# Latent Semantic Indexing: A Probabilistic Analysis

**Christos H. Papadimitriou**
Computer Science Division, U. C. Berkeley
christos@cs.berkeley.edu

**Prabhakar Raghavan**
IBM Almaden Research Center
pragh@almaden.ibm.com

**Hisao Tamaki**
Computer Science Department, Meiji University
tamaki@cs.meiji.ac.jp

**Santosh Vempala**
Department of Mathematics, M.I.T.
vempala@math.mit.edu

December 30, 1998

**Abstract**

Latent semantic indexing (LSI) is an information retrieval technique based on the spectral analysis of the term-document matrix, whose empirical success had heretofore been without rigorous prediction and explanation. We prove that, under certain conditions, LSI does succeed in capturing the underlying semantics of the corpus and achieves improved retrieval performance. We also propose the technique of *random projection* as a way of speeding up LSI. We complement our theorems with encouraging experimental results. We also argue that our results may be viewed in a more general framework, as a theoretical basis for the use of spectral methods in a wider class of applications such as collaborative filtering.

## 1 Introduction

The field of information retrieval has traditionally been considered outside the scope of database theory. While database theory deals with queries that are precise predicates (the so-called "employee-manager-salary paradigm"), in information retrieval we have the rather nebulous and ill-defined concept of "relevance", which depends in intricate ways on the intent of the user and the nature of the corpus. Evidently, very little theory can be built on this basis. (See [11, 22] for surveys on information retrieval, including discussions of the technique that is the focus of this paper, from database and theoretical points of view, respectively; [23, 24] are classical texts on the subject of information retrieval.)

However, the field of information retrieval has been evolving in directions that bring it closer to databases. Information retrieval systems are increasingly being built on relational (or object-relational) database systems, rather than on flat text and index files. Another important change is the dramatic expansion of the scope of information retrieval, with the advent of multimedia, the internet, and globalized information; database concepts and some theory have started to find fertile ground there (see for example [10, 3, 19], as well a record number of information retrieval papers in the 1997 SIGMOD Proceedings)[1] Secondly, the techniques employed in information retrieval have

[1]And, perhaps as importantly, the stakes have become too high for database theory to lightly pass this field by.

become more mathematical and sophisticated, more plausibly amenable to analytical treatment. The present paper is an attempt to treat rigorously one such technique, *latent semantic indexing (LSI),* introduced next. Thirdly, information retrieval systems are increasingly being built on relational (or object-relational) database systems (rather than on flat text and index files). Finally, the advent of the web has enabled powerful new applications such as *collaborative filtering* (also known as target or personalized recommendation systems) that can be tackled using techniques inspired in part by information retrieval [4]; more on this in Section 6.

### IR and LSI

The complexity of information retrieval is best illustrated by the two nasty classical problems of *synonymy* (missing documents with references to "automobile" when querying on "car") and *polysemy* (retrieving documents about the internet when querying on "surfing"). To deal with these two and other similar problems, we would ideally like to represent documents (and queries) not by terms (as in conventional vector-based methods), but by the *underlying (latent, hidden) concepts* referred to by the terms. This hidden structure is not a fixed many-to-many mapping between terms and concepts, but depends critically on the *corpus* (document collection) in hand, and the term correlations it embodies.

*Latent Semantic Indexing* [6] is an information retrieval method which attempts to capture this hidden structure by using techniques from linear algebra. Briefly (see the next section for a more detailed description), vectors representing the documents are projected in a new, low-dimensional space obtained by *singular value decomposition* of the term-document matrix $A$ (see the next subsection). This low-dimensional space is spanned by the eigenvectors of $A^T A$ that correspond to the few largest eigenvalues —and thus, presumably, to the few most striking correlations between terms. Queries are also projected and processed in this low-dimensional space. This results not only in great savings in storage and query time (at the expense of some considerable preprocessing), but also, according to empirical evidence reported in the literature, to *improved information retrieval* [1, 8, 9]. Indeed, it has been repeatedly reported that LSI outperforms, with regard to precision and recall in standard collections and query workloads, more conventional vector-based methods, and that it does address the problems of polysemy and synonymy.

There is very little in the literature in the way of a mathematical theory that predicts this improved performance. An interesting mathematical fact due to Eckart and Young (stated below as Theorem 1), often cited as an explanation of the improved performance of LSI, states, informally, that LSI retains as much as possible the relative position of the document vectors. This may only provide an explanation of why LSI *does not deteriorate too much* in performance over conventional vector-space methods; it fails to justify the observed improvement in precision and recall.

This paper is a first attempt at using mathematical techniques to rigorously explain the empirically observed improved performance of LSI. Since LSI seems to exploit and reveal the statistical properties of a corpus, we must start with a rigorous probabilistic model of the corpus (that is to say, a mathematical model of how corpora are generated); we do this in Section 3. Briefly, we model *topics* as probability distributions on terms. A *document* is then a probability distribution that is the convex combination of a small number of topics. We also include in our framework *style* of authorship, which we model by a stochastic matrix that modifies the term distribution. A *corpus* is then a collection of documents obtained by repeatedly drawing sample documents. (In

Section 6 we briefly discuss an alternative probabilistic model, motivated in part by applications to collaborative filtering.)

Once we have a corpus model, we would like to determine under what conditions LSI results in enhanced retrieval. We would like to prove a theorem stating essentially that *if the corpus is a reasonably focused collection of meaningfully correlated documents, then LSI performs well.* The problem is to define these terms so that (1) there is a reasonably close correspondence with what they mean intuitively and in practice, and (2) the theorem can be proved. In Section 4 we prove results that, although not quite as general as we would have liked, definitely point to this direction. In particular, we show that in the special case in which (a) there is no style modifier; (b) each document is on a single topic; and (c) the terms are partitioned among the topics so that each topic distribution has high probability on its own terms, and low probability on all others; *then* LSI, projecting to a subspace of dimension equal to the number of topics, will discover these topics exactly, with high probability assuming that the length of each document in the corpus is large enough.

In Section 5 we show that, if we project the term-document matrix on a *completely random* low-dimensional subspace, then with high probability we have a distance-preservation property akin to that enjoyed by LSI. This suggests that random projection may yield an interesting improvement on LSI: we can perform the LSI precomputation not on the original term-document matrix, but on a low-dimensional projection, at great computational savings and no great loss of accuracy (Theorem 5).

Random projection can be seen as an alternative to (and a justification of) *sampling* in LSI. Reports on LSI experiments in the literature seem to suggest that LSI is often done not on the entire corpus, but on a randomly selected subcorpus (both terms and documents may be sampled, although it appears that most often documents are). There is very little non-empirical evidence of the accuracy of such sampling. Our result suggests a different and more elaborate (and computationally intensive) approach —projection on a random low-dimensional subspace— which can be *rigorously proved* to be accurate.

## 2 LSI background

A *corpus* is a collection of documents. Each document is a collection of *terms* from a universe of $n$ terms. Each document can thus be represented as a vector in $\Re^n$ where each axis represents a term. The $i$th coordinate of a vector represents some function of the number of times the $i$th term occurs in the document represented by the vector. There are several candidates for the right function to be used here (0-1, frequency, etc.), and the precise choice does not affect our results.

Let $A$ be an $n \times m$ matrix of rank $r$ whose rows represent terms and columns represent documents. Let the singular values of $A$ (the eigenvalues of $AA^T$) be $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ (not necessarily distinct). The *singular value decomposition* of $A$ expresses $A$ as the product of three matrices $A = UDV^T$, where $D = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is an $r \times r$ matrix, $U = (u_1, \ldots, u_r)$ is an $n \times r$ matrix whose columns are orthonormal, and $V = (v_1, \ldots, v_r)$ is an $m \times r$ matrix which is also column-orthonormal.

LSI works by omitting all but the $k$ largest singular values in the above decomposition, for some appropriate $k$; here $k$ is the dimension of the low-dimensional space alluded to in the informal description on page 2. It should be small enough to enable fast retrieval, and large enough to adequately capture the structure of the corpus (in practice, $k$ is in the few hundreds, compared with $r$ in the many thousands). Let $D_k = diag(\sigma_1, \ldots, \sigma_k)$, $U_k = (u_1, \ldots, u_k)$ and $V_k = (v_1, \ldots, v_k)$. Then

$$A_k = U_k D_k V_k^T$$

is a matrix of rank $k$, which is our approximation of $A$. The rows of $V_k D_k$ above are then used to represent the documents. In other words, the column vectors of $A$ (documents) are projected to the $k$-dimensional space spanned by the column vectors of $U_k$; we sometimes call this space the LSI space of $A$. How good is this approximation? The following well-known theorem gives us some idea (the subscript $F$ denotes the Frobenius norm).

**Theorem 1** *(Eckart and Young, see [16].) Among all $n \times m$ matrices $C$ of rank at most $k$, $A_k$ is the one that minimizes $\|A - C\|_F^2 = \sum_{i,j}(A_{i,j} - C_{i,j})^2$.*

Therefore, LSI preserves (to the extent possible) the relative distances (and hence, presumably, the retrieval capabilities) in the term-document matrix while projecting it to a lower-dimensional space. It remains to be seen in what way it *improves* these retrieval capabilities.

## 3 The probabilistic corpus model

There are many useful formal models of IR in the literature, and probability plays a major role in many of them —see for instance the surveys and comparisons in [14, 23, 25]. The approach in this body of work is to formulate information retrieval as a problem of learning the concept of "relevance" that relates documents and queries. The corpus and its correlations plays no central role. In contrast, our focus is on the probabilistic properties of the corpus.

Since LSI is supposed to exploit and bring out the structure of the corpus, it will fare well in a meaningful collection of strongly correlated documents, and will produce noise in a random set of unrelated documents. In order to study the dependence of the performance of LSI on the statistical properties of the corpus, we must start with a probabilistic model of a corpus. We state now our basic probabilistic model, which we will use for much of this paper.

Let the universe of all terms be $U$. A *topic* is a probability distribution on $U$. A meaningful topic is very different from the uniform distribution on $U$, and is concentrated on terms that might be used to talk about a particular subject. For example, the topic of "space travel" might favor the terms "galaxy" and "starship", while rarely mentioning "misery" or "spider". A possible criticism against this model is that it does not take into account correlations of terms within the same topic (for example, a document on the topic "internet" is much more likely to contain the term "search" if it also contains the term "engine").

The structure of documents is also heavily affected by *authorship style.* We model style as a $|U| \times |U|$ stochastic matrix (a matrix with nonnegative entries and row sums equal to 1), denoting the way whereby style modifies the frequency of terms. For example, a "formal" style may map "car" often to "automobile" and "vehicle," and seldom to "car" — and almost never to "wheels." Admittedly, this is not a comprehensive treatment of style; for example, it makes the assumption – not always valid – that this influence is independent of the underlying topic.

A *corpus model* $\mathcal{C}$ is a quadruple $\mathcal{C} = (U, \mathcal{T}, \mathcal{S}, D)$, where $U$ is the universe of terms, $\mathcal{T}$ is a set of topics, and $\mathcal{S}$ a set of styles, and $D$ a probability distribution on $\hat{\mathcal{T}} \times \hat{\mathcal{S}} \times \mathbf{Z}^+$, where by $\hat{\mathcal{T}}$ we denote the set of all convex combinations of topics in $\mathcal{T}$, by $\hat{\mathcal{S}}$ the set of all convex combinations of styles in $\mathcal{S}$, and by $\mathbf{Z}^+$ the set of positive integers (the integers represent the lengths of documents). That is, a corpus model is a probability distribution on topic combinations (intuitively, favoring combinations of a few related topics), style combinations, and document lengths (total number of term occurrences in a document).

A document is generated from a corpus model $\mathcal{C} = (U, \mathcal{T}, \mathcal{S}, D)$ through the following two-step sampling process. In the first step, a convex combination of topics $\hat{T}$ from $\hat{\mathcal{T}}$, a convex combination of styles $\hat{S}$ from $\hat{\mathcal{S}}$, and a positive integer $\ell$ are sampled according to distribution $D$. Then terms are sampled $\ell$ times to form a document, each time according to distribution $\hat{T}\hat{S}$. A *corpus of size* $m$ is a collection of $m$ documents generated from $\mathcal{C}$ by repeating this two-step sampling process $m$ times.

# 4 An analysis of LSI

Does LSI indeed bring together semantically related documents? And does it deal effectively with the problem of synonymy? We present below theoretical evidence that it does. Our results assume the corpus model has a particularly simple structure. We show that, in this case, LSI does discover the structure of the corpus, and handles synonymy well. These results should be taken only as *indications* of the kinds of results that can be proved; our hope is that the present work will lead to more elaborate techniques, so that LSI can be proved to work well under more realistic assumptions.

We will first need a useful lemma, which formalizes the following intuition: if the $k$ largest singular values of a matrix $A$ are well-separated from the remaining singular values, then the subspace spanned by the corresponding singular vectors is preserved well when a small perturbation is added to $A$.

**Lemma 1** *Let $A$ be an $n \times m$ matrix of rank $r$ with singular value decomposition*

$$A = UDV^T,$$

*where $D = \mathsf{diag}(\sigma_1, \ldots, \sigma_r)$. Suppose that, for some $k$, $1 \leq k < r$, $\sigma_k/\sigma_{k+1} > c\sigma_1/\sigma_k$ for sufficiently large constant $c$. Let $F$ be an arbitrary $n \times m$ matrix with $\|F\|_2 \leq \epsilon$, where $\epsilon$ is a sufficiently small positive constant. Let $A' = A + F$ and let $U'D'V'^T$ be its singular-value decomposition. Let $U_k$ and $U'_k$ be $n \times k$ matrices consisting of the first $k$ columns of $U$ and $U'$ respectively. Then, $U'_k = U_k R + G$ for some $k \times k$ orthonormal matrix $R$ and some $n \times k$ matrix $G$ with $\|G\|_2 \leq O(\epsilon)$.*

The proof of this lemma, given in the appendix, relies on a theorem of Stewart [17] about perturbing a symmetric matrix.

Let $\mathcal{C} = (U, \mathcal{T}, D)$ be a corpus model. We call $\mathcal{C}$ *pure* if each document involves only a single topic. We call $\mathcal{C}$ *$\epsilon$-separable*, where $0 \leq \epsilon < 1$, if a set of terms $U_T$ is associated with each topic $T \in \mathcal{T}$ so that (1) $U_T$ are mutually disjoint and (2) for each $T$, the total probability $T$ assigns to the terms in $U_T$ is at least $1 - \epsilon$. We call $U_T$ the *primary set of terms* of topic $T$.

The assumption that a corpus model is style-free and pure is probably too strong and its elimination should be addressed in future investigations. On the other hand, the assumption that a corpus is $\epsilon$-separable for some small value of $\epsilon$ may be reasonably realistic, since documents are usually preprocessed to eliminate commonly-occurring stop-words.

Let $\mathcal{C}$ be a pure corpus model and let $k = |\mathcal{T}|$ denote the number of topics in $\mathcal{C}$. Since $\mathcal{C}$ is pure, each document generated from $\mathcal{C}$ is in fact generated from some single topic $T$: we say that the document *belongs to* the topic $T$. Let $C$ be a corpus generated from $\mathcal{C}$ and, for each document $d \in C$, let $v_d$ denote the vector assigned to $d$ by the rank-$k$ LSI performed on $C$. We say that the rank-$k$ LSI is *$\delta$-skewed* on the corpus instance $C$ if, for each pair of documents $d$ and $d'$, $v_d \cdot v_{d'} \leq \delta \|v_d\| \|v_{d'}\|$ if $d$ and $d'$ belong to different topics and $v_d \cdot v_{d'} \geq 1 - \delta \|v_d\| \|v_{d'}\|$ if they belong to the same topic. Informally, the rank-$k$ LSI is $\delta$-skewed on a corpus (for small $\delta$), if it assigns nearly orthogonal vectors to two documents from different topics and nearly parallel vectors to two documents from a single topic: LSI does a particularly good job of classifying documents when applied to such a corpus. We would like to show that the rank-$k$ LSI is $O(\epsilon)$-skewed on a corpus generated from an $\epsilon$-separable corpus model with $k$ topics. We start from the case where $\epsilon = 0$.

**Theorem 2** *Let $\mathcal{C}$ be a pure, 0-separable corpus model with $k$ topics such that the probability each topic assigns to each term is at most $\tau$, where $\tau > 0$ is a sufficiently small constant. Let $C$ be a corpus of $m$ documents generated from $\mathcal{C}$. Then, the rank-$k$ LSI is 0-skewed on $C$ with probability $1 - O(m^{-1})$.*

*Proof.* Let $C_i$ denote the subset of the generated corpus $C$ consisting of documents belonging to topic $T_i$, $1 \leq i \leq k$. Each document of $C_i$ consists only of terms in $U_i$, the primary set of terms associated with topic $T_i$. Thus, the term-document matrix $A$ representing corpus $C$ consists of blocks $B_i$, $1 \leq i \leq k$: the rows of $B_i$ correspond to terms in $U_i$ and columns of $B_i$ correspond to documents in $C_i$; the entire matrix $A$ can have non-zero entries in these rows and columns only within $B_i$. Therefore, $A^T A$ is block-diagonal with blocks $B_i^T B_i$, $1 \leq i \leq k$. Now focus on a particular block $B_i^T B_i$ and let $\lambda_i$ and $\lambda_i'$ denote the largest and the second largest eigenvalues of $B_i^T B_i$. Intuitively, the matrix $B_i^T B_i$ is essentially the adjacency matrix of a random bipartite multigraph and then, from the standard theory of spectra of graphs[5], we have that $\lambda_i'/\lambda_i \to 0$ with probability 1 as $\tau \to 0$ and $|C_i| \to \infty$. Below we give a formal justification of this by showing that a quantity that captures this property, the *conductance* [21] (equivalently, *expansion*) of $B_i^T B_i$ is high. The conductance of an undirected edge-weighted graph $G = (V, E)$ is

$$\min_{S \subset V} \frac{\sum_{i \in S, j \in \overline{S}} wt(i,j)}{\min\{|S|, |\overline{S}|\}}$$

Let $x^1, x^2, \ldots, x^t$ be random documents picked from the topic $T_i$. Then we will show that the conductance is $\Omega(\frac{|t|}{|T_i|})$, where $|T_i|$ is the number of terms in the topic $T_i$. Let $G$ be the graph induced by the adjacency matrix $B_i^T B_i$. For any subset $S$ of the vertices (documents),

$$\begin{aligned} \sum_{i \in S, j \in \overline{S}} wt(i,j) &= \sum_{i \in S, j \in \overline{S}} x^i \cdot x^j \\ &= (\sum_{i \in S} x^i) \cdot (\sum_{j \in \overline{S}} x^j). \end{aligned}$$

Assume w.l.o.g. that $|S| \leq |\overline{S}|$. Let $p_s$ be the probability of the $s^{th}$ term in $T_i$. Then we can estimate, for each term, $\sum_{j \in \overline{S}} x_s^j \geq \min\{p_s/2, p_s - \epsilon\}$ with probability at least $1 - \frac{1}{2^t}$ using the independence of the $x^j$'s via a simple application of Chernoff-Hoeffding bound [18]. Using this we lower bound the weight of the cut $(S, \overline{S})$:

$$(\sum_{i \in S} x^i) \cdot (\sum_{j \in \overline{S}} x^j) \geq \sum_{i \in S} x_s^i (\min\{p_s/2, p_s - \epsilon\})$$

which is $\Omega(\frac{|S|}{|T_i|})$ with high probability by a second application of the Chernoff-Hoeffding bound. The desired bound on the conductance follows from this.

Thus, if the sample size $m = |C|$ is sufficiently large, and the maximum term probability $\tau$ is sufficiently small (note this implies that the size of the primary set of terms for each topic is sufficiently large), the $k$ largest eigenvalues of $A^T A$ are $\lambda_i$, $1 \leq i \leq k$, with high probability. Suppose now that our sample $C$ indeed enjoys this property. Let $\hat{u}_i$ denote the eigenvector of $B_i^T B_i$ corresponding to eigenvalue $\lambda_i$ (in the space where coordinates are indexed by the terms in $T_i$) and let $u_i$ be its extension to the full term space, obtained by padding zero entries for terms not in $T_i$. Then, the $k$-dimensional LSI-space for corpus $C$ is spanned by the mutually orthogonal vectors $u_i$, $1 \leq i \leq k$. When a vector $v_d$ representing a document $d \in C_i$ is projected into this space, the projection is a scalar multiple of $u_i$, because $v_d$ is orthogonal to $u_j$ for every $j \neq i$. Therefore, the rank-$k$ LSI is 0-skewed on $C$ as claimed. □

The following theorem generalizes the above result to the corpus which is slightly perturbed from the one generated from a 0-separable corpus model.

**Theorem 3** *Let $\mathcal{C}$ be a pure, $\epsilon$-separable corpus model with $k$ topics such that the probability each topic assigns to each term is at most $\tau$, where $\tau > 0$ is a sufficiently small constant. Let $B$ be the term-document matrix of a corpus of $m$ documents generated from $\mathcal{C}$ and let $C$ be a corpus whose term-document matrix is $A = B + F$ where $||F||_2 \leq \epsilon$. Then, the rank-$k$ LSI is $O(\epsilon)$-skewed on $C$ with probability $1 - O(m^{-1})$.*

*Proof.* The term document matrix $B$ consists of blocks $B_i$ as in the proof of Theorem 2.

Let $W_k$ and $W'_k$ denote the basis matrix of the $k$-dimensional space the rank-$k$ LSI applied to $B$ and $A$, respectively, identifies. Since the two-norm of the perturbation $F = A - B$ is at most $\epsilon$, Lemma 1 implies that the $W'_k$ is obtained from $W_k$ applying an orthogonal transformation and then adding a perturbation with a two-norm of $O(\epsilon)$. Roughly speaking, since the space spanned by $W'_k$ is a small perturbation of that spanned by $W_k$, projecting a vector representing a document in $C_i$ into $W'_k$ yields a vector close, in its direction, to $u_i$ (the dominating eigenvector of $B_i^T B_i$). Therefore, the LSI representations of two documents are almost in the same direction if they belong to the same topic and are nearly orthogonal if they belong to different topics. The analysis below shows that the rank-$k$ LSI is indeed $O(\epsilon)$-skewed on $C$ with high probability.

Let $A^{(i)}$ denote the vector corresponding to the $i^{th}$ document, i.e. the $i^{th}$ row of $A$. Similarly $B^{(i)}$ denotes the $i^{th}$ row of the unperturbed matrix $B$. For any pair of documents $i, j$ we will show that

$$|A_k^{(i)} \cdot A_k^{(j)} - B_k^{(i)} \cdot B_k^{(j)}| = O(\epsilon |B_k^{(i)}||B_k^{(j)}|) \tag{1}$$

Recall that for any $i$,

$$A_k^{(i)} = A^{(i)} W_k' \quad \text{and} \quad B_k^{(i)} = B^{(i)} W_k$$

Thus

$$\begin{aligned}
|A_k^{(i)} - B_k^{(i)}| &= |A^{(i)} W_k' - B^{(i)} W_k| \\
&= |A^{(i)}(W_k' - W_k) + (A^{(i)} - B^{(i)}) W_k| \\
&\leq |A^{(i)}(W_k' - W_k)| + |(A^{(i)} - B^{(i)}) W_k| \\
&\leq |A^{(i)}| \|(W_k - W_k')\|_2 + |A^{(i)} - B^{(i)}| \|W_k\|_2 \\
&= O(\epsilon)(|A^{(i)}| + \|W_k\|_2) \\
&= O(\epsilon |A^{(i)}|) \\
&= O(\epsilon |A_k^{(i)}|)
\end{aligned}$$

Note that the conclusion above remains true with $B_k^{(i)}$ in place of $A_k^{(i)}$.

Now for a pair $i, j$,

$$\begin{aligned}
|A_k^{(i)} \cdot A_k^{(j)} - B_k^{(i)} \cdot B_k^{(j)}| &= |(A_k^{(i)} - B_k^{(i)}) \cdot A_k^{(j)} + B_k^{(i)} \cdot (A_k^{(j)} - B_k^{(j)})| \\
&= O(\epsilon |B_k^{(i)}| |A_k^{(j)}|) \\
&= O(\epsilon |B_k^{(i)}| |B_k^{(j)}|)
\end{aligned}$$

For a pair $i, j$ from the same topic, $B_k^{(i)} \cdot B_k^{(j)} = |B_k^{(i)}| |B_k^{(j)}|$, and for $i, j$ from different topics, $B_k^{(i)} \cdot B_k^{(j)} = 0$. Using (1) this implies that the rank-$k$ LSI is $O(\epsilon)$-skewed. $\square$

If an $\epsilon$-separable corpus model $\mathcal{C}$ satisfies the additional conditions in Theorem 2 and moreover is such that the length of each document is large enough, then a corpus generated from $\mathcal{C}$ satisfies the condition of Theorem 3 and therefore is $O(\epsilon)$-skewed.

## Experiments

Even though Theorems 2 and 3 give asymptotic results and only claim that the probability approaches 1 as the size parameters grow, the phenomenon they indicate can be observed in corpora of modest sizes, as is seen in the following experiment. We generated 1000 documents (each 50 to 100 terms long) from a corpus model with 2000 terms and 20 topics. Each topic is assigned a disjoint set of 100 terms as its primary set. The probability distribution for each topic is such that 0.95 of its probability density is equally distributed among terms from the primary set, and the remaining 0.05 is equally distributed among all the 2000 terms. Thus this corpus model is 0.05-separable. We measured the angle (*not* some function of the angle such as the cosine) between all pairs of documents in the original space and in the rank 20 LSI space. The following is a typical result; similar results are obtained from repeated trials. Call a pair of documents *intra-topic* if the two documents are generated from the same topic and *inter-topic* otherwise.

| | intra-topic | | | |
|---|---|---|---|---|
| | min | max | average | std |
| original space | 0.801 | 1.39 | 1.09 | 0.079 |
| LSI space | 0 | 0.312 | 0.0177 | 0.0374 |

| | inter-topic | | | |
|---|---|---|---|---|
| | min | max | average | std |
| original space | 1.49 | 1.57 | 1.57 | 0.00791 |
| LSI space | 0.101 | 1.57 | 1.55 | 0.153 |

Here, angles are measured in radians. It can be seen that the angles of intra-topic pairs are dramatically reduced in the LSI space. Although the minimum inter-topic angle is rather small, indicating that some inter-topic pairs can be close enough to be confused, the average and the standard deviation show that such pairs are extremely rare. Results from experiments with different size-parameters are also similar in spirit. In this and the other experiments reported here, we used SVDPACKC [2] for singular value decomposition.

### Synonymy

We end this section with a brief discussion of synonymy in the context of LSI. Let us consider a simple model in which two terms have identical co-occurrences (this generalizes synonymy, as it also applies to pairs of terms such as *supply–demand* and *war–peace*). Furthermore, these two terms have each a small occurrence probability. Then, in the term-term autocorrelation matrix $AA^T$, the two rows and columns corresponding to these terms will be nearly identical. Therefore, there is a very small eigenvalue corresponding to this pair — presumably the smallest eigenvalue of $AA^T$. The corresponding eigenvector will be a vector with 1 and $-1$ at these terms — that is to say, the *difference* of these terms. Intuitively then, this version of LSI will "project out" a very weak eigenvector that corresponds to the presumably insignificant semantic differences between the two synonymous terms. This is exactly what one would expect from a method that claims to bring out the hidden semantics of the corpus.

## 5 LSI by random projection

A result by Johnson and Lindenstrauss [12, 20] states that if points in a vector space are projected to a random subspace of suitably high dimension, then the distances between the points are approximately preserved. Although such a random projection can be used to reduce the dimension of the document space, it does not bring together semantically related documents. LSI on the other hand seems to achieve the latter, but its computation time is a bottleneck. This naturally suggests the following *two-step* approach:

1. Apply a random projection to the initial corpus to $l$ dimensions, for some small $l > k$, to obtain, with high probability, a much smaller representation, which is still very close (in terms of distances and angles) to the original corpus.
2. Apply rank $O(k)$ LSI (because of the random projection, the number of singular values kept may have to be increased a little).

In this section we establish that the above approach works, in the sense that the final representation is very close to what we would get by directly applying LSI. Another way to view this result is that random projection gives us a fast way to *approximate* the eigenspace (eigenvalues, eigenvectors) of a matrix.

We first state the Johnson-Lindenstrauss lemma.

**Lemma 2** *(Johnson and Lindenstrauss, see [12, 20].) Let $v \in R^n$ be a unit vector, let $H$ be a random $l$-dimensional subspace through the origin, and let the random variable $X$ denote the square of the length of the projection of $v$ onto $H$. Suppose $0 < \epsilon < \frac{1}{2}$, and $24 \log n < l < \sqrt{n}$. Then, $\mathbf{E}[X] = l/n$, and*

$$Pr(|X - l/n| > \epsilon l/n) < 2\sqrt{l} e^{-(l-1)\epsilon^2/4}.$$

Using the above result, we can infer that with high probability, all pairwise Euclidean distances are approximately maintained under projection to a random subspace. By choosing $l$ to be $\Omega(\frac{\log m}{\epsilon^2})$ in Lemma 2, with high probability the projected vectors, after scaling by a factor $\sqrt{n/l}$, $\{v_i'\}$, satisfy

$$\|v_i - v_j\|_2 (1 - \epsilon) \leq \|v_i' - v_j'\|_2 \leq \|v_i - v_j\|_2 (1 + \epsilon).$$

Similarly inner products are also preserved approximately: $2 v_i \cdot v_j = v_i^2 + v_j^2 - (v_i - v_j)^2$. So the projected vectors satisfy

$$2 v_i' \cdot v_j' \leq (v_i^2 + v_j^2)(1 + \epsilon) - (v_i - v_j)^2 (1 - \epsilon)$$

Therefore, $v_i' \cdot v_j' \leq v_i \cdot v_j (1 - \epsilon) + \epsilon (v_i^2 + v_j^2)$. In particular, if the $v_i$'s are all of length at most 1, then any inner product $v_i \cdot v_j$ changes by at most $2\epsilon$.

Consider again the term-document matrix $A$ generated by our corpus model. Let $R$ be a random column-orthonormal matrix with $n$ rows and $l$ columns, used to project $A$ down to an $l$-dimensional space. Let $B = \sqrt{\frac{n}{l}} R^T A$ be the matrix after random projection and scaling, where,

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \quad B = \sum_{i=1}^{t} \lambda_i a_i b_i^T$$

are the SVD's of $A$ and $B$ respectively.

**Lemma 3** *Let $\epsilon$ be an arbitrary positive constant. If $l \geq c \frac{\log n}{\epsilon^2}$ for a sufficiently large constant $c$ then, for $p = 1, \ldots, t$*

$$\lambda_p^2 \geq \frac{1}{k} [(1 - \epsilon) \sum_{i=1}^{k} \sigma_i^2 - \sum_{j=1}^{p-1} \lambda_j^2].$$

The proof of this lemma is given in the appendix. As a corollary we have:

**Corollary 4**

$$\sum_{p=1}^{2k} \lambda_p^2 \geq (1 - \epsilon) ||A_k||_F^2.$$

Now our rank $2k$ approximation to the original matrix as

$$B_{2k} = A \sum_{i=1}^{2k} b_i b_i^T$$

. From this we get the main result of this section:

**Theorem 5**

$$||A - B_{2k}||_F^2 \leq ||A - A_k||_F^2 + 2\epsilon ||A||_F^2$$

The measure $||A - A_k||_F$ tells us how much of the original matrix is recovered by direct LSI. So, in other words, the theorem says that the matrix obtained via random projection followed by LSI (expanded to twice the rank) recovers almost as much as the matrix obtained by direct LSI.

*Proof of theorem 5.* We have,

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \quad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

And also,

$$B = \sum_{i=1}^{l} \lambda_i a_i b_i^T, \quad B_{2k} = A \sum_{i=1}^{2k} b_i b_i^T$$

Since $b_1, \ldots, b_n$ are an orthonormal set of vectors,

$$||A - B_{2k}||_F^2 = \sum_{i=1}^{n} |(A - B_{2k}) b_i|^2$$

But, for $i = 1, \ldots, 2k$,

$$(A - B_{2k}) b_i = A b_i - A b_i = 0$$

And for $i = 2k + 1, \ldots, n$

$$(A - B_{2k}) b_i = A b_i$$

Hence,

$$||A - B_{2k}||_F^2 = \sum_{i=2k+1}^{n} |A b_i|^2 = \sum_{i=1}^{n} |A b_i|^2 - \sum_{i=1}^{2k} |A b_i|^2$$

$$= ||A||_F^2 - \sum_{i=1}^{2k} |A b_i|^2$$

On the other hand,

$$||A - A_k||_F^2 = \sum_{i=k+1}^{n} \sigma_i^2 = ||A||_F^2 - ||A_k||_F^2$$

Hence

$$||A - B_{2k}||_F^2 - ||A - A_k||_F^2 = ||A_k||_F^2 - \sum_{i=1}^{2k} |Ab_i|^2$$

That is,

$$||A - B_{2k}||_F^2 = ||A - A_k||_F^2 + (||A_k||_F^2 - \sum_{i=1}^{2k} |Ab_i|^2)$$

Next, we will show that

$$(1+\epsilon)\sum_{i=1}^{2k} |Ab_i|^2 \geq \sum_{i=1}^{2k} \lambda_i^2$$

We have

$$\begin{aligned} \sum_{i=1}^{2k} \lambda_i^2 &= \sum_{i=1}^{2k} |Bb_i|^2 \\ &= \sum_{i=1}^{2k} |\sqrt{\frac{n}{l}} R^T (Ab_i)|^2 \end{aligned}$$

Now from lemma 2 we have that for $l$ large enough, i.e., $l = \Omega(\frac{\log n}{\epsilon^2})$, with high probability,

$$(1-\epsilon)|Ab_i|^2 \leq \frac{n}{l}|R^T(Ab_i)|^2 \leq (1+\epsilon)|Ab_i|^2$$

for each $i$. Therefore with high probability,

$$\sum_{i=1}^{2k} \lambda_i^2 \leq (1+\epsilon)\sum_{i=1}^{2k} |Ab_i|^2$$

From corollary 4

$$\sum_{i=1}^{2k} \lambda_i^2 \geq (1-\epsilon)||A_k||^2$$

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq \frac{1}{(1+\epsilon)} \sum_{i=1}^{2k} \lambda_i^2 \geq \frac{1-\epsilon}{1+\epsilon} ||A_k||_F^2 \geq (1-2\epsilon)||A_k||_F^2$$

Substituting this above, we have,

$$||A - B_{2k}||_F^2 \leq ||A - A_k||_F^2 + 2\epsilon ||A_k||_F^2$$

Hence

$$||A - B_{2k}||_F^2 \leq ||A - A_k||_F^2 + 2\epsilon ||A||_F^2$$

■

What are the computational savings achieved by the two-step method? Let $A$ be an $n \times m$ matrix. The time to compute LSI is $O(mnc)$ if $A$ is sparse with about $c$ non-zero entries per column (i.e., $c$ is the average number of terms in a document). The time needed to compute the random projection to $l$ dimensions is $O(mcl)$. After the projection, the time to compute LSI is $O(ml^2)$. So the total time is $O(ml(l+c))$. To obtain an $\epsilon$ approximation we need $l$ to be $O(\frac{\log n}{\epsilon^2})$. Thus the running time of the two-step method is asymptotically superior: $O(m(\log^2 n + c \log n))$ compared to $O(mnc)$.

In [13], Frieze, Kannan and Vempala propose an alternative way to speed up LSI. They give fast algorithms for finding low rank approximations to an $m \times n$ matrix $A$. They compute an approximate Singular Value Decomposition from a randomly chosen submatrix of $A$. For any given $k, \epsilon, \delta$, their Monte Carlo algorithm finds *the description of* a matrix $D$ of rank at most $k$ so that

$$||A - D||_F \leq ||A - A_k||_F + \epsilon ||A||_F$$

holds with probability at least $1 - \delta$. ($||A||_F^2$ is the sum of squares of all the entries of the matrix, and $A_k$ is the best rank $k$ approximation with this measure.) The algorithm takes time polynomial in $k, 1/\epsilon, \log(1/\delta)$ only, i.e. *independent of* $m, n$. In [7] it is shown how to find such a low-rank approximation explicitly constructed in $O(mk^2)$ time.

# 6 Conclusion and further work

Our corpus model is, from a practical standpoint, rather weak; what would it take to establish theorems for "more realistic" models? Our present work has relied on studying the spectra of the perturbations of block matrices; more generally, we require tools to study the eigenvectors of the linear superposition of several matrices (under some assumptions about the individual matrices). Alternatively, we may think of the spectral properties of random graphs with (a small set of) varying edge probabilities. This appears be the technical bottleneck to furthering our results.

A theoretician's first reaction to an unexpected (positive *or* negative) empirical phenomenon is to understand it in terms of mathematical models and rigorously proved theorems; this is precisely what we have tried to do, with substantial if partial success. What we have been able to prove should be seen as a mere *indication* of what might hold; we expect the true positive properties of LSI to go far beyond the theorems we are proving here.

There are several specific technical issues to be pursued. Can Theorem 2 be extended to a model where documents could belong to several topics, or to one where term occurrences are not independent? Also, does LSI address polysemy (as spectral techniques of slightly different kind to, see [19])? We have seen some evidence that it does handle synonymy.

Theory should ideally go beyond the *ex post facto* justification of methods and explanation of positive phenomena, it should point the way to new ways of exploiting them and improving them. Section 5, in which we propose a random projection technique as a way of speeding up LSI (and possibly as an alternative to it), is an attempt in this direction.

Another important role of theory is to *unify and generalize.* Spectral techniques are not confined to the vector-space model, neither to the strict context of information retrieval. Furthermore, spectral analysis of a similar graph-theoretic model of the world-wide web has been shown experimentally to succeed in identifying topics and to substantially increase precision and recall in web searches [19], as well as in databases of law decisions, service logs and patents [4]. Finally, it is becoming clear that spectral techniques and their theoretical analysis may prove to be key methodologies in many other domains of current interest, such as data mining (using spectral techniques to discover correlations in relational databases [15]) and collaborative filtering (personalizing subscriber preferences and interests) [4]. The rows and columns of $A$ could in general be, instead of terms and documents, consumers and products, viewers and movies, or components and systems.

We conclude this section with a brief description of a promising alternative, *graph-theoretic*, corpus model. Suppose that documents are nodes in a graph, and weights on the edges capture conceptual proximity of two documents (for example, this distance matrix could be derived from, or in fact coincide with, $AA^T$). Then a *topic* is defined implicitly as a subgraph with high *conductance* [21], a concept of connectivity which seems very appropriate in this context. We can prove that, under an assumption similar to $\epsilon$-separability, spectral analysis of the graph can identify the topics in this model as well (proof omitted from this abstract):

**Theorem 6** *If the corpus consists of $k$ disjoint subgraphs with high conductance, and joined with edges with total weight bounded from above by an $\epsilon$ fraction of the distance matrix, then rank-$k$ LSI will discover the subgraphs.*

## References

[1] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. SIAM Review, 37(4), 1995, 573-595, 1995.

[2] M. W. Berry, T. Do, G. W. O'Brien, V. Krishna, and S. Varadhan. SVDPACKC (Version 1.0) User's Guide. University of Tennessee, April 1993.

[3] E. Brewer. Invited talk, 1997 PODS/SIGMOD, 1997.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan and S. Rajagopalan. "Mining information networks through spectral methods". In preparation, IBM Almaden Research Center, 1997.

[5] D.M. Cvetković, M. Doob, and H. Sachs, Spectra of Graphs, Academic Press, 1979.

[6] S. Deerwester, S. T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. Journal of the Society for Information Science, 41(6), 391-407, 1990.

[7] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. To appear, *Proceedings of the ACM-SIAM SODA*, 1999.

[8] S.T. Dumais, G.W. Furnas, T.K. Landauer, and S. Deerwester. Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285, 1988.

[9] S.T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers, 23(2), 229-236, 1991.

[10] R. Fagin. Combining fuzzy information from multiple sources. Proc. 1996 PODS, pp. 216–223, 1996.

[11] C. Faloutsos and D. Oard. A survey of information retrieval and filtering methods. Technical Report, University of Maryland Computer Science Dept., College Park, MD.

[12] P. Frankl and H. Maehara. The Johnson-Lindenstrauss Lemma and the Sphericity of some graphs, J. Comb. Theory B **44** (1988), 355-362.

[13] A. Frieze, R. Kannan and S. Vempala, "Fast Monte-Carlo Algorithms for finding low-rank approximations," Proc. 1998 FOCS, pp. 370-378, 1998.

[14] N. Fuhr "Probabilistic models of information retrieval," *Computer Journal, 35*, 3, pp. 244–255, 1992.

[15] D. Gibson, J.M. Kleinberg and P. Raghavan. "Using nonlinear dynamical systems to mine categorical data". Submitted for publication, 1997.

[16] G. Golub and C. Reinsch. Handbook for matrix computation II, Linear Algebra. Springer-Verlag, New York, 1971.

[17] G. H. Golub and C. F. Van Loan. Matrix computations. Johns Hopkins University Press, London, 1989.

[18] W. Hoeffding. Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58** 13–30, 1963.

[19] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998. Also available as IBM Research Report RJ 10076(91892) May 1997.

[20] W. B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space, Contemp. Math. **26** (1984), 189–206.

[21] M. Jerrum and A. Sinclair. Approximating the permanent. *Siam J. Comp.* **18**, pp. 1149-1178, (1989).

[22] P. Raghavan. Information retrieval algorithms: a survey. *Proceedings of the ACM Symposium on Discrete Algorithms*, 1997.

[23] C. J. van Rijsbergen *Information Retrieval* Butterworths, London 1979.

[24] G. Salton and M. McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.

[25] H. R. Turtle and W. B. Croft "A comparison of text retrieval methods," *The Computer Journal, 35*, 3, pp. 279–289, 1992.

[26] J.H. Wilkinson. The Algebraic Eigenvalue Problem. Oxford University Press, London 1965.

## 7 Appendix: Proofs of Lemmas

In the following version of Lemma 1, we take some specific values for some constants to facilitate the proof.

**Lemma 4** *Let $A$ be an $n \times m$ matrix of rank $r$ with singular value decomposition*

$$A = UDV^T,$$

*where $D = \mathsf{diag}(\sigma_1, \ldots, \sigma_r)$. Suppose that, for some $k$, $1 \leq k < r$, $21/20 \geq \sigma_1 \geq \ldots \geq \sigma_k \geq 19/20$ and $1/20 \geq \sigma_{k+1} \geq \ldots \geq \sigma_r$. Let $F$ be an arbitrary $n \times m$ matrix with $\|F\|_2 \leq \epsilon \leq 1/20$. Let $A' = A + F$ and let $U'D'V'^T$ be its singular-value decomposition. Let $U_k$ and $U'_k$ be $n \times k$ matrices consisting of the first $k$ columns of $U$ and $U'$ respectively. Then, $U'_k = U_k R + G$ for some $k \times k$ orthonormal matrix $R$ and some $n \times k$ matrix $G$ with $\|G\|_2 \leq 9\epsilon$.*

The proof of this lemma relies on a theorem of Stewart [17] about perturbing a symmetric matrix.

**Theorem 7** *Suppose $B$ and $B + E$ are $n \times n$ symmetric matrices and*

$$Q = [\underset{k}{Q_1} \quad \underset{n-k}{Q_2}]$$

*is an $n \times n$ orthogonal matrix such that range$(Q_1)$ is an invariant subspace for $B$. Partition the matrices $Q^T BQ$ and $Q^T EQ$ as follows, where $B_{11}$ and $E_{11}$ are $k \times k$ matrices:*

$$Q^T BQ = \begin{bmatrix} B_{11} & 0 \\ 0 & B_{22} \end{bmatrix}$$

$$Q^T EQ = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

*If*

$$\delta = \lambda_{min} - \mu_{max} - \|E_{11}\|_2 - \|E_{22}\|_2 > 0,$$

*where $\lambda_{min}$ is the smallest eigenvalue of $B_{11}$ and $\mu_{max}$ is the largest eigenvalue of $B_{22}$, and $\|E_{12}\|_2 \leq \delta/2$ then there exists an $(n-k) \times k$ real matrix $P$ such that*

$$\|P\|_2 \leq \frac{2}{\delta}\|E_{21}\|_2$$

*and the columns of $Q'_1 = (Q_1 + Q_2 P)(I + P^T P)^{-1/2}$ form an orthonormal basis for a subspace that is invariant for $B + E$.*

*Proof of Lemma 4.* We apply Theorem 7 to $B = AA^T$, $E = A'(A')^T - B$. We choose the block-diagonalizing matrix $Q$ in the theorem to be $U$ followed by $n - r$ zero-columns. Thus, when we write $Q = [Q_1 Q_2]$, $Q_1 = U_k$, the first $k$ columns of $U$, and $Q_2$ consists of remaining columns of $U$ followed by zero-columns. Since $U^T BU$ is a diagonal matrix, $Q^T BQ$ is also a diagonal matrix.

Let $Q^T E Q$ be decomposed into blocks $E_{ij}$, $1 \leq i,j \leq 2$, as in Theorem 7. To apply the theorem, we need to bound $\|E_{i,j}\|_2$. We do this simply by bounding $\|E\|_2$. Since $E = (A+F)(A+F)^T - AA^T = AF^T + FA^T + FF^T$, we have $\|E\|_2 \leq 2\|A\|_2\|F\|_2 + \|F\|_2{}^2 \leq 2(21/20)\epsilon + \epsilon^2 < (43/20)\epsilon$. Therefore, $\|E_{ij}\|_2 \leq (43/20)\epsilon$, $1 \leq i,j \leq 2$. The non-zero eigenvalues of $B$ are $\sigma_1^2, \ldots, \sigma_r^2$. Of these, $\sigma_1^2, \ldots, \sigma_k^2 \geq 361/400$ and $\sigma_{k+1}^2, \ldots, \sigma_r^2 \leq 1/400$. Hence $\delta = \lambda_{min} - \mu_{max} - \|E_{11}\|_2 - \|E_{22}\|_2$ is positive: $\delta > 361/400 - 1/400 - (43/10)\epsilon \geq 137/200$. Also we have $\|E_{12}\|_2 \leq (43/20)\epsilon \leq 43/400 < \delta/2$ and all the assumptions of Theorem 7 are satisfied. It follows that there exists an $((n-k) \times k$ matrix $P$ satisfying $\|P\|_2 \leq \frac{2}{\delta}\|E_{21}\|_2 \leq 7\epsilon$ such that

$$Q_1' = (Q_1 + Q_2 P)(I + P^T P)^{-1/2} \tag{2}$$

forms an orthonormal basis for a subspace that is invariant for $B+E$. This invariant subspace corresponds to the $k$ largest singular values of $A+F$. Therefore, the column vectors of $U_k'$, (the first $k$ eigenvectors of $B+E$) span the same invariant subspace as that spanned by the column vectors of $Q_1'$. In other words, there is a $k \times k$ orthonormal matrix $R$ such that $U_k' = Q_1' R$.

Since $\|Q_1\|_2 \leq 1$, $\|Q_1\|_2 \leq 1$, and

$\|P\|_2 \leq 7\epsilon$, it follows from (2) that $Q_1' = Q_1 + H$ for some $H$ with $\|H\|_2 \leq 9\epsilon$. Therefore, $U_k' = U_k R + HR$, with $\|HR\|_2 \leq 9\epsilon$, as claimed. ■

*Proof of Lemma 3:* The $p^{th}$ eigenvalue of $B$ can be written as

$$\lambda_p^2 = max_{|v|=1} v^T [B - \sum_{j=1}^{p-1} a_j b_j^T]^T [B - \sum_{j=1}^{p-1} a_j b_j^T] v$$

Consider the above expression for $v_1, \ldots, v_k$, the first $k$ eigenvectors of $A$. For the $i^{th}$ eigenvector $v_i$ it can be reduced to

$$v_i^T (B^T B - \sum_{j=1}^{p-1} \lambda_j^2 b_j b_j^T) v_i$$

$$v_i^T B^T B v_i - \sum_{j=1}^{p-1} \lambda_j^2 (b_j \cdot v_i)^2$$

$$\sigma_i^2 |u_i^T R|^2 - \sum_{j=1}^{p-1} \lambda_j^2 (b_j \cdot v_i)^2$$

$$\geq (1-\epsilon)\sigma_i^2 - \sum_{j=1}^{p-1} \lambda_j^2 (b_j \cdot v_i)^2$$

Summing this up for $i = 1, \ldots, k$,

$$\sum_{i=1}^{k} v_i^T B^T B v_i \geq (1-\epsilon) \sum_{i=1}^{k} \sigma_i^2 - \sum_{j=1}^{p-1} \lambda_j^2 \sum_{i=1}^{k} (b_j \cdot v_i)^2$$

Since the $v_i$'s are orthogonal and the $b_j$'s are unit vectors,

$$\geq (1-\epsilon)\sum_{i=1}^{k}\sigma_i^2 - \sum_{j=1}^{p-1}\lambda_j^2$$

Hence

$$\lambda_p^2 \geq max_{v_i} v_i^T B^T B v_i \geq \frac{1}{k}[(1-\epsilon)\sum_{i=1}^{k}\sigma_i^2 - \sum_{j=1}^{p-1}\lambda_j^2]$$

∎