STATISTICAL MODELING TO INFORMATION RETRIEVAL FOR SEARCHING FROM BIG TEXT DATA AND HIGHER ORDER INFERENCE FOR RELIABILITY

XIAOFENG ZHOU

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE
STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

JULY 2014

Abstract

This thesis examined two research projects: probabilistic information retrieval modeling and third-order inference on reliability.

In the first part of this dissertation, two research topics in the information retrieval are carried out and experimented on large-scale text data set. First, we conduct an in-depth study of relationship between information of document length and document relevance to user need. Two statistical methods are proposed which incorporates document length as a substantial weighting factor to achieve higher retrieval performance. Second, we utilize the property of survival function to propose a cost-based re-ranking method to promote ranking diversity for biomedical information retrieval, and to model the proximity between query terms to improve retrieval performance. Through extensive experiments on standard TREC collections, our proposed models perform significantly better than the classical probabilistic information retrieval models.

In the second part of this dissertation, a small sample asymptotic method is proposed for higher order inference in the stress-strength reliability model, $R = P(Y < X)$, where X and Y are independently distributed. A penalized likelihood method is proposed to handle the numerical complications of maximizing the constrained likelihood model. Simulation studies are conducted on two distributions: Burr type X distribution and exponentiated exponential distribution. Results from simulation studies show that the proposed method is very accurate even when the sample sizes are small.

Acknowledgements

I would like to express my sincerest appreciation to my thesis supervisors, Professor Augustine Wong and Professor Jimmy Huang, whose persistent encouragement, patience, guidance and continuous support to lead me into a promising area of research. I am deeply grateful to them, their enthusiasm, inspirations and great efforts to guide and help me to approach research projects throughout my Ph.D study.

I am heartily thankful to my supervisory committee member, Professor Steven Wang, who is always generously offered his knowledge and help and make my doctoral study smooth and rewarding.

A special thank to all my colleagues from the department of mathematics and statistics and information retrieval and knowledge management research lab for their personal support, interesting discussions and valuable suggestions.

Finally, I wish to express my love and gratitude to my husband, my parents and my brother's family for their support, care and understanding. To them I dedicate this dissertation.

Table of Contents

List of Tables

List of Figures

1 Introduction and Literature Survey in Information Retrieval

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. Generally, an IR system receives a query from user and returns the supposedly relevant documents, where a query is a statement of an information need. A crucial issue underlying an IR system is to rank the returned documents by decreasing order of relevance. In most traditional retrieval systems, queries are translated into query representations. Similarly, documents are converted into document representations. Figure 1.1 describe a basic IR system, in which the IR model proposes to match the query representation against the document representation and computes a numeric score on how relevant each document representation satisfies the user's query, and then ranks the documents according to their scores. Finally, the document representations are recovered to be the original documents for users reading.

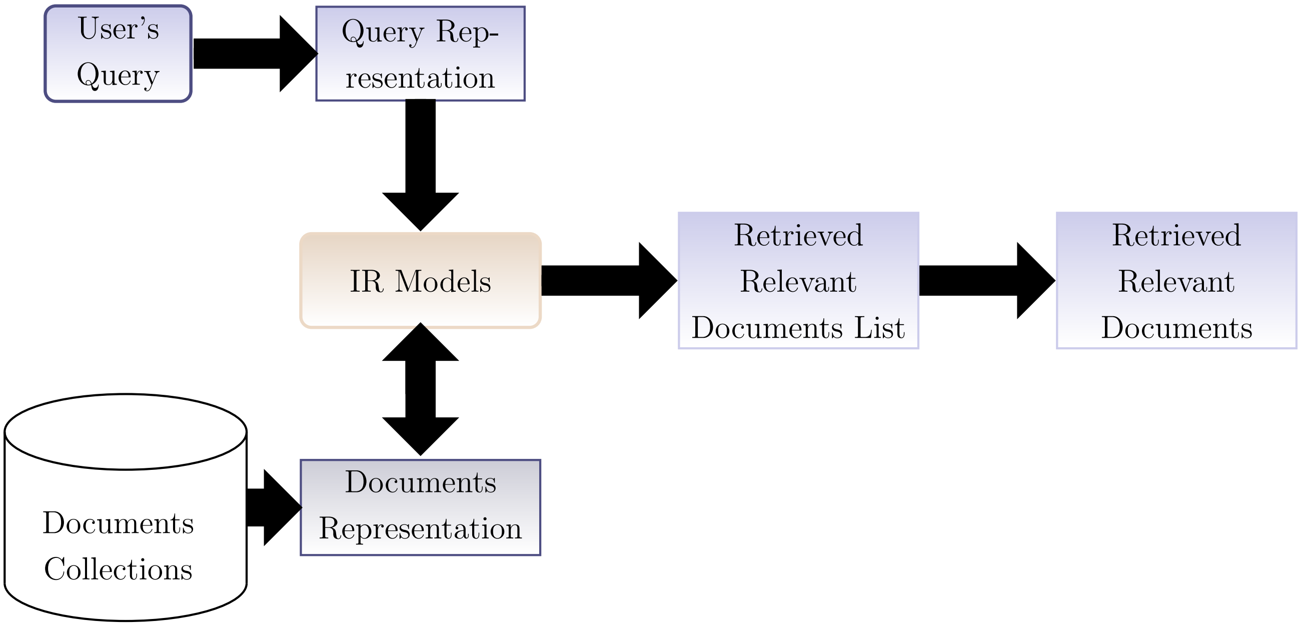


Figure 1.1: A basic IR system.

In general, ranking is based on a weighting model. In Section 1.1, we review the basic probabilistic model which is one of the most popular weighting models in modern IR systems and its development. Retrieved documents are ranked in the order of relevance to the query. A good IR system should return ranked lists that respect both the query-relevance and the breadth of available information. Section 1.2 introduce the most recent research on promoting search result diversification in ranked document lists. Utilizing term proximity to improve retrieval performance is given in Section 1.3. Finally some concluding remarks are given in 1.4.

1.1 Background of Probabilistic Modeling

The probabilistic approach in IR is extensively studied in the literature, this family of IR models is developed by Robertson (1977) from the Probability Ranking Principle(PRP). Cooper (1976) gives the formal statement of PRP as follows:

Probability Ranking Principle : If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

The basic weighting function proposed by Robertson et al. (1981) is one of the most popular weighting models in modern IR systems, and it can be expressed as follows:

$$w(\underline{X}) = \log \frac{P(\underline{X}|R)\ P(\underline{0}|\overline{R})}{P(\underline{X}|\overline{R})\ P(\underline{0}|R)}, \tag{1.1}$$

where $\underline{X}$ is an information vector about the document, $\underline{0}$ is a reference vector representing a zero-weighted document, and R and $\overline{R}$ are relevance and non-relevance respectively. This model considers the independent assumption of the query term within documents made by Robertson and Jones (1976). At this stage, $\underline{X}$ contains the information about term frequency within documents only. So it is so natural to incorporate the document length into this consideration. Robertson and Walker (1994) introduced the document length d into the basic probabilistic weighting model in Equation (1.1) and refined it based on the two hypotheses: Verbosity and Scope hypotheses. Verbosity hypothesis given by Robertson and Walker (1994) states that the document length is independent from its relevance. In other words, long documents simply use more words than short documents to cover similar scope. An opposite assumption about document length is the so-called Scope hypothesis, which states that some documents may contain more material than others if longer, more details can be found at (Robertson and Walker 1994). That is, long documents are more likely to be retrieved. In practice, a document may be considered as a trade-off between the Verbosity hypothesis and the Scope hypothesis. How to balance between these two hypotheses by modeling document length within the basic probabilistic weighting paradigm remains a challenging research issue. The impact of document length on relevance is particularly important for ad-hoc retrieval, where relevance is defined in a binary or graded manner. Compared to a short document, a long document is likely to be relevant if it contains paragraphs that meet the information need of the query, even if a large part of the document is in fact non-relevant. Figure 1.2 elaborates the idea of two hypotheses with the three extreme cases. For example, if the

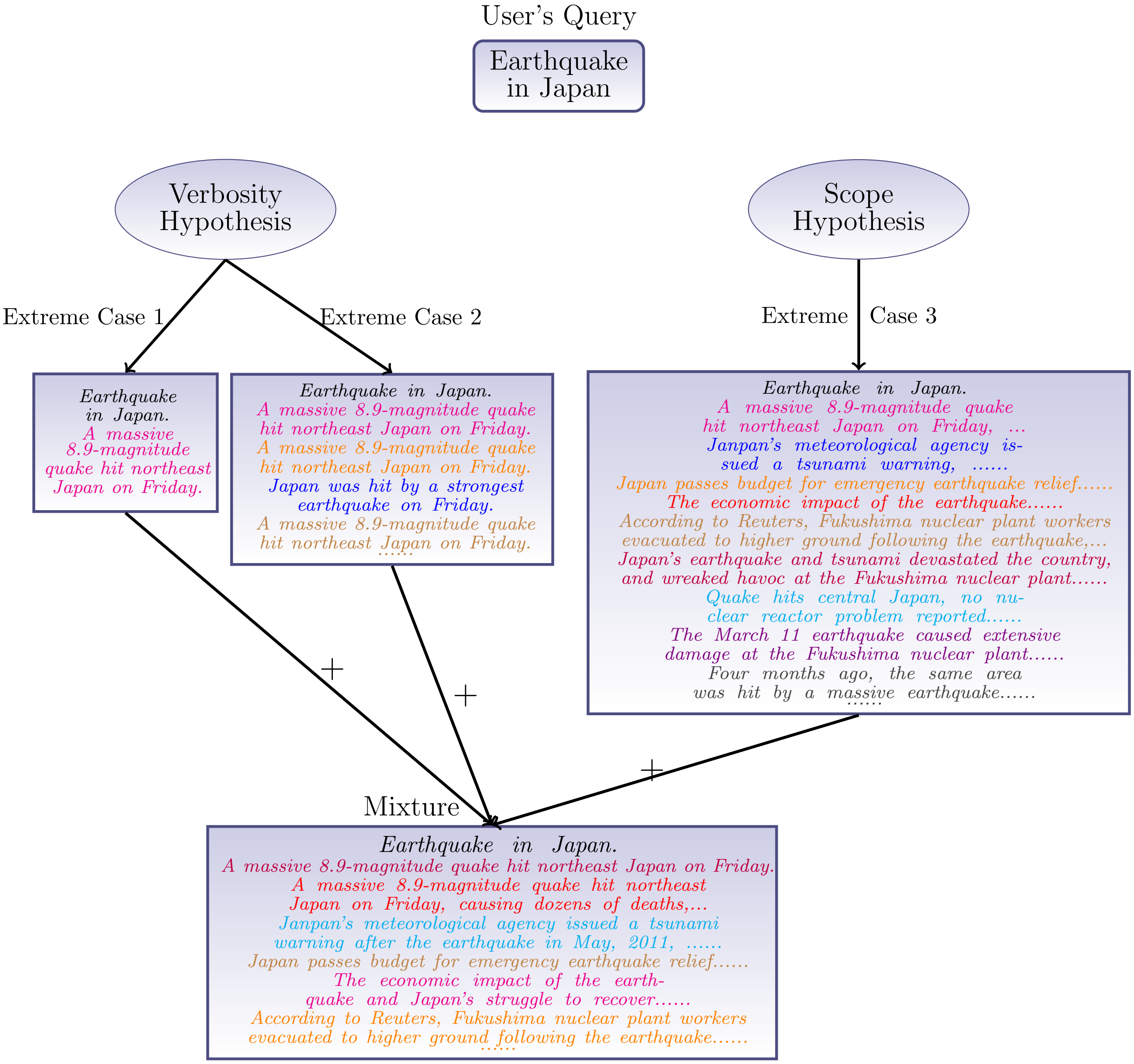

User's Query
Earthquake in Japan
Verbosity Hypothesis
Scope Hypothesis
Extreme Case 1
Extreme Case 2
Extreme Case 3
Earthquake in Japan. A massive 8.9-magnitude quake hit northeast Japan on Friday.
Earthquake in Japan. A massive 8.9-magnitude quake hit northeast Japan on Friday. A massive 8.9-magnitude quake hit northeast Japan on Friday. Japan was hit by a strongest earthquake on Friday. A massive 8.9-magnitude quake hit northeast Japan on Friday. ......
Earthquake in Japan. A massive 8.9-magnitude quake hit northeast Japan on Friday, ... Janpan's meteorological agency issued a tsunami warning, ... Japan passes budget for emergency earthquake relief...... The economic impact of the earthquake...... According to Reuters, Fukushima nuclear plant workers evacuated to higher ground following the earthquake,... Japan's earthquake and tsunami devastated the country, and wreaked havoc at the Fukushima nuclear plant...... Quake hits central Japan, no nuclear reactor problem reported...... The March 11 earthquake caused extensive damage at the Fukushima nuclear plant...... Four months ago, the same area was hit by a massive earthquake...... ......
+
+
+
Mixture
Earthquake in Japan. A massive 8.9-magnitude quake hit northeast Japan on Friday. A massive 8.9-magnitude quake hit northeast Japan on Friday, causing dozens of deaths,... Janpan's meteorological agency issued a tsunami warning after the earthquake in May, 2011, ... Japan passes budget for emergency earthquake relief...... The economic impact of the earthquake and Japan's struggle to recover...... According to Reuters, Fukushima nuclear plant workers evacuated to higher ground following the earthquake...... ......

Figure 1.2: Illustrate two hypotheses using three extreme cases.

user's query is "Earthquake in Japan", we can find a short document with only one sentence "A massive 8.9-magnitude quake hit northeast Japan on Friday"(Extreme Case 1) and a long document with repeated sentences "A massive 8.9-magnitude quake hit northeast Japan on Friday" or similar scope but other words like "Japan was hit by a strongest earthquake on Friday"(Extreme Case 2). Verbosity hypothesis says the document length is independent from its relevant, so these two documents should be weighted the same, but clearly document in extreme case 2 is longer than the one in extreme case 1. Figure 1.2 gives a typical example of Scope hypothesis, extreme case 3. A document with a lot of more and different information related to user's query: Earthquake in Japan but has same document length as extreme case 2. It may repeated same piece of information several times, but it do contain more information than the previous two cases. Scope hypothesis says that we should weight it more than the previous two cases, but Verbosity hypothesis says opposite. Robertson and Walker (1994) recite that real document collections contains a mixture of effects from two hypothesis, and the individual document may be at either extreme case or the combination. Normally, in a collection of a documents, the document length is varying, may from couple of words to thousands, or millions words. One may think the more likely the document contain the query term if the document length is more longer. It has been recognized as an important factor for adjusting the IR system to avoid the length bias, but the impact of document length on relevance remains challenging.

To address the effect of document length on relevance, the basic probabilistic weighting function with document length been taken into account by Robertson and Walker (1994) becomes

$$w(\underline{X}, d) = \log \frac{P(\underline{X}, d|R)}{P(\underline{X}, d|\overline{R})} \frac{P(\underline{0}, \Delta|\overline{R})}{P(\underline{0}, \Delta|R)}, \tag{1.2}$$

where $w(\underline{X}, d)$ is the relevance weight of a given document. d is the document evidence for relevance, which is given by document length. Δ denotes the average document length of the reference vector $\underline{0}$, and $\underline{X}$ represents all other information about the document. $\overline{R}$ and R stand for the non-relevance and relevance, respectively. This function measures the difference between the probabilities of document length and all other information we have for the document when it is relevant and when it is not relevant, respectively, in log scale. The above equation also implies a relevant document should receive a higher weight than a non-relevant document in order to achieve a satisfying retrieval performance.

Equation (1.2) can be further decomposed into the three components as follows Robertson and Walker (1994):

$$w(\underline{X}, d) = w(\underline{X}, d)_1 + w(\underline{0}, d)_{21} + w(d, \Delta)_{22}, \tag{1.3}$$

where

$$w(\underline{X}, d)_1 = \log \frac{P(\underline{X}, d|R)}{P(\underline{X}, d|\overline{R})} \frac{P(\underline{0}, d|\overline{R})}{P(\underline{0}, d|R)},$$

$$w(\underline{0}, d)_{21} = \log \frac{P(\underline{0}|d, R)}{P(\underline{0}|d, \overline{R})} \frac{P(\underline{0}|\Delta, \overline{R})}{P(\underline{0}|\Delta, R)},$$

and

$$w(d, \Delta)_{22} = \log \frac{P(d|R)P(\Delta|\overline{R})}{P(d|\overline{R})P(\Delta|R)}.$$

Under the Verbosity hypothesis, document length has been considered as independent evidence of relevance. This hypothesis nullifies the component $w(d, \Delta)_{22}$ in Equation (1.3), which as a consequence is set to zero in (Robertson and Walker 1994). Thus, the weighting function becomes

$$w(\underline{X}, d) = w(\underline{X}, d)_1 + w(\underline{0}, d)_{21}. \tag{1.4}$$

The classical BM25 weighting model derived by Hancock-Beaulieu et al. (1996) from Equation (1.4), more specifically, $w_{BM25} = w(\underline{X}, d)$, where w_{BM25} is the relevance score of BM25, given by the following weighting function

$$w_{BM25} = \frac{(k_1 + 1)tf}{K + tf} \cdot w^{(1)} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf} \oplus L, \tag{1.5}$$

where

$$K = k_1 \cdot \left((1 - b) + b \cdot \frac{dl}{avdl}\right),$$

$$w^{(1)} = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)},$$

$$L = k_2 \cdot nq \cdot \frac{avdl - dl}{avdl + dl}.$$

N is the number of indexed documents in the collection, n is the number of documents containing the query term, R is the number of known relevant documents to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the document length (i.e. the document evidence d in Equation (1.3)), $avdl$ is the average document length, nq is the number of query terms, k_i's and b are tuning constants (whose setting depends on the dataset used and is usually empirically determined), and $\oplus$ indicates that its following component is added only once per document. Particularly, b functions as a justification factor that adjusts the relative importance between the two hypotheses given by Robertson et al. (1996).

The classical probabilistic models for IR rank documents according to their relevance scores, assigned by matching the query terms with adjustment for the relationship between document length and term frequency. This approach is developed based on the Verbosity hypothesis which assumes the document's relevance is independent of its length. However, in practice, the impact of document length on relevance may be a mixture of both the Scope hypothesis and the Verbosity hypothesis brought by Robertson and Walker (1994). Many previous studies have been conducted to investigate the impact of document length on relevance. Singhal et al. (1996a) suggested that long documents tend to have more unique terms, and consequently, long documents have a better chance to be retrieved than short documents. As the document length increases, the number of times the query terms occur in the documents also increases, which in turn increases the matching score. For instance, Singhal et al. (1996a,b) illustrated that the probability of a document's relevance increases proportionally with document length in the early TREC test collections and showed that better retrieval performance can be achieved with normalization techniques. They also reported that documents retrieved by a model produce a retrieval pattern by the distribution of the document length. Moreover, a number of empirical studies have provided statistical evidence supporting that the probability of a document's relevance to an information need is considered to be correlated with the length of the document. Kraaij et al. (2002) showed that the probability of relevance is positively correlated with document length on a number of TREC ad-hoc and Web collections. Furthermore, Voorhees et al. (2005) found that proper term weighting strategies based on document length can also improve retrieval performance. For example, normalization techniques have been applied for each term in the query through the length adjustment to avoid the bias introduced by document length. Losada and Azzopardi (2008) applied Jelinek-Mercer and Dirichlet prior to two-stage smoothing strategies on document length in language modeling to show the significant impact of document length on the information retrieval performance.

1.2 Diversification in Biomedical Search

Traditional retrieval models assume that the relevance of a document is independent of the relevance of other documents as stated in Section 1.1. However, in reality, this assumption may not hold. This assumption may result in high redundancy and low diversification in a ranked document list, since documents that are similar in content tend to appear over and over again. That is novel information is needed.

The usefulness of retrieving a document usually depends on previous ranked documents, since a user may want to see the top ranked documents concerning different aspects of his/her information need instead of reading relevant documents that only deliver redundant information. A better information retrieval system thus should return ranked lists that represent both the query-relevance and the breadth of available information. For example, in biomedical domain, the desired information of a question (query) asked by biologists usually is a list of a certain type of entities covering different aspects that are related to the question, such as genes, proteins, diseases, mutations, etc. Hence, it is important for a biomedical IR system to be able to provide comprehensive and diverse answers to fulfill biologists' information needs. The dissimilarity between documents has to be considered. In the TREC 2006 and 2007 Genomics tracks, the "aspect retrieval" was investigated. Its purpose was to study how a biomedical IR system can support a user gather information about the different aspects of a topic. In the Genomics tracks, biomedical IR systems were required to return relevant information at the passage level, while relevance judges not only rated the passages, but also grouped them by aspect. Aspects of a retrieved passage could be a list of named entities or MeSH terms, representing answers that cover different portions of a full answer to the query. Aspect Mean Average Precision (Aspect MAP) was defined in the Genomics tracks to capture similarities and differences among retrieved passages. It indicates how comprehensive the questions are answered. Hersh et al. (2006a) found that relevant passages that do not contribute any new aspects to the aspects retrieved by higher ranked passages will not be used to accumulate Aspect MAP. Therefore, Aspect MAP is a measurement for redundancy and diversity of the IR ranked list.

Our work is inspired by several recent papers that concerned with promoting diversity and novelty in the IR ranked list. Carbonell and Goldstein (1998) introduced the maximal marginal relevance (MMR) method, which attempted to maximize relevance while minimizing similarity to higher ranked documents. In order to measure the redundancy between documents, Zhang et al. (2002) presented four redundancy measures, which were "set difference", "geometric distance", "distributional similarity" and "a mixture model". They modeled relevance and redundancy separately. Since they focused on redundant document filtering, experiments in their study were conducted on a set of relevant documents. However, in reality, non-relevant documents are always returned by IR systems along with relevant documents. Redundancy and relevance should both be considered. Zhai et al. (2003) validated a subtopic retrieval method based on a risk minimization framework. Their subtopic retrieval method combined the mixture model novelty measure with the query likelihood relevance ranking. More recently, a new diversity task of Web retrieval was defined in the TREC 2009 Web track by Clarke et al. (2009a). Two evaluation measures, α-nDCG by Clarke et al. (2008a) and an intent-aware version of precision (IA-P) by Agrawal et al. (2009a), both of which reward novelty and diversity, were validated in the diversity task of the 2009 Web track. Top diversity task results showed that re-ranking methods based on anchor text, sites of search results, link filtering, clustering and sub-queries suggestion were effective in Web retrieval result diversification. More details can be found at(Song et al. 2004, Craswell et al. 2009, Kaptein et al. 2009 and McCreadie et al. 2009). Santos et al. (2010b) proposed a novel framework, namely xQuAD, for search result diversification that builds such a diversified ranking by explicitly accounting for the relationship between documents retrieved for the original query and the possible aspects underlying this query, in the form of sub-queries. Their study showed that the sub-query generation step plays a fundamental role in the method. When they use the ground-truth sub-topics provided by the collection of TREC Web track as input to their proposed diversification models, performance improvements can be obtained. However,when they use a clustering-based query expansion technique, in an attempt to uncover terms representative of different aspects underlying a query from a clustering of the top retrieved results for the query, no consistent performance improvements can be observed. Santos et al. (2010a) also used three major web search engines (WSE) to generate the sub-queries and investigated the impact of sub-query importance. In general, using the related or suggested queries from WSEs, improvements over the initial ranking are obtained in most settings.

In biomedical information retrieval, the Genomics aspect retrieval was firstly proposed in the TREC 2006 Genomics track and further investigated in the 2007 Genomics track. Many research groups joined these annual campaigns to evaluate their systems and methodologies. However, to the best of our knowledge, there is not too much previous work conducted on the Genomics aspect retrieval for promoting diversity in the ranked list.

1.3 Term Proximity

In literature, term proximity is interpreted as query term co-occurrences, or phrases. Proximity searches for documents where two or more separately matching term occurrences are within a specified distance, where distance is the number of intermediate words or characters. Proximity among query terms has been found to be useful for improving retrieval performance. The intuition behind the development of the proximity-based models is that documents in which query terms co-occur in a close proximity, or within the same phrases, tend to be highly relevant.

Term proximity is particularly useful in Web environment, where documents are highly diverse and heterogeneous. Indeed, some previous studies have demonstrated the effectiveness of using term proximity for improving the retrieval performance. Such research has resulted in the development of the so-called n-gram models, in contrast to the classical "unigram" models which assume term independence. An n-gram is referred to as a subsequence of n terms from a given sequence, or a given window, of terms. In addition, the n-gram terms are not necessarily adjacent to each other. For example, if an n-gram "modeling proximity" appears in "Modeling term proximity for probabilistic models", the two n-gram terms have a distance of 1 since "*term*" appears between "*modeling*" and "*proximity*" in the text.

In the past decades, there have been attempts to develop more sophisticated IR models by employing the term proximity information. Fagan (1987) empirically identified phrases using features such as the frequency of the phrase in the collection and the proximity of the phrase terms, and concludes that the tested phrase methods do not perform consistently well across a variety of collections. Croft et al. (1991) automatically extracted phrases from natural language queries to form structured queries for a probabilistic ranking model. van Rijsbergen (1977) proposed a theoretical model for incorporating term co-occurrence information to a binary independence model. Losee and Jr. (1994) applied the Bahadur Lazarsfeld expansion to identify term dependence between more than two terms. In addition, Yu et al. (1983) generalized both the tree dependence model and the Bahadur Lazarsfeld expansion. Although the two methods are of high complexity in theory, their evaluation showed little improvement over a unigram model argued by Metzler and Croft (2005a).

Recently, the application of term proximity has achieved a certain degree of success in the context of the language modeling approach. Gao et al. (2004) incorporated the term dependence based on a link structure for each query, which could be time-consuming in practice. Metzler and Croft (2005a) proposed a Markov Random Field model to estimate term dependencies in the context of language modeling approach. A similar idea was presented by Mishne and de Rijke (2005). Lv and Zhai (2009) applied a list of kernel functions to estimate a language model for every position in a document. Zhao and Yun (2009) constructed the query term proximity as the Dirichlet hyper-parameter that weights the parameters of the unigram document language model. The application of term proximity to the probabilistic models has little success so far. Plachouras and Ounis (2007a) conducted an initial study on incorporating term proximity in the Divergence From Randomness models. They assume a Poisson distribution of n-gram co-occurrence in the elite set, i.e., the set of retrieved documents. However, the approach proposed by Plachouras and Ounis (2007a) only leads to a moderate improvement over the unigram baseline, possibly due to the lack of the collection model for the n-gram frequency distribution. Bttcher and Clarke (2005), Rasolofo and Savoy (2003) also attempted to incorporate the term proximity into the classical BM25 model, which define the bi-gram co-occurrences as an inverse function of the square distance between two query terms. However, the retrieval performance reported by Bttcher and Clarke (2005), Rasolofo and Savoy (2003) is not as good as expected, possibly due to the naive bi-gram occurrence function.

1.4 Summary

I briefly reviewed the basic weighting model in this Chapter. With its recent development in IR field, my research project focus on three parts. In Chapter 2, I study the impact of document length on its relevance in the context of the Scope hypothesis. In Chapter 3, I will show that retrieval performance can be improved by applying survival function on aspect and term proximity. Finally, in Chapter 4, I will discuss possible future direction of this research.

2 Impact of Document Length on Relevance

The main focus of this research project is to study the impact of the document length on its relevance. The Scope hypothesis presented by Robertson and Walker (1994) suggests the existence of a relationship between document length and relevance. It implies that the component $w(d, \Delta)_{22}$ in Equation (1.3) may not be zero. In this chapter, we consider document length itself as a direct predictor of relevance. The study of document length is based on the intuition that long documents tend to have high retrieval probabilities, since long documents usually have a large number of unique terms, which are likely to be picked up by the query term matching Singhal et al. (1996a). In Section 2.1, we study the pattern of the document length and its relevance by exploring a list of probability density distribution investigate the behavior of $w(d, \Delta)_{22}$ in Equation (1.3), a new weighting function incorporating this relationship proposed in Section 2.2, which is the result of a mixture of the two hypotheses. Our results presented in Section 2.3 show that the retrieval performance of BM25 can be markedly improved over different settings of the parameter b in

Equation (1.5) by exploiting the document length evidence. More details can be founded in Zhou et al. (2011).

2.1 Density Analysis and Length Relevance Weighting

Under the Scope hypothesis, $w(d,\Delta)_{22}$ in Equation (1.3) is no longer zero since a dependence of relevance on document length is assumed. To add the length information into the weighting function $w(x,d)$, we decompose the $w(d,\Delta)_{22}$ further into

$$w(d,\Delta)_{22} = \log\frac{P(d|R)}{P(d|\overline{R})} + \log\frac{P(\Delta|\overline{R})}{P(\Delta|R)}. \tag{2.1}$$

The second component of Equation (2.1) is constant over a given document collection. This is because the average document length Δ for the reference vector $\underline{0}$ in a document collection is known and fixed. Therefore, for each document in a collection, the second component of Equation (2.1) above is the same across the whole document collection and does not affect the document ranking. For simplicity, we refer $w(d,\Delta)_{22}$ to as the first component in the Equation (2.1). Thus, the relevance weight $w(d,\Delta)_{22}$ is given by the log odds ratio of the relevance and non-relevance probabilities $P(d|R)$ and $P(d|\overline{R})$. In other words, $w(d,\Delta)_{22}$ measures the difference between the probabilities of given document length condition on relevance and non-relevance in log scale. We name Equation (2.1) as *length relevance weighting.* Our ultimate goal is to calculate the $w(d,\Delta)_{22}$ in Equation (2.1), this needs a way to estimate the probabilities $P(d|R)$ and $P(d|\overline{R})$. By adding the measurement of document length itself into the basic weighting function, the retrieval system is expected to achieve high accuracy since the length information brings more evidence of relevance. The estimation of probability distribution function [1] of document length will be discussed in the next subsections.

For the rest of this chapter, we use $\mathbf{d}$ to denote the document length. As a general rule, we usually make an assumption about observed d's, *i.e.* $d_1, d_2, ..., d_N$ are independent and identically distributed, N is the number of documents in the collection. For example, N is 741,856 in test collection disk1&2. The proposed method will be examined on the four standard TREC test collections: disk1&2, WT10G, .GOV2 and ClueWeb B. The detailed introduction on four test collections will be given in the Section 2.3.

2.1.1 Kernel Density Analysis

Kernel density estimation (or Parzen window method) is a non-parametric way of estimating the probability density function of a random variable. It can be used to

[1] Probability distribution function is a general concept in probability theory, it could refer to probability mass function on discrete random variable, probability density function on continuous random variable or cumulative distribution function based on context. So we use probability distribution function and probability density function interchangeable in the rest of this paper.

extrapolate data to the entire population. In particular, Silverman (1986) defines

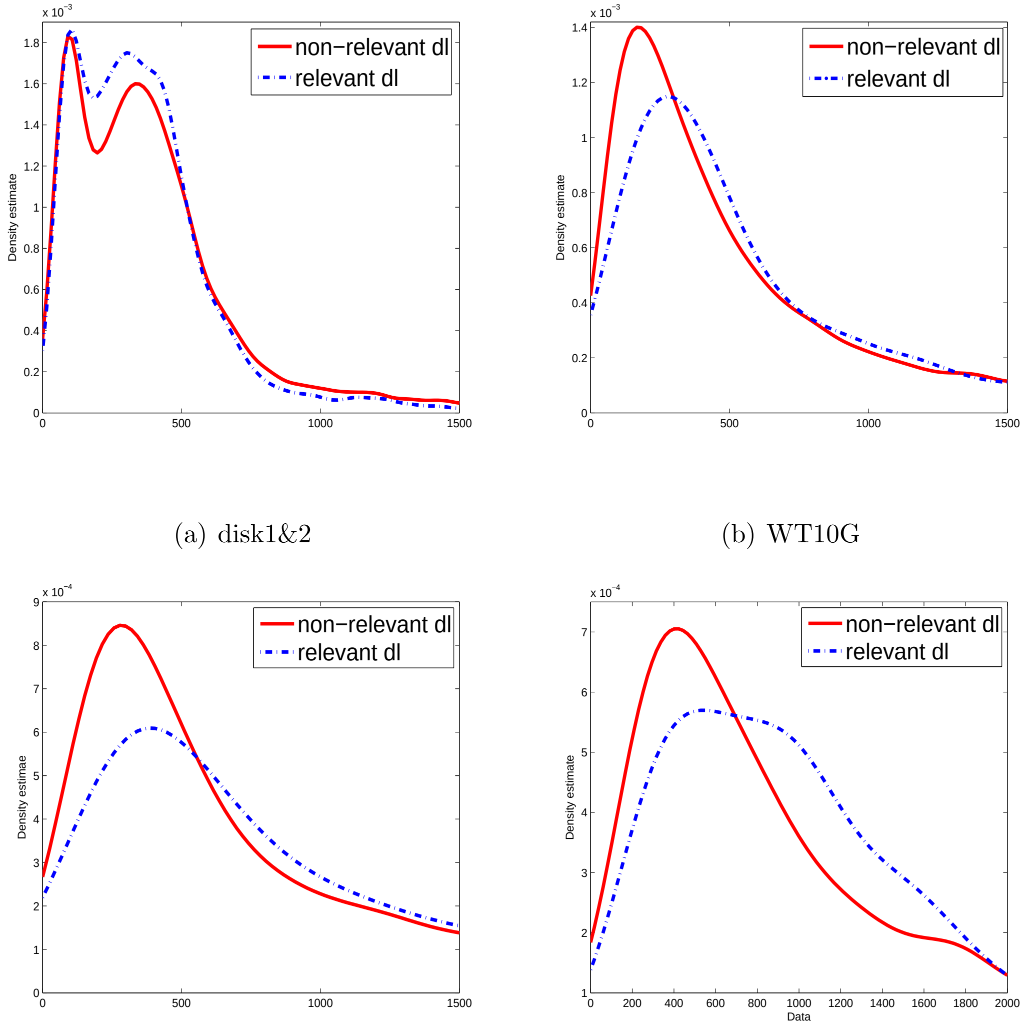
$$\widehat{f_h}(d) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{d - d_i}{h}\right),$$

where $d_i, i = 1, \ldots, n$ is the independent and identically-distributed sample from some unknown distribution, n is the number of samples we draw from the population, K is the kernel function and h is the bandwidth (also known as the smoothing parameter). We can obtain the smoothing curve by adjusting the parameter h. Usually K is a standard Gaussian function with a mean of zero and a variance of 1:

$$K\left(\frac{d - d_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(d - d_i)^2}{2h^2}\right).$$

Kernel density estimation gives us a global picture of the given dataset.

Figure 2.1 shows the distributional pattern of relevant and non-relevant document length for the test collections disk1&2, .GOV2, WT10G and ClueWeb B respectively using kernel density estimation with the standard normal kernel function. In Figure 2.1, the length curves of first three subfigures have been substract when the document length is large than 1500 in $x-$axis in order to visualize the difference between the relevant and non-relevant document length because the long documents have low frequencies to be emerged in test collections. The length of non-relevant and relevant documents from four test collections are both positively skewed and have a long tail with different tail shapes. Non-relevant document has higher frequency



(a) disk1&2

(b) WT10G

(c) .GOV2

(d) ClueWeb B

Figure 2.1: Kernel density estimate constructed from the document length on four test collections. “dl” stands for document length.

than relevant documents when the document length is relatively short. The curve on disk1&2 appears to have two distinct peaks called bimodality. In contrast, there is only one mode that arises on WT10G, .GOV2 and ClueWeb B.

2.1.2 Mixture Model

The mixture model is a probabilistic model for representing the presence of more than one sub-populations within an overall population. Gaussian Mixture Models (GMMs) are among the most statistically mature methods for density estimation. A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by

$$f(\mathbf{d}|\mu_1, \ldots, \mu_M, \Sigma_1, \ldots, \Sigma_M, \lambda_1, \ldots, \lambda_M) = \sum_{m=1}^{M} \lambda_m \, g(\mathbf{d}|\mu_m, \Sigma_m),$$

where λ_m is the weight of each component and $\sum_{m=1}^{M} \lambda_m = 1$, μ_m, Σ_m are the parameters of Gaussian distribution g_m which takes form of

$$g(\mathbf{d}|\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{V/2} |\Sigma_m|^{1/2}} \, \exp\left\{ -\frac{1}{2} (\mathbf{d} - \mu_m)' \, \Sigma_m^{-1} \, (\mathbf{d} - \mu_m) \right\}.$$

Each Gaussian has its own mean μ_m and covariance matrix Σ_m, V is the dimension of $\mathbf{d}$. For mixture model, data within same group assumed be homogeneous tends to tight, data between groups are heterogeneous. All component Gaussian are acting together to model the overall feature density, so mixture model covers the data well.

The GMMs not only provide a smooth overall distribution fit, its components also clearly detail the multi-modal nature of the density. Figure 2.1 exhibits that there are two modes on disk1&2. The parameters of GMMs can be estimated by the maximum likelihood (ML) criterion using the iterative Expectation-Maximization (EM) algorithm.

2.1.3 Data Transformation

The data transformation technique is widely used in data processing or pre-processing for stabilizing the variance and make the data more normal distribution-like. In our case, all document lengths are positive, whose distribution is skewed to the right as described in Figure 2.1, and document length cannot be described by standard statistical methods because of the skewness. Therefore, data transformation is required to extract a better characteristic of the data. We start with standardization: the transformation I. By standardizing the data, it forces the data to locate on the common scales to be compared. Secondly, the Box-Cox transformation is from the family of functions that are applied to create a rank-preserving transformation of data which improves the correlation between variables and for other data stabilization procedures Box and Cox (1964). Box-Cox transformation is commonly used to alleviate heteroscedasticity when the distribution of the variable of interest is not known, *i.e* transformations II and III. We also transform the document length to be within the scale of $0 - 1$ using transformation IV. The four types of transformation is described as follows:

- Transformation I: Standardization

$$z = \frac{d - \bar{d}}{s_d} \tag{2.2}$$

- Transformation II: Log transformation. It is the special case of following transformation III when $\theta = 0$.

$$z = \log(d) \tag{2.3}$$

- Transformation III: Box-Cox transformation

$$z = \frac{d^\theta - 1}{\theta} \tag{2.4}$$

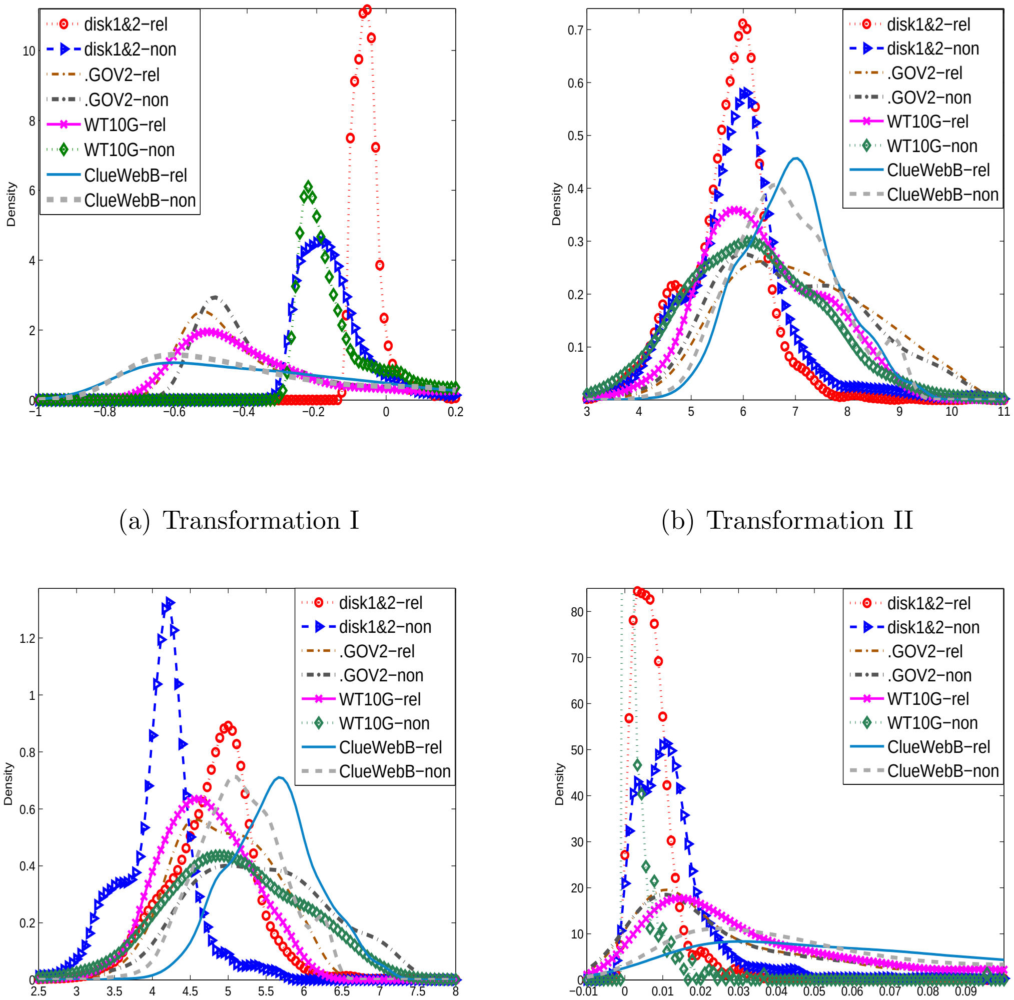
  Box-cox transformation is a parametric power transformation technique in order to reduce anomalies such as non-additivity, non-normality and heteroscedasticity, $\theta \neq 0$ is the transformation power.

- Transformation IV: $0 - 1$ normalization

$$z = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{2.5}$$

where z is the document length after the transformation, $\bar{d}$ is the average document length, s_d is the standard deviation of document length, d_{min} denotes the minimum document length and d_{max} is the maximum document length.

Figure 2.2 plots the distributional pattern of transformed relevant and non-relevant document length on four test collections. In Figure 2.2, we examine length distribution patterns of relevant and non-relevant documents on the four test collections used. A major observation is that the the curves of the transformed document length distribution have similar shapes before and after the transformation. That is, the curves on disk1&2 remain bimodal, while the curves on the other three test collections are still left-skewed, but not as much as those of the original document length distribution, thanks to the data transformation. Moreover, the curves on .GOV2 and ClueWeb B become more symmetric after the transformation. On disk1&2, .GOV2 and WT10g, the center of the non-relevant document length distribution shifts far away to the right of the relevant document length distribution. From Figure 2.2(d), we can clearly see that relevant and non-relevant document length on Clueweb B can be distinguished from their distributional frequencies. Note that similar observations can also be drawn from other three collections used, but the difference between relevant and non-relevance document length distribution is not as obvious as on ClueWeb B. This is an encouraging finding as it gives us clue of differentiating



(a) Transformation I

(b) Transformation II

(c) Transformation III

(d) Transformation IV

Figure 2.2: Kernel density estimates constructed from the transformed document length of four test collections. ”rel” stands relevant document length and ”non” stands non-relevant document length.

between relevance and non-relevant documents based on their length distribution. In the next subsection, we propose to fit the document length distribution with a list of statistical distributions, in order to find the distributions that can match the characteristics of relevance and non-relevant documents.

2.1.4 Distribution of Document Length

The criterion of selecting distributions is that the distribution must be positive skewed with shape and rate parameters. It is impossible to use one single distribution to capture all kinds features of all different document collections. The commonly used distributions we applied to fit the transformed document length are as follows:

- Gamma distribution with $(\gamma > 0, \beta > 0)$

  $$f(y) = \frac{y^{\gamma-1}e^{-y/\beta}}{\beta^{\gamma}\Gamma(\gamma)}$$

  for $y \geq 0$, where γ and β are shape and scale parameters respectively. Varying setting of γ can lead to symmetrical or skewed figures.

- Normal distribution with (μ, σ^2), which is symmetric with respect to its mean value (μ), and the variance (σ^2) measures the width of the distribution. A Normal distribution is bell shaped and the shape is independent of its distribution parameters. The reason of choosing normal distribution is that transformation

I try to standardize the document length. The Normal distribution density function is given as follows:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

where σ is the standard deviation.

- Lognormal distribution with (μ, σ^2)

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}}\, exp\left\{-\frac{(\log y-\mu)^2}{2\sigma^2}\right\}$$

for $y \geq 0$, where if Z is distributed lognormally with parameters μ and σ, $\log(Y)$ is distributed normally with a mean of μ and a standard deviation of σ. Lognormal and gamma distribution can produce similar graphs, but the curvature of lognormal distribution is more steep than gamma distribution.

- Inverse Gaussian distribution (IGD) with (μ, λ)

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} exp\left\{-\frac{\lambda}{2\mu^2 y}(y-\mu)^2\right\}$$

for $y > 0$, where $\mu > 0$ is the mean and $\lambda > 0$ is the shape parameter, changing λ changes the level of the skewness for the IGD.

- Weibull distribution with (a, b)

$$f(y) = \frac{b}{a}\left(\frac{y}{a}\right)^{b-1} exp\left\{-(\frac{y}{a})^b\right\}$$

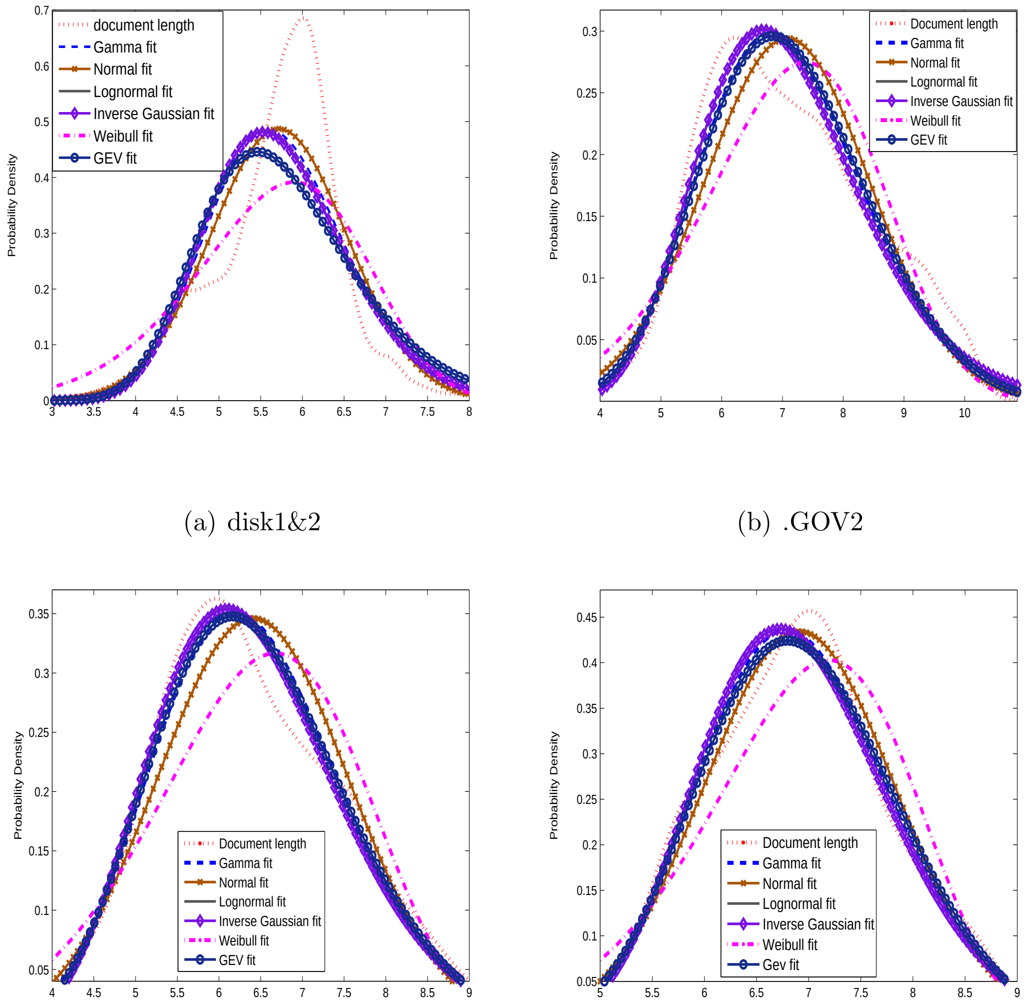
where $a > 0$ is the scale parameter and $b > 0$ is the shape parameter. Weibull distribution can produce the graph similar to Gamma distribution but with less steep curve.

- Generalized Extreme Value distribution (GEV) (κ, μ, σ),

$$f(y) = \frac{1}{\sigma} exp \left\{ -(1 + \kappa \frac{(y-\mu)}{\sigma})^{-\frac{1}{\kappa}} \right\} \left(1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{(-1-\frac{1}{\kappa})}$$

  where $\kappa \neq 0$ is the shape parameter, μ is the location parameter and $\sigma > 0$ is the scale parameter. Compared to the statistical distributions mentioned above, GEV is a complicated distribution developed within the extreme value theory introduced by Gumbel (1958).

Figure 2.3 illustrates the six distribution fittings for the relevant document length of four test collections using Transformation I. Similar plots can be obtained for the relevant and non-relevant document length of all four test collections using Transformation I, Transformation III and Transformation IV respectively. All six distributions fit the .GOV2, WT10G and ClueWebB well, not disk1&2 since the bimodality. Inverse Gaussian and GEV distribution fit the data best on all test collections, Weibull distribution can preserve the skewness better than the Lognormal, Gamma distribution, but normal distribution performs very badly in this case since skewness of the data. After the density functions are fit to the actual length distribution, it



(a) disk1&2

(b) .GOV2

(c) WT10G

(d) ClueWebB

Figure 2.3: Distribution fittings for the relevant document length using transformation two $z = log(d)$.

is necessary to use goodness of fit test to determine how well the distributions fit to the actual data.

Bootstrapping is a resampling method to learn about the sample characteristics to infer the population. It has been proved effective in reducing the bias of samples by Gentle (2002). Adèr et al. (2011) recommend to use bootstrapping when the sample size is insufficient for straightforward statistical inference. The bootstrapping procedure is described as follows:

1. Construct an empirical probability distribution Ω from the sample by placing a probability of $1/n$ at each point, $z_1, z_2, \ldots, z_n$ of the sample. This is the empirical distribution function of the sample, which is the nonparametric maximum likelihood estimate of the population distribution, ω. Now, each sample's element has the same probability of being drawn.

2. From the empirical distribution function, Ω, draw a random sample of size n with replacement. This step is called resampling.

3. Calculate the statistic of interest, θ, for this resample, yielding $\hat{\theta}^*$.

4. Repeat steps 2 and 3 for B times, where B is a large number, in order to create B resamples. The setting of B depends on the tests to be run on the data.

5. Compute $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}^*_j$.

2.2 Length Relevance Weighting

Based on the discussion above, we first derive the length relevance weighting function using Equation (1.3) for disk1&2 when the distribution function obtained from mixture model. We named it 'Mixture'.

$$w(d,\Delta)_{22} = \log \frac{\frac{\lambda_{R1}}{\sigma_{R1}}\, exp\left\{-\frac{(d-\mu_{R1})^2}{2\sigma_{R1}^2}\right\} + \frac{\lambda_{R2}}{\sigma_{R2}}\, exp\left\{-\frac{(d-\mu_{R2})^2}{2\sigma_{R2}^2}\right\}}{\frac{\lambda_{NR1}}{\sigma_{NR1}}\, exp\left\{-\frac{(d-\mu_{NR1})^2}{2\sigma_{NR1}^2}\right\} + \frac{\lambda_{NR2}}{\sigma_{NR2}}\, exp\left\{-\frac{(d-\mu_{NR2})^2}{2\sigma_{NR2}^2}\right\}} \tag{2.6}$$

The subscript $R1, R2, NR1, NR2$ represent the relevant and non-relevant for Gaussian component 1 and 2 respectively, λ, μ, σ are the parameters of GMMs.

Second, when the distribution function for transformed document length Z is obtained, we apply variable change technique in Shao (2003) to obtain the distribution function for the original document length D, *i.e.* the document length before transformation. The theorem of change variable we used is as follows

Theorem 2.2.1 *Let Z be a random variable with probability density function (pdf) $f_Z(z)$ and support $\mathcal{S}_Z$. Let $D = g(Z)$, where $g(z)$ is one to one differentiable function, on the support of Z, $\mathcal{S}_Z$. Denote the inverse of g by $z = g^{-1}(d)$ and let $\frac{\partial z}{\partial d} = \frac{\partial[g^{-1}(d)]}{\partial d}$. Then the pdf of D is given by*

$$f_D(d) \;=\; f_Z(z^{-1}(d)) \left|\frac{\partial z}{\partial d}\right|, \quad \text{for } d \in \mathcal{S}_D. \tag{2.7}$$

with the support of D which is the set $\mathcal{S}_D = \{d = g(z) : z \in \mathcal{S}_Z\}$.

Where $z^{-1}(d)$ is equivalent to Equations (2.2), (2.3), (2.4) and (2.5) when transforming the document length, and $|\frac{\partial z}{\partial d}|$ is the determinant of Jacobian of the transformation in Hogg et al. (2004).

Based on the discussion above, we initiate the pattern of the document length by kernel density estimation. Based on the findings in step one, second, we apply data transformation and the change variable techniques to find the distribution functions of relevant and non-relevant document length and use maximum likelihood estimation(MLE) to obtain the parameter estimators. Two statistical methods, EM and bootstrapping are exploited to prevent potential bias during parameter estimation, such as incomplete test collection, randomness of sampling. Hypothesis test employees to eliminate the distributions at 5% significance. Finally Equation (1.3) is used to construct the following seven models:

1. In this model, with standardization transformation I in Equation (2.2), transformed relevant and non-relevant document length follow the Normal distribution. Using Equation (1.3) and change of variable technique in Equation (2.7), the length relevance weighting function, named 'Normal', is as follows

$$w(d, \Delta)_{22} \propto -\frac{1}{2\sigma_1^2}\left[\left(\frac{d-\bar{d}}{s_d}\right) - \mu_1\right]^2 + \frac{1}{2\sigma_2^2}\left[\left(\frac{d-\bar{d}}{s_d}\right) - \mu_2\right]^2$$

where subscript 1 indicates that it is the estimates of distribution for relevant document length, 2 is the estimates of non-relevant document length distribution.

2. In this model, with log transformation II in Equation (2.3), transformed relevant and non-relevant document length follow the Gamma distribution. Similarly, the length relevance weighting function, named 'Log-Gamma', is as follows

$$w(d)_{22} \propto (\gamma_1 - \gamma_2)\log d - \frac{\log d}{\beta_1} + \frac{\log d}{\beta_2}$$

3. In this model, with log transformation, transformed relevant and non-relevant document length follow the IGD distribution. We call the length relevance weighting function 'Log-IGD'.

$$\begin{aligned} w(d)_{22} \propto \; & \log\sqrt{\frac{\lambda_1}{2\pi(\log d)^3}} - \log\sqrt{\frac{\lambda_2}{2\pi(\log d)^3}} \\ & - \frac{\lambda_1(\log d - \mu_1)^2}{2\mu_1^2 \log d} + \frac{\lambda_2(\log d - \mu_2)^2}{2\mu_2^2 \log d} \end{aligned}$$

4. In this model, with Box-Cox transformation in Equation (2.4), transformed relevant and non-relevant document length follow the Inverse Gaussian distribution distribution. Then the length relevance weighting function is called 'Box-Cox-IGD'.

$$\begin{aligned} w(d)_{22} \propto \; & \log\sqrt{\frac{\lambda_1}{2\pi(\frac{d^\theta-1}{\theta})^3}} - \frac{\lambda_1(\log(\frac{d^\theta-1}{\theta}) - \mu_1)^2}{2\mu_1^2 \log(\frac{d^\theta-1}{\theta})} \\ & - \log\sqrt{\frac{\lambda_2}{2\pi(\frac{d^\theta-1}{\theta})^3}} + \frac{\lambda_2(\log(\frac{d^\theta-1}{\theta}) - \mu_2)^2}{2\mu_2^2 \log(\frac{d^\theta-1}{\theta})} \end{aligned}$$

5. In this model, using Box-Cox transformation, transformed relevant and non-relevant document length follow the GEV distribution. Then the length relevance weighting function is called ‘Box-Cox-GEV’.

$$\begin{aligned} w(d)_{22} \propto & \left(-1-\frac{1}{\kappa_1}\right)\log(1+\kappa_1\frac{V}{\sigma_1}) - \left(1+\kappa_1\frac{V}{\sigma_1}\right)^{-\frac{1}{\kappa_1}} \\ & - \left(-1-\frac{1}{\kappa_2}\right)\log(1+\kappa_2\frac{V}{\sigma_2}) + \left(1+\kappa_2\frac{V}{\sigma_2}\right)^{-\frac{1}{\kappa_2}} \end{aligned}$$

where $V = (d^{\theta} - 1)/\theta$.

6. In this model, using normalization transformation in Equation (2.5), transformed relevant and non-relevant document length follow the Lognormal distribution. Then the length relevance weighting function is called ‘Lognormal’.

$$w(d)_{22} \propto -\frac{1}{2\sigma_1^2}\left[\log L - \mu_1\right]^2 \;+\; \frac{1}{2\sigma_2^2}\left[\log L - \mu_2\right]^2$$

where $L = (d - d_{min})/(d_{max} - d_{min})$.

7. In this model, using normalization transformation, transformed relevant and non-relevant document length follow the Weibull distribution. Then the length relevance weighting function is called ‘Weibull’.

$$w(d,\Delta)_{22} \propto (b_1 - b_2)\log L - (\frac{L}{a_1})^{b_1} + (\frac{L}{a_2})^{b_2}$$

where $L = (d - d_{min})/(d_{max} - d_{min})$.

Based on the discussion above, we propose a new length-based weighting function BM25L as follows:

$$w(\underline{x}, d) = (1 - \beta)\ w_{BM25}\ \oplus\ \beta\ *\ w(d, \Delta)_{22} \tag{2.8}$$

where w_{BM25} is the relevance score of BM25, $\oplus$ indicate that the term $w(d, \Delta)_{22}$ is added only once for each document, β is not only the interpolation factor which is empirically determined and highly depends on the dataset used, but also an adjust factor of the mixture of two hypotheses: Verbosity and Scope hypothesis. A document could be either extreme or of mixture of these two hypotheses as discussed in Robertson and Walker (1994). More over, the reason of adding β here is that we ignore the constant term in the calculation of $\log \frac{P(\Delta|\overline{R})}{P(\Delta|R)}$, we need to adjust the scale for the weights between w_{BM25} and $w(\underline{x}, d)$, and the weights between two hypotheses because BM25L consider the situation when both Verbosity and Scope hypothesis are presented in the same document. For a given query, each of the w_{BM25} or $w(\underline{x}, d)$ scores is normalized by the maximum w_{BM25} or $w(\underline{x}, d)$ score. The parameter β is obtained by Simulated Annealing in Kirkpatrick et al. (1983) over a set of training topics.

2.3 Evaluation and Experiment Results

We first give detail information about the four test collections in subsection 2.3.1, introduