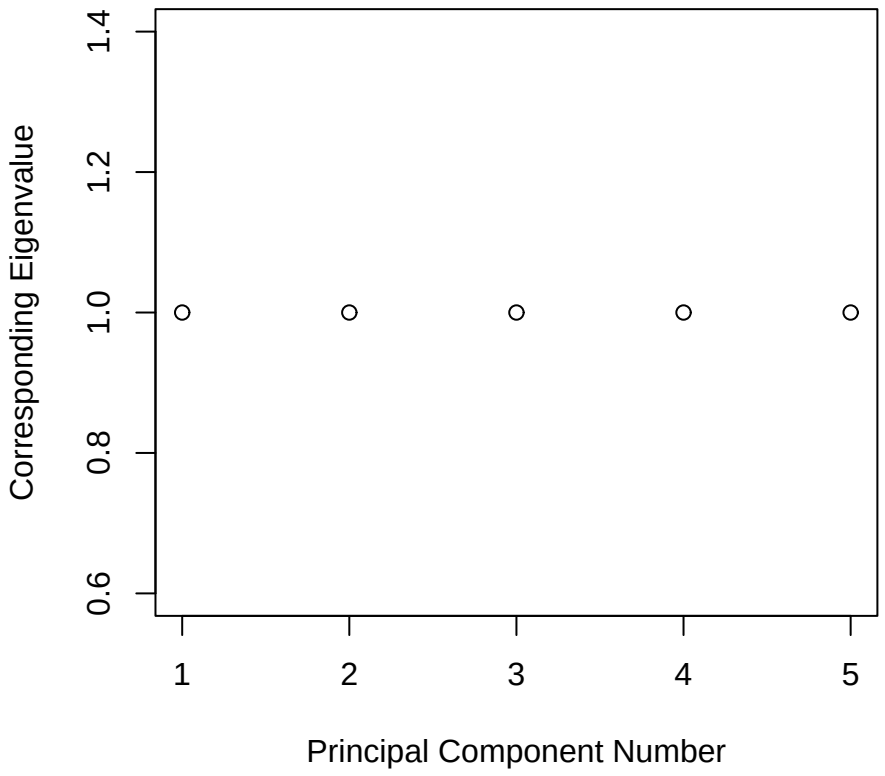

FIGURE 3.3.1. Scree plot for orthogonal data

of **A** yields Figure 3.3.1. This scree plot shows that the eigenvalues of **S** divide the variance of the matrix equally. Because the columns of **A** are orthogonal to begin with, there is no redundancy among the data. Thus the principal components are equally significant, and there is no rationale for discarding any of them.

In practice, such a matrix as **A** is unlikely to exist. Even a matrix generated from an independent probability distribution will display some opportunistic correlation due to sampling error. It is the degree to which such an independent matrix deviates from true orthogonality that bears on the slope of its scree plot. To examine this phenomenon, consider Figure 3.3.2, which displays scree plots for three simulated 10-variate data sets. Each data set was drawn from a distribution with mean vector equal to 0 and covariance matrix $\mathbf{I}_{10}$.

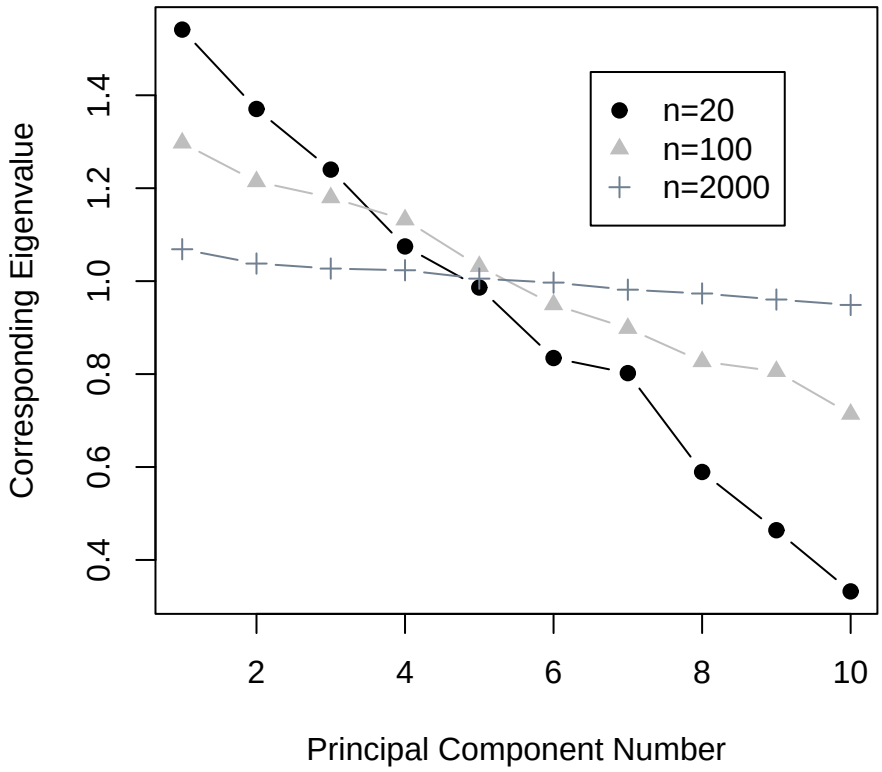

FIGURE 3.3.2. Scree plots of simulated independent data

The black points in Figure 3.3.2 are the eigenvalues from a data set with only 20 observations. On the other hand, the grey triangles derive from 100 observations. Finally, the dark grey pluses were obtained from a sample of 2000 data points. For each sampled data set let **S** be the observed covariance matrix. Because of the law of large numbers, as the number of observations in the sample grows, **S** will approach the identity matrix that parameterized the data. Thus Figure 3.3.2 shows that as the data approach independence, their scree plot will become less steep. Figure 3.3.1 marks the extreme case of this behavior. Amended parallel analysis operates by assuming that we desire to reduce dimensionality to account for term correlations, and that we may estimate the degree of term correlation in the data by recourse to analysis of the slope of the corpus' scree plot.

Assuming that dimensionality reduction is warranted due to correlation among variables, APA operates by rejecting principal components whose eigenvalues are significantly smaller than the corresponding eigenvalues expected under the null hypothesis: *the variables of* $\mathbf{A}$ *are independent.* As in Horn's parallel analysis, amended PA uses the fact that the eigenvalues of a matrix with independent variates will decrease in magnitude at a different rate than the eigenvalues of a matrix with significant inter-variable correlation (as shown above in Figures 3.3.1 and 3.3.2). As described in Section 2.3.3 the point where the observed eigenvalues become less than the null eigenvalues gives an estimate of the amount of error incurred by application of the VSM's assumption of term independence.

To estimate the magnitude of eigenvalues under the null hypothesis, parallel analysis uses a statistical simulation. Let $\boldsymbol{\mu}$ be the p-dimensional mean vector of $\mathbf{A}$ and $\mathbf{s}$ be the vector of variances for each column of $\mathbf{A}$. Also let $\boldsymbol{\sigma} = \mathbf{s}'\mathbf{I}_p$ be a $p \times p$ diagonal matrix with the variances of $\mathbf{A}$ on the main diagonal. We thus generate the $n \times p$ matrix $\mathbf{A}_0^*$ by sampling n p-vectors from the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma}$. The matrix $\mathbf{A}_0^*$ is drawn from a distribution very similar to the estimated distribution of $\mathbf{A}$, with the caveat that the distribution of $\mathbf{A}_0^*$ has no parameterized covariance among the data (sampling error will generate some erroneous covariance). Having obtained the matrix drawn from an independent distribution, we calculate $\mathbf{S}_0^*$ the sample covariance matrix of $\mathbf{A}_0^*$. Next we find $\boldsymbol{\lambda}_0^*$, the vector of eigenvalues of $\mathbf{S}_0^*$. This vector comprises an estimate of the eigenvalues expected if $\mathbf{A}$ had independent variables.

Parallel analysis proceeds by calculating $\boldsymbol{\lambda}_0^*$ many times, across a number of generations of $\mathbf{A}_0^*$. Having calculated $\boldsymbol{\lambda}_0^*$ B times, we finally calculate $\widehat{\boldsymbol{\lambda}}_0^*$, the average of our B observations of $\boldsymbol{\lambda}_0^*$. Thus $\widehat{\lambda}_{01}^*$ is the average magnitude of the first eigenvalue of $\mathbf{S}_0^*$ across B samples.

Horn defines significant eigenvalues to be those that are greater than their corresponding entry in $\widehat{\boldsymbol{\lambda}}_0^*$. As discussed above, Figure 2.3.5 gives a visual representation of traditional parallel analysis. Shown in black are the eigenvalues derived from the athletic physiology data, while in grey are the values of $\widehat{\boldsymbol{\lambda}}_0^*$ after $B = 100$ simulations. Parallel analysis instructs us to discard all eigenvalues to the right of the point where the two scree plots in Figure

2.3.5 cross each other. Horn's approach retains the first and second principal components because they are greater than the first and second entries of $\widehat{\boldsymbol{\lambda}}_0^*$.

Parallel analysis is a powerful method for dimensionality selection. It has been shown to perform well on a variety of data [**1, 76, 142**]. However, Horn's method is prone to error due to the somewhat arbitrary nature of the cutoff point that it defines. The method defines as crucial the point where the scree plot of observed data crosses the scree plot of the simulated, independent data. Another way to understand this is to consider that parallel analysis is concerned with the distance between each λ_k and $\widehat{\lambda}_{0k}^*$. Under Horn's method, we retain all principal components where the corresponding $\lambda_k - \widehat{\lambda}_{0k}^* > 0$. Let $G_k = \lambda_k - \widehat{\lambda}_{0k}^*$ be a random variable of unspecified distribution with mean $\lambda_k - \widehat{\lambda}_{0k}^*$ and variance σ^2. Horn's method gives a point estimate of G. Traditional parallel analysis retains all principal components with corresponding $G = g > 0$. This is unsatisfying insofar as basing our decision on g, a point estimate, fails to account for the variability of G.

The proposed method of amended parallel analysis answers this problem by deriving confidence intervals for $\widehat{\lambda}_{0k}^*$. By extension this gives a confidence interval for G. APA thus operates by retaining all principal components whose λ_k is significantly less than the corresponding $\widehat{\lambda}_{0k}^*$, in the traditional statistical sense. Thus the researcher may derive, say, a 95% confidence interval on G, which results in APA rejecting those principal components for whom the corresponding λ_k lies below 95% confidence interval for $\widehat{\lambda}_{0k}^*$. This technique is useful insofar as it applies a more rigorous rationale to the prediction of the intrinsic dimensionality than one based on a point estimate. It is also helpful because it leads to higher estimations of k_{opt} than does traditional, point-based parallel analysis. I anticipate that in the highly complex arena of IR, a less aggressive dimensionality is merited than traditional PA is prone to give. Thus the confidence interval-based approach described here constitutes a well motivated amendment to horn's technique that will improve its applicability to the task at hand.

3.3.1.1. *The Idea of Confidence Intervals.* Confidence intervals (CI's) are integral to the field of inferential statistics [**11**]. Closely related to hypothesis testing, CI's bring the theory of probability to bear on the matter of predicting a population parameter based on a statistic

calculated on a random sample. For example, consider a researcher who wishes to estimate the average height of Americans. Let X be the random variable height, which we shall assume is normally distributed $N(\mu\,\sigma)$. Thus μ is the true average height of Americans, while σ is the true standard deviation of X. To estimate μ we might take a random sample of, say, 1000 Americans. For each member of the sample, we calculate $X = x$, his or her height. In such an experiment we derive a point estimate of μ by:

$$\overline{x} = \frac{\sum_{i=1}^{1000} x_i}{1000}$$

which is simply the sample mean height. Due to the distribution of X and the unbiasedness of the statistical average, $E(X) = E(\overline{x}) = \mu$. Inferential statistics allows us to estimate the accuracy of our estimate $\overline{x}$. That is, having estimated μ by calculating $\overline{x}$, we now wish to know something about the relationship between our statistic and the parameter it estimates.

Crucial to understanding confidence intervals is the notion of a statistic's standard error. To find this, we consider our calculated average $\overline{x}$ to be a sample from the random variable $\overline{X}$, the population of means calculated on samples from X. What we wish to know, then, is how likely is it that our observed $\overline{x}$ is close to the population parameter μ. By treating $\overline{X}$ as a random variate in its own right with $\overline{x}$ as a point estimate of $\overline{X}$, we can use probability theory to estimate their relation.

The central limit theorem of univariate statistical theory **[11, 44]** states that the distribution of the statistic $\overline{X}$ is normal:

$$\overline{X} \propto N(\mu\,\frac{\sigma}{\sqrt{n}})$$

The expected value of $\overline{X}$ is thus the population parameter μ. However, the standard deviation of $\overline{X} = \frac{\sigma}{\sqrt{n}}$. Thus the expected values of X and $\overline{X}$ are the same, μ. However, the standard deviation of $\overline{X}$ is smaller than the deviation of X. This is why averages are useful; they use our data to tell us more about μ by virtue of having less spread around the true parameter value. In this example, the standard deviation of $\overline{X} = \frac{\sigma}{\sqrt{n}}$; thus as n becomes large, observations $\overline{x}$ from $\overline{X}$ will be centered more closely around the population mean μ. The standard deviation of a statistic is called the *standard error* of the statistic.

Speaking more formally, let μ be an unknown parameter of the population X, from which we have a random sample $X_1, X_2, \cdots X_n$. A confidence interval for μ is an interval $C = (L, U)$ that includes the unknown true value of μ with a pre-specified probability $1 - \alpha$. Formally, we have:

$$P[L < \mu < U] = 1 - \alpha.$$

Thus $C = (L, U)$ gives a $100(1 - \alpha)\%$ confidence interval for μ.

To construct a 95% confidence interval we begin with an unbiased point estimate of μ, $\overline{x}$. We also consider the standard error of $\overline{X} = \frac{\sigma}{\sqrt{n}}$. Due to the central limit theorem we know that $\overline{X}$ will tend to be centered around μ. The definition of the Normal distribution states that with probability 0.95 a normal random variable will lie within 1.96 standard deviations of its mean. Thus for $\overline{X}$:

$$P[\mu - 1.96\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}] = 0.95$$

which is equivalent to:

$$P[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}] = 0.95$$

which in turn leads to the following confidence interval for μ:

$$(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

where 1.96 is the so-called *z-score* for the 95% confidence level. This interval informs us that with probability 0.95, the population mean μ lies within 1.96 standard errors of our observed point estimate $\overline{x}$.

Returning to the matter of amended parallel analysis, we wish to obtain confidence intervals on statistics other than average height. In this case we wish to estimate the accuracy of the statistic $\widehat{\lambda}^*_{0k}$, the k^{th} element of the vector of eigenvalues averaged across simulations of independent data.

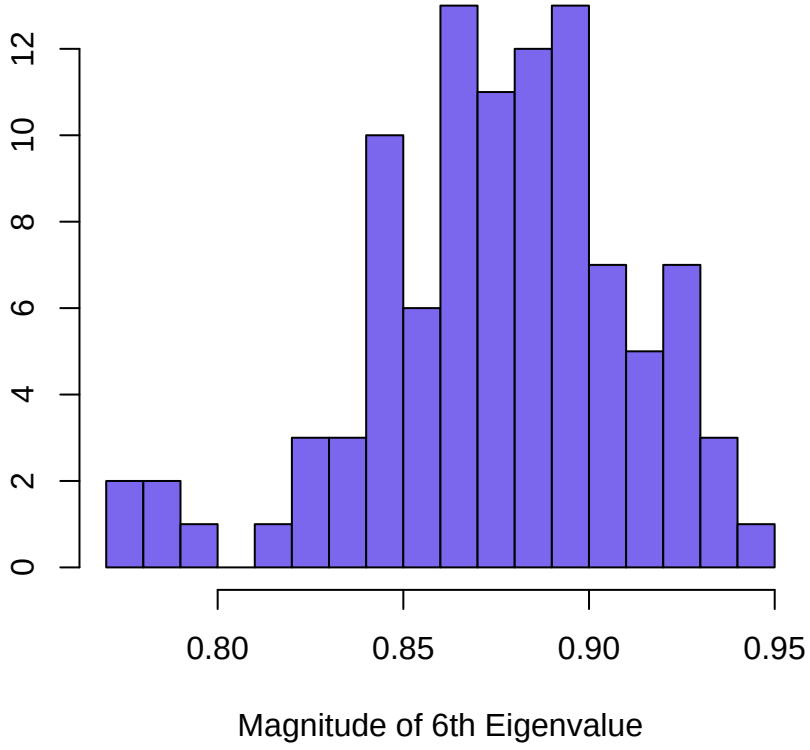

FIGURE 3.3.3. Histogram of λ_{06} ($B = 100$) for athletic data

It would be tempting to fashion a 95% confidence interval for $\widehat{\lambda}^*_{0k}$ via:

$$(\widehat{\lambda}^*_{0k} - 1.96\frac{s}{\sqrt{B}},\ \widehat{\lambda}^*_{0k} + 1.96\frac{s}{\sqrt{B}})$$

where s is the standard deviation of the k^{th} eigenvalues across B simulations. However, such a confidence interval is unjustified in the case of parallel analysis because we have little reason to believe that $\widehat{\lambda}^*_{0k}$ is normally distributed. Consider Figure 3.3.3, which shows a histogram of the magnitude of the $\widehat{\lambda}^*_{06}$ after 100 simulations on the athletic physiology data. The distribution appears skewed to the right and possibly bimodal. This informal check for normality is in agreement with a large body of literature that suggests that eigenvalues cannot be assumed to follow a Gaussian distribution [**105, 80, 147, 42**]. Without the

Symbol	*Meaning*
$\mathbf{X}^*$	Boostrap sample from $\mathbf{X}$
θ^*	Statistic θ calculated from bootstrap sample
$\theta^*(b)$	Caclulation of θ from the *bth* iteration of $\mathbf{X}^*$

TABLE 3.3.1. Bootstrap confidence interval notation

assumption of normality, adopting the 95% Z-score 1.96 is unfounded and is thus likely to give an improper confidence bound.

3.3.1.2. *Bootstrap Confidence Intervals.* Without some idea of the distribution of our eigenvalues, we turn to non-parametric methods of deriving confidence intervals. In particular, APA makes use of the bootstrap [**43, 44**], a computer-intensive resampling method that allows us to estimate the standard deviation of an eigenvalue λ_k without recourse to assumptions about its distribution. Resampling techniques estimate the variability of a statistic θ by drawing observations (with replacement) from a sample $\mathbf{X}$ to derive a simulated sample $\mathbf{X}^*$. Having constructed $\mathbf{X}^*$, bootstrap analysis proceeds by calculating θ on $\mathbf{X}^*$ to find θ^*. By repeating this process B times, one can achieve a quantifiable appreciation of the variability of θ without depending on parametric assumptions.

Before proceeding I offer a brief summary of notation related to the bootstrap. Table 3.3.1 shows several symbols that inform the following analysis. Thus, for instance, $\mathbf{X}^*(b)$ is the b^{th} bootstrap sample taken by selecting n rows at random (with replacement) from the $n \times p$ matrix $\mathbf{X}$.

Bootstrapping is common in many areas of inferential statistics, and it is not surprising that it has found its way into the problem of selecting significant principal components. For instance Lambert *et al.* use bootstrapping of eigenvalues to augment the eigenvalue-one criterion [**88**]. They suggest that the stopping point of $\lambda = 1$ is arbitrary, and that our cutoff should account for the variability of each λ. Thus they use the bootstrap to derive a confidence interval for each λ_k, suggesting that those principal components should be retained whose corresponding λ_k has a confidence interval that is entirely greater than 1. Likewise Jackson uses bootstrap analysis of λ to augment the hypothesis test approach described earlier as Bartlett's test of isotrapy [**76**]. Arguing that we should discard all eigenvalues that are not appreciably different from their neighbors, Jackson suggests that

we retain those principal components whose eigenvalues' confidence intervals lie outside the confidence intervals of their neighbors. Thus Jackson's approach amounts to a non-parametric hypothesis test of equality between λ_k and λ_{k+1}.

Under parallel analysis we wish to derive confidence intervals for $\boldsymbol{\lambda_0}$, the eigenvalues of independent data as specified by Horn's $\mathbf{A}_0$. To begin building the confidence interval, we define the following variables. Let $\mathbf{A}_0^*(b)$ be the b^{th} bootstrap sample of $\mathbf{A}_0$. Let $\widehat{\lambda}_{0k}^*$ be the average magnitude of the k^{th} eigenvalue across our B samples. Also define the standard error of $\widehat{\lambda}_{0k}^*$:

$$\widehat{se}_k = \{\sum_{b=1}^{B}[\lambda_{0k}^*(b) - \widehat{\lambda}_{0k}^*]^2/(B-1)\}^2 \tag{3.3.1}$$

where $\lambda_k^*(b)$ is the k^{th} eigenvalue of the b^{th} bootstrap sample.

To construct the confidence interval we use the so-called bootstrap-t approach, discussed in [**44**, ch. 12] (on which the following discussion draws heavily). In the discussion above, I noted that its lack of normality prohibits us from adopting the standard Z-score in our estimation of the variability of the statistic $\overline{\lambda}_{0k}^*$. Under the bootstrap-t method, we instead calculate $t^*(b)$, a non-parametric estimate of the likelihood of observing the b^{th} observation of λ_k^*:

$$t^*(b) = \frac{\lambda_{0k}^*(b) - \widehat{\lambda}_{0k}^*}{\widehat{se}_k}. \tag{3.3.2}$$

We thus find the α^{th} percentile of $t^*(b)$ by the value $\widehat{t}^{(\alpha)}$ (i.e. the bootstrap-t) such that

$$\#\{t^*(b) \leq \widehat{t}^{(\alpha)}\}/B = \alpha. \tag{3.3.3}$$

In other words if we have $B = 100$ bootstrap iterations, the estimate of the fifth percentile point is the fifth largest value of $t^*(b)$ and the 95th percentile is given by the 95th largest $t^*(b)$. This approach essentially allows us to construct a pseudo-probability table, tailored to the distribution of the observed data. In other words, we observe the variability of our test statistic over a wide number of iterations, generating $t^*(b)$ for each of our $B = b$ samples. Based on these calculations we derive probability estimates.

Having thus used our pseudo-table of $t^*(b)$ values to derive an appropriate $\widehat{t}^{(\alpha)}$, our bootstrap-t confidence interval is given by Equation 3.3.4.

$$(\widehat{\lambda}^*_{0k} - \widehat{t}^{(\alpha)},\ \widehat{\lambda}^*_{0k} - \widehat{t}^{(1-\alpha)}). \tag{3.3.4}$$

So with probability $1-\alpha$ we state that if given infinite data from the same distribution that gives $\mathbf{A}$, the k^{th} null eigenvalue λ_{0k} would lie within the interval specified by Equation 3.3.4.

Performing APA involves rejecting those principal components whose associated eigenvalues λ_k lie below the confidence interval of the corresponding λ_{0k}. Whereas Horn's traditional parallel analysis gives us a point estimate of the cutoff between meaningful and non-meaningful principal components, amended parallel analysis supplements this point estimate by accounting for the sampling distribution of the null eigenvalues. Thus APA operates by rejecting those principal components whose variance—i.e. whose eigenvalue—is *significantly smaller* than the variance of principal components expected under the assumption of term independence.

3.3.2. An Alternate understanding of APA. It may be objected that the distribution of two null eigenvalues λ_{0k} and $\lambda_{0k'}$ $k \neq k'$, are not independent. Therefore taking separate confidence intervals on each statistic is inappropriate. However, I argue that the effect of this problem is mitigated by the fact that the variables of $\mathbf{A}_0$ are by definition independent, and therefore any correlational structure in $\mathbf{A}_0$ is due to sampling error. Thus the sampling distribution of a given null eigenvalue λ_{0k} will be negligibly affected by the distribution of the remaining null eigenvalues.

To demonstrate that this is the case, I offer an alternative definition of the APA procedure that obviates the problem of null eigenvalue correlation. Applications of this alternate version of APA yield approximately the same dimensionality estimate as the articulation given above, modulo a small implementation-specific difference.

Instead of calculating a confidence interval on each null eigenvalue, one may estimate the data's intrinsic dimensionality by calculating $D^*(b)$, the value of k at which the observed eigenvalues become smaller than the null eigenvalues generated by the b^{th} bootstrap sample. We may think of this as the distance along the x-axis of a scree plot between the origin and

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MEDLINE*
Standard APA	698	76	45	80	94	87
Alternate APA	698	75	45	80	88	83

TABLE 3.3.2. Estimates derived by two implementations of APA

the point of crossing. Calculating D^* over B replications gives a vector of estimates of D, the true crossing point of observed and null eigenvalues. Using the bootstrap-t approach described above, then, one can construct a $1-\alpha\%$ confidence interval on D. That is, the standard error of D^* is given by:

$$\widehat{se}_D = \{\sum_{b=1}^{B}[D^*(b) - \overline{D}^*]^2/(B-1)\}^2 \tag{3.3.5}$$

where $\overline{D}^*$ is the mean of the B observations of D^*. Given this standard error, confidence intervals are based on the bootstrap-t score given in Equation 3.3.6.

$$t^*(b) = \frac{D^*(b) - \overline{D}^*}{\widehat{se}_D} \tag{3.3.6}$$

And a $1-\alpha\%$ confidence interval is computed by:

$$(\overline{D}^* - \widehat{t}^{(\alpha)}, \overline{D}^* - \widehat{t}^{(1-\alpha)}) \tag{3.3.7}$$

where $\widehat{t}^{(.)}$ is given above in Equation 3.3.3.

In terms of a scree plot, the original definition of APA given above is concerned with the relationship between the observed eigenvalues and the null eigenvalues on the y-axis. This alternate definition of APA is concerned with the x-axis. This approach obviates the the matter of null eigenvalue correlation insofar as it takes only a single measurement D^* at each of the B bootstrap replications.

The single-measure approach to APA yields a nearly identical solution to the method proposed above. Table 3.3.2 shows the estimates of intrinsic dimensionality for each our the six corpora tested in this experiment. Clearly the two definitions of APA yield very similar results. Moreover, I argue that the results could be made even more similar by means of interpolation. That is, the alternate definition of APA measures D^*, which is an integer-valued variable, corresponding to a particular dimensionality. On the other hand, the original

definition of APA is concerned with λ^*_{0k}, which is real-valued. This difference in measurement adds to the divergence of estimates given by each of the methods. If instead of individual points corresponding to each eigenvalue, D were measured on a continuous function defining the distribution of observed and null eigenvalues in a scree plot, the actual point of crossing between real and null eigenvalues might occur between two values of k. By interpolating the slope of each function it would thus be possible to improve the correspondence between each approach to APA. Even without such interpolation, however, the methods yield very similar answers. A paired t-test on equality of treatments of the values given in table3.3.2 yielded $p = 0.14$, suggesting that the two methods are statistically indistinguishable for our data. For the simulated data described in Chapter 5 both definitions of APA gave identical results for all parameterizations.

Although the alternate definition of APA is attractive insofar as it obviates the problem of null eigenvalue correlation, in the remainder of this dissertation I report results obtained by pursuing the first articulation of the method. I prefer the original definition of APA because it does not necessitate any interpolation. As noted in the previous paragraph, neither method is statistically different from the other; thus I retain the original definition in the remainder of this study.

3.3.3. An Example of Amended Parallel Analysis. Figure 2.3.5 shows a visual representation of traditional parallel analysis as performed on the athletic physiology data. In this example, Horn's method dictates that we retain the first two principal components because the scree plots of the observed data and the data generated under the null hypothesis of independence overlap after eigenvalue number two.

To perform amended parallel analysis on these data, we begin with the same scree plot as Horn's. Thus we note that the observed eigenvalues are:

$$\boldsymbol{\lambda}' = \begin{pmatrix} 1.6 & 1.17 & 0.97 & 0.82 & 0.57 & 0.36 \end{pmatrix}.$$

The null eigenvalues appear in Table 3.3.3, along with corresponding 95% confidence intervals. These confidence intervals were obtained via the bootstrap-t method described above, after 100 resampling iterations.

95% CI	λ_{01}	λ_{02}	λ_{03}	λ_{04}	λ_{05}	λ_{06}
Lower Bound	1.10	0.98	0.98	0.92	0.81	0.69
Observed eigenvalue	1.17	1.11	1.06	0.96	0.86	0.79
Upper Bound	1.27	1.16	1.08	0.99	0.95	0.86

TABLE 3.3.3. Confidence intervals on simulated null data

Under amended parallel analysis, we begin inspecting eigenvalues at the small end and work backwards. Thus we note that with 95% confidence $0.69 \leq \lambda_{06} \leq 0.86$. Because the corresponding observed eigenvalue is smaller than the lower bound on λ_{06} we reject λ_6. We continue in this fashion, rejecting each eigenvalue that lies below the corresponding null eigenvalue, until we reach λ_3. Note that $\lambda_3 = 0.97$ lies inside the confidence interval for λ_{03}. Thus we conclude that the third eigenvalue is not significantly smaller than the corresponding null eigenvalue. Moreover, all the remaining eigenvalues follow suit. Thus we conclude that the remaining eigenvalues are significant and should be retained. Under APA we would therefore conclude that the fourth through sixth principal components are not significantly different in magnitude than what we would expect under the null case of independence. As such we reject them, concluding that only the first three components of the athletic physiology data are significant, and set $k = 3$.

Figure 3.3.4 depicts the amended parallel analysis process. As in Figure 2.3.5, eigenvalues for the observed data appear as black dots, while the simulated null eigenvalues appear as grey triangles. Unlike in Horn's analysis, however, Figure 3.3.4 also shows the confidence intervals for each eigenvalue. Figure 3.3.4 thus re-enforces the point discussed above; APA leads us to estimate the intrinsic dimensionality of the athletic data as 3, a figure that is greater than the estimate of 2 provided by Horn's method. I argue that a dimensionality truncation less aggressive than Horn's is in order due to the variability of eigenvalues derived during the bootstrap resampling phase of this analysis. Thus I argue that APA improves upon traditional parallel analysis by accounting for null-eigenvalue sampling error.

3.4. Methods of Data Analysis

This section offers a brief overview of the methods of data analysis that were undertaken during the experimental phase of this study. A complete report is given in Chapter 4. My

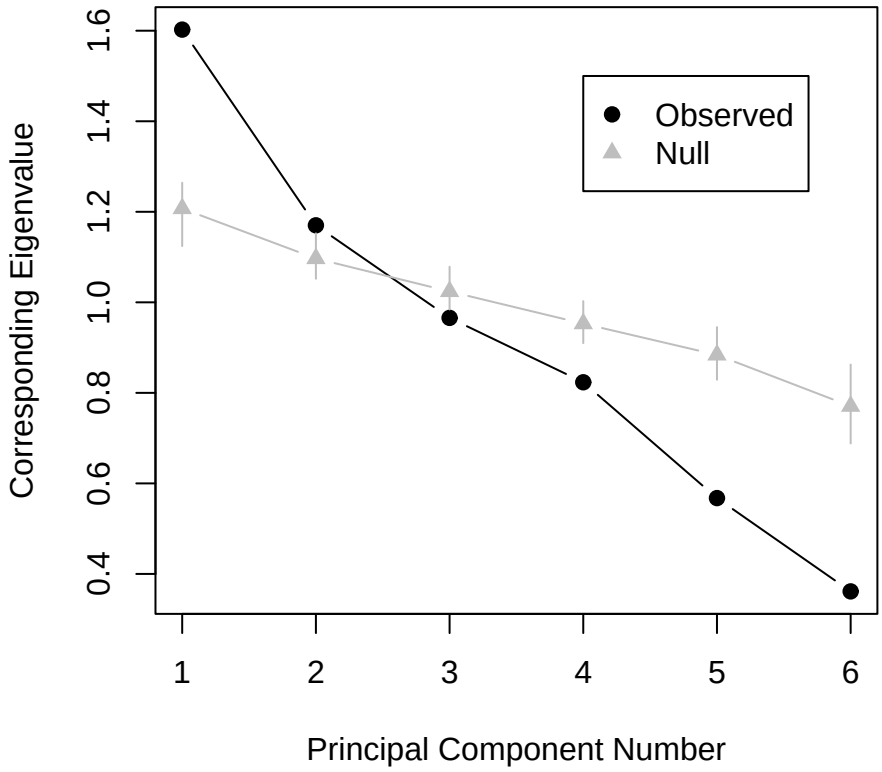

FIGURE 3.3.4. Amended parallel analysis on the athletic physiology data

data analysis involved two major efforts, which were mutually reinforcing. First, I analyzed LSI's performance at a range of dimensionalities on six standard test collections. This yielded an estimate of each collection's optimal dimensionality for IR. Next I attempted to determine how each eigenvalue-based predictor correlated with the performance-based analysis of optimal k. In other words, if average precision, ASL, and optimal F all suggest that $k_{opt} \approx 200$ for a given data set, my analysis judged how close to 200 were the predictions derived from the eigenvalue-one criterion, the percent-of-variance rule, Bartlett's test of isotrapy, parallel analysis, and APA.

My analysis employed the six test collections described in Table 3.1.1. For each collection I estimated the intrinsic dimensionality by employing the eigenvalue-based predictors shown

Name	*Abbreviation*
eigenvalue-one	EV1
Bartlett's test of isotrapy ($\alpha = 0.01$)	Bartlett's
percent of variance=80	85% Var
parallel analysis	PA
amended parallel analysis ($\alpha = 0.05$)	APA

TABLE 3.4.1. Selected eigenvalue-based metrics for experimentation

in table 3.4.1. Details about each of these metrics are discussed above in Section 2.3.3. With regard to parameterization of these methods, I adopted the following rationale. Bartlett's test is known to admit spurious principal components. As such, I used it only with a very high confidence level, $\alpha = 0.01$. At the outset of this research I anticipated testing the percent-of-variance criterion at a variety of parameterizations. However, preliminary research suggested that an 85% approach yielded results that were consistently superior to other likely percentages, such as 70% or 95%. Thus in my description of the percent-of-variance criterion's performance (Section 4.2.2.3) I describe how other parameterizations would have worked for each data set, but I omit extended consideration of these alternate parameterizations.

APA also requires that the researcher choose a confidence level prior to analysis. To motivate my analysis, I define $\alpha = 0.05$ as a baseline throughout this experiment. In the following data analysis, unless otherwise specified, descriptions of APA assume a 95% confidence interval. However, APA is a new approach to dimensionality estimation, and so in Section 4.2.2.1 I offer an analysis of how choices of α affect APA's accuracy.

This dissertation is concerned with the suitability of eigenvalue-based dimensionality estimators for information retrieval applications. To address my research question, the study used the six test collections, the three performance metrics, and five dimensionality estimation techniques described in this chapter. As described in Chapter 4, the goal of my data analysis was twofold. First, I analyzed each corpus with respect to its observed optimal dimensionality. The second stage of analysis involved a comparison of each eigenvalue analysis technique. If a given corpus evinced a clearly optimal semantic subspace of k_{opt} dimensions, I attempted to judge how well each eigenvalue selection criterion estimated this intrinsic dimensionality? In light of the complex data derived from this process, Chapter 5 offers a

more simplified analysis of the dimensionality estimation problem, focusing attention on the performance of eigenvalue-based estimators on simulated data.

3.5. Computational Tools

Before turning to the results of these experiments, however, a discussion of the computational resources that enabled this research is in order. A variety of hardware and software was used to perform the described analyses. This section reviews salient infrastructural choices that I have made, with an eye toward offering help to future researchers interested in pursuing the matter of dimensionality estimation. I first describe the hardware that was used to conduct the experiments. Next I introduce the software, describing the programming environments, pre-existing software, and my own scripts that were used to conduct this study. Finally, I address several choices related to the indexing of documents, such as the use of stop-lists, term weighting schemes, morphological analysis, etc.

All of the experimentation was performed on a Sun 280R Sun Fire server. The machine is equipped with 4 gigabytes of memory, 600 gigabytes of disk, and dual 770 Sparc III processors each of which uses a sparcv9 floating point processor. This configuration allowed the system to perform a singular value decomposition (using software described below) on a 5831-term by 1033-document matrix (retaining all 1033 factors) in under an hour of CPU time.

To perform the singular value decomposition and attendant information retrieval tasks, I used the LSI software package available from Telcordia Industries. Although LSI is a proprietary product, the software is available under a limited license for research purposes. Licenses may be requested via the Internet at `http://lsi.research.telcordia.com`. The LSI software contains several modules, all of which I make ample use. First, the suite contains a robust program called *pindex* that converts a corpus of documents into a term-document matrix. Given a properly formatted test collection, *pindex* generates a representation of the collection in the Harwell-Boeing sparse matrix format. *Pindex* uses this sparse matrix to perform the singular value decomposition via the single-vector Lanczos method (cf. [**9**, Section 3.4]), storing the resulting singular vectors and singular values in a compressed format.

Pindex is written in C. However, the limited research distribution does not provide source files, instead giving only pre-compiled binaries for the Solaris operating system.

The Telcordia LSI software also includes shell scripts for performing Cranfield-type IR experiments on a database derived by *pindex*. The script called *runeval* invokes several C programs that project stored queries into the LSI space derived by *pindex*. The script then ranks all documents against the query. As its output, *runeval* creates a three-column file containing the ranked output for each query (i.e. a separate file for each query). The first column of row i in this file is the document number of the i^{th} most similar document to the query. The second column shows the cosine similarity in k-space between the query and document i. The third column contains the relevance judgement for the i^{th} document against this query, given *a priori* by the test collection.

To interpret the output of *runeval* I have written Perl scripts to compute the performance measures discussed in Section 3.2. I have also written a wrapper script that allows me to invoke *runeval* many times, increasing the value of k used for each round of retrieval. In this way I was able to perform the analysis of observed optimal dimensionality discussed in Section 3.4. For a given data set, I thus perform Algorithm 1 to observe performance quality across a range of dimensionalities.

Algorithm 1 Algorithm for calculating observed optimal dimensionality

(1) For $k = 1...k_{max}$ repeat:
(2) For each query q: repeat:
(3) Calculate precision at four levels of recall for query q.
(4) Calculate ASL for query q.
(5) Calculate F for query q.
(6) Average each metric across all queries to derive a single value for $K = k$.

All of the software used for predicting k_{opt} based on eigenvalue analysis has been written using the R programming language. R is an open source statistical scripting language, available for download from `http://www.r-project.org` . It is an object-oriented programming environment that is based closely on the S language from Bell Labs. Although R lacks the speed and efficiency of compiled code, I have used it in my research due to its powerful features for conducting statistical analysis. Figure 3.5.1 shows an example R script. This program is used to calculate the bootstrap-t confidence interval described in Section 3.3.

```
# returns a vector of the upper and lower bounds for a 1-alpha% CI
   # on the eigenvalues specified in input vector thetaStar
   #
   # usage: bootstrapCI(lStar[,1], 0.05)

   bootstrapCI_function(thetaStar, alpha)  {
     # number of bootstrap samples
     b <<- length(thetaStar);
     cutoffDown <<- b * (1-alpha) - 1;
     cutoffUp <<- b * alpha - 1;

     thetaStarHat <<- mean(thetaStar);

     # get our deviations from the mean of thetaStar
     zStar <<- thetaStar - thetaStarHat;

     # get our estimate of the standard error of the stat thetaStar
     seStar <<- sqrt(sum((zStar)^2)/(nrow(lStar)-1));

     # get our vector of pseudo z-scores
     zStarVec <<- zStar / seStar;

     pctileDown <<- sort(zStarVec)[length(thetaStar) - cutoffDown];
     pctileUp <<- sort(zStarVec)[length(thetaStar) - cutoffUp];

     down <<- thetaStarHat - (pctileDown * seStar);
     up   <<- thetaStarHat - (pctileUp * seStar);

     return(c(up, down));
   }
```

FIGURE 3.5.1. Sample R code

Despite being an interpreted language (and despite its lack of sparse matrix optimization), R is useful for my analysis because of its native ability to treat statistical objects such as linear models, confidence intervals, and probability distributions.

When preparing each document collection for indexing and latent semantic analysis, a number of operational choices needed to be made. Table 3.5.1 lists these choices, along with the values that I selected for each parameter. The values shown for term length criteria were simply the defaults for the *pindex* software, and I saw no reason to alter them. Although it is likely that better retrieval is possible with another term weighting scheme I chose $tf \times idf$

Parameter	*Chosen value*	*Description*
min term length	2	minimum length (in characters) for an indexing term
max term length	1000	number of characters after which a term is truncated
local term weight	*tf*	weighting for the ith term in the jth document
global term weight	*idf*	overall weight for the ith term in the database
stop list	SMART	list of stop-words

TABLE 3.5.1. Text processing parameters for the study

due to its wide use in the IR literature. In a future study it would be interesting to compare predictions and observations of k_{opt} using various weighting methods, but here my goal is simply to keep this variable constant across data sets. To remove common words from the analysis, I employed the stop-list that comes standard with the SMART system. Like my adoption of $tf \times idf$, this choice stems not from any conviction that the SMART stop list is better than any other, but because it is commonly used in IR experiments and will thus improve the comparability between my study and other research.

3.6. Final Methodological Discussion

The experiments described here treated six standard IR test collections. The study measured both the observed optimal dimensionality (via Cranfield-style evaluation) and the optimal dimensionality predicted by five eigenvalue-based estimators. The goal of my analysis was to describe the suitability of each eigenvalue analysis technique for parameterizing operational IR systems. Thus the discussion in the following chapters begins with an attempt to ascertain the dimensionality of each corpus' optimal semantic subspace retrospectively. If such a subspace is in evidence, I measure the relationship between its optimal dimensionality and the estimates afforded by each eigenvalue-based predictor. I hypothesize that eigenvalue-based dimensionality estimators will provide strong evidence for parameterizing LSI models in the absence of *a priori* relevance judgements. In particular, I am eager to assess the performance of the proposed novel method of amended parallel analysis. Based on the already successful parallel analysis, APA promises even greater accuracy and stronger theoretical motivation due to its reliance on non-parametric hypothesis testing to derive an estimate of a given corpus' k_{opt}.

This is an ambitious agenda, but it marks only an initial step in a longer study. The research described here is concerned with finding optimal models of IR. Operating on a variety of corpora under a variety of performance criteria allows this research to address this question explicitly. However, LSI has implications for many other unsupervised learning applications such as synonym identification, word sense disambiguation, information filtering, and automatic classification. By way of full disclosure, then, before I discuss the outcome of this research, I remind the reader that the following analysis can hope to illuminate only a portion of its problem domain. In Section 6.4, I offer a more detailed account of exactly what this experiment speaks to and what my future efforts will address.

CHAPTER 4

Results and Analysis

The experiments described in Section 3 yielded a large, complex body of data. In this chapter I analyze these data to judge the utility of eigenvalues for estimating the intrinsic dimensionality of IR systems. In response to my initial research question, I found that a statistical analysis of a corpus' co-occurrence matrix eigenvalues provides useful information for optimizing the dimensionality of LSI systems. In particular, the proposed method of amended parallel analysis (APA) gave many accurate estimates of optimal model dimensionality while implying a strong theoretical basis for LSI. Overall, the family of dimensionality estimators comprised by APA, PA, and the EV1 criterion fared especially well. However, I also discovered that different corpora respond to dimensionality reduction in complex and idiosyncratic ways. Several corpora (e.g. *MEDLINE)* saw great improvement in retrieval performance after a 90% dimensionality reduction. On the other hand, corpora such as *CF,* performed best under the full-rank model. Still other data sets such as *CACM* and *CISI* sent conflicting messages. For *CACM* and *CISI* the ASL metric was optimized under a low-rank model, while average precision and the optimal F measure required models of higher dimensionality.

Because dimensionality reduction met with mixed success in my experiments, judging dimensionality estimators via retrospective, Cranfield-style analysis demands nuanced attention. To apply such attention, this chapter begins with a general comparison of IR performance under reduced-rank and full-rank models for each of the six test corpora described in Section 3.1. I address the ability of LSI to improve retrieval by discovering a low-dimensional model of each corpus, giving special attention to the way in which Cranfield-style performance evaluation bears on the ability to judge an LSI model's goodness of fit. In Section 4.2 I move to a systematic comparison of each eigenvalue analysis technique's dimensionality

estimates. Following this high-level comparison, Section 4.2.2 offers more detailed analyses of the strengths and weaknesses of each dimensionality estimation method. Finally, Section 4.3 summarizes the data presented here, articulating how my findings bear on the research questions articulated in Section 1 and on the theory of LSI more generally.

4.1. Evidence of Optimal Semantic Subspaces for IR

This section considers the effect of dimensionality reduction on IR performance for each of the test corpora. Overall I found evidence that dimensionality reduction improved retrieval moderately. Yet the dynamics of this improvement—and the concomitantly implied dimensionality of each corpus' optimal semantic subspace—was complex. This complexity is to be expected. Despite the intuitive and theoretical appeal of LSI, its ability to improve retrieval over the classic vector space model has never been conclusively proven. Instead, Deerwester *et al.* suggest that dimensionality reduction is well suited to certain corpora, while offering little benefit for others [**32**]. Parry Husbands, for instance, has argued that in its most basic articulation LSI is poorly equipped to model extremely large corpora [**73**]. On the other hand, Susan Dumais reports excellent performance in her application of LSI on several TREC experiments [**38, 39, 40**].

Confronted with these conflicting reports, it is not surprising that LSI's performance on the six corpora described in Section 3.1 ran a wide gamut. I found that aggressive dimensionality truncation was merited in the case of the *MEDLINE* data, for instance; while the full-text *CF* data brooked relatively little dimensionality reduction. In still other cases—e.g. *CACM* and *CRAN*—retrospective, Cranfield-style performance measures disagreed on optimal k, offering widely divergent interpretations of the system's intrinsic dimensionality.

4.1.1. Overview of Observed Optimal Dimensionality Findings. In this section I analyze the ability of LSI to discover what Ding terms an optimal semantic subspace of each corpus [**36, 37**]. At issue here is whether, for a given corpus, LSI's low-rank approximation of the term-document matrix $\mathbf{A}$ improved the similarity model over the full-rank model. If LSI did improve a system's similarity function, I attempt to ascertain which value of k led to the most marked improvement.

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
Docs	3204	1239	392	1460	1398	1033
Terms	5831	5116	9549	5615	4612	3204
k_{opt} *(ASL)*	271	1067	212	751	121	91
ASL at k_{opt}*(ASL)*	386.61	345.34	171.41	385.24	329.03	60.43
$k_{max} - k_{opt}$*(ASL)*	2933	172	180	709	1277	942
var at k_{opt}*(ASL)*	0.4	0.95	0.64	0.73	0.19	0.16
overfit (ASL)	-71.96	-6.65	-1.44	-4.52	-24.58	-58.33
k_{opt} *(Pr)*	1936	872	257	1276	661	151
$k_{max} - k_{opt}$*(Pr)*	1268	367	135	184	737	882
Pr at k_{opt}*(Pr)*	0.1375	0.0838	0.0446	0.1302	0.136	0.5599
var at k_{opt}*(Pr)*	1	0.87	0.74	0.96	0.71	0.25
overfit (Pr)	0.002	0.0003	0.012	0.0006	0.003	0.0366
k_{opt}*(F)*	2660	872	257	1411	811	151
$k_{max} - k_{opt}$*(F)*	544	367	135	49	587	882
F at k_{opt}*(F)*	0.2171	0.1391	0.068	0.2092	0.2289	0.5823
var at k_{opt}*(F)*	1	0.87	0.74	0.99	0.8	0.25
overfit (F)	0	0.0008	0.0038	0.0004	0.001	0.029

TABLE 4.1.1. Evidence of optimal semantic subspaces

Using Cranfield-style evaluation to gauge the utility of dimensionality reduction suggests that overall, reduced-rank models improved retrieval over full-rank models. However, the amount of improvement afforded by LSI and the amount of dimensionality reduction needed to obtain an optimal model varied widely across corpora. Table 4.1.1 summarizes my findings with respect to observed k_{opt}. For each of the three performance metrics detailed in Section 3.2 Table 4.1.1 reports five statistics: the value of k that led to optimal performance with respect to the measure, the amount of dimensionality reduction called for by the metric (i.e. $k_{max} - k_{opt}$), the actual value of the performance metric observed for $k = k_{opt}$, the proportion of total variance captured by this k-dimensional model, and the difference between performance at $k = k_{opt}$ and performance at $k = k_{max}$. Thus $k_{opt}(m)$ gives the observed optimal dimensionality for corpus c with respect to measure m. The row labeled *var at* $k_{opt}(m)$ gives the percent of total variance captured by measure m's optimal model. And *overfit(m)* approximates the strength of overfitting effect seen for measure m on corpus c.

Two important results are clear in Table 4.1.1. First, the rank of a data set and its observed optimal dimensionality appear to be approximately linear in their relationship.

The correlation between each corpus' rank and the optimal dimensionality according to average precision was 0.9. For optimal F the correlation was 0.94. The second important result evident in Table 4.1.1 is disagreement among the evaluation metrics. In many cases the three performance metrics were optimized at widely different dimensionalities. Overall, ASL calls for models of lower dimensionality than do average precision and optimal F. Thus no linear relationship was evident between matrix rank and ASL's optimal dimensionality. While the close agreement between average precision and optimal F might tempt us to discount ASL because of its divergence from their values, the ASL metric has proven robust in several analyses (cf. **[98, 97]**). Before turning to a general analysis of Table 4.1.1, then, I argue that our attention will need to seek a balance between the models called for by each performance metric.

Among the data of Table 4.1.1, the most aggressive dimensionality reduction was merited for the *MEDLINE* corpus. Using the ASL measure, *MEDLINE*'s observed optimal dimensionality was 91, merely 8.8% of the total possible dimensions. Similar results were obtained via the precision and optimal F metrics, with $k_{opt}(Pr) = k_{opt}(F) = 151$, or 14% of the full-rank model. Figure 4.1.1 shows interpolated recall/precision graphs of three retrieval models. Performance of the full-rank model is shown with black dots, while ASL's optimal model appears as dark grey triangles, and performance for $k = k_{opt}(Pr)$ is light grey pluses. While the full-rank model appears to have some advantage at very low levels of recall, the results suggest that LSI's reduced model improves retrieval for *MEDLINE* across a broad spectrum of recall levels. This is in step with suggestions by Deerwester *et al.* that *MEDLINE* is especially amenable to dimensionality reduction because it was constructed by a series of keyword queries. This implies that a set of well-defined concepts may be evident in the *MEDLINE* data, a fact that works in LSI's favor **[32]**. Evidence that *MEDLINE* is well-suited for dimensionality reduction is also manifest in the fact that the precision/recall curves obtained by setting $k = k_{opt}(Pr)$ are nearly identical to those for $k = k_{opt}(ASL)$. The agreement between the performance metrics suggests that in the case of *MEDLINE,* Cranfield-style analysis detects the optimal semantic subspace in the vicinity of $k \approx 100$.

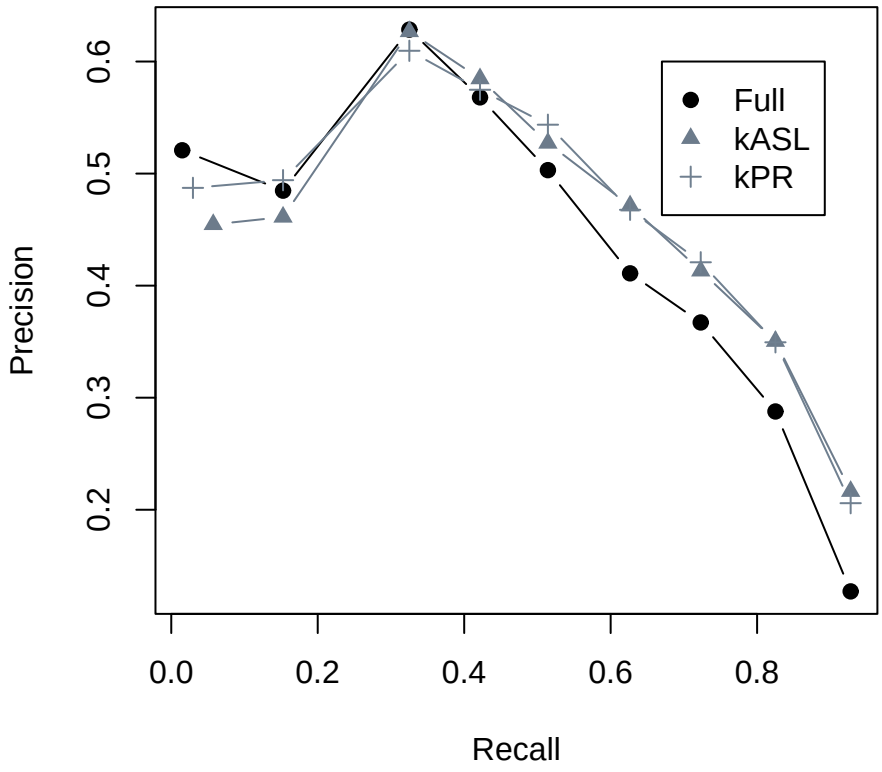

FIGURE 4.1.1. Precision and recall for *MEDLINE* data

Dimensionality reduction also yielded a large improvement in retrieval for the full-text CF data. Overall, this corpus yielded very poor performance for all models. However, as seen in Figure 4.1.2, LSI's reduced rank approach improves recall and precision dramatically over the full-rank model. Again, setting $k = k_{opt}(ASL)$ and $k = k_{opt}(Pr)$ yields very similar performance. Based on retrospective analysis, then, it seems that there is a pronounced subspace of approximately 200 dimensions that comprises a good model for retrieval on the *CF_FULL* data. Yet as can be seen from Table 4.1.1, in contrast to the *MEDLINE* data, an optimal model for *CF_FULL* retains about 70% of the total possible variance—a modest dimensionality reduction. Although the notion of an optimal subspace appears to be valid

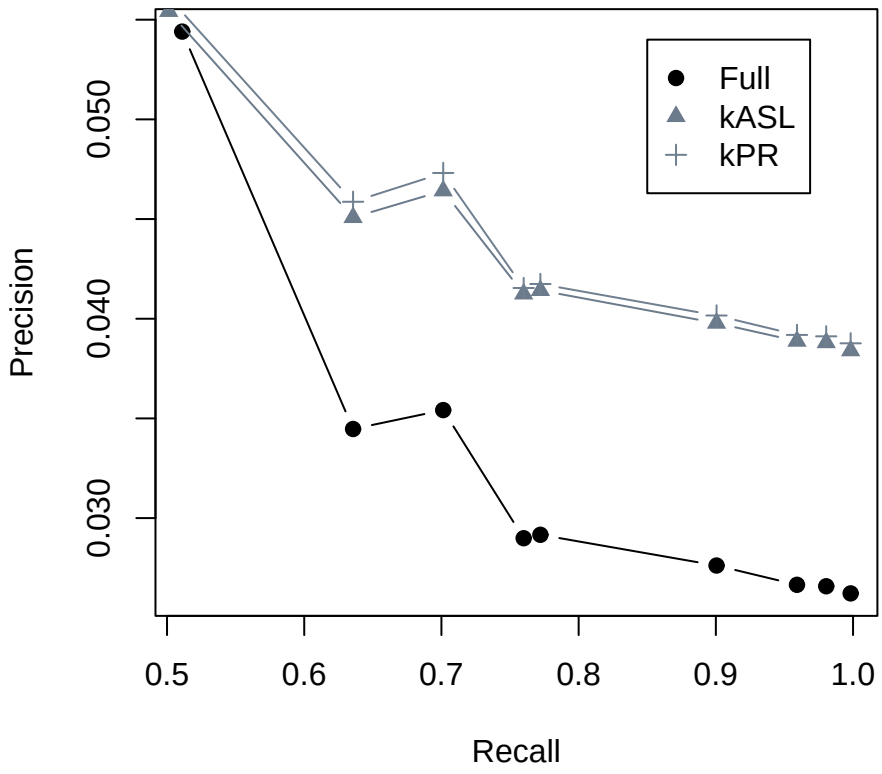

FIGURE 4.1.2. Precision and recall for the *CF_FULL* data

for these data, then, it is important to note that the dimensionality of this space is quite different across data sets.

Furthermore, in analyzing performance on the *CF_FULL* data, we see a wholesale improvement over full-rank retrieval when using dimensionality reduction, suggesting that LSI's benefit comes to the fore given a larger vocabulary. This is especially interesting insofar as no comparable benefit was seen on the standard *CF* data. When the *CF* data were represented only by titles and abstracts, I found no appreciable difference in performance between the full-rank and reduced-rank models. Insofar as the subject matter and query types were constant across these corpora, this finding suggests that characteristics of the termspace bear significantly on the suitability of dimensionality reduction.

In these experiments, corpora with redundant vocabularies reaped the greatest benefit from dimensionality reduction. The data in Table 4.1.1 show a tendency towards smaller models for corpora with greater amounts of term repetition. Across the six tested corpora, the correlation between the median frequency of terms and the overfitting effect according ASL was 0.71. Thus collections with greater repetition of terms benefited more from dimensionality reduction. Likewise, I found a 64% negative correlation between the magnitude of overfitting effect (measured by optimal F) and the median number of documents that a corpus' terms appear in (measuring performance improvement via average precision and ASL yielded correlations of -59% and -22%, respectively). Collections with low median term-document frequency benefited the most from dimensionality reduction. Consider, for instance the *CF* and *CRAN* corpora. Although the number of terms and documents are similar for each of these data sets, they responded to dimensionality reduction in distinct ways; the optimal LSI model of *CF* improved ASL by only 6.65 documents, while LSI improved ASL for *CRAN* by 24.5 documents. As shown in Table 3.1.1, median term frequency in *CRAN* is higher than in *CF*. Likewise, CRAN's median term-document frequency is 18, versus *CF*'s term-document frequency of 12. In other words, *CRAN* displays more term repetition than does *CF*. LSI's advantage on the *CRAN* data may thus be attributable in part to the redundancy of *CRAN*'s terms.

I also found that large termspaces lent advantage to dimensionality reduction. The data show a 45% correlation between the number of terms in a corpus and the amount of benefit (measured as ASL improvement) afforded by LSI. This measure is quite distinct from the correlation between matrix rank and LSI advantage reported above, insofar as for all collections, rank was determined by the number of documents. The correlations between term-count and dimensionality reduction's improvement in average precision and the optimal F measure were similar (30%, and 48%, respectively). For example, *CF_FULL* had by far the largest termspace. It also evinced a large improvement (relative to its overall poor performance) by use of dimensionality reduction. On the other hand, *CACM* had a small termspace (especially considering its large document space). Yet according to average

precision and optimal F, the $CACM$ data benefited very little, if at all from dimensionality reduction.

In this study, then, corpora with large, or redundant term spaces were especially ripe for LSI. To understand how the size and distribution of the termspace bears on retrieval, we may think of LSI's effect as entailing two parts. Given a corpus C and a query q, traditional keyword retrieval will derive one document ranking, R_1, while LSI will give another ranking R_2. The differences between R_1 and R_2 will be due to two mechanisms. First, LSI alters the weights of each cell in the term-document matrix described by C, which will change the document ordering. I call this LSI's "mechanism-one" effect. Mechanism one improves retrieval in the presence of large amounts of term redundancy. Second, LSI creates dense representational vectors for each document. Thus, as I mention in Section 2.2, queries and documents may match under LSI despite sharing no indexing terms. By inferring term-document associations, then, LSI may rank documents highly that the classic VSM omits from ranking entirely. I call this "mechanism-two." In a corpus with a large vocabulary, this second mechanism appears to come into play. For example, given its 9549-term vocabulary, the opportunity for a CF_FULL query term to "miss" an exact match with relevant documents is increased above, say, the 5116-term CF corpus. A larger termspace, then, implies a greater opportunity for synonymy to impede retrieval.

Regardless of these generalizations, my goal is to gauge the intrinsic dimensionality of each corpus. The fact that the Cranfield-style performance metrics were not always in agreement about k_{opt} complicates this task. Consider the $CACM$ corpus. As mentioned above, $CACM$ yielded little benefit from dimensionality reduction, at least so far as average precision and optimal F were concerned. But from Table 4.1.1 it can be seen that according to ASL, dimensionality reduction did improve retrieval performance for $CACM$. To help us understand the dynamics of dimensionality reduction, Figure 4.1.3 compares reduced-rank and full-rank retrieval performance as measured by precision and recall for the $CACM$ data. Here dimensionality provides no discernible advantage, with the precision-optimized LSI model and the full-rank solution showing nearly identical behavior. On the other hand the ASL-optimized LSI model gives significantly worse results than the full-rank model for low

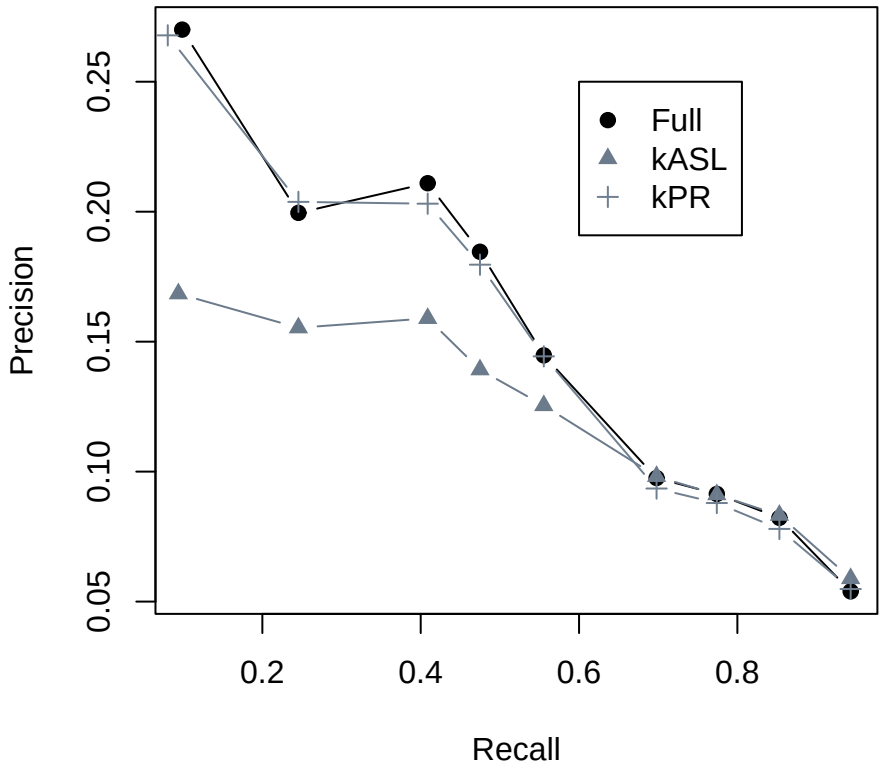

FIGURE 4.1.3. Precision and recall for the *CACM* data

recall levels. All three models converge on comparable performance for high levels of recall, with the ASL-optimized representation offering a very slight benefit. Figure 4.1.3 shows that for the *CACM* data, a heavy dimensionality reduction deprives the model of important discriminatory power, while failing to offset this deprivation with a comparable improvement based on mechanism-two effect.

The optimal LSI models discerned via the three retrospective metrics were widely divergent for *CACM*. Average precision and optimal F called for very little dimensionality reduction ($k_{opt}(Pr) = 1936$, and $k_{opt}(F) = 2660$), while ASL called for a much more drastic dimensionality reduction, of $k_{opt}(ASL) = 271$. Because Figure 4.1.3 plots performance for *CACM* in terms of precision and recall, it gives the ASL-optimized model an inherent

disadvantage. This suggests that one must be circumspect in identifying the dimensionality of *CACM*'s optimal semantic subspace. In fact, the disagreement between performance metrics and the failure of any of them to demonstrate a convincing improvement over the full-rank model by means of dimensionality reduction implies that Cranfield-style analysis may not be sufficient for identifying the intrinsic dimensionality of this corpus. This is not surprising, as *CACM* only uses 64 queries, the second-lowest count among the set of test collections. And *CACM*'s median number of relevant documents per query is only 12, in the face of the largest number of total documents among the tested corpora. Thus it seems likely that the supplied queries were not adequate to gather a complete picture of *CACM*'s intrinsic dimensionality.

A similarly complex portrait emerged as I analyzed the observed optimal dimensionality for the *CISI* and *CRAN* databases. In both cases, the three performance measures disagree on how many dimensions an optimal model should retain. As seen in Table 4.1.1, *CISI*'s best model in terms of ASL is 751-dimensional, while average precision is optimized for $k = 1276$, and optimal F performs best with 1411 dimensions. Likewise, the *CRAN* data require 121, 661, and 811 dimensions to optimize ASL, precision, and optimal F, respectively. However, this disagreement is understandable if we compare the precision/recall behavior for each of these corpora. In both cases, dimensionality reduction appears to yield almost no benefit or detriment to retrieval performance. That is, the difference in document rankings afforded by each measure's optimal model is very slight. For *CRAN* and *CISI*, then, two things are apparent. First, query-document matching (for the supplied queries) appears to be resistant to an overfitting effect. Secondly, and somewhat confusingly, these corpora also appear resistant to degradation of performance by selecting parsimonious models[1].

This apparent idiosyncrasy is perhaps ascribable to problems native to Cranfield-style evaluation. In Table 3.1.2 we see that of *CRAN*'s 225 supplied queries, the median number of relevant documents per query is only 7, by far the lowest number among the corpora tested

[1]It bears mentioning that *CRAN* contains a large number of documents that are not relevant to any queries. In many studies researchers remove these documents prior to analysis. However, I did not remove these universally non-relevant documents. In retrospect I believe that including these documents increased the observed optimal dimensionality of *CRAN*, thus penalizing PA and APA's low-dimensional models. The findings reported here, for instance, report an observed k_{opt} *vis a vis* average precision that is significantly higher than that reported by Jiang and Littman [**77**].

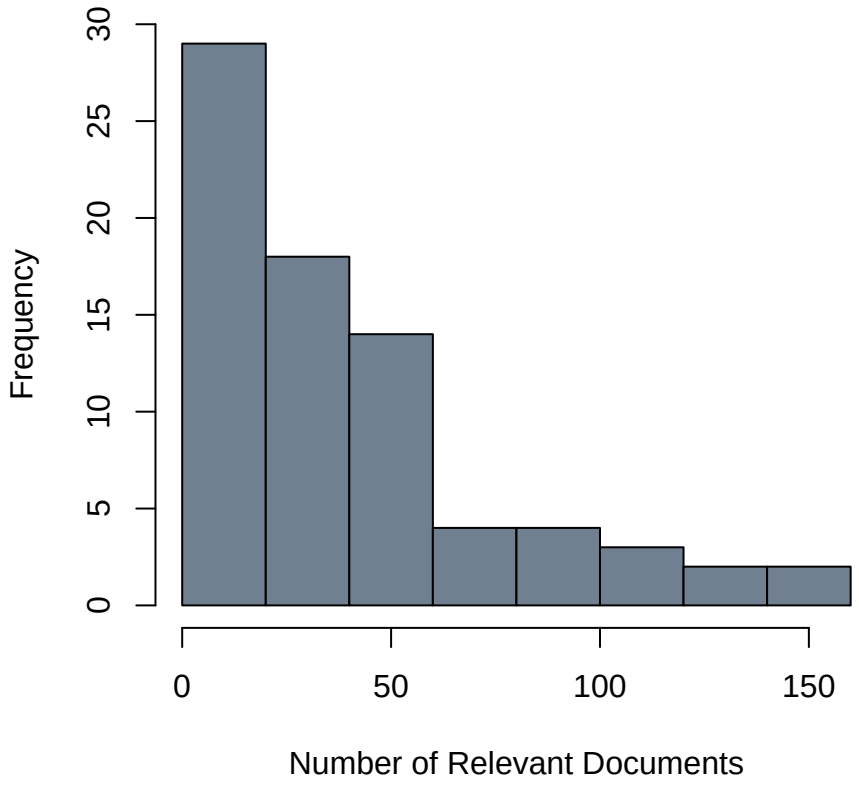

FIGURE 4.1.4. Distribution of relevant documents per query (*CISI*)

here. Thus Cranfield-style performance metrics for this corpus will be heavily influenced by the location in a given ranking of only a few documents, leading to possibly erratic measurement, and a high degree of randomness in the results. Likewise, the *CISI* database is prone to noisy analysis via Cranfield-style evaluation. *CISI* contains 76 queries, with the median number of relevant documents per query of 30.5. But the variance of the number of relevant documents per query is 1308.373, over six times the next largest variance observed. In fact a small number of *CISI's* queries have many relevant documents, while many have only a handful of positive hits, as is evident in Figure 4.1.4.

My suspicion that the Cranfield methodology provides a very noisy portrait of the intrinsic dimensionality of corpora is borne out by an interesting result. The correlations between

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MEDLINE*
ASL % retained	0.085	0.861	0.541	0.514	0.087	0.088
ASL % improved	0.157	0.019	0.008	0.117	0.07	0.491
PR % retained	0.604	0.704	0.656	0.874	0.473	0.146
PR % improved	0.016	0.0036	0.365	0.0046	0.022	0.699
F % retained	0.83	0.704	0.656	0.966	0.58	0.146
F % improved	0.0	0.0058	0.06	0.0019	0.0044	0.05

TABLE 4.1.2. Summary of observed optimal dimensionality findings

the strength of each performance metric's observed overfitting effect and the number of queries in a given corpus were high. For instance, across the six corpora, the correlation between the overfitting effect noted under ASL and the number of queries was 0.31. Likewise, collections with fewer queries showed less improvement under LSI as measured by average precision and optimal F. These correlations were both -0.48. Thus the amount of improvement afforded by dimensionality reduction in my data appears to owe almost as much to the number of supplied queries as it does to the size of the termspace. This suggests that for corpora with skewed distributions of relevant documents and queries Cranfield-style analysis may not yield an accurate picture of intrinsic dimensionality. Comparing the accuracy of each dimensionality estimation technique thus demands an admission that my instrument of measurement is inherently noisy.

4.1.2. Summary of Observed Optimal Dimensionality Findings. Table 4.1.2 provides a digest of my findings on the matter of observed optimal dimensionality, as discussed in detail above (cf. Section 4.1.1). For each corpus, Table 4.1.2 reports six statistics, two per performance metric. Rows labeled *m % retained* show the percentage of possible dimensions retained under the optimal model given by metric m. The rows named *m % improved* reports the percent of improvement over the full-rank model afforded by the metric m's optimal model. In general, ASL calls for more drastic dimensionality reduction than average precision or optimal F. Also, the percentage of total dimensions retained across corpora varies widely.

LSI's dimensionality reduction gives the most conclusive improvement for the *MEDLINE* data, where low-rank models improved performance greatly, and where all three performance metrics were in close agreement about the dimensionality of the optimal model, concurring

that $k_{opt} \approx 100$. The *CF_FULL* data also benefited decisively from dimensionality reduction, as seen in Figure 4.1.2, with the concomitant region of observed optimality near $k \approx 250$. The ASL metric indicated the the *CRAN* data merited a 92% dimensionality reduction, yielding a 7% improvement over the full-rank model. On the other hand, the *CF* data appear to respond poorly to dimensionality reduction; none of the performance metrics noted strong evidence of a low-rank optimal semantic subspace for *CF*. Finally, *CACM* and *CISI* appear to give fairly noisy data when I applied Cranfield-style analysis to them. Due to the unusual distributions of relevant documents and queries for these corpora, I suspect that *CACM* and *CISI* may well possess low-rank semantic subspaces, but that the dimensionality of these spaces is obscured by the performance measurement process. That said, *CACM* appeared to benefit from a 92% dimensionality reduction with respect to ASL, yielding $k_{opt}(ASL) \approx 270$. *CISI*'s $k_{opt}(ASL)$ of 751 (a 49% reduction) yielded an 11% benefit in observed ASL over the full-rank model.

Overall, Cranfield-style evaluation suggests that the notion of a corpus' optimal semantic subspace is valid, and that it is partially observable by retrospective performance analysis of *ad hoc* retrieval runs. However, there appears to be no simple way of choosing the dimensionality of such a subspace *a priori*. Although I found a strong relationship between the rank of a matrix and its optimal dimensionality via average precision, ASL frequently disagreed with these findings. Some corpora required only 8% of their eigenvectors to yield optimal performance, while others tolerated no significant dimensionality reduction. Thus query-independent dimensionality estimation methods such as those described in Table 3.4.1 appear crucial for applying LSI. However, due to the inherent noisiness of Cranfield-style analysis, evaluating the performance of eigenvalue-based dimensionality estimators will be non-trivial. In Section 4.2 I use the performance metrics reported here to undertake such an evaluation. To supplement these findings with less complex analysis, Chapter 5 uses simulated data of known dimensionality to validate my empirical findings.

4.2. Performance of Eigenvalue-Based Dimensionality Estimators

As noted in Section 4.1, my experimental results show that discerning the dimensionality of a corpus' optimal semantic subspace via Cranfield-style performance analysis is a complex task. However, I found enough consistency among the results of such analysis to suggest that retrospective performance evaluation does provide useful, if incomplete, information about the intrinsic dimensionality of the test collections described in Section 3.1. In this section, I use ASL, average precision, and optimal F to compare the quality of dimensionality estimates afforded by the eigenvalue analysis techniques outlined in Table 3.4.1. I argue that my proposed method—amended parallel analysis—provides the best dimensionality estimate for those cases where Cranfield-style analysis yields the most decisive picture of a low-rank optimal semantic subspace. Moreover the family of estimators based on an error correction rationale—APA, PA, and EV1—form a bloc of decisively good performers.

I begin my data analysis with a high-level discussion in Section 4.2.1. This section forms the bulk of the treatment, and concerns the conditions under which one estimation technique appears to outperform the others. Next I turn to a more detailed analysis of amended parallel analysis, noting the differences and similarities in its estimates and those afforded by traditional parallel analysis. I then pursue an individual analysis of the other proposed estimation techniques in Sections 4.2.2.2 through 4.2.2.4.

4.2.1. Quality and Suitability of Eigenvalue analysis Techniques. Across the six corpora, the five tested eigenvalue analysis techniques yielded distinct estimates of corpus dimensionality. Despite this variation, however, these methods also exhibited some consistency in their predictions. Thus parallel analysis always yielded the most parsimonious model, followed by amended parallel analysis. On the other extreme, Bartlett's test of isotrapy was effectively a non-performer, always delivering models of near-full complexity. Between these extremes, EV1 and 85% Var gave models of middling size, with EV1 predicting lower dimensionalities than 85% Var. This behavior led to a complex portrait of estimator accuracy. To begin my comparison of each eigenvalue analysis technique's performance, consider Figure 4.2.1, which shows estimation accuracy for the *MEDLINE* data. The x-axis is k, the number of dimensions included in the LSI model. In all of the exper-

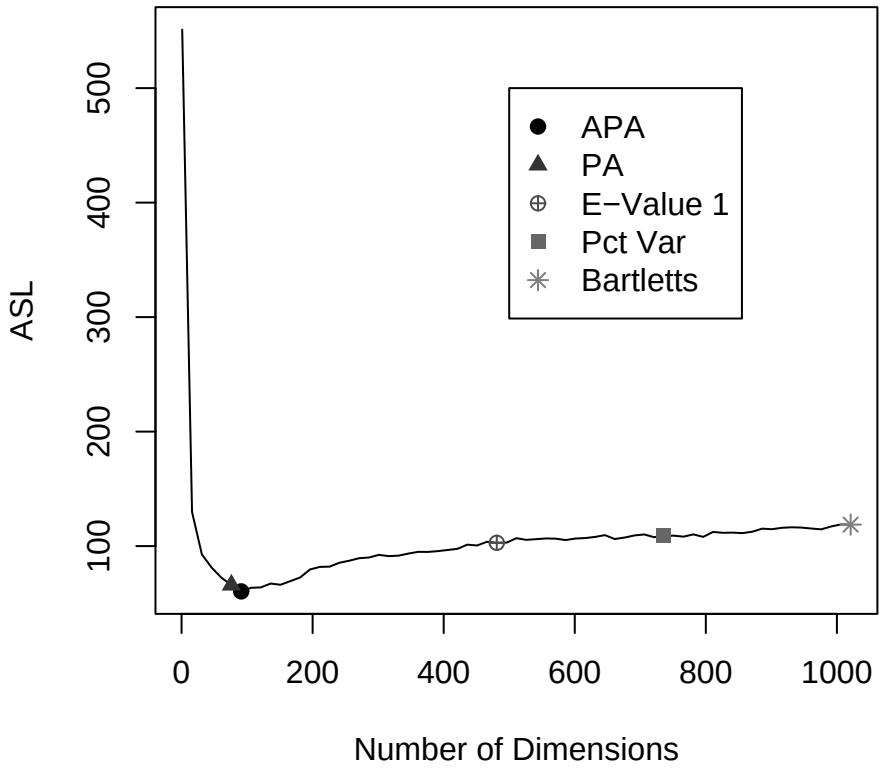

FIGURE 4.2.1. ASL versus k for *MEDLINE* data

iments reported in this chapter, k was measured from $k = k_{min} \ldots k_{max}$ in increments of fifteen[2]. On the y-axis I have plotted the value of ASL observed at k. Not surprisingly, a two-dimensional model provides inadequate information for retrieval on the *MEDLINE* data. But as k increases we see a dramatic improvement in retrieval performance. As shown in Table 4.1.1, ASL on the *MEDLINE* data is optimized for $k = 91$. After this point on the x-axis, a marked overfitting effect appears, causing ASL to increase (i.e. degrade) as

[2]The 15-dimensional increment was chosen primarily for practical reasons. While it would have been feasible to test every possible dimensionality, iterating over the domain of k by increments of fifteen was much more efficient, and still yielded a nuanced picture of each corpus' dimensional profile. That is, I found that performance changed as a function of k in a stable fashion, suggesting that the chances of missing important dynamics between k and $k+15$ were slim. Moreover, in **[90]** Landauer and Dumais perform a similar analysis using 30-dimensional increments. Thus I hypothesized that a 15-dimensional increment provided suitable granularity of analysis.

I add more dimensions. Thus the full-rank model offers ASL performance inferior to the 91-dimensional model.

Superimposed on this performance plot, Figure 4.2.1 shows the dimensionality estimate afforded by each eigenvalue analysis technique described in Table 3.4.1. Amended parallel analysis (APA) yields the estimate closest to $k_{opt}(ASL)$, followed closely by traditional parallel analysis. The eigenvalue-one EV1 criterion offers the next-best estimate, though it overestimates by a wide margin. The 85% Var rule is slightly higher than EV1. Finally Bartlett's recommended almost no dimensionality reduction for *MEDLINE*. Thus APA appears to yield the best estimate of the optimal dimensionality for the *MEDLINE* data, an impression borne out in Figure 4.2.2, where I have measured performance by average precision instead of ASL. Figure 4.2.2 shows an analogous region of optimal dimensionality near $k = 150$. The plot for optimal F is nearly identical to Figure 4.2.2, strengthening my conviction that APA yields the best estimate of *MEDLINE*'s intrinsic dimensionality.

A similar, though slightly more complex, picture emerges in Figure 4.2.3, which shows performance (measured by ASL) as a function of k for the *CRAN* data. Again, we see a stark improvement in performance as I add the first singular vectors to the model, followed by a slow decay after $k \approx 150$. And as in Figure 4.2.1, for the *CRAN* database APA and PA yield the best dimensionality estimates, with the simpler *EV1* and *85% Var* criteria overestimating the optimal dimensionality.

However, in Figure 4.2.4 the more complex models recommended by the EV1 and 85% Var approaches appear to have some merit. Figure 4.2.4 plots average precision as a function of k for the *CRAN* data. It is clear from Tables 4.1.1 and 4.1.2 that ASL and average precision disagree on the dimensionality of an optimal LSI model for the *CRAN* data. If we consult only average precision, then, the EV1 and 85% Var criteria are more accurate than PA or APA. But we must also keep in mind that the picture afforded by average precision suggests that almost no overfitting effect is incurred by moving from $k = k_{opt}(Pr)$ to $k = k_{max}$. Without a strong case for dimensionality reduction's actual utility, we should be skeptical about the optimality of the EV1 or 85% Var estimates.

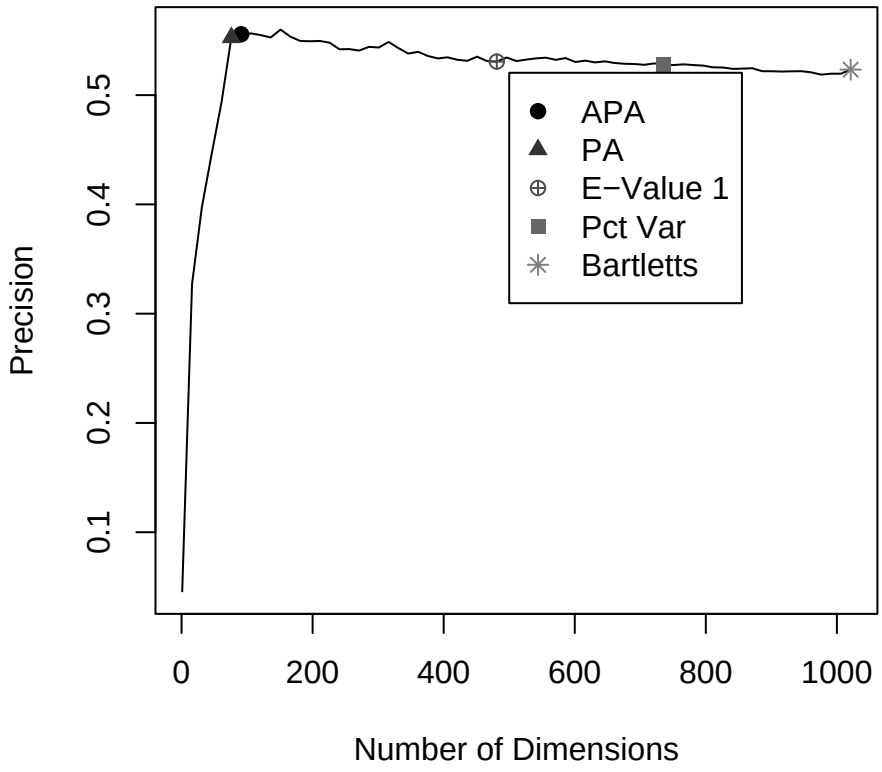

FIGURE 4.2.2. Precision versus k for *MEDLINE* data

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	459	-981	-167	-671	**-27**	**-4**
PA	**426**	-996	-170	-679	-40	-16
EV1	1017	-513	**-41**	**-100**	492	380
85% Var	719	-457	96	231	789	643
Bartlett's	3729	**170**	178	709	1275	939

TABLE 4.2.1. Raw dimensionality estimates (ASL)

Tables 4.2.1 and 4.2.2 summarize the quality of each eigenvalue analysis technique's dimensionality estimates, with respect to ASL performance.

Table 4.2.1 contains the directed distance of each eigenvalue analysis technique's dimensionality estimate from the observed optimal dimensionality afforded by ASL. Thus the first

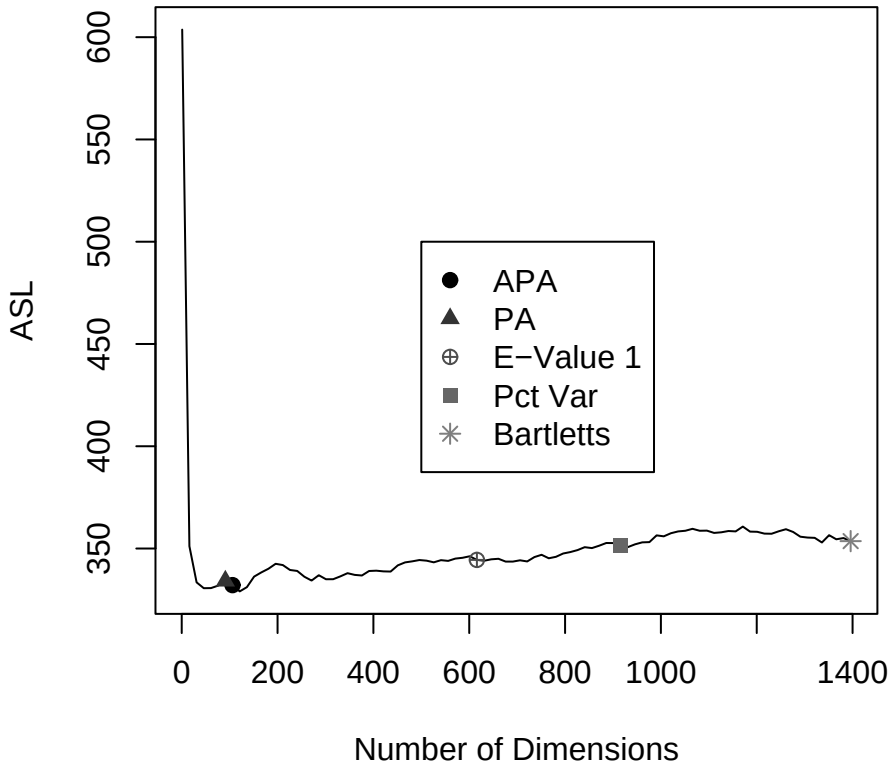

FIGURE 4.2.3. ASL versus k for the *CRAN* data

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	0.154	-0.796	-0.426	-0.461	**-0.019**	**-0.004**
PA	**0.142**	-0.808	-0.434	-0.466	-0.029	-0.016
EV1	0.34	-0.416	**-0.105**	**-0.069**	0.352	0.372
85% Var	0.24	-0.371	0.245	0.159	0.565	0.63
Bartlett's	0.915	**0.138**	0.454	0.487	0.913	0.92

TABLE 4.2.2. Normalized dimensionality estimates (ASL)

cell contains $k_{opt}(ASL) - k_{opt}(APA) = 459$, indicating that APA overestimated ASL's optimal model by 459 dimensions. Table 4.2.2 contains the same information as Table 4.2.1, only in this case the cells have been normalized to fall between -1 and 1, thus making cross-corpus comparison more meaningful. In both tables, values near 0 indicate good performance (i.e.

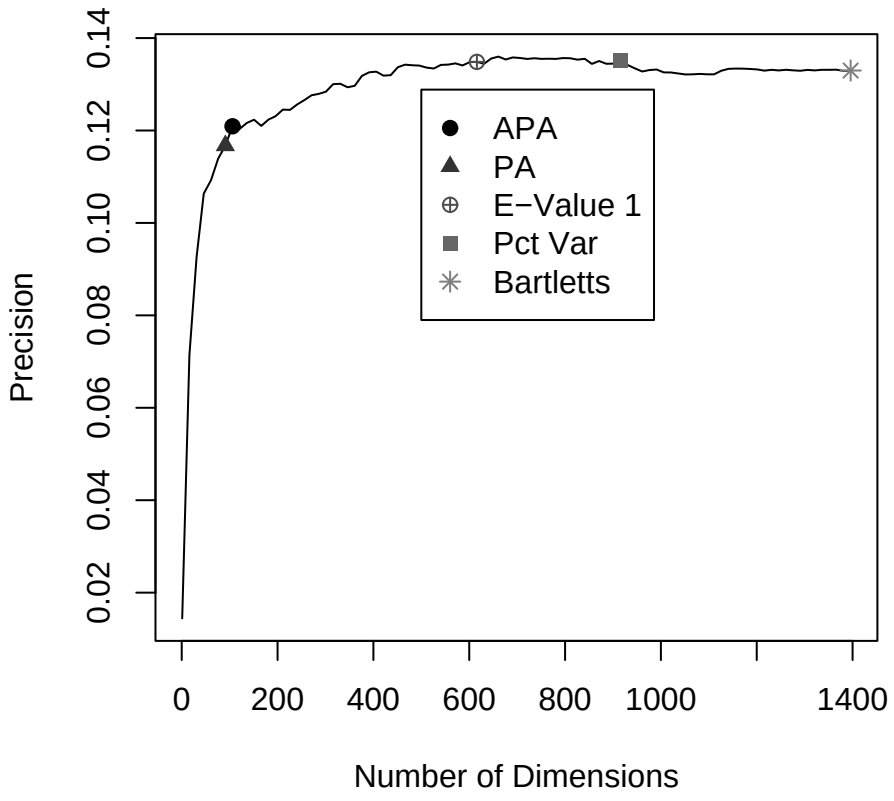

FIGURE 4.2.4. Precision versus k for the *CRAN* data

little estimation error). The best dimensionality estimate (in terms of absolute distance) is shown in boldface.

Similar information appears in Tables 4.2.3 through 4.2.6. These tables provide raw and normalized distances between dimensionality estimates and the observed optimal dimensionality given by the average precision and optimal F metrics. An initial inspection of these tables shows that no single eigenvalue analysis technique offers the best dimensionality estimates for all corpora across all metrics. However, we can gain a high-level appreciation of these findings from Table 4.2.9, which counts how many times each dimensionality estimator performed best or worst among the five techniques. The 85% Var approach was best five times, and was never worst. APA, EV1, and Bartlett's were all best four times.

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	-1206	-786	-212	-1196	-567	**-64**
PA	-1239	-801	-215	-1204	-580	-76
EV1	**-648**	-318	-86	-625	**-48**	320
85% Var	-946	**-30**	**-51**	-294	249	583
Bartlett's	2064	365	133	**184**	735	879

TABLE 4.2.3. Raw dimensionality estimates (Pr)

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	-0.403	-0.638	-0.541	-0.821	-0.406	**-0.063**
PA	-0.414	-0.65	-0.548	-0.827	-0.415	-0.074
EV1	**-0.217**	-0.258	-0.219	-0.429	**-0.034**	0.313
85% Var	-0.316	**-0.24**	**-0.13**	-0.202	0.178	0.571
Bartlett's	0.69	0.296	0.339	**0.126**	0.527	0.861

TABLE 4.2.4. Normalized dimensionality estimates (Pr)

However, Bartlett's was also worst nine times, while APA never gave the worst estimate. Traditional parallel analysis offered the best estimate once, but gave the worst answer nine times, consistently underestimating the intrinsic dimensionality of several corpora.

4.2.2. Analyses of Each Dimensionality Estimator's Performance. Having pursued a broad comparison of the five dimensionality estimators of interest, I now turn to an analysis of the strengths and weaknesses of each estimation technique on its own merits. For instance, the percent of variance approach appeared to perform favorably. However, I argue that its success is due more to chance than to a systematic advantage over more rigorously motivated techniques. Likewise, I note that Bartlett's test of isotrapy performed well on several occasions. However, its success may be understood as an indicator that for the tested corpora, dimensionality reduction was not always merited. In general, each eigenvalue analysis technique excelled in certain respects and failed in others. The following sections articulate these strengths and weaknesses.

4.2.2.1. *Performance of PA and APA.* In Tables 4.2.7 and 4.2.8 it may be seen that APA afforded the best dimensionality estimates on four of the eighteen pairings of a given corpus and performance metric. Traditional PA performed best once. On the other hand, PA often provided the worst dimensionality estimate (on nine observations). APA was never the worst performer. But on the occasions where PA was worst, APA ranked second-to-worst. This

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	-1930	-786	-212	-1331	-717	**-64**
PA	-1963	-801	-215	-1339	-730	-76
EV1	-1372	-318	-86	-760	-198	320
85% Var	-1670	**-30**	**-5**	-429	**99**	583
Bartlett's	**1340**	365	133	**49**	585	879

TABLE 4.2.5. Raw dimensionality estimates (opt F)

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	-0.645	-0.638	-0.541	-0.914	-0.514	**-0.063**
PA	-0.657	-0.65	-0.548	-0.92	-0.523	-0.074
EV1	-0.459	-0.258	-0.219	-0.522	-0.142	0.313
85% Var	-0.559	**-0.24**	**0.13**	-0.295	**-0.071**	0.571
Bartlett's	**0.448**	0.296	0.454	**-0.034**	0.419	0.861

TABLE 4.2.6. Normalized dimensionality estimates (opt F)

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MEDLINE*
ASL	PA	Bartlett's	EV1	EV1	APA	APA
PR	EV1	85% Var	85% Var	Bartlett's	EV1	APA
F	Bartlett's	85% Var	85% Var	Bartlett's	85% Var	APA

TABLE 4.2.7. Best dimensionality estimates

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MEDLINE*
ASL	Bartlett's	PA	Bartlett's	Bartlett's	Bartlett's	Bartlett's
PR	Bartlett's	PA	PA	PA	Bartlett's	Bartlett's
F	PA	PA	PA	PA	PA	Bartlett's

TABLE 4.2.8. Worst dimensionality estimates

seemingly paradoxical behavior—best and near-worst performance by a single estimator—can be addressed to a large extent by considering *which* observations APA excelled at, and which it was ill-suited for. PA consistently gave the lowest model dimensionalities among the five analysis techniques tested here. As discussed in Section 4.1, however, Cranfield-style evaluation failed to discern convincing benefits from dimensionality reduction for several corpora. In these cases, then, PA failed *de facto*.

APA provided a decisively superior estimate for the *MEDLINE* data, giving a value for k that was closest to the optimal value according to all three performance metrics. Given my discussion in Section 4.1 I feel especially confident that the Cranfield-style analysis

	APA	*PA*	*EV1*	*85% Var*	*Bartlett's*
Best	4	1	4	5	4
Worst	0	9	0	0	9

TABLE 4.2.9. Best and worst dimensionality estimates (counts)

undertaken here was able to discern the intrinsic dimensionality of the *MEDLINE* data. Moreover, I suspect that the *MEDLINE* data are especially amenable to LSI (due at least in part to their concept-driven construction, cf. **[32]**). APA's accuracy with respect to *MEDLINE*, then, is especially promising, suggesting that the proposed approach is adept at intuiting a well-defined intrinsic dimensionality.

APA also gave the best estimate for the *CRAN* data, with respect to ASL performance. However, the performance metrics were widely divergent on this corpus. Thus APA and PA drastically underestimated the best dimensionality with respect to average precision and optimal F (PA was actually the worst performer in this case). However, as mentioned above, in many IR experiments, researchers remove universally non-relevant documents from the *CRAN* data. Had that been done during the current experiment, I suspect that the observed optimal dimensionality would have been reduced greatly, and by extension, PA and APA would have fared much better. Due to the discrepancy between performance metrics, then, I argue that the *CRAN* data offer a less compelling base for dimensionality estimator comparison than does *MEDLINE*. A similar dynamic emerged for the *CACM* data, where PA performed best with respect to ASL, but worst *vis a vis* optimal F. Again, I note a wide divergence among performance metrics' observed k_{opt} for *CACM*.

PA was the worst performer for the *CF* data. This is due to the fact that *CF* brooked no substantial dimensionality reduction; its low-rank models were inferior to the keyword approach according to all three performance metrics. Thus it appears to be a serious defect in the application of PA that it fails to react to circumstances when no dimensionality reduction is merited. PA's inherent tendency to deliver parsimonious models emphasizes the need for the confidence-interval based amendment utilized by APA. APA's moderating effect on PA assuaged the under-estimation problem to some extent, insofar as APA was

consistently better than PA for all corpora and all performance metrics, save one ($CACM$ measured by ASL).

Overall, PA appears prone to under-estimation of intrinsic dimensionality for IR applications. I return to the question of whether this error is systematic in Chapter 5. But PA's poor performance is somewhat unexpected insofar as early research into the application of PA found that it tended to over-estimate the true dimensionality (cf. **[57, 71, 48]**). This suggests that scaling the unsupervised learning task into highly complex environments such as IR changes the problem qualitatively. That is, given the large number of variables native to IR problems, the mean null eigenvalue $\widehat{\lambda}^{*}_{0k}$ from B samples, is not necessarily the best estimator of the corresponding population null eigenvalue.

As described in Section 3.3 APA takes account for this fact. Instead of testing $\lambda_k > \widehat{\lambda}^{*}_{0k}$, APA is concerned with the $1 - \alpha\%$ confidence interval on $\widehat{\lambda}^{*}_{0k}$. This appears to improve dimensionality estimation. To gauge the significance of this moderating effect, I performed two statistical tests. First, a paired t-test was performed, testing equality of the actual dimensionality estimates afforded by PA and APA, where the null hypothesis was H_0 : $k_{opt}(APA) = k_{opt}(PA)$. The unit of analysis in this test was an individual corpus. Thus my sample size was very small ($n = 6$), and one must be cautious when drawing conclusions from it. However, APA's improvement over PA did appear to be highly significant. The p-value for this test was $p = 0.04$. Next I performed an identical test, this time testing the equality of PA's and APA's estimates, normalized by the rank of each data set (i.e. analyzing the percentage of total eigenvectors retained by each method). In this case, $p = 0.001$. The difference between APA and PA is thus statistically significant. And insofar as APA appears to outperform PA with respect to standard IR test metrics, I conclude that the difference implies benefit in favor of APA. However, given the small sample size here, we must wonder about the validity of the t-distribution. To address these misgivings, I examine the matter of APA's relation to PA further in Chapter 5.

APA appears to improve the dimensionality estimates afforded by PA by supplementing estimation based on point estimates with a confidence interval-based approach. Theoretically, then, APA should be able to moderate to an arbitrary degree PA's tendency toward

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$k_{opt}(APA)$	79	80	104

TABLE 4.2.10. Confidence intervals for the *CISI* data

under-estimation by setting an appropriate α. However, my results suggest that in practice this is not feasible. In fact, for large IR data sets APA's simulated null eigenvalues exhibit such small variance that the derived confidence intervals change very little across standard values of α. For example, consider Table 4.2.10, which shows APA's dimensionality estimates for the *CISI* data using various levels of confidence. According to the ASL metric, optimal k for the *CISI* data is 751. Thus setting $\alpha = 0.01$ would have improved APA's estimate over the reported 95% confidence interval. But APA's estimate would still be far too low. Changing the confidence level yielded similarly subtle changes to the other corpora, as well.

Figure 4.2.5 plots the width of *MEDLINE*'s 95% null eigenvalue confidence intervals against k. In black we see the width of confidence intervals generated from $B = 50$ bootstrap simulations. The grey triangles are the corresponding intervals for $B = 1000$. Two phenomena are notable in Figure 4.2.5. First is that the confidence intervals for the first few eigenvalues are much wider than the intervals for large values of k. This suggests that APA's divergence from PA will be most dramatic if the null eigenvalues outsize the observed eigenvalues at a relatively low value of k. It also suggests that considering the width of APA's confidence intervals might provide useful evidence in further amendments to parallel analysis. Thus in forthcoming research, I anticipate using null eigenvalue variance as an indicator of a corresponding eigenvalue's validity.

The second phenomenon seen in Figure 4.2.5 is that as B is increased, the average width of a $1 - \alpha\%$ confidence interval decreases. This phenomenon is further shown in Figure 4.2.6, which plots the variance of confidence interval width against k for the *MEDLINE* data. Although setting $B = 1000$ gives us more data than the $B = 50$ case, the simulated null eigenvalues are centered very closely around their mean. By Equation 3.3.1 we note that as B grows, the standard error of $\widehat{\lambda}^*_{0k}$ shrinks. In other words, we gain confidence in our point estimate with a larger sample and require a narrower confidence interval to

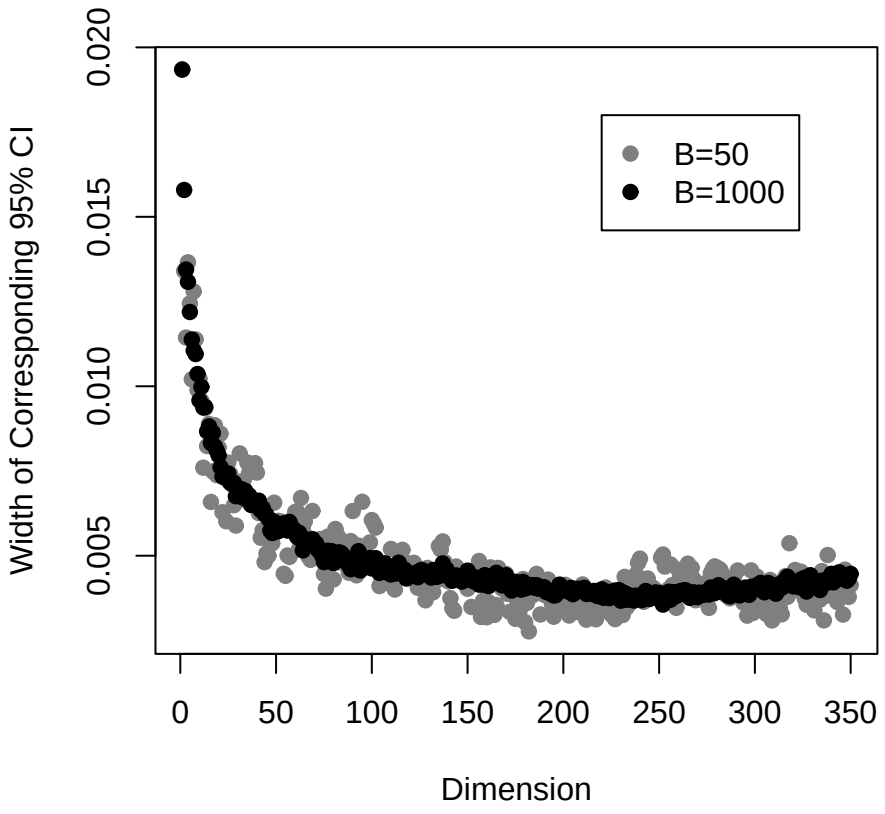

FIGURE 4.2.5. Widths of 95% null eigenvalue confidence intervals

satisfy our confidence requirements. This is not at all surprising. However, insofar as the confidence intervals for $B = 50$ appear quite narrow in their own right, From Figure 4.2.6 it is evident that we gain little information about the distribution of $\widehat{\lambda}^*_{0k}$ by undertaking many more simulations. Thus setting $B = 100$, as reported in this study, appears to give a good estimate of the true null eigenvalue confidence interval. Setting $B = 1000$, for example, did not improve APA's dimensionality estimates in this experiment to a significant degree, and required a ten-fold increase in processing time.

Overall, amended parallel analysis appears to improve dimensionality estimation for IR over Horn's parallel analysis. Confronted with the vastly complex models native to

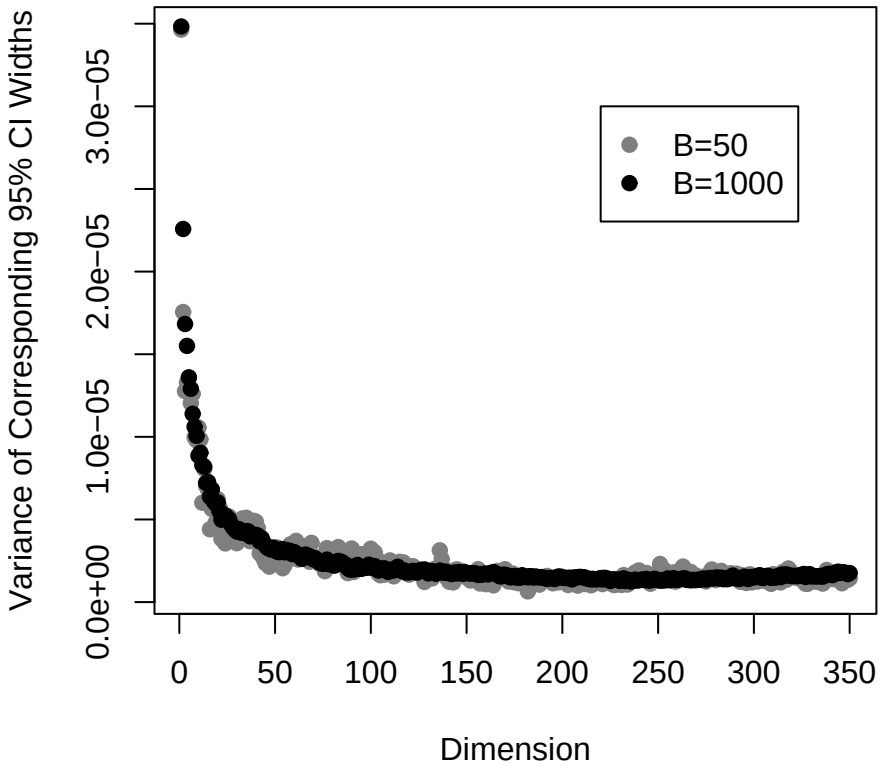

FIGURE 4.2.6. Variance of confidence interval width

retrieval, PA consistently underestimated the optimal dimensionality discerned by Cranfield-style performance evaluation techniques. By replacing PA's point estimate of each null eigenvalue λ_{0k} with a $1-\alpha\%$ confidence interval, APA improved dimensionality estimates at a level that was statistically significant for the six corpora tested here. However, even this amendment appears to be too little to offset PA's problems when confronted with a data set that merits no significant dimensionality truncation. The *CF* database, for instance, did not appear to benefit from dimensionality reduction. Thus APA's failure to derive a high-rank model for *CF* must be interpreted as a defect. However, this judgement must also be tempered by the fact that the ability of Cranfield-style evaluation to discern the intrinsic dimensionality of a data set is itself suspect. APA, then, appears to improve

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	730	76	45	80	94	87
PA	697	71	42	72	81	75
EV1	1288	554	171	651	613	471

TABLE 4.2.11. EV1, PA, and APA dimensionality estimates

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MED*
APA	0.228	0.061	0.115	0.055	0.067	0.084
PA	0.218	0.057	0.107	0.049	0.058	0.073
EV1	0.402	0.447	0.436	0.446	0.438	0.456

TABLE 4.2.12. EV1, PA, and APA dimensionality estimates (Normalized)

upon traditional parallel analysis. However, it is more difficult to say categorically whether APA offers superior dimensionality estimates to the remaining three eigenvalue analysis techniques.

4.2.2.2. *Performance of the Eigenvalue-One Criterion.* As noted in Sections 2.3.3 and 3.3, parallel analysis and the eigenvalue-one criterion share basic mathematical and statistical assumptions. In fact the EV1 approach to dimensionality estimation is the parallel analysis approach, with the exception that EV1 treats the observed correlation matrix as if it were the population correlation matrix. Thus APA may be understood as yet another refinement on the EV1 procedure. Like PA, APA takes account of the fact that the observed correlation matrix is a sample. But unlike traditional PA, APA also recognizes that the derived null eigenvalues have a sampling distribution of their own. All three approaches—EV1, PA, and APA—share the notion that an LSI model should retain as many eigenvectors as there are independent variables in the PDF that generated the term-document matrix **A**.

Table 4.2.11 shows the dimensionality estimates afforded by PA, APA, and EV1. Table 4.2.12 shows the same data, with each estimate normalized by the rank of the term-document matrix to lie between 0 and 1.

Interestingly, the EV1 approach always retained between 40% and 45% of the eigenvectors, a fairly narrow window. This is not at all surprising, insofar as it has been shown (cf. [**105**]) that the eigenvalues of large, term-document matrices follow very consistent power-law distributions. Thus it seems that the estimates afforded by the EV1 criterion are

heavily driven by this tendency. That is, given Mihail's demonstration that scree plots of IR problems tend to look the same, the EV1 criterion offers a similar dimensionality estimate (with regard to the proportion of eigenvalues retained) for any corpus. On the other hand the parallel analysis-based techniques derived models of widely divergent complexity, calling for between about 5% and 22% retention.

The difference between APA and EV1 was statistically significant. As described above, I tested for equality of estimates across corpora. A paired t-test comparing APA's and EV1's estimates yielded $p = 0.001$. Tables 4.2.11 and 4.2.12 suggest two important differences between the EV1 and APA approaches to dimensionality estimation:

(1) APA yields models of fewer dimensions than the EV1 criterion.

(2) APA's sensitivity to the sampling error in the observed correlation matrix makes its dimensionality estimates more specific to the data at hand than the EV1 criterion's models.

Whether APA's greater sensitivity to the observed data entails an improvement over EV1 is difficult to say, given the noisiness of IR evaluation and the small sample of corpora studied here. APA gave the best estimate on four occasions, while EV1 was best only three times. Neither EV1 nor APA was ever the worst performer, although traditional PA often gave the worst estimate. More damning for APA, perhaps, was its tendency to underestimate model dimensionality. That is, in the cases where APA appeared to fare poorly (for example, on the *CF* data), it retained far too few eigenvectors. Retaining too many dimensions is apt to incur a relatively mild overfitting error in retrieval. On the other hand, rejecting too few dimensions will rob the model of important discriminatory power. Thus PA's and APA's tendency to under-estimate is worrisome.

However, the degree of advantage enjoyed by APA versus EV1 appears to be related to the applicability of dimensionality reduction itself to a given corpus. In the case of the *MEDLINE* data, and for the *CRAN* data's ASL measurements, a strong semantic subspace was evident. In these cases, APA gave the best estimates among all tested statistics. EV1 gave the best performance on three corpora (*CACM, CF_FULL,* and *CISI*) that saw broad

divergence among the Cranfield-style analysis techniques. In these cases, the notion of optimality is therefore somewhat suspect.

In sum, it appears that using the EV1 criterion offers a very conservative, but also effective, approach to dimensionality estimation. Like parallel analysis, EV1 begins with the assumption that dimensionality reduction is merited to the extent that the observed variables depart from independence. If the term-document matrix **A** were orthogonal, all eigenvalues would equal 1. In such a case no dimensionality reduction is merited according to EV1, PA, or APA. Under all three criteria, we reject eigenvalues that are smaller than the eigenvalues predicted if the indexing features were independent. The difference between these approaches lies in their notion of what "independent features" actually means and in what it means for data to deviate from independence. Whereas EV1 rejects eigenvalues less than 1, PA and APA reject eigenvalues that are "significantly less" than 1, where "significantly less" is defined by the distribution of null eigenvalues, as described in Section 3.3.

EV1 tended to over-estimate the optimal dimensionality of models for the corpora tested here, according to the three tested performance metrics. APA and PA entail an attempt to mitigate this over-estimation by making inferences about the deviation of the observed terms from independence by recourse to statistical simulation. According to my results, the differences between APA and EV1 are statistically significant. While it is difficult to say whether this difference implies an advantage, it appears that EV1 is a safer approach to dimensionality estimation than APA. However, APA appears to offer superior estimates in cases when dimensionality reduction itself is obviously appropriate.

4.2.2.3. *Performance of the Percent-of-Variance Approach.* The percent-of-variance approach to dimensionality estimation has seen broad criticism in the statistical literature (cf. [**76**]) due to its inherently *ad hoc* character. Critics of this approach argue that selecting m, the percentage of total variance that the final model should account for involves poor theoretical and empirical motivation. That is, choosing to retain, say, 95% of the total variance does nothing to help us understand the relationship between the reduced and full-rank models. Moreover, statisticians such as Jackson have argued that no universally suitable value for m

	CACM	*CF*	*CF_FULL*	*CISI*	*CRAN*	*MEDLINE*
APA	0.74	0.14	0.18	0.12	0.15	0.16
EV1	0.87	0.66	0.54	0.66	0.68	0.63

TABLE 4.2.13. Amount of variance retained in APA and EV1 models

is forthcoming. That is, one cannot choose a value of m and apply it in good conscience to all data sets.

Despite these criticisms, using an 85% of variance criterion for dimensionality estimation yielded good results in this study. While APA was the best performer for four observations, the percent-of-variance approach performed best on six occasions, making it the most frequent best dimensionality predictor. Moreover, the 85% Var approach never fared worst among the dimensionality estimation techniques that I tested.

However, I believe that the observed success of the 85% Var technique is rather misleading, and I argue against its application for dimensionality estimation in IR problems. Table 4.1.1 above suggests why I am skeptical about the value of a percent-of-variance approach. The problem lies in the fact that the observed optimal dimensionalities of the six test corpora varied widely with regard to their cumulative variance. The rows of Table 4.1.1 labeled *var at* $k_{opt}(.)$ show the percent of total variance accounted for by the optimal LSI model with regard to a given performance metric. The values for this measure vary tremendously. For example, consider the ASL measure. The optimal model of *MEDLINE* for ASL retained only 16% of the total variance, while the optimal model for *CF* accounted for 95% of the total variance. The distribution for average precision and optimal F were also wide. For both optimal F and average precision the optimal *MEDLINE* model retained 25% of the initial variance, while *CACM* demanded a full-rank representation for each of these metrics to be optimized. Given such disparity, it is difficult to justify adopting an across-the-board rule for optimizing LSI models. No value of m allows us to optimize models of all six corpora with an m% of variance dimensionality retention criterion.

Further evidence that no value of m exists that will be generally optimal for IR problems is presented in Table 4.2.13, which shows the percent of variance accounted for by the models selected via the APA and EV1 criteria. As discussed above in Section 4.2.2.2, the EV1 criterion yielded models of very consistent size, insofar as it retained 40%-45% of the total

eigenvectors for all six corpora. However, this consistency did not translate into models that account for consistent amounts of variance. Although EV1 retained approximately 65% of the initial variance for the *CF, CISI, CRAN,* and *MEDLINE,* it retained much more than this for the larger *CACM* corpus, and significantly less than 65% for the full-text cystic fibrosis data. The variance accounted for by APA's models are even less predictable, with optimal models retaining between 12% and 74% of the total variance.

In light of this discussion it seems unlikely that any systematic rule governs the amount of variance that should comprise an optimal LSI model. Via retrospective performance analysis we have seen that neither ASL, average precision nor optimal F is optimized for the tested corpora at any given level of m. Nor did I find a consistent amount of variance accounted for via either of the top-performing eigenvalue analysis techniques. Thus I consider the apparent success of the 85% Var approach to be an artifact of the noisy portrait of the data sets' semantic subspaces discussed in Section 4.1.1. That is, given that many observations appeared to be optimized near full rank, with only minor evidence of overfitting beyond this point, the 85% Var approach succeeded on several occasions by virtue of offering consistently high dimensionality estimates.

It should be noted further that an 85% Var approach is widely out of step with the mainstream of multivariate statistical theory. That is, more usual values for m are 95% or perhaps 90%, which have been defended as outgrowths of traditional hypothesis testing with its corresponding levels of confidence. Thus an 85% of variance rule pushes the heuristic approach to dimensionality estimation past even these contentious statistical bounds. Perhaps the best thing to be said for this approach, then, is that it appears to work well in some cases. However, its improvement over the EV1 approach was negligible; a paired t-test on the equality of estimates provided by EV1 and 85% Var yielded $p = 0.14$. Thus it seems more desirable to use the theoretically sound EV1 criterion in place of a more arbitrary percent of variance solution.

4.2.2.4. *Performance of Bartlett's Test of Isotrapy.* As mentioned in Section 2.2 I included consideration of Bartlett's test of isotrapy largely in the interests of completeness. It is widely known that this method tends to over-estimate the number of dimensions. In the

case of IR, this tendency surfaces writ large. In fact, for all six corpora, Bartlett's test only rejected the last two eigenvalues, leading to a nearly full-rank model. At first glance this implies that Bartlett's approach has no practical or theoretical merits for IR applications. Insofar as IR's models are more complex than those for which Bartlett's technique was developed, this may be accurate.

However, we should bear in mind that the evidence collected here suggests that no dimensionality reduction is merited for several corpora, at least according to a given performance metric. Performance on the *CF* data appeared to suffer whenever a significant number of its eigenvectors was removed from an LSI model. Likewise, *CACM* performed best according to average precision and optimal F under a full-rank model. Before discounting Bartlett's test out of hand, it is important to consider the larger question of whether its high-dimensional models were in fact erroneous, or whether they reflect a valid argument that all dimensions should be retained.

Bartlett's test provided the best dimensionality estimate for the *CF* data according to the ASL metric. It was also the best predictor for the *CACM* corpus, *vis a vis* optimal F. However, Bartlett's provided the worst estimates on nine occasions. While the agreement of all three performance metrics on a high-dimensional model for *CF* suggests that Bartlett's might be correct in its estimation for that corpus, there was strong inter-metric disagreement for *CACM*, which calls into question the merit of its estimate.

But perhaps most damning was the inability of Bartlett's to tailor its estimates at all for different corpora. That the technique rejected two principal components for all six corpora despite their emphatically different statistical properties (cf. Table 3.1.1) and different observed optimal dimensionality profiles (cf. 4.1.1) smacks of a glaring defect. This implies that the χ^2 distribution of the Bartlett's test statistic is simply ill suited to the over-sized models native to IR. Given many references in the literature to Bartlett's failures in the face of high-dimensional data (cf. **[3, 76, 116]**) this is not surprising. Thus I attribute the technique's successes here more to failures either in the suitability of LSI to the test data or to shortcomings in the Cranfield paradigm's ability to address the intrinsic dimensionality of corpora.

4.2.3. Overview of Results for each Corpus. This study has analyzed the ability of five dimensionality estimation techniques to discern the intrinsic dimensionality of six IR test collections. In the previous section I outlined the results of my experiments, organizing my discussion around a treatment of each estimator's accuracy. In this section, I summarize these results, this time organizing my findings by corpus.

The most conclusive results were obtained for the *MEDLINE* data. This collection evinced an obvious semantic subspace of approximately 100 dimensions; all three Cranfield-style performance metrics agreed on this. For the *MEDLINE* collection, APA gave the most accurate estimate of the intrinsic dimensionality among the five tested eigenvalue analysis techniques. Traditional parallel analysis yielded the second-best estimate, while Bartlett's egregiously overestimated the intrinsic dimensionality. As in all six experiments, EV1 and 85% Var delivered models of middling complexity for *MEDLINE*.

Dimensionality reduction was also highly successful for the *CRAN* data. It must be admitted that ASL disagreed with average precision and optimal F on the intrinsic dimensionality of this corpus. However, this disagreement may have been assuaged if I had removed universally non-relevant documents from the analysis. Had I done this, average precision may well have approached ASL in its estimation of *CRAN's* intrinsic dimensionality. Despite this interference, however, APA and PA again performed well on *CRAN*, offering the best estimate with respect to ASL. The moderately sized EV1 and 85% Var estimates were best according to average precision and optimal F.

A similarly complex picture emerged from my analysis of *CACM*, which showed a strong semantic subspace under the lens of ASL, but not under average precision or optimal F. Thus parallel analysis provided *CACM's* best estimate *vis a vis* ASL while simultaneously giving the worst estimate according to optimal F. Though PA's delivery of the worst estimate in this and several other cases is worrisome, in Chapter 5 I argue that parallel analysis does not appear to underestimate model dimensionality systematically. The conflicted case of *CACM*—where PA delivered both best and worst estimates simultaneously—suggests that PA's defects may be distorted by artifacts of the Cranfield-style evaluation reported here.

The *CISI* database appeared to merit little dimensionality reduction. In fact Bartlett's nearly full-rank model provided the best estimate for *CISI* with respect to average precision and optimal *F*. The EV1 approach gave the best estimate according to ASL. Moreover, the amount of benefit yielded by the optimal LSI models was very small for *CISI,* reinforcing the contention that this database responded poorly to dimensionality reduction.

Of particular interest were the results for *CF* and *CF_FULL*. As described in Section 3.1, *CF_FULL* is a subset of *CF*. Whereas *CF* represents each document by its title and abstract, *CF_FULL* provides a full-text representation of each document. Thus although these corpora treat the same general subject matter, they do so under very different representations. The full-text representation responded well to dimensionality reduction. Yet perhaps because of the large termspace of *CF_FULL,* this data set required a relatively complex model to perform optimally under LSI. Thus EV1 and 85% Var excelled for *CF_FULL*. On the other hand, *CF* did not benefit significantly from dimensionality reduction; all models except Bartlett's underestimated the observed optimal dimensionality of *CF*.

4.3. Concluding Remarks

Returning to the initial research question of Section 1, it appears that a statistical analysis of co-occurrence matrix eigenvalues yields useful but not infallible evidence for parameterizing k, the dimensionality of an LSI model. In the reported experiments, choosing the dimensionality of an LSI system proved to be both important and difficult. Of the six corpora tested, each appeared to have a unique observed optimal dimensionality. Thus the *ad hoc* approach to dimensionality estimation that has been common in LSI implementations appears to be ill advised. Because each corpus appeared to have a unique optimal dimensionality, there seems little motivation for a heuristic approach to parameterizing k; retaining, say, 100 eigenvectors (as is the default for the Bellcore LSI software) would have led to severely sub-optimal results for the *CF* data. Furthermore, because the tested performance metrics often disagreed about the best dimensionality for a given corpus, I argue that selecting k by finding a value that leads to good performance is also risky. My results

suggest that Cranfield-style analysis gives *some* information about the intrinsic dimensionality of a corpus, but that this information is often noisy. Some corpora (e.g. *CF*) appear to benefit from little or no dimensionality reduction, while others (e.g. *MEDLINE*) respond well to a 90% dimensionality reduction. Measuring precision and recall at various levels of k can help system designers select an appropriate model. However, the relevance judgements necessary for such an analysis are often lacking in practical IR applications. And even in the presence of pre-existing relevance information, my results suggest that without a large number of queries, dimensionality estimation by Cranfield-style analysis is hardly foolproof.

Given the difficulties inherent in dimensionality estimation, then, I consider the performance of eigenvalue-based predictors to be highly encouraging. In particular, the family of predictors comprised by PA, APA, and EV1 performed especially well. Although PA often suffered due to its tendency to underestimate model dimensionality, APA's confidence-intervals assuaged this defect to a statistically significant degree. APA gave the best estimate on four occasions. EV1 was the best predictor on three observations. Altogether, these approaches to dimensionality estimation yielded the best results for eight of the eighteen observations.

Although the five dimensionality estimators tested here varied in their actual estimates, my results imply that the best estimates come from APA, EV1 and PA, all of which share a common theoretical motivation. The rationale that underpins APA, PA and EV1 is that LSI's dimensionality reduction is tantamount to an error correction procedure. Each of these criteria rejects those principal components whose corresponding eigenvalues are less than what we would expect given independent terms. For the EV1 approach, this rationale is taken to the extreme. Rejecting all eigenvalues less than 1 admits no distinction between the correlation matrix of the term-document matrix $\mathbf{A}$ and the correlation matrix of the multivariate PDF that generated $\mathbf{A}$. Thus the only way to achieve a full-rank model under the EV1 criterion is if $\mathbf{A}$ is an orthogonal matrix. PA relaxes this stringency. Instead of demanding numerical orthogonality among the columns of $\mathbf{A}$ for eigenvector retention, PA implies that we should retain as many dimensions as there are statistically independent variables in the PDF that generated $\mathbf{A}$. Thus PA accounts for the fact that the observed

correlation matrix is only a sample from a larger population. Finally, APA takes this approach one step farther, resting its dimensionality estimate on confidence intervals derived from resampling "null eigenvalues" based on **A**.

APA, PA, and EV1 constitute a family of dimensionality estimation techniques that view dimensionality reduction as a means of correcting the VSM similarity function's erroneous assumption of term orthogonality. The traditional vector space model assumes term independence. Wong's generalized vector space model accounts for this error by introducing the term correlation matrix into similarity judgements. I argue that LSI improves retrieval by building on Wong's approach. LSI's dimensionality reduction derives a low-rank approximation of the term correlation matrix. This low rank approximation aids retrieval by improving the system's estimate of the population correlation matrix. In other words, an optimally parameterized LSI system provides the best model, in the least squares sense, of the relationships that obtain between terms and documents in the population. The success of the three error-correction-based dimensionality estimators lends credence to this argument.

CHAPTER 5

Dimensionality Estimates for Simulated Data

To compare the quality of each eigenvalue-based dimensionality estimator in an idealized environment, I conducted a series of data simulations. This chapter describes these simulations and their results. Data gleaned from the retrospective IR performance evaluation described in Section 4.1 suggest that analyzing term and document co-occurrence matrix eigenvalues yields useful information for intuiting the intrinsic dimensionality of a corpus. In particular, my analysis of Section 4.2 showed that the family of dimensionality estimators based on an error-correction rationale—APA, PA, and EV1—were especially well suited to this task. However, I also noted that discerning k_{opt} by tracking ASL, average precision, and optimal F across increasingly complex models carries some risk of error. Do the supplied queries adequately demonstrate the dimensional structure of the corpus? What are we to think of corpora whose observed optimal dimensionality varies across performance metrics? Are an estimator's observed successes due to real advantage, or are they merely byproducts of the vagaries of Cranfield-style evaluation? Without *a priori* knowledge of a corpus' intrinsic dimensionality, stating conclusively how well an eigenvalue analysis technique discerns the dimensionality of data's optimal semantic subspace is problematic.

The simulations undertaken in this chapter address three broad questions:

(1) Given a corpus of rank r and intrinsic dimensionality $k_{opt} \ll r$, How well does each dimensionality estimator perform as we increase or decrease the noise in the system?

(2) how does each dimensionality estimation technique fare when presented with a corpus $\mathbf{A}$ where dimensionality reduction is inappropriate, i.e. $k_{opt} = rank(\mathbf{A})$?

(3) Are the dimensionality estimates afforded by each eigenvalue analysis technique self-consistent and mutually distinct? In other words, will an estimator e yield the

same estimate when applied to similar problems? And in general, how different are the estimates afforded by two estimators e and e'?

As opposed to the corpus-based analysis reported in Chapter 4, simulations allow us to attack these questions in a more direct fashion. Instead of inferring the "right answer" based on possibly noisy Cranfield-style evaluation, simulating the problem allows us to begin from the solution and work backwards. While this is desirable insofar as it lends our evaluation a precise instrument for gauging an estimator's accuracy, simulation raises a host of questions in its own right. Section 5.1 discusses my rationale in designing the simulations, detailing the mathematics that governs my approach. In Section 5.2 I discuss the actual parameters and data sets that were generated in the simulations, along with an account of the methods of analysis that I bring to these data. Section 5.3 relates the outcome of the simulation experiments. Finally I summarize the implications of my findings in Section 5.4.

5.1. Construction of the Simulations

Constructing simulations to test criteria for principal component retention is a non-trivial task. While a fairly *ad hoc* approach to simulation has been presented in the literature, I pursue a different method in this study. This section begins by describing the most common technique for simulation in dimensionality estimation problems, outlining some of its deficiencies. I then turn to a description of my own technique, using the deficiencies of the prior approach as a motivation.

5.1.1. Past Approaches to Simulations for Dimensionality Estimation . A wide body of literature describes the behavior of number-of-factors rules with simulated data. For instance, Hakstian (cf. [**63**]), Zwick and Velicer ([**153**]), and Jackson (cf. [**76**]) use eigenvalue analysis methods such as those described in Table 3.4.1 to estimate the number of significant principal components in simulated correlation matrices. Glorfeld ([**57**]) and Linn ([**95**]) use simulated data to test the performance of parallel analysis, in particular. Thus there is ample precedent for using a simulation-based approach to dimensionality estimation.

In previous simulation studies, researchers have defined the "true" dimensionality of a data set by constructing a population covariance matrix with a highly controlled structure.

In [**76**], Jackson calls this a matrix with "block" structure. The block structure approach to simulation involves choosing p (the number of variables), k_{opt} (the true dimensionality), r (the degree of factor loading), and f (the amount of background noise). An $n \times p$ data set is then drawn from a multivariate distribution (usually Gaussian) with a zero mean vector and this structured covariance matrix. For example, consider the population covariance matrix $\mathbf{\Sigma}$, with $p = 9$, $k = 3$, $r = .8$, and $f = 0$:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & .8 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\ .8 & 1 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\ .8 & .8 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & .8 & .8 & 0 & 0 & 0 \\ 0 & 0 & 0 & .8 & 1 & .8 & 0 & 0 & 0 \\ 0 & 0 & 0 & .8 & .8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & .8 & .8 \\ 0 & 0 & 0 & 0 & 0 & 0 & .8 & 1 & .8 \\ 0 & 0 & 0 & 0 & 0 & 0 & .8 & .8 & 1 \end{pmatrix} \tag{5.1.1}$$

Despite its $p = 9$ variables, this matrix appears to be of significantly lower dimensionality. In Jackson's analysis, the three blocks of correlated variables imply that data generated from a distribution with covariance matrix $\mathbf{\Sigma}$ will be three-dimensional.

However, it remains unclear in Jackson's approach (and that of the other authors cited above) exactly what it means to call $\mathbf{\Sigma}$ three-dimensional. Consider the eigenvalues of $\mathbf{\Sigma}$:

$$\boldsymbol{\lambda}' = \begin{pmatrix} 2.6 & 2.6 & 2.6 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix} \tag{5.1.2}$$

Matrix $\mathbf{\Sigma}$ has nine linearly independent rows and columns, and is thus of full rank. Obviously the magnitude of the first three eigenvalues dwarfs the rest, implying that eigenvalues 4 through 9 are minor. But the rationale for discarding these "minor" eigenvalues is no more founded than our attempt to discern an elbow in a scree plot.

Creating k-dimensional data by recourse to so-called "block-structured" covariance matrices deprives simulations of the clarity that researchers desire from them. To exemplify the

shortcomings of simulations based on blocked covariance matrices, consider LSI's rationale. Given a $n \times p$ document-term matrix $\mathbf{A}$, LSI operates on $\widehat{\mathbf{A}}_k = \mathbf{T}_k \mathbf{S}_k \mathbf{D}_k'$, where $\mathbf{T}_k$ and $\mathbf{D}_k$ contain the first k singular vectors of $\mathbf{A}$, and $\mathbf{S}_k$ has the k largest singular values on its main diagonal. LSI's proponents argue that $\widehat{\mathbf{A}}_k$ provides a more accurate basis for a VSM similarity model than the full-rank $\mathbf{A}$ can, due to overspecification error in the sample matrix. As argued in Section 1.1.2, dimensionality reduction is motivated by the notion that the first k principal components of $\mathbf{A}$ provide a superior model of the population term correlation matrix than the observed, full-rank covariance matrix $\widehat{\mathbf{\Sigma}}$ can.

However, under the block structure rationale, this benefit by matrix approximation is frustrated. As an example of why, let $\mathbf{\Sigma}$ be defined as in Equation 5.1.1. Let $\boldsymbol{\mu}$ be a nine-dimensional zero vector. I created the 500×9 matrix $\mathbf{A}$ (i.e. 500 observations on 9 variables) by sampling 500 vectors from the multivariate normal distribution, $N(\boldsymbol{\mu}, \mathbf{\Sigma})$. Based on $\mathbf{A}$ I calculated the 9×9 observed covariance matrix $\widehat{\mathbf{\Sigma}}$. Due to the distribution of $\mathbf{A}$, we expect $\widehat{\mathbf{\Sigma}}$ to be similar to $\mathbf{\Sigma}$. But sampling error will introduce noise between the population covariance matrix $\mathbf{\Sigma}$ and the sample $\widehat{\mathbf{\Sigma}}$. The intuition behind LSI suggests that if $\mathbf{\Sigma}$ is truly $k = 3$-dimensional, then we should obtain a better estimate of $\mathbf{\Sigma}$ by retaining only the first three eigenvectors of $\widehat{\mathbf{\Sigma}}$. If dimensions 4 through 9 are noise, $\widehat{\mathbf{\Sigma}}_k = \mathbf{V}_k \boldsymbol{\lambda}_k \mathbf{V}_k'$ (where $\mathbf{V}_k$ is the first k eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\boldsymbol{\lambda}_k$ is the diagonal matrix of its first k eigenvalues) should be closer to $\mathbf{\Sigma}$ than $\widehat{\mathbf{\Sigma}}$, in the least-squares sense.

Despite our best efforts, Figure 5.1.1 shows that under the block structure model, the $k = 3$-dimensional model is not the best approximation of $\mathbf{\Sigma}$. Let Equation 5.1.3 define a loss function for an LSI model of given dimensionality k:

$$\ell(k) = \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_k \right\| \tag{5.1.3}$$

where $\|\cdot\|$ denotes the L_2 norm. Thus the value of k that minimizes $\ell(k)$ provides the closest approximation of the population covariance matrix. The x-axis of Figure 5.1.1 is k, the number of eigenvectors included in our model of the population covariance matrix $\mathbf{\Sigma}$. On the y-axis we plot $\ell(k)$. If $\mathbf{\Sigma}$ were truly three-dimensional, we would expect $min(\ell(k)) = \ell(k = 3)$. This is clearly not the case. Instead, we see a tremendous improvement in model

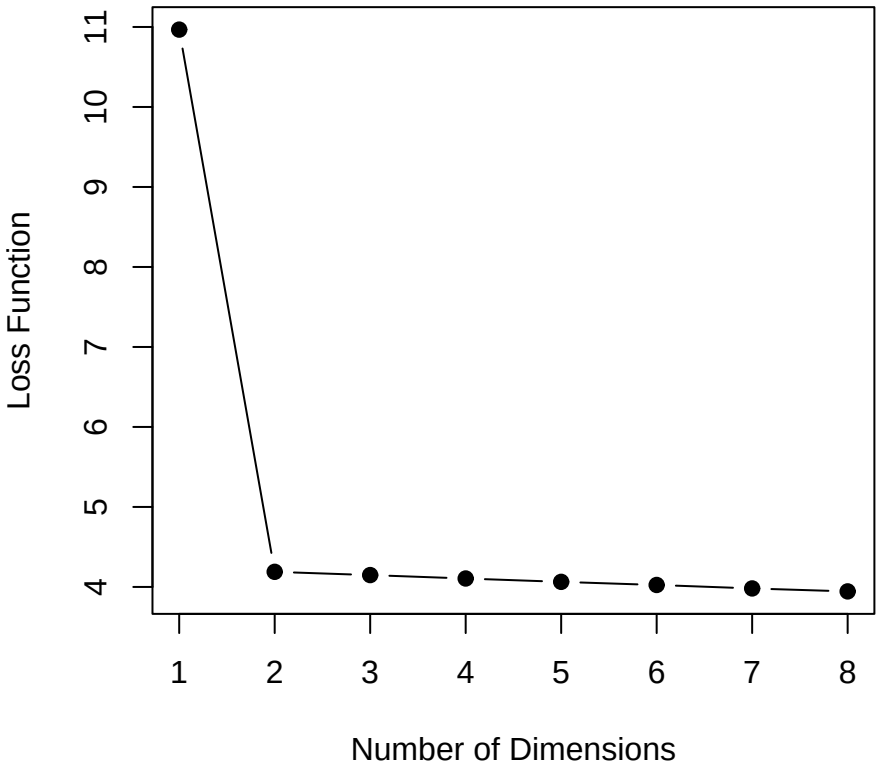

FIGURE 5.1.1. Loss function on $\mathbf{\Sigma}$

fit by including the second dimension. But after that, each additional dimension improves the fit by a negligible amount, ending in a full-dimensional model as the optimum.

While it is clear that the block structure approach to simulation creates data whose scree plots suggest an obvious intrinsic dimensionality (cf. Equation 5.1.2), the previous example shows that these data do not directly address the problem of LSI. In the example above, a model of very low dimensionality ($k = 2$) provides a good approximation of the true covariance matrix. But the 2-dimensional model is not optimal; instead the model is optimized for $k = k_{max}$. Using the sum of squared error as a goodness of fit criterion, then, $k = 3$ never enters the scene as a possible value for the intrinsic dimensionality, despite its obvious appeal on inspection of the eigenvalues. Thus it is problematic to base

our simulations on this approach, insofar as we desire to simulate data whose intrinsic dimensionality is decisively known.

5.1.2. Simulations based on an Explicit Model of the Eigenvalues. Instead of the block structure approach to simulation, I propose an agenda based on an explicit model of the population eigenvalues. Consider the matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{5.1.4}$$

with eigenvalues:

$$\boldsymbol{\lambda}_c' = \begin{pmatrix} 3 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.1.5}$$

Despite its nine rows and columns, matrix $\mathbf{C}$ is only of rank three, as evidenced by its three non-zero eigenvalues. My simulations begin by considering $\boldsymbol{\lambda}_c$ to be the eigenvalues of the true population covariance matrix for our data. Thus our population covariance matrix

contains only three linearly independent variables, which I model by matrix $\mathbf{\Sigma}_c$:

$$\mathbf{\Sigma}_c = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.1.6}$$

Before generating data based on this population covariance matrix, I subject it to a perturbation. To accomplish this let f be a positive-valued number describing the amount of noise we wish to introduce into the system. I thus define the perturbed population eigenvalues as $\boldsymbol{\lambda}' = \boldsymbol{\lambda}'_c + \mathbf{f}$, where $\mathbf{f}$ is a 9-vector with each element equal to f. Thus if $f = 1$ we have the final population covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.1.7}$$

which gives eigenvalues:

$$\boldsymbol{\lambda}' = \begin{pmatrix} 4 & 4 & 4 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}. \tag{5.1.8}$$

The implication is that there are only k variables at work in the data. However, due to our perturbation, we have introduced randomness and redundancy into the system, thus giving the appearance of p dimensions. Hence we have k large eigenvalues and $p - k$ small eigenvalues. As in LSI, the goal is to discover which eigenvalues are so small that they correspond to zero-elements in $\mathbf{\Sigma}_c$. The goal of model fitting, then, is not to approximate $\mathbf{\Sigma}$ (as in the block structure approach to simulation discussed above). Rather, the point is to derive the best estimate of $\mathbf{\Sigma}_c$.

Before leaving our discussion of the simulation design, consider momentarily how this design relates to my theorized rationale for LSI's dimensionality reduction. In Sections 1.1.2 and 3.3 I argued that *APA* is well-suited to the dimensionality estimation problem due to its orientation towards error correction. I suggested that *APA*'s merits would lend evidence to my theory that dimensionality reduction in IR amounts to a correction applied to the GVSM similarity function. At that point I described the correction in terms of term-term correlation, suggesting that dimensionality reduction is merited to the extent that the indexing features depart from orthogonality. However, in the simulation approach just described, the population covariance matrix is diagonal; where is the error in need of correction, then? Consider matrices $\mathbf{\Sigma}_c$ (Equation 5.1.6) and $\mathbf{\Sigma}$ (Equation 5.1.7). The error lies in the perturbation that transforms $\mathbf{\Sigma}_c$ into $\mathbf{\Sigma}$. These matrices comprise the population eigenvalues of the PDF that generates our simulated data. Thus the appearance of $p - k$ non-zero eigenvalues in $\mathbf{\Sigma}$ implies the addition of spurious correlations among the variables.

Whereas the block structure simulation approach advocated by Jackson models the correlational structure explicitly, my approach models the eigenvalues explicitly. Under my approach, the data are truly k dimensional, and the inclusion of $k' > k$ dimensions in the model incurs model error by overspecification. As in real-world LSI, then, dimensionality reduction for our simulated data acts as an error correction procedure, removing spurious correlational data from the estimation of the population covariance matrix by recourse to an analysis of the eigenvalue distribution.

5.1.2.1. *Steps in the generation of simulated data.* Performing a simulation under the explicit model of population eigenvalues demands that we parameterize five variables, shown in

Symbol	*Description*
p	The number of variables
k	The intrinsic dimensionality
λ	The magnitude of the true eigenvalues
f	The noise coefficient
n	The sample size for the simulated data set

TABLE 5.1.1. Simulation parameters

Parameter	*Value*
p	9
k	3
λ	2
f	1
n	1000

TABLE 5.1.2. Example simulation parameters

Table 5.1.1. Next we define $\boldsymbol{\lambda}$, a p-vector with the first k elements equal to λ, and elements $(p-k)\cdots p = 0$. To this we add the noise factor $\mathbf{f}$, a p-vector with all values equal to f, to get $\boldsymbol{\lambda}_f = \boldsymbol{\lambda} + \mathbf{f}$. Thus we have the $p \times p$ population covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\lambda}_f \mathbf{I}_p$. Based on this we draw n samples from $N(\boldsymbol{\mu} = 0,\ \boldsymbol{\Sigma})$ to derive the $n \times p$ data matrix $\mathbf{A}$.

Having obtained a simulated data set, each simulation involves the operations of Algorithm 2. Following Algorithm 2 allows us to track the goodness of fit obtained by each value

Algorithm 2 Simulation procedure

(1) Obtain $\widehat{\boldsymbol{\Sigma}}$, the sample covariance matrix of $\mathbf{A}$.
(2) Compute $\widehat{\mathbf{V}}$ and $\widehat{\boldsymbol{\lambda}}$, the eigenvectors and eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, respectively.
(3) for $k = 1 \cdots p$
(4) Compute $\widehat{\boldsymbol{\Sigma}}_k = \widehat{\mathbf{V}}_k \widehat{\boldsymbol{\lambda}}_k \mathbf{I}_k \widehat{\mathbf{V}}'_k$
(5) Compute $\ell(k)$ as described in Equation 5.1.3
(6) Compute dimensionality estimates by each dimensionality estimation technique described in Table 3.4.1.

of k, as well as showing us how well each eigenvalue analysis technique correlates with this goodness of fit data.

For example I chose the parameters described in Table 5.1.2. Using these parameters, I iterated through the simulation process to derive Figure 5.1.2. The figure shows a clear optimum at $k = 3$, exactly as we desire. Moreover, as k increases toward k_{max}, we see an overfitting effect. Because the last $p - k$ eigenvalues correspond to variables that are not

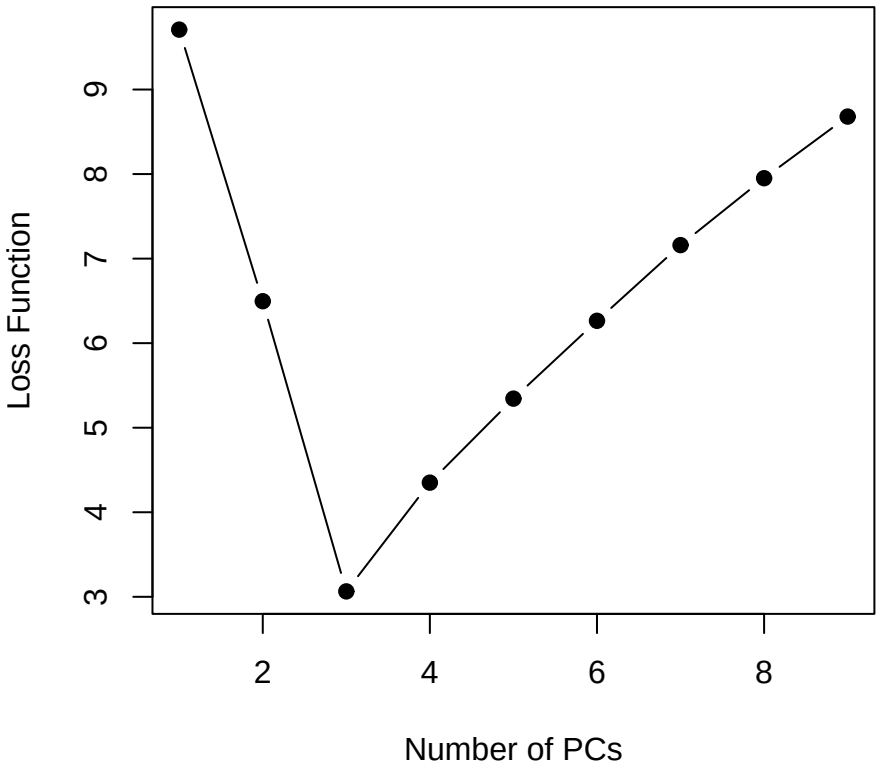

FIGURE 5.1.2. Simulation goodness of fit

present in the unperturbed population covariance matrix, adding them to the model simply introduces noise into the system.

5.2. Data Generation and Methodological Approach

As mentioned earlier, the simulations undertaken in the current study address three questions:

(1) Given a corpus of rank r and intrinsic dimensionality $k_{opt} \ll r$, How well does each dimensionality estimator perform as we increase or decrease the noise in the system?

	LRLN	*LRBN*	*LRHN*	*FRLN*	*FRBN*	*FRHN*
p (variables)	100	100	100	100	100	100
k (true dims.)	15	15	15	100	100	100
λ *(true eigenvals)*	2	2	2	2	2	2
f (noise factor)	0.5	1	1.5	0.5	1	1.5
n (sample size)	1000	1000	1000	1000	1000	1000

TABLE 5.2.1. Parameter Settings for Simulations

(2) How does each dimensionality estimation technique fare when presented with a corpus $\mathbf{A}$ where dimensionality reduction is inappropriate, i.e. $k_{opt} = rank(\mathbf{A})$?

(3) Are the dimensionality estimates afforded by each eigenvalue analysis technique self-consistent and mutually distinct? In other words, will an estimator e yield the same estimate when applied to similar problems? And in general, how different are the estimates afforded by two estimators e and e'?

To address these questions I ran a set of simulations whose parameters are shown in Table 5.2.1. The column headings of Table 5.2.1 refer to the rank of the data's unperturbed covariance matrix and the amount of noise in the system. Thus *LRLN* refers to "low-rank, low noise," while *FRHN* means "full-rank, high noise." I also define low-rank and high-rank baseline noise runs, *LRBN* and *FRBN,* each with moderate noise coefficients.

The parameters shown in Table 5.2.1 were chosen to provide a broad spectrum of dimensionality estimation problems. That is, I chose to produce data that were variously amendable to dimensionality reduction by producing low-rank and high-rank runs. I also desired to create estimation problems of varying difficulty; hence the three levels of system noise. Figures 5.2.1 through 5.2.4 visualize the simulations that I undertook.

Each figure contains two sub-figures. On the left is the scree plot derived from a simulation run at a given parameterization. The right panel shows the loss function $\ell(k)$ across all possible k values for the data.

The scree plots are intended to convey the difficulty of a given simulation's dimensionality estimation problem. Thus the *LRLN* situation (Figure 5.2.2) shows a clear demarcation in eigenvalue magnitude between the true dimensions and the noise variables. This is an easy problem, and most eigenvalue analysis techniques should discern a qualitative difference

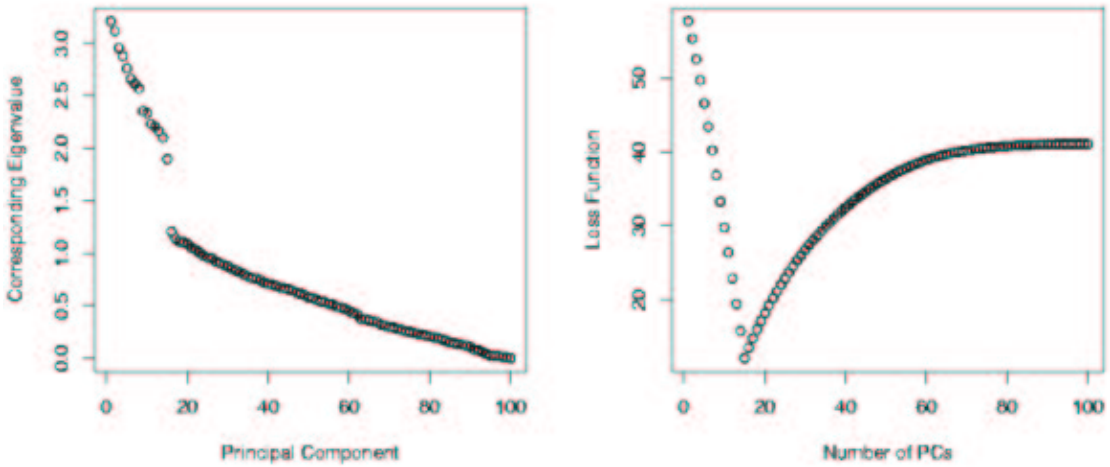

FIGURE 5.2.1. *LRBN* simulation overview

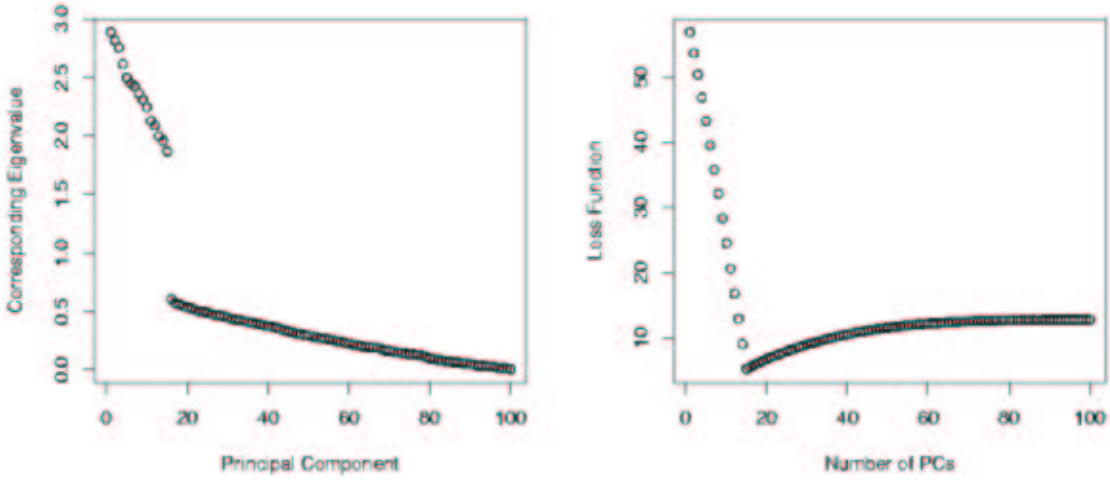

FIGURE 5.2.2. *LRLN* simulation overview

between the first 15 eigenvalues and the remaining 85. On the other hand, the *LRHN* simulation (Figure 5.2.3) presents a more difficult challenge. While an elbow is visible in the scree plot near $k = 15$, these eigenvalues lack the precipitous phase transition seen under the low-noise simulation. Thus we suspect that the high-noise simulations provide more challenge for a dimensionality estimation technique.

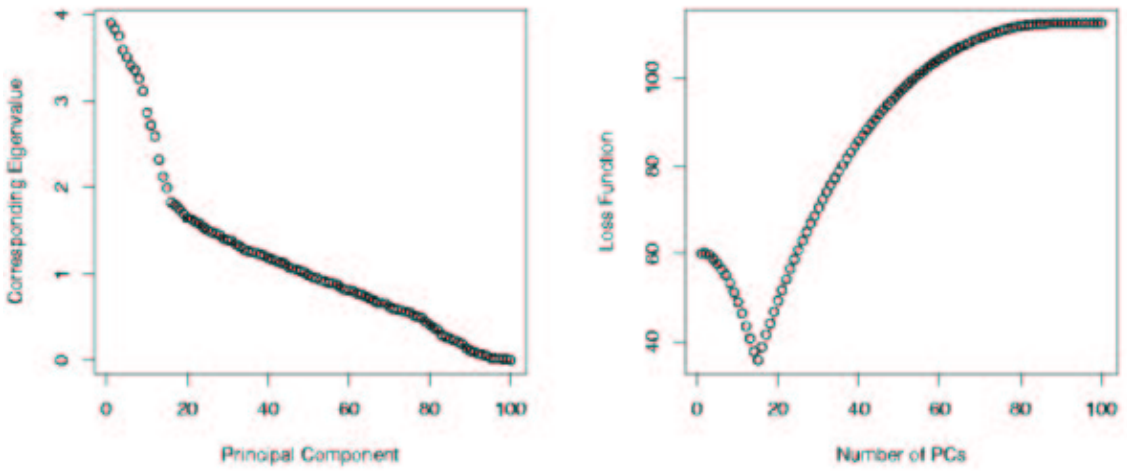

FIGURE 5.2.3. *LRHN* simulation overview

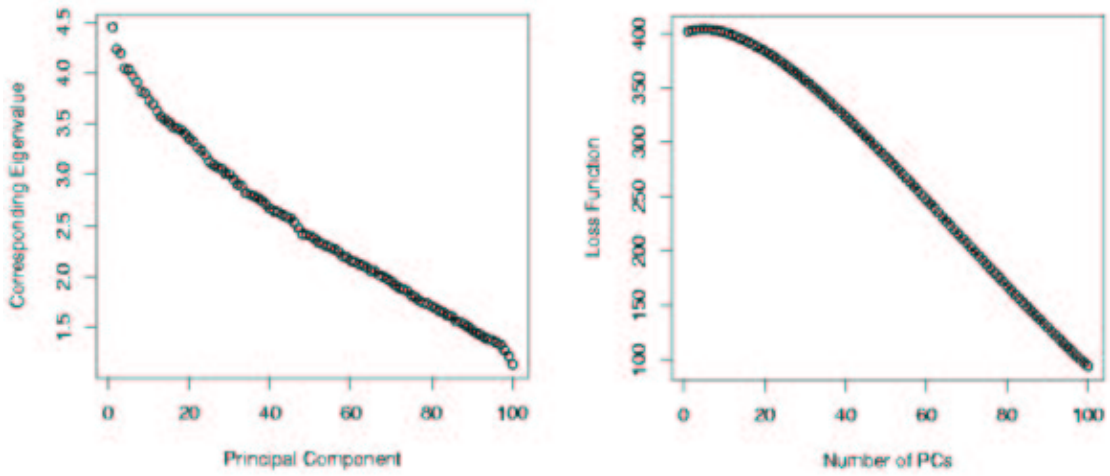

FIGURE 5.2.4. *FRBN* simulation overview

Whereas the scree plots show the difficulty of a given problem, the right-most plots (i.e. the loss plots) of Figures 5.2.1 through 5.2.4 give a sense of what is at stake at each parameterization. Each of these sub-figures shows $\ell(k)$ for $k = 1 \cdots p$. In other words, it shows the sum of squared error for each model. Low values of $\ell(k)$ imply that the k-dimensional

model is a close approximation of the true covariance matrix. In the 15-dimensional baseline simulation (Figure 5.2.1), for example, the 15-dimensional model provides the best fit. Choosing to retain too few principal components—e.g. $k = 1$—entails a large loss, while setting $k = 100$ incurs a moderate overfitting effect. On the other hand, the 15-dimensional high noise model (Figure 5.2.3) involves a very high penalty for choosing an overfitted model.

Having defined the six parameterizations shown in Table 5.2.1, I generated 50 data sets for each parameterization, for a total of 300 simulations. I chose to repeat each simulation 50 times in order to derive adequate power for the statistical tests reported in Section 5.3. In this respect my methodology followed Hakstian [**63**] and Jackson [**76**]. In each of these studies, simulations were repeated several times in order to compare the consistency of each estimation technique. However, these earlier studies relied on fewer repetitions of each simulation. A sample of $n = 50$ was chosen in the current study due to improvements in computational power since Hakstian and Jackson ran their experiments; i.e. a larger number of simulations is easy to implement now, and yields a more complete statistical picture than would arise from a smaller sample.

In the discussion that follows, I analyze the performance of each of the five dimensionality estimators given in Table 3.4.1 when they were applied to these simulated data. I address each of the three questions presented at the beginning of this section, while giving special attention to the questions that were raised in my data analysis of Chapter 4.

5.3. Results of the Simulations

Overall the simulations showed that PA and APA offer dimensionality estimates that are decisively more accurate than the other tested methods. Table 5.3.1 summarizes the results from the simulations. As in Table 4.2.1, the data here are the directed distance between each eigenvalue analysis technique's dimensionality estimate and the true dimensionality of a simulated data set. The individual values shown are the averaged errors across all 50 simulations. Thus for the first cell, we see that in the *LRLN* simulation, on average, APA over-estimated the true dimensionality by one. In other words, on average, APA's estimate was 16, in the face of a 15-dimensional data set. Especially desirable in our analysis will

	LRLN	*LRBN*	*LRHN*	*FRLN*	*FRBN*	*FRHN*
APA	1.00	1.00	23.12	0.00	0.00	0.00
PA	1.00	1.00	21.72	0.00	0.00	0.00
EV1	2.94	16.88	23.18	-53.64	-53.98	-54.32
85% Var	30.18	38.20	41.10	-24.00	-24.76	-25.02
Bartletts	83.00	83.00	83.00	-2.00	-2.00	-2.00

TABLE 5.3.1. Summary of simulation error

be weighing the merits of one dimensionality estimator versus the others. This judgement is simplified by the fact that for each simulation, the mean errors of all five dimensionality approaches lie in the same direction. For instance, for the low-rank models, all six estimation techniques over-estimated the true dimensionality, though by varying degrees. On the other hand, the full-rank case obviously provides no room for over-estimation. Thus all errors for the full-rank simulations are less than or equal to zero. This outcome allows us to compare the dimensionality estimators' accuracy simply by noting their errors' absolute value.

A number of facts are immediately apparent from inspection of Table 5.3.1. First, the *LRHN* problem appeared to be especially difficult, insofar as all six dimensionality estimators fared poorly on that series of simulations. Conversely the *LRLN* example appears to have provided a fairly easy problem. Thus, as we desire, adding noise to the low-rank models appears to change the difficulty of the dimensionality estimation problem. To test this hypothesis, I performed a Welch, two-sample t-test on the 5×50 matrices containing the errors of each method's dimensionality estimates from each of the 50 simulations under the *LRLN* and *LRHN* simulations. In other words, I tested the null hypothesis that adding noise to the low-rank model did not change the accuracy of the dimensionality estimates. This test gave $p \approx 0$, suggesting that the amount of noise in the low-rank simulations is a significant factor in the accuracy of the dimensionality estimators.

On the other hand, I note that during the full-rank simulations, adding noise to the system yielded very little variation in estimation quality. In the full-rank case, APA and PA were accurate across all noise parameterizations, while EV1 and 85% Var fared poorly for all full-rank simulations. Bartlett's continued to demonstrate its poor applicability to large-scale dimensionality estimation problems, behaving for all simulated data as it did for the real corpora, consistently rejecting only two eigenvalues. Testing $H_0 : FRLN = FRHN$

(where $FRLN$ and $FRHN$ are the 5×50 matrices of dimensionality estimate errors for each type of simulation) gave $p = 0.86$. Thus there is no statistical difference (with regard to estimation accuracy) between the full-rank models. This is completely understandable insofar as adding noise to the full-rank model only changes the magnitude of all the true eigenvalues. Because the full-rank simulations have no spurious eigenvalues, adding noise to the system merely amplifies the true eigenvalues symmetrically, a change that does not impact the problem of dimensionality estimation[1]. Thus in the following discussion, I omit comparison between the full-rank simulations, using the $FRBN$ simulation for all full-rank simulation analysis.

Figures 5.3.1 and 5.3.2 depict the outcome of the simulations graphically.

Each figure plots $\ell(k)$ versus k, with the output of each dimensionality estimation technique (from a single run) superimposed as various characters. As I describe in the sections below, Figure 5.3.1 shows that APA and PA[2] provided the best dimensionality estimates in both the low-rank and full-rank simulations. The EV1 criterion is second-best for the low-rank data, with Bartlett's offering the second-best estimate for the full-rank data, by virtue of its preference for high-dimensional models. In contrast to the real corpora analyzed in Chapter 4, the simulated data confounded the 85% Var criterion, suggesting that its success in my previous analysis was, as I suggested, a methodological artifact rather than a function of its own merits.

5.3.1. Performance of Parallel Analysis and APA on Simulated Data. Parallel analysis and amended parallel analysis yielded superior results for all of the simulations. It is evident from Table 5.3.1 that the parallel analysis-based methods yielded much more accurate dimensionality estimations than the other eigenvalue analysis techniques for all low-rank simulations except the high-noise iteration, where parallel analysis was only moderately

[1]It is worth noting that I considered several other approaches to noise introduction during the design of these simulations. For instance I considered adding uniformly distributed noise vectors to the vector of true eigenvalues. I also experimented with adding normally distributed matrices of noise to the population covariance matrix. In the case of low-rank data, these alterations yielded no substantive difference in the estimation problem, and thus I chose the simpler model of a constant noise factor. However, I did not test these alternative noise models on full-rank data, an avenue I will explore in upcoming research.

[2]The actual estimates of APA and PA were identical for this figure. I have thus shown PA's estimate skewed slightly to the left to include all five estimation techniques on the plot.

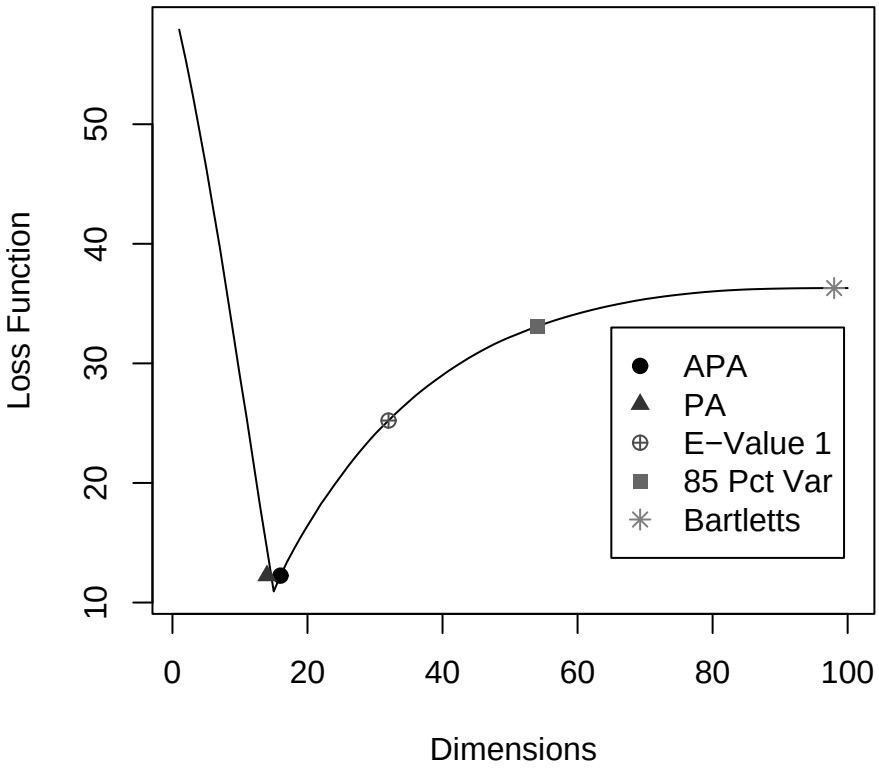

FIGURE 5.3.1. Accuracy of dimensionality estimators ($LRBN$)

superior to other techniques. Likewise, PA's performance on the full-rank data was decisively better than the other methods, except for Bartlett's whose tendency to give nearly full-rank models ceased to be a liability here, but whose performance otherwise suggests stark inadequacy to IR dimensionality estimation problems. Thus PA and APA appear to be by far the best methods of dimensionality estimation for the type of simulated data treated here.

In the case of simulated data, PA and APA offered nearly identical estimates. In fact, both methods yielded identical estimates for all simulations except for the $LRHN$ runs. Testing the null hypothesis of equality of means for each method's accuracy on the $LRHN$

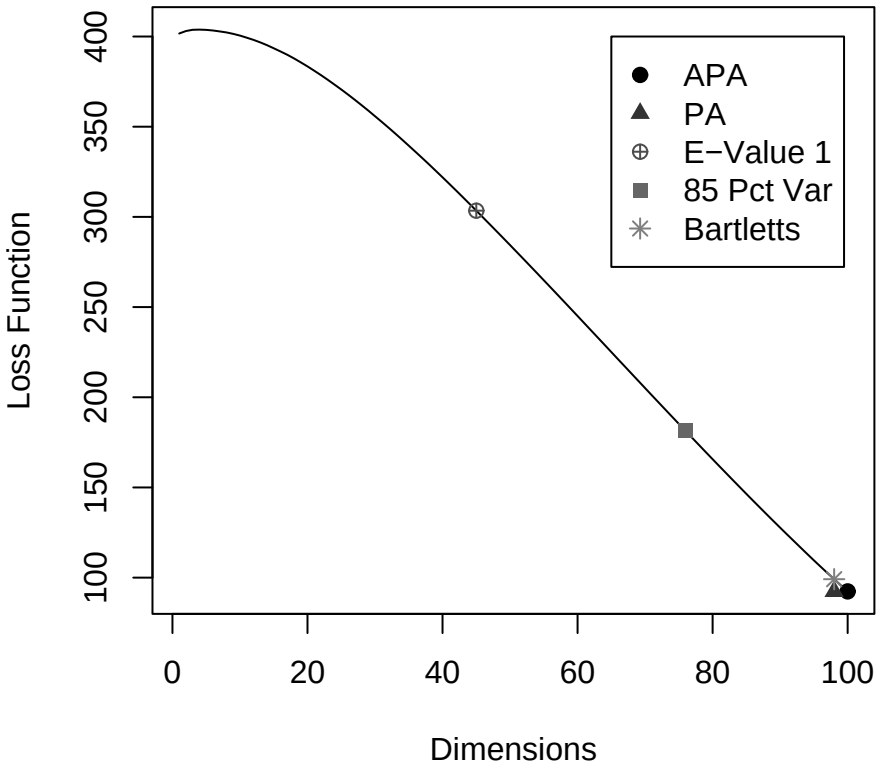

FIGURE 5.3.2. Accuracy of dimensionality estimators (*FRBN*)

runs by means of a standard t-test yielded $p = 0.22$. This test suggests that the difference between APA and PA accuracy on the simulated data was not significant.

To understand why APA and PA gave identical solutions for simulated data, consider the scree plot shown in Figure 5.2.1. Here we see the eigenvalues obtained from an iteration of the *LRBN* simulation. There is a clear gap in eigenvalue magnitude between $k = 15$ and $k = 16$, indicating where the true dimensions yield to noise dimensions. In contrast to Figure 5.2.1's scree plot, consider Figure 5.3.3, which visualizes the operation of APA on the same data. The black line traces the observed eigenvalues, while the light line shows the null eigenvalues obtained after $B = 100$ bootstrap replications. The vertical hash marks are the 95% confidence intervals on the null eigenvalues. By comparing Figures 5.2.1 and 5.3.3

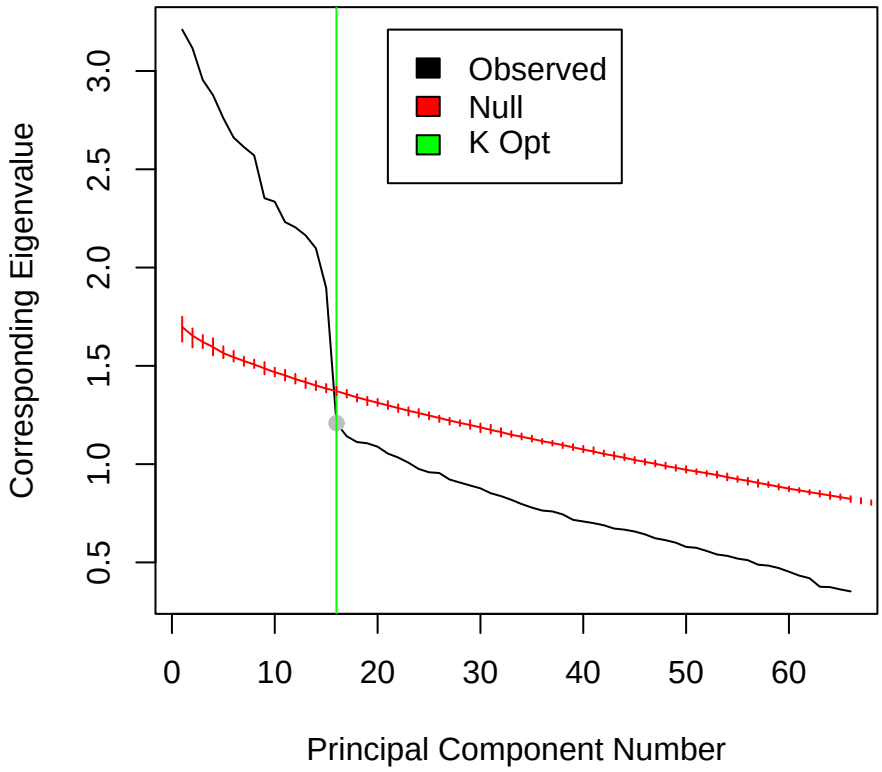

FIGURE 5.3.3. *APA* applied to simulated *LRBN* data

one may note that the gap between the 15^{th} and 16^{th} observed eigenvalues is wider than the corresponding null eigenvalue confidence interval. Because of this gap, then, the relatively subtle amendment entailed by APA does not alter the dimensionality estimate, a fact which is for the best, as APA's tendency to give a larger model than PA would lead to incorrect results here. Thus the fact that APA and PA are qualitatively similar for our simulated data suggests that APA's amendment does not lead to degraded performance in the presence of a problem ideally suited for traditional PA. In other words, when PA was presented with an easy problem (and got the answer right), the APA solution correctly converged on the PA solution.

Because PA and APA are statistically indistinguishable *vis a vis* the simulated data, I shall only consider the performance of PA during the remainder of this discussion. Figure 5.3.1 suggests that PA and APA provided dimensionality estimates that were superior to the four other eigenvalue analysis techniques pursued in this study. This theory was borne out by a series of hypothesis tests[3]. For each eigenvalue analysis technique—EV1, 85% Var, and Bartlett's—I performed four hypothesis tests, one for each simulation round (*LRLN, LRBN, LRHN, FRBN*). During each test, the null hypothesis was that the mean error of PA (i.e. the absolute value of PA's estimate minus the true dimensionality) was greater than or equal to the error of the other estimation technique in question. Rejecting the null hypothesis thus means that PA's mean error was lower than the error rate of the other estimation technique for the given simulation.

Comparing the accuracy of PA to the other eigenvalue analysis techniques demonstrated the superiority of the parallel analysis approach. Among the tested methods PA's accuracy and the accuracy of EV1 on the low-rank, high noise data were the closest. For the null hypothesis $H_0 : \mu_{PA(LRHN)} > \mu_{EV1(LRHN)}$ I obtained $p = 0.04$. All other comparisons across simulation rounds and dimensionality estimation techniques yielded $p \approx 0$. Thus PA's benefit over the other dimensionality estimators was statistically significant at the 95% level for all simulation parameterizations. And for all simulations other than the *LRHN* round, PA*'s* benefit was significant above the 99% level.

Clearly PA provides dimensionality estimates for simulated data that are superior to the other studied estimation techniques. However, the question remains, are PA's estimates statistically distinct from the data's intrinsic dimensionality? Upon inspecting the simulation results I noted that my implementation of the PA method defines k_{opt} to be the eigenvalue immediately after the point where the null line crosses the observed eigenvalues (I defined it in this way in service to my implementation of APA, which judges eigenvalues starting from k_{max} and working left). It would have been more strictly correct to re-state the rule to define k_{opt} as the exact point of this crossing. In the case of the real corpora described

[3]These were standard t-tests. This test was chosen after inspecting the density of the errors obtained from the various dimensionality estimation techniques. While PA showed very low standard deviations for several of the simulations, its error rate on on the *LRHN* simulations, and the error rates of other estimation techniques suggested that the t-distribution was appropriate.

in Chapter 4 the effect of this phenomenon is negligible. However, in the case of our simulations, PA over-estimated the intrinsic dimensionality by 1 in several cases. Under the alternative definition of PA these over-estimations would have been correct. I let these data stand as calculated during my analysis of the simulation results. However, for the sake of argument I subsequently corrected this error. Under the corrected implementation, APA's error was 0 for all 50 iterations of all simulations except *LRHN*. Thus with the exception of the very difficult high noise parameterization, it appears that APA was able to find the correct answer.

PA's performance on the full-rank simulations assuages some of the worry over systematic underestimation described in Chapter 4. In Section 4.2.1, it was noted that PA consistently gave the lowest estimates among the eigenvalue analysis techniques tested in this study. This led to PA giving the worst performance for several corpora. PA's errors raised two worrisome considerations. First, does PA systematically underestimate model dimensionality for IR problems? Second, could PA perform well in the case where no dimensionality reduction is merited?

The results of my simulations suggest that these worries, while still worth pursuing, are less vexing than the analysis in Chapter 4 suggested. As regards the concern over PA*'s* consistently low-dimensional models, these simulations suggest that the technique has no inherent inability to deliver models of full-rank. When the data were 15-dimensional, PA did indeed provide the lowest-rank models, which were also the most accurate models. And when the data were 100-dimensional (i.e. of full-rank), PA gave a 100-dimensional model. It thus appears that PA is well-suited to applications where data are of high- or low-dimensionality. That PA often under-estimated the observed optimal dimensionality of test collections *vis a vis* a given performance metric may still prove an indictment to its applicability to IR problems. However, the results of these simulations also suggest that the Cranfield-style analysis used to judge PA's accuracy in Chapter 4 may have obscured the merits of severe dimensionality truncation for several corpora.

5.3.2. Performance of the Other Dimensionality Estimators on Simulated Data . PA and APA were decisively superior to the other three dimensionality estimation techniques—EV1, 85% Var, and Bartlett's—for almost all simulations. However, the behavior of these other estimators proved interesting in several ways. In this section I discuss these points of interest. First, I describe the performance of the EV1 criterion. Not only did EV1 provide the best estimates (after PA and APA) on the simulated data, it is also theoretically similar to parallel analysis. Thus its behavior will help us understand the behavior of PA, especially in the face of high-noise problems (cf. question number 3 at the outset of this chapter). Second, I turn to a discussion of how the 85% Var rule fared. In Section 4.2 I noted that the percent-of-variance approach gave seemingly good estimates on the six test collections. However, I also questioned these results, suggesting that their accuracy had more to do with shortcomings inherent in using retrospective performance analysis to measure a corpus' intrinsic dimensionality than it had to do with any native superiority of the 85% Var rule.

During the simulations the EV1 criterion—the theoretical cousin of PA and APA—was the third-best performer. In the case of the *LRHN* simulation, its accuracy was close to that of PA. As reported above, the null hypothesis $H_0 : \mu_{PA(LRHN)} > \mu_{EV1(LRHN)}$ gave $p = 0.04$. However, replacing this with a two-sided test (i.e. $H_0 : \mu_{PA(LRHN)} = \mu_{EV1(LRHN)}$) gave $p = 0.09$. Thus EV1 and PA were statistically indistinguishable at the 95% level under the *LRHN* parameterization. Yet the EV1 criterion appeared to be better overall than the 85% Var rule. Applying the Welch two-sample t-test to the estimation errors afforded by each criterion (i.e. $H_0 : \mu_{EV1} > \mu_{85\%Var}$) yielded, $p \approx 0$. Although 85% Var was more accurate than EV1 for the *FRBN* simulation, EV1 was much more accurate than 85% Var on all low-rank data. Thus it seems that APA, PA, and EV1 do share a basic affinity, which is not shared by the other estimation procedures tested here.

The affinity between PA, APA, and EV1 came to the fore as I added noise to low-rank data. Consider Table 5.3.1. During the high-noise simulation, EV1 and *PA* converge on the same answer. However, at lower noise parameterizations (*LRLN,* and *LRBN*) they behave quite differently. The accuracy of EV1 degrades linearly with the introduction of

noise. On the other hand, PA performs with near perfect accuracy until the high-noise simulation. As is evident from Figures 5.2.1 through 5.2.4, PA's resistance to degradation in the face of increased noise has to do with the gap in eigenvalue magnitude under various data parameterizations. That is, so long as there is a significant gap between the true eigenvalues and the noise eigenvalues, PA is quite robust against noise effects. But when the system becomes so noisy that the scree plot basically shows linear descent of eigenvalues (suggesting no obvious elbow in the scree plots), PA shows its relation to the EV1 criterion. Because each simulation has $n = 1000$ with only $p = 100$ variables, the estimate afforded by PA and EV1 under the high-noise condition converge, as is expected in the context of our discussion in Section 3.3. There I stated that as the number of observations grows, the PA solution and EV1 solution will become increasingly similar. These results suggest that this is true, but that PA also maintains a sensitivity to the latent structure of a data set that EV1 lacks. Only in the case of a very difficult estimation problem, where the distribution of eigenvalues is highly unstructured–does PA actually offer the same estimate as EV1.

Another important outcome of the simulations involves the demonstrated inaccuracy of the 85% Var criterion. Whereas retaining 85% of the total variance yielded a surprisingly accurate model selection rule for the test collections discussed in Section 4.2, such was not the case for simulated data. As seen in Table 5.3.1, only Bartlett's performed worse than the percent-of-variance approach, and Bartlett's virtues were strictly a matter of its retention of near-full-rank models across the board. Thus 85% Var appears to have benefited from good luck in the empirical results of Section 4.2.2.3. It's apparent accuracy in the face of real-world corpora, I contended in Section 4.3, was an artifact of the necessarily blunt instrument (i.e. retrospective performance analysis) used to gauge corpus dimensionality. These results bear out my contention. A percent-of-variance approach to dimensionality estimation is necessarily *ad hoc* and inflexible. Thus, in the low-rank simulations, retaining 85% of the total variance simply overfitted the model, while in the 100-dimensional simulations, 85% Var underestimated the intrinsic dimensionality. The failure of 85% Var to predict optimal model dimensionality for simulated data points to its deficiencies in the real-world, too. As seen in Table 4.1.1, real corpora appear to demand a wide variety of dimensionalities (and

a wide range of percentage ratios) to attain their optimal model. Lacking an effective apparatus to read these demands, a percent-of-variance approach to dimensionality estimation is necessarily *ad hoc*.

5.4. Implications of the Results for Simulated Data

Applying eigenvalue analysis techniques to simulated data shows the appeal of parallel analysis and amended parallel analysis. Over the course of 300 simulations, these techniques demonstrated a decisive superiority to the other proposed dimensionality estimators—EV1, 85% Var, and Bartlett's. I found that PA and APA were significantly more accurate (above the 99% confidence level) than the other approaches to dimensionality estimation. Not only did PA and APA outperform the other tested estimators, they appeared to discern the data's true dimensionality; for all simulations except the *LRHN* iterations, PA estimated the intrinsic dimensionality perfectly.

I also found that APA's alterations to PA do not lead to significant degradation in prediction accuracy when PA finds the correct answer. In Section 4.2.2.1 I noted that APA constituted a significant improvement over traditional PA. Here I complement that assertion. Parallel analysis is a robust, flexible approach to dimensionality estimation. In the presence of highly complex data sets (such as the corpora discussed in Section 4.1) the confidence-interval-based APA provides a more accurate estimate of optimal dimensionality. But in the presence of a simpler problem (such as the simulations discussed in this chapter), PA's point-estimate-based approach estimated the intrinsic dimensionality accurately. In this case APA's amendment became negligible; for simulated data, the estimates afforded by PA and APA were statistically identical.

At the outset of this chapter I posed three questions that data simulations would address. The first question addressed the question of dimensionality estimation in the face of varying noise coefficients. I found that APA and PA performed very well for the low-rank data with a variety of noise parameterizations. These techniques offered near-perfect accuracy for the *LRLN* and *LRBN* simulations. In contrast the other dimensionality estimation techniques—EV1 and 85% Var—yielded linearly decreasing quality in the face of increased

noise. I also noted that the family of statistically related estimation procedures comprised of APA, PA, and EV1, provided mutually-indistinguishable results for the *LRHN* simulation. These results were statistically superior to those obtained via 85% Var and Bartlett's. Thus APA and PA appear to converge on the answer provided by EV1 in the presence of decreasing correlational structure in the data. Taken as a whole, these data suggest that the family of estimators comprised by APA, PA, and EV1 use the distribution of eigenvalues to estimate the number of linearly independent variables in the population covariance matrix that generated a given data set. Moreover, the data suggest that this approach—inferring the number of independent variables based on eigenvalue distribution—is robust against the introduction of noise, and that it gives dimensionality estimates with consistently low error rates.

After the empirical analysis of Section 4.2.2 I was concerned about the behavior of eigenvalue analysis techniques when applied to data that merit no dimensionality reduction. However, these simulations suggest that APA and PA can in fact excel at identifying such situations. The full-rank simulated data sets yielded zero error rates from APA and PA. On the other hand EV1, PCTVAR, and Bartlett's evidenced much less flexibility. EV1 and PCTVAR provided dimensionality estimates near the middle of the range of possible dimensionalities. While this proved useful in some of the real-world corpora discussed earlier, it became a liability for the simulated data. Both of these techniques consistently overestimated model dimensionality for the low-rank simulations, and underestimated the number of factors in the full-rank cases. On the other hand Bartlett's test of isotrapy was essentially a non-performer, never advocating anything but a 2-dimensional reduction in k over k_{max}, as in the empirical data analysis of Chapter 4.

On simulated data APA and PA were statistically identical with regard to estimation accuracy. They were also tremendously self-consistent. Only in the case of the *LRHN* simulation did their estimates have a standard deviation above zero across the 50 runs. APA and PA showed categorical improvement over all other estimators. Only for the *LRHN* data was the superiority of PA over another method (EV1) significant below the 99% level. In this case all three methods all gave a statistically indistinguishable best estimate on all

50 simulations. Moreover, EV1 was significantly superior to 85% Var in all instances except the full-rank case, where 85% Var gave a superior showing.

Overall, then, only the parallel analysis-based methods showed both accuracy and flexibility across a range of simulation parameters. APA and PA gave accurate estimates of intrinsic dimensionality for both low-rank and full-rank data, showing that it is applicable to problems where dimensionality reduction is merited, or where dimensionality reduction should be avoided. I have also shown that APA and PA provide the best estimate of model dimensionality (among the techniques pursued in this study) across a range of structural configurations. Adding noise to a simulation made the problem more difficult for all of the dimensionality estimators (except for Bartlett's, which was almost always a poor performer). But APA and PA addressed this challenge by giving consistently good estimates (at *LRLN* and *LRBN*) before finally converging on the EV1 estimate at *LRHN*. Admittedly, the answer that EV1, APA, and PA converge on in the high-noise case is far afield of the true dimensionality, leaving ample room for future improvements to the technique. However, the collective behavior of these methods supports my original contention that detecting the number of independent variables in the population covariance matrix by eigenvalue analysis provides a strong approach to dimensionality estimation for IR applications.

CHAPTER 6

Concluding Remarks

This chapter revisits my research question, articulated in Chapter 1. In light of the research described in the previous three chapters, I argue that eigenvalue-based analysis offers a useful, but not categorically accurate means of gauging the optimal dimensionality of an LSI system. In addition to assessing the general utility of eigenvalues in service to dimensionality estimation, my initial research question sought to identify which methods of analysis provide the best estimates of data's intrinsic dimensionality. The experiments reported in this dissertation suggest that the family of dimensionality estimators that operate on an error-correction premise—APA, PA, and EV1—were especially effective in problems involving real-world corpora and simulated data. In particular, amended parallel analysis has proven that it merits future research, as its performance was both compelling and revealing during my experimentation. The performance of APA and related estimation methods supports my theoretical argument from Chapter 1: dimensionality reduction for IR is merited to the extent that the indexing features depart from statistical independence.

LSI entails an important and effective elaboration of Salton's vector space model of information retrieval. As discussed in Chapters 1 and 2, dimensionality reduction extends the standard VSM to account for the correlational structure among terms. Despite its intuitive and practical appeal, however, LSI's dimensionality reduction has remained poorly theorized. In particular, k_{opt}, the optimal dimensionality of an LSI system has traditionally been the domain of *ad hoc* approaches and un-analyzed assumptions. The research presented here has attempted to speak to these assumptions. In this chapter I review the outcome of my research, contextualizing my findings and detailing their theoretical and practical significance.

The chapter begins with a re-statement of my research question, and a review of its grounding in the theory of IR models. In Section 6.2 I revisit the empirical findings of my experimentation, summarizing the strengths and weaknesses of each dimensionality estimation technique tested, and offering suggestions about research methodology. Section 6.3 contextualizes these findings, pursuing their implications for IR theory and practice. After this, Section 6.4 describes important shortcomings in this study, suggesting room for future work on dimensionality estimation for IR. Finally, I conclude with several reflections on the significance of the research reported here.

6.1. Dimensionality Estimation and the Vector Space Model

As described in Sections 1.3 and 2.1, Salton's vector space model theorizes information retrieval as a geometrical problem. Under the VSM, inter-object similarity is a function of vector orientation, measured with respect to a given set of dimensions. Each document in a traditional VSM-based IR system is represented as a vector in the vector space spanned by the corpus' indexing terms. Documents with similar distributions of terms thus lie near each other in the information space, and query-document matching simply involves ranking each document by its proximity to a given query vector.

Despite its intuitive appeal, however, the traditional VSM suffers from serious theoretical shortcomings. Most notably, Salton's approach assumes orthogonality of the indexing terms, despite ample evidence to the contrary. Insofar as it defines similarity as a linear function on the dimensions of its vector space, the VSM treats each term as a statistically independent variable. This introduces error into the VSM similarity function that manifests most notably as the *synonymy* problem; queries about *cars* fail to retrieve documents about *automobiles* despite an intuitive correlation between these terms.

Wong's generalized vector space model addresses the VSM's assumption of term independence. Under Wong's approach, the term-term correlation matrix supplements the traditional VSM similarity function. Thus a high observed correlation between *cars* and *automobiles* provides evidence for the GVSM that documents with either of these terms may be describing a single concept. However, Wong's model is concerned only with the sample

correlation matrix. The GVSM similarity function is given by Equation 6.1.1:

$$\mathbf{s} = \mathbf{qRA}' \tag{6.1.1}$$

where $\mathbf{A}$ is the $n \times p$ document-term matrix of rank r, $\mathbf{R}$ is the $p \times p$ term-term correlation matrix computed from $\mathbf{A}$, $\mathbf{s}$ is the $1 \times n$ vector of similarity scores, and $\mathbf{q}$ is the $1 \times p$ query vector. For the GVSM, then, the observed correlation matrix describes the model of relationships among the corpus terms.

If the GVSM extends Salton's model, LSI entails still further extension. Under LSI, we have a similarity defined by Equation 6.1.2:

$$\mathbf{s} = \mathbf{qR}_k\mathbf{A}' \tag{6.1.2}$$

where $\mathbf{R}_k$ is the best rank-k approximation of the observed correlation matrix. Under LSI we approximate the correlation matrix by using the first k eigenvalues and eigenvectors of $\mathbf{R}$. If $\mathbf{D}$ contains the eigenvectors of the correlation matrix $\mathbf{R}$ on the columns, and $\mathbf{\Sigma}$ has the eigenvalues on the main diagonal, then $\mathbf{R}_k = \mathbf{D}_k\mathbf{\Sigma}_k\mathbf{D}_k'$, where $\mathbf{D}_k$ is the first k columns of $\mathbf{D}$. Proponents of LSI argue that removing the last $r - k$ eigenvectors from $\mathbf{R}$ improves the system's similarity function by removing overspecification error from the model. The matrix $\mathbf{R}_k$, I have argued, constitutes a better statistical model of the population correlation matrix $\mathbf{P}$ than does the full-rank matrix $\mathbf{R}$.

While LSI's dimensionality reduction has shown good performance in empirical studies, its motivation has remained largely un-formalized in the research literature. Why should a reduced-rank approximation provide a superior estimate of the population correlation matrix? How aggressively should we reduce the dimensionality to derive the optimal model? The notion of k_{opt}, the best number of dimensions for a given corpus, is thus key to the theoretical tenability of LSI. Without an overt notion of model goodness of fit, optimality has been difficult to define. Traditional approaches to balancing the bias-variance trade-off in statistical models do not translate easily to the unsupervised learning environment presented by information retrieval. Thus LSI has often been guided by *ad hoc* approaches

to dimensionality estimation, approaches that are of questionable practical utility and that carry almost no theoretical weight.

This study has pursued the notion that a statistical analysis of the eigenvalues that arise during LSI can provide a definition for optimal dimensionality. Each of the five estimation techniques studied here—APA, PA, EV1, 85% Var, and Bartlett's—defines its own notion of optimality. Concomitantly, each estimator implies a different theory for dimensionality reduction. Three of the estimation methods—APA, PA, and EV1—constitute a family insofar as they rest on similar assumptions. Each of these three criteria argues that dimensions should be rejected if their eigenvalues are smaller than we would expect to see if the data were independent. On the other hand, the percent-of-variance approach assumes that a fairly constant noise factor has interfered with the data. Retaining enough eigenvalues to account for, say, 85% of the total variance is the intellectual kin to the common notion that LSI entails a noise reduction procedure. Finally, Bartlett's test of isotrapy suggests that we should retain a given dimension if its corresponding eigenvalue λ_k is significantly greater than λ_{k+1}. This is a very conservative approach to dimensionality estimation, and its failure in the face of IR data suggests that it does not address the dynamics of LSI.

Dimensionality estimation is crucial to the viability of LSI in two senses. First, empirical studies (cf. Section 2.3) have shown that finding k_{opt} has strong ramifications for retrieval performance under LSI. Models of insufficient dimensionality are impoverished, lacking the expressive power to discriminate between relevant and non-relevant documents. On the other hand, including too many dimensions has been shown to incur an overfitting effect, leading to familiar problems in handling *synonymy* and *polysemy*. In addition to these practical considerations, dimensionality estimation is crucial to the theory that underpins LSI. As I have argued here, each dimensionality estimation technique implies a notion of model goodness-of-fit. Thus each model selection criterion implies a notion of optimality. For the theory of LSI, choosing a model selection criterion, then, is as important as choosing a model.

6.2. Eigenvalue Analysis for Dimensionality Estimation in IR

This section attempts to synthesize the major results from my experiments, taking pains to detail the dynamics of k_{opt} and the strengths and weaknesses of each tested dimensionality estimation technique. The data analysis of Chapters 4 and 5 paints a complex picture of the competing imperatives that inform model selection under LSI. It was no surprise that during my research, LSI dimensionality proved itself to be an important parameter. Any benefits afforded by dimensionality reduction evidenced themselves only in a narrow range of values for k. Outside of this range, performance was consistently lower than performance seen under the full-rank model. However, optimal dimensionality also appeared to be corpus-dependent. Thus my experimentation strongly evidenced the need for robust dimensionality estimation techniques.

6.2.1. Findings from Empirical Data. As described in Section 4.1 the six tested corpora showed widely different behavior with respect to dimensionality reduction. For instance, *MEDLINE* demonstrated strong evidence of a semantic subspace of approximately 100 dimensions according to all three IR performance metrics. On the other hand, *CF* appeared to tolerate almost no dimensionality reduction. None of the Cranfield-style performance metrics saw a significant advantage via dimensionality reduction on *CF*. For other corpora the three performance metrics were less unanimous. For instance, ASL demonstrated a pronounced semantic subspace of about 200 dimensions for the *CRAN* data, *CF_FULL,* and the *CACM* data. But dimensionality reduction was less helpful in improving average precision or optimal F scores for these corpora. Thus I questioned the ability of Cranfield-style analysis to indicate these corpora's intrinsic dimensionality. Inadequacies of the supplied queries, I argued, may have frustrated attempts to gauge the intrinsic dimensionality of these corpora.

By the account of all performance metrics, however, optimal dimensionality appears to be highly corpus-specific. As shown in Table 4.1.1, the amount of variance accounted for by ASL-optimized models ran the gamut from 16% (*MEDLINE*) to 95% (*CF*). For other metrics, the spread was even wider. Thus the number of eigenvectors and the proportion of variance described by an optimal model resisted any one-size-fits-all summarization. This

suggests that rigorously motivated and highly sensitive dimensionality estimation techniques are crucial for the successful application of LSI.

Given the corpus-specificity of k_{opt}, it is somewhat surprising that the 85% Var criterion performed well in our experiments. On five of eighteen observations (three metrics applied to six corpora), retaining enough eigenvalues to account for 85% of the total variance yielded the best results. However, I argue that this success is misleading. As mentioned above, several corpora appeared to be optimized at fairly high values of k. On these occasions, 85% Var's tendency to produce large models paid off. However, in Section 4.2.2.3 I showed that the situations in which 85% Var excelled were among those for which the IR performance metrics were least in agreement. Moreover, in cases where low-dimensional models were called for, 85% Var showed its native inflexibility, overestimating the dimensionality. These detriments were writ large when I applied the 85% Var criterion to simulated data, where it consistently over-estimated the intrinsic dimensionality. Thus I argue that a percent-of-variance approach to dimensionality estimation is ill-advised for LSI insofar as evidence in its favor is weak in our data and insofar as it lacks sensitivity to a given corpus' distribution of terms across documents.

Bartlett's test of isotrapy demonstrated a categorically poor fit to the LSI problem. I included it in these experiments out of a desire to consider an approach to dimensionality estimation based on traditional, parametric hypothesis testing. However, Bartlett's is known to over-estimate model dimensionality, and this tendency is magnified in the case of IR, which takes Bartlett's far afield from the conditions for which it was developed. Under Bartlett's we retain an eigenvalue λ_k if it is significantly greater than the next eigenvalue λ_{k+1}. In my experiments, however, Bartlett's always rejected two eigenvalues and accepted the remaining $r-2$. In those cases where dimensionality reduction did not appear to be merited (e.g. *CF*), Bartlett's approach appeared accurate *de facto*. However, this modest success came at the price of intolerable rigidity. Bartlett's demonstrated itself to be even less flexible than the 85% Var criterion. Its performance on both the real data and our simulations demonstrated that the χ^2 distribution assumed by Bartlett's does not hold when its test statistic is applied to the complex data sets native to IR.

Whereas evidence for the utility of Bartlett's and 85% Var was weak, the family of dimensionality estimators based on error correction evinced good performance in these experiments. These estimation techniques—PA, APA, and EV1—rest on the assumption that the distribution of eigenvalues derives from the degree of inter-term correlation. Under these methods, the deviation of observed eigenvalues from the eigenvalues expected under term independence is used as evidence for the corresponding dimension's significance. Members of this family provided the best dimensionality estimates on nine of the eighteen observations of IR performance on real corpora. APA performed best on four occasions, including *CRAN* and *MEDLINE*, where evidence of a semantic subspace was especially strong. PA gave the best estimate for the *CACM* data with respect to average precision. With its identification of moderately complex models, EV1 gave the best estimate on four occasions.

A serious failure, PA also gave the worst estimate on nine observations. However, I argue that it must be understood in the context of the data analysis undertaken in Section 4.1. Retrospective performance metrics disagreed on the optimal dimensionality of several corpora, especially *CACM, CISI,* and *CF_FULL*. In these cases, ASL called for fairly low-dimensional models, while average precision and optimal F needed more factors for optimal performance. Because PA consistently returned low-dimensional models, it appeared as a worst-performer *vis a vis* precision and F. On the *CACM* data, for instance, PA was the worst performer for the F measure, but the best performer for ASL. In Section 4.2.2.1 I suggest that PA's frequently poor performance is likely to be an artifact of the Cranfield analysis used to gauge intrinsic dimensionality.

Despite some equivocation, it must be admitted that parallel analysis was a compelling worst-performer on the *CF* data. In this case, PA under-estimated the intrinsic dimensionality according to all three performance metrics. The *CF* database was best represented with no dimensionality reduction, a fact that PA seemed ill-equipped to recognize. However, it is important to note that APA's moderating effect on PA improved dimensionality estimations for all corpora on all performance metrics except for *CACM* measured by ASL. In Section 4.2.2.1 I note that APA's improvement over traditional parallel analysis was significant above the 95% level. Thus APA's confidence interval-based approach to estimation

entails a statistically significant improvement over Horn's method, which relies only on point estimates.

6.2.2. Findings from Simulated Data. The empirical analysis of Chapter 4 left several questions unanswered. Among these were, do the tested estimators display systematic error? How do our dimensionality estimation techniques fare when applied to full-rank data where dimensionality reduction is not merited? If the EV1 criterion is the theoretical cousin of parallel analysis, why did each method give such different results in our experiments? To address these questions I conducted the simulations described in Chapter 5. This section reviews the major outcomes of the data simulations.

Most significantly, the simulations suggested that parallel analysis does not systematically under-estimate the intrinsic dimensionality of data. Across the simulation parameters described in Table 5.1.1 PA and APA were consistently the most accurate performers of all eigenvalue analysis techniques tested here. With confidence above 99% their estimates were superior to all other estimator predictions (except for EV1 under the high-noise parameterization, where confidence was above 95%). Given low-rank data with various noise factors, PA always performed best. Likewise, PA and APA were the only techniques that recognized the full-rank data as such.

APA's moderating effect on PA does not introduce systematic error into the estimation process. In my simulations, traditional parallel analysis consistently found the right answer. Given this accuracy, amended parallel analysis gave the same solution. I find it especially encouraging, then, that APA yielded a significant improvement over PA in the face of complex, real-world data while converging on the PA solution when confronted with simpler, more structured simulations.

Systematic error was, however, evident in the other surveyed eigenvalue analysis techniques. EV1, 85% Var, and Bartlett's all overestimated the dimensionality of the low-rank simulations, while under-estimating the dimensions for the full-rank data. EV1 and 85% Var consistently delivered models of middling complexity. Regardless of the amount of noise in the system, or the number of non-zero eigenvalues, these estimators predicted that the

optimal dimensionality was near the middle of the possible range. This behavior was especially egregious for 85% Var and Bartlett's, each of which evidenced the inflexibility that led us earlier to caution against their use for IR.

The relationship between PA, APA, and EV1 came to the fore when I admitted large amounts of noise into the simulated data sets. As the noise coefficient was increased, parallel analysis converged on the EV1 solution. These techniques—PA, APA, and EV1—reject eigenvalues that are smaller than those expected under the condition of term independence. They differ with respect to how they model the so-called null eigenvalues. EV1 treats the sample data as if it were a population. PA is more realistic, admitting into the analysis that fact that there are only $n < \infty$ observations. APA admits a still more realistic model of the null eigenvalues insofar as it accounts for their sampling distribution. The convergence of these methods on a single solution given noisy data provides more evidence in favor of the parallel analysis approach to dimensionality estimation. That is, given a well-structured data set whose intrinsic dimensionality is easy to ascertain, PA and APA effectively exploit that structure to derive an accurate prediction. However, if the data are relatively unstructured, the methods appear to concede as much, turning to the more conservative EV1 solution.

6.3. Implications of The Findings

Overall, eigenvalue analysis gave useful information about the intrinsic dimensionality of the datasets tested here. However, no viable alternative to dimensionality estimation via eigenvalue analysis exists, and so demonstrating the categorical superiority of eigenvalue-based evidence is impossible. It is possible—and correct—however, to note that the retrospective approach to judging intrinsic dimensionality appears flawed. In my experiments, ASL, average precision and optimal F frequently disagreed about the optimal dimensionality of an LSI model, often by a wide margin. This suggests that approaches to dimensionality estimation based on retrospective, *ad hoc* judgment of "what works best" carry high risks. What appears to work best under one lens of performance analysis may be grievously suboptimal in another context. More damning still, exhaustive relevance judgements are rarely

present in IR applications outside the laboratory. Without relevance judgements, even the dubious prospect of selecting k by consulting Cranfield-style performance metrics is not feasible.

It is no accident that eigenvalues yield good evidence for dimensionality estimation. The term and document co-occurrence matrix eigenvalues provide a basis for Ding's probabilistic model of LSI. Under Ding's theory, the magnitude of an eigenvalue λ_k is directly proportional to the increase in model likelihood gained by adding the k^{th} LSI dimension. We may also understand the role that eigenvalues play in LSI in terms of principal component analysis. Assuming that the columns of the term-document matrix $\mathbf{A}$ have been centered and scaled to unit length, then LSI gives the principal components of the terms and documents. Likewise, in this case, the singular values are the positive square roots of each principal component's variance. Thus the k^{th} dimension describes λ_k units of variance, where λ_k is the k^{th} eigenvalue. Given the intimate relationship between LSI and the eigenvalue-eigenvector decomposition, it is natural to use eigenvalue magnitude as evidence of dimensional validity.

In particular, I argue that the optimal dimensionality of an LSI system is given by the value of k at which point the observed eigenvalues become smaller than the eigenvalues expected under term independence. Different analysis techniques—PA, APA, and EV1—disagree on how to estimate the null case, but whatever their statistical differences their motivation is the same. The value of k whose corresponding point on a scree plot is where the observed eigenvalues cross the so-called "null eigenvalues" is a measure of the strength of inter-term correlation in a corpus. This suggests that LSI's dimensionality reduction is in essence an error correction mechanism. The vector space model assumes term orthogonality, an oversimplification addressed by Wong's GVSM. LSI carries the GVSM error correction one step further. Whereas Wong's approach uses $\mathbf{R}$, the sample term correlation matrix, to supplement the VSM similarity model, LSI uses the best rank-k approximation of $\mathbf{R}$. The rationale behind LSI is that our interest lies in the *population* correlation matrix, not the *sample* correlation matrix. Due to inter-term correlation, the number of non-zero eigenvalues in the population is $k_{opt} < p$, as showed by Lederman [**93**]. In other words, by retaining k dimensions, we assume that the population correlation matrix contains k

non-zero eigenvalues, and that the remaining $p - k$ sample eigenvalues derive from sampling error. Rejecting the smallest $p - k$ eigenvalues thus removes sampling error from the GVSM similarity function. The degree of this error correction—i.e. the amount of dimensionality reduction—is proportional to the degree to which the data depart from independence.

I promote this error correction-based theory in light of my experimental results. The data analyses reported here suggest that the 85% Var and Bartlett's approach to dimensionality estimation are ill-suited to IR applications. Though 85% Var did perform well during the empirical study, I found evidence that its virtues were inflated by artifacts of the experimental methodology. Instead, 85% Var appears to be an inflexible and poorly motivated approach to dimensionality estimation. Likewise, Bartlett's—a statistically based technique—fared poorly, implying that the χ^2 distribution of its test statistic does not hold on large data sets.

On the other hand, the error correction-based methods—APA, PA, and EV1—performed well in general. Between the empirical analysis of Chapter 4 and the simulations of Chapter 5 I found that parallel analysis and amended parallel analysis comprise a compelling model for dimensionality estimation under LSI. The parallel analysis approach has, in previous studies, proved its utility for traditional multivariate statistical applications. In the context of IR, however, it was unclear at the outset how PA would fare. While PA did under-estimate the intrinsic dimensionality for some real-world data, the extension entailed by APA mitigated this tendency to a significant extent. Thus I have found compelling evidence to suggest that APA is well suited to the task of parameterizing k during LSI.

In sum, of the three error correction-based techniques, PA offers the most aggressive approach to dimensionality reduction. Though my simulations evinced no systematic tendency toward under-estimation, PA did under-estimate the best dimensionality for several test corpora. APA's amendment to PA mitigated this aggressiveness to a limited, but statistically significant degree. Finally, EV1 constitutes the most conservative approach to error correction-based estimation. In my simulations, EV1 evinced a disappointing lack of flexibility, always returning models of middling dimensionality. However, its inherent bias toward

moderately sized models did keep its error rate relatively low for the real-world data. Given these results, researchers and practitioners in IR may consider the following suggestions.

Using the eigenvalue-one criterion appears to be a safe and effective approach to dimensionality estimation. In my analysis retaining eigenvalues larger than the average eigenvalue often provided a good estimate of the optimal dimensionality, although the technique only occasionally provided the best estimate. We may thus consider EV1 to provide a rough estimate of the solution given by APA and PA. It's merits lie in its ease of calculation and in its inherent conservatism. In situations where under-estimating model dimensionality would lead to egregious error, EV1's estimate may in fact be a safe choice. However, in such cases, LSI itself is probably ill-advised.

That said, APA and PA appear to be much more sensitive than EV1 to the correlational structure of a given corpus. For the simulated data, APA and PA were more accurate than EV1. And in my empirical studies they were more accurate than EV1 for the corpora that provided the strongest evidence of a semantic subspace. Thus I argue that APA and PA provide the best estimate of a corpus' intrinsic dimensionality among the estimators tested here. Moreover, my analyses suggest that APA's improvement over traditional PA is statistically significant. Since APA requires very little computation in addition to PA, I argue that it is important in IR applications to take advantage of APA's confidence interval-based approach to dimensionality estimation.

The upshot of this dissertation's findings, however, is that the matter of dimensionality estimation is still unsolved. I have produced compelling evidence that the APA approach is well suited for optimizing LSI models. However, the difficulty in ascertaining intrinsic dimensionality via retrospective IR performance evaluation prohibits categorical statements of estimator superiority or inferiority. From a practical standpoint, researchers would be well advised to apply all of the estimation techniques tested here, comparing their individual solutions and weighing the relationships among them.

6.4. Study Limitations and Future Work

No research study is without its limitations, and this dissertation is no exception. In this section I discuss these limitations in detail, offering suggestions for eliminating their influence in future work. Although I have answered my research question, I have not done so unequivocally. Three major issues have left open questions that future research will need to address:

(1) The size and number of test corpora

(2) The number of tested dimensionality estimators and the number of dimensionality reduction techniques

(3) The method of intuiting the intrinsic dimensionality of each corpus

In fact these three shortcomings are not independent of each other. My results suggest strongly that APA and related estimation techniques provided the best dimensionality estimates among the tested eigenvalue analysis methods. However, my assertions in this regard must be qualified in a number of ways, each of which pertain to one (or more) of the issues enumerated above.

Perhaps most importantly, when comparing dimensionality estimators we desire a means of assessing the "true" dimensionality of a given test corpus. In this study I approximated such an assessment by recourse to three IR performance metrics: ASL, average precision, and optimal F. While this approach yielded useful evidence about each corpus' intrinsic dimensionality, in several cases these performance metrics were in disagreement. As described in Section 4.1, I suspect that the Cranfield-style analysis introduced unwanted artifacts into the evaluation process. For instance, I speculated that the supplied queries for several corpora (e.g. *CACM* and *CISI*) may have been inadequate for the task at hand. Due to the evident noisiness of retrospective performance evaluation, this research would benefit from an alternative means of assessing intrinsic dimensionality. An unbiased, highly accurate knowledge of the intrinsic dimensionality would provide a much firmer basis for comparing eigenvalue analysis techniques.

The simulations undertaken in Chapter 5 addressed this problem by constructing data of known dimensionality, an approach that yielded highly informative results. However,

simulations lack the richness of empirical data, and thus I would still welcome an alternate methodology for analyzing real-world corpora. Of course our interest in eigenvalue-based dimensionality estimators stems from the hypothesis that they provide the best estimate of a corpus' intrinsic dimensionality. I know of no other technique for assessing k_{opt}, and thus have had to tolerate the somewhat messy data analysis that attends all Cranfield-style IR evaluation.

To address this issue, in future work I plan to translate the dimensionality estimation problem into the domain of supervised learning. In [**136**] Schütze, Hull, and Pedersen use the first k principal components as inputs to automatic classification systems. Following this approach, one could estimate the intrinsic dimensionality as that value of k that leads to the best classifier. The literature of supervised learning has much more thoroughly developed notions of model optimality than are common in unsupervised learning. In a supervised learning environment, one could employ, for instance, information-theoretic measures such as the Akaike Information Criterion (AIC) to select the optimal model. APA and associated analysis techniques could then be compared against the model dimensionality selected in this fashion. The supervised learning approach will give us a stronger grounding in statistical model building. But it is important to stress that supervised learning is significantly different from IR, and thus results obtained under a supervised model will not be unambiguously interpretable in the context of retrieval.

A related problem in this dissertation's research was the sample of test corpora that was analyzed. I selected the six corpora treated here due to their distinct statistical qualities, as detailed in Tables 3.1.1 and 3.1.2, and due to their use in previous studies. Although these corpora did evince a wide variety of observed optimal dimensionalities, the relationship between their statistical characteristics and their optimal dimensionalities was not obvious. Although I found evidence that corpora with large term spaces tend to benefit from LSI more than other data sets, other rules of thumb are elusive. In future work it will be desirable to undertake experiments like those described in this dissertation on more IR test collections. Having $n = 6$ in this study made statistical inference about the performance of eigenvalue analysis techniques a tenuous proposition. But inferential statistics formed only a part of

my analysis. As described in Chapter 4, much of my analysis involved nuanced interrogation of the results on a case-by-case basis. Thus I selected $n = 6$ because it provided an ample base of comparison, without generating unmanageable quantities of data. However, in future work it will be desirable to undertake less detailed but more statistically inclined analyses on larger samples of corpora in order to improve the generalizability of my statements.

Likewise, it will be essential to repeat the experiments undertaken here on much larger corpora. In contemporary IR, the TREC data comprise a gold standard of experimentation. At the time that this study was undertaken, resampling techniques such as APA are computationally unmanageable on data sets requiring multiple gigabytes of storage. However, the inexorable development of processing power and computer memory guarantees that this infeasibility will fall to the wayside. I thus anticipate analyzing the matter of optimal dimensionality for LSI for much larger corpora in future research.

Finally, this study has treated only five dimensionality estimation techniques. Moreover, I have applied them only to models based on LSI. I chose these estimators because they are representative of several families of eigenvalue analysis techniques. APA, PA, and EV1 are all based on the notion that dimensionality reduction is merited to the extent that the data violate the assumption of term independence. On the other hand, 85% Var is a relatively *ad hoc* approach. But its underlying rationale is that a Gaussian noise factor imposes itself on observed data. Finally, I included Bartlett's as the most widely studied example of a statistically-based, parametric dimensionality estimator. However, other means of analyzing eigenvalues to assess intrinsic dimensionality do exist, and in future research I anticipate testing APA against their solutions.

I chose to study LSI due to its wide deployment in the IR literature. Though many other methods of dimensionality reduction have been proposed (cf. Section 2.2), LSI retains an important visibility in the IR community. In particular, I have worked with LSI due to its status as an extension of Salton's VSM and Wong's GVSM. Nonetheless, it will be of great interest to note how matters of dimensionality estimation may be resolved under other methods of dimensionality reduction such as Independent Component Analysis and Probabilistic LSI.

In addition to the directions outlined so far in this discussion, my future work will be concerned with improving the APA technique. Experimental results reported here suggest that APA's moderating effect over PA is significant and useful. However I suspect that APA can be made more accurate in the future. To improve the APA model, I anticipate two lines of research. First, APA's representation of null eigenvalues was derived in this study by taking the mean of B eigenvalue simulations. However, since these null eigenvalues do not show a distribution marked by strong central tendency, the motivation for using the 50^{th} quantile was weak. Thus I plan to analyze the probability density of the null eigenvalues more thoroughly in efforts to select a more overtly motivated (and probably lower) quantile with which to estimate the "true" null eigenvalues. Secondly, the analysis of Section 4.2.2.1 suggested that we may derive useful information about the validity of a dimension based on the variability of the simulated null eigenvalues. Future articulations of the APA procedure may account for null eigenvalue variance. Both of these proposed lines of research stand to improve the accuracy of APA by improving its sensitivity to the null eigenvalue distribution.

6.5. Conclusion

This study has engaged the problem of dimensionality estimation experimentally. At the outset I asked whether co-occurrence matrix eigenvalues are useful for parameterizing k, the number of dimensions in an LSI system. I approached this question by observing the quality of estimates derived by five eigenvalue analysis techniques for six IR test collections and 200 simulated data sets. My results argue that analyzing co-occurrence matrix eigenvalues can lead to very good or very poor estimates of the intrinsic dimensionality. Thus a method such as Bartlett's test of isotrapy appears ill-suited to deployment in large-scale IR problems. Likewise, the percent-of-variance approach occasionally works well, but its successes in my experiments appeared to be an artifact of noise in the methodology rather than a valid perception of the optimal semantic subspace. On the other hand, APA, PA, and EV1 fared very well, overall. Though it is difficult to quantify their accuracy on the real-world corpora, the performance of these methods gave the best estimate on nine of the

eighteen corpus/performance metric pairings. Moreover, they excelled dramatically on simulated data, delivering a statistically correct answer on all but the most difficult estimation problems. The performance of APA was especially encouraging. Its improvement over PA was statistically significant for the empirical data, offering the best estimate on four observations. APA converged on the PA solution for all simulated data, where PA was often able to discover the right answer. Thus APA demonstrated impressive flexibility and accuracy during both the empirical and simulated data analyses.

The success of APA, PA, and EV1 supports my argument that LSI's dimensionality reduction functions as a form of error correction. LSI comprises an extension of Wong's GVSM. The difference between LSI and the GVSM lies in the fact that LSI uses a low-rank approximation of the term correlation matrix to inform its similarity function, while the GVSM uses the full-rank correlation matrix. An LSI system with $k = k_{max}$ thus converges on the GVSM solution. This yields a system that treats the sample correlation matrix as if it were the population correlation matrix. In light of this we may understand an LSI system where $k = k_{opt}$ as optimized insofar as it operates on the best approximation of the population term correlation matrix, in the least-squares sense. Thus reducing k to k_{opt} removes error from the LSI model by excluding eigenvalues that are likely to have arisen in the observed data due to sampling error. By applying APA, PA, and EV1 we assume that the difference between k_{opt} and k_{max} is proportional to the degree of correlation among the terms.

I cannot—and do not—hope that this study offers the last word on dimensionality estimation for IR. The multivariate statistical literature is replete with research on choosing the number of principal components to retain, a problem that is identical to estimating k_{opt} for LSI. Likewise, work in the design of dimensionality reduction-based IR systems continues to proceed apace. If this dissertation has succeeded in its goals it will encourage future collaboration between researchers in the IR and statistical communities. By adapting standard statistical models (i.e. Horn's parallel analysis) I have lent a practical basis and theoretical credibility to LSI's dimensionality reduction. Likewise, the failure of methods

such as percent-of-variance and Bartlett's test suggest important new challenges for statisticians. Like all research, then, this dissertation has answered some modest questions while introducing a host of new ones.

Bibliography

[1] S. J. Allen and R. Hubbard. Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research*, 21:393–398, 1986.

[2] R. A. Amsler. Machine–readable dictionaries. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 19, pages 161–209. Knowledge Industry Publication, Inc., 1984.

[3] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 1984.

[4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[5] M. S. Bartlett. The statistic conception of mental factors. *The Journal of the Royal Statistical Society*, 2:248–252, 1937.

[6] M. S. Bartlett. Methods of estimating mental factors. *Nature*, 141:609–610, 1938.

[7] M. S. Bartlett. The standard errors of discriminant function coefficients. *The Journal of the Royal Statistical Society*, 6:169–173, 1939.

[8] J. M. Ten Berge and H. A. Kiers. Retrieving the correlation matrix from a truncation pca solution. *Psychometrika*, 64(3):317–324, 1999.

[9] M. W. Berry. Svdpackc (version 1.0) User's Guide, University of Tennesse tech. report cs–93–194. Technical report, University of Tennessee, 1993 (Revised October 1996). Available at http://www.netlib.org/svdpack/index.html.

[10] M. W. Berry, S. T. Dumais, and G. W. OBrien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[11] G. K. Bhattacharyya and R. A. Johnson. *Statistical Concepts and Methods*. Wiley, 1977.

[12] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford, 1995.

[13] K. Bollacker, S. Lawrence, and C. L. Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In K. P. Sycara and M. Wooldridge, editors, *Proceedings of the Second International Conference on Autonomous Agents*, pages 116–123, New York, 1998. ACM Press.

[14] A. Bookstein and D. R. Swanson. A decision theoretic foundation for indexing. *Journal of the American Society for Information Science*, 26(1):45–50, 1975.

[15] B. C. Brookes. The measure of information retrieval effectiveness proposed by swets. *Journal of Documentation*, 24(1):41–54, 1968.

[16] K. P. Burnham and D. R. Anderson. *Model Selection and Inference: A Practical Information–Theoretic Approach*. Springer, New York, 1998.

[17] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:140–161, 1966.

[18] R. B. Cattell and J. Jaspars. A general plasmode for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, 67(3):1–212, 1967.

[19] V. S. Cherkassky and F. M. Mulier. *Learning from data: concepts, theory, and methods*. Wiley Interscience, 1998.

[20] K. W. Church and W. A. Gale. Poisson mixtures. *Natural language engineering*, 1:163–190, 1995.

[21] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[22] C. W. Cleverdon. The cranfield tests on index language devices. *ASLIB Proceedings*, 19:173–192, 1967.

[23] C. W. Cleverdon and J. Mills. The testing of index language devices. *ASLIB Proceedings*, 15:106–130, 1963.

[24] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, 19(1):30–41, 1968.

[25] W. S. Cooper. On selecting a measure of retrieval effectiveness, part 1. *Journal of the American Society for Information Science*, 39:87–100, 1975.

[26] W. S. Cooper. Getting beyond Boole. *Information Processing and Management*, 24:243–248, 1988.

[27] W. S. Cooper. Inconsistences and misnomers in probabilistic IR. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–62, 1991.

[28] W. S. Cooper and M. E. Maron. Foundation of probabilistic and utility theoretic indexing. *Journal of the ACM*, 25(1):67–80, 1978.

[29] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Number 88 in Monographs on Statistics and Probability. Wiley, 2nd edition, 2001.

[30] W. B. Croft. User–specified domain knowledge for document retrieval. In *ACM Conference on Research and Development in Information Retrieval*, pages 201–206. Association for Computing Machinery, 1986.

[31] W. B. Croft. Approaches to intelligent information retrieval. *Information Processing and Management*, 23(4):249–254, 1987.

[32] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of the American Society for Information Science*, 41(6):391–407, 1990.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Association*, 39(B):1–38, 1977.

[34] K. W. Dickman. *Factorial validity of a rating intrument*. PhD thesis, University of Illinois, 1960.

[35] W. R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley, 1984.

[36] C. H. Q. Ding. A similarity–based probability model for latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[37] C. H. Q. Ding. A probabilistic model for dimensionality reduction in information retrieval and filtering. Read at 1st SIAM Computational Information Retrieval Workshop, October 2000.

[38] S. T. Dumais. LSI meets TREC: A status report. In *Proceedings of the First Text Retrieval Conference (TREC1)*, pages 137–152, 1992.

[39] S. T. Dumais. Latent semantic indexing (LSI) and TREC–2. In *Proceedings of the Second Text Retrieval Conference (TREC 2)*, 1993.

[40] S. T. Dumais. Latent semantic indexing (LSI): TREC–3 report. In *Proceedings of the Third Text Retrieval Conference (TREC 3)*, 1994.

[41] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the Conference on Human Factors in Computing Systems CHI'88*, 1988.

[42] A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal of Matrix Analysis and Applications*, 9(4):543–560, 1988.

[43] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[44] B. Efron. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[45] M. Efron and G. Geisler. Using dimensionality reduction to improve similarity judgements for recommendation. In *Proceedings of the Second DELOS Network of Excellence Workshop on 'Personalisation and Recommender Systems in Digital Libraries'*, number No. 01/W03 in ERCIM Workshop Proceedings, 2001. http://www.ercim.org/publication/ws–proceedings/DelNoe02/index.html.

[46] D. Ellis. The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science*, 47(1):23–36, 1996.

[47] A. A. Cota et al. Interpolating 95th percentile eigenvalues from random data: an empirical example. *Educational and Psychological Measurement*, 53:585–595, 1993.

[48] R. S. Longman et al. A regression equation for the parallel analysis criterion in principal components analysis. *Multivariate Behavior Research*, 24(1):56–69, 1989.

[49] C. Fellbaum. *Wordnet: An electronic lexical database*. MIT Press, 1998.

[50] R. A. Fisher. *Collected papers of R. A. Fisher*. University of Adelaide, 1971–1974.

[51] P. Foltz, W. Kintsch, and T. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307, 1998.

[52] G. Forsythe, M. Malcom, and C. Moler. *Computer Methods for Mathematical Computations*. Prentice Hall, 1977.

[53] E. A. Fox. Lexical relations: Enhancing effectiveness of information retrieval systems. *ACM SIGIR Forum*, 15(3):5–36, 1980.

[54] K. Fukunaga. Intrinsic dimensionality extraction. In *Handbook of Statistics: Classification, Pattern Recognition, and Reduction of Dimensionality*, volume 2, pages 347–360. North Holland, 1982.

[55] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human–system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[56] P. Gardenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[57] L. W. Glorfeld. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55:377–393, 1995.

[58] G. H. Golub and C. F. van Loan. *Matrix Computations*. Baltimore, The Johns Hopkins University Press, 1989.

[59] M. D. Gordon and S. T. Dumais. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, 49(8):674–685, 1998.

[60] P. E. Green. *Analyzing Multivariate Data*. Dryden, 1978.

[61] L. Guttman. Some necessary conditions for common factor analysis. *Psychometrika*, 19(2):149–161, 1954.

[62] L. Guttman. To what extent can communalities reduce rank? *Psychometrika*, 23(3):297–308, 1958.

[63] A. R. Hakstian and W. T. Rogers. The behavior of number–of–factors rules with simulated data. *Multivariate Behavioral Research*, 17:193–219, 1982.

[64] D. K. Harman. The TREC conferences. In R. Kuhlen and M. Rittberger, editors, *Hypertext, Information Retrieval, Multimedia: Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95*, pages 9–28, 1995.

[65] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[66] S. P. Harter and C. A. Hert. Evaluation of information retrieval systems: Approaches, issues, and methods. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 32, pages 3–94. American Society for Information Science, 1997.

[67] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2001.

[68] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[69] T. Hofmann. Learning probabilistic models of the web. In *Research and Development in Information Retrieval*, pages 369–371, 2000.

[70] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[71] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–186, 1965.

[72] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 2000.

[73] P. Husbands, H. Simon, and C. Ding. The use of singular value decomposition for text retrieval, 2000.

[74] W. J. Hutchins. The concept of 'aboutness' in subject indexing. *ASLIB Proceedings*, 30:172–181, 1978.

[75] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*, pages 337–354. Prentice Hall, 1971.

[76] J. E. Jackson. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74:2204–2214, 1993.

[77] F. Jiang and M. L. Littman. Approximate dimension equalization in vector–based information retrieval. In *Proc. 17th International Conf. on Machine Learning*, pages 423–430. Morgan Kaufmann, San Francisco, CA, 2000.

[78] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML–98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[79] J. D. Jobson. *Applied Multivariate Data Analysis.* Springer, 1991.

[80] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.

[81] I. T. Jolliffe. *Principal Component Analysis.* Springer, 2nd edition, 2002.

[82] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[83] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35:30–48, 1979.

[84] K. Sparck Jones and K. Tait. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66, 1984.

[85] D. Jurafsky and J. H. Martin. *Speech and Language Processing.* Prentice Hall, 1999.

[86] H. Kaiser. The application of electronic computers to factor analysis. Read at the Meeting of the American Psychological Assocation, 1959.

[87] W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective.* Oxford University Press, 1988.

[88] Z. V. Lambert, A. R. Wildt, and R. M. Durand. Assessing sampling variation relative to number–of–factors criteria. *Educational and Psychological Measurement*, 50:365–395, 1990.

[89] T. K. Landauer. On the computational basis of learning and cognition: Arguments from LSA. *The Psychology of Learning*, 41:43–84, 2002.

[90] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[91] T. K. Landauer, D. Laham, and P. Foltz. Learning human–like knowledge by singular value decomposition: A progress report. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[92] S. Laurence and E. Margolis. Concepts and cognitive science. In *Concepts: Core Readings*, pages 3–82. MIT Press, 1999.

[93] W. Lederman. On the rank of the reduced correlational matrix in multiple factor analysis. *Psychometrika*, 2(2):85–93, 1937.

[94] M. E. Lesk. Word–word associations in document retrieval systems. *American Documentation*, 20(1):27–38, January 1969.

[95] R. L. Linn. A monte carlo approach to the number of factors problem. *Psychometrika*, 33:37–71, 1968.

[96] R. M. Losee. Term dependence: Truncating the bahadur lazarsfeld expansion. *Information Processing and Management*, 30(2):293–303, 1994.

[97] R. M. Losee. *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer, Boston, 1998.

[98] R. M. Losee. When information retrieval measures agree about the relative quality of document rankings. *Journal of the American Society for Information Science*, 51(9):834–840, 2000.

[99] H. P. Luhn. A new method of recording and searching information. *American Documentation*, 4(1):14–16, 1955.

[100] H. P. Luhn. A statistical approach to the mechanized encoding and searching of literary information. *IBM Journalof Research and Development*, 1(4):309–317, October 1957.

[101] H. P. Luhn. The automatic derivation of information retrieval encodements from machine–readable texts. In A. Kent, editor, *Information Retrieval and Machine Translation*, volume 3, pages 1021–1028. Interscience Publication, 1961.

[102] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press,, Cambridge, 1999.

[103] M. T. Maybury, editor. *Intelligent Multimedia Information Retrieval*. AAAI Press, 1997.

[104] P. McCullagh and Nelder J. A. *Generalized Linear Models*. Chapman and Hall, Boca Raton, second edition, 1989.

[105] M. Mihail and C. H. Papadimitriou. On the eigenvalue power law. Read at RANDOM 2002, 2002.

[106] T. M. Mitchell. *Machine Learning*. McGraw–Hill, 1997.

[107] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1982.

[108] A.K. Jain N. Wyse, R. Dubes. A critical evaluation of intrinsic dimensionality algorithms. In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, pages 415–425. North-Holland, 1980.

[109] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. Irwin, Chicago, 1996.

[110] G. B. Newby. Cognitive space and information space. *Journal of the American Society for Information Science*, 52(12):1026–1048, 2001.

[111] M. P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, 1998.

[112] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 159–168, 1998.

[113] M. F. Porter. An algorithm for suffix stripping. In *Program*, pages 130–137, 1980.

[114] K. Prey, J. C. French, A. L. Powell, and C. L. Viles. Inverse document frequency and web search engines. Technical Report CS–2001–07, University of Virginia, 5 2001.

[115] W. V. O. Quine. Two dogmas of empiricism. In *Concepts: Core Readings*, pages 153–170. MIT Press, 1999.

[116] A. C. Rencher. *Methods of Multivariate Analysis*. Wiley–Interscience, 1995.

[117] C. H. Van Rijsbergen. A theoretical basis for the use of cooccurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, June 1977.

[118] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

[119] S. E. Robertson. The parametric description of retrieval tests, part 1. *Journal of Documentation*, 25(1):1–27, 1969.

[120] S. E. Robertson. The parametric description of retrieval tests, part 2. *Journal of Documentation*, 25(2):93–107, 1969.

[121] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System–Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[122] E. Rosch. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.

[123] E. Rosch. Principles of categorization. In *Concepts: Core Readings*, pages 189–206. MIT Press, 1999.

[124] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–349, 1976.

[125] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[126] G. Salton and C. Buckley. Term–weighting approaches in automatic text retireval. *Information Processing and Management*, 24:513–523, 1988.

[127] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[128] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

[129] G. Salton and M. McGill. *Introduction into Modern Information Retrieval*. McGraw–Hill, 1983.

[130] G. Salton and M. J. McGill. *The SMART and SIRE Experimental Retrieval System*. McGraw–Hill, 1983.

[131] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[132] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[133] T. Saracevic. Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[134] L. Schamber. Relevance and information behavior. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 29, pages 3–48. American Society for Information Science, 1994.

[135] L. Schamber, M. Eisenberg, and M. S. Nilan. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26(6):755–776, 1990.

[136] H. Schutze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Research and Development in Information Retrieval*, pages 229–237, 1995.

[137] W. M. Shaw, R. Burgin, and P. Howell. Performance standards and evaluation in ir test collections: Cluster–based retrieval models. *Information Processing and Management*, 33(1):1–14, 1997.

[138] W. M. Jr. Shaw, J. B. Wood, R. E. Wood, and H. R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.

[139] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Blackwell, 2 edition, 1995.

[140] R. E. Story. An explanation of the effectiveness of latent semantic indexing by means of a bayesian regression model. *Information Processing and Management*, 32(3):329–344, 1996.

[141] G. Strang. *Linear Algebra and its Applications*. International Thompson Publishing, 1988.

[142] S. Subhash. *Applied Multivariate Techniques*. Wiley, 1996.

[143] J. A. Swets. Effectiveness of information retrieval methods. *American Documentation*, 20(1):72–89, 1969.

[144] R. Tang and P. Solomon. Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing and Management*, 34(2/3):237–256, 1998.

[145] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[146] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, 1998.

[147] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

[148] G. Williams. *Linear Algebra with Applications*. Jones and Bartlett Mathematics, 2001.

[149] L. Wittgenstein. *Philosophical Investigations*. Prentice-Hall, 3 edition, 1953.

[150] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong. On modeling of information retrieval concepts in vector space. *TODS*, 12(2):299–321, 1987.

[151] G. K. Zipf. Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 40:1–95, 1929.

[152] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

[153] W. R. Zwick and W. F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99:432–442, 1986.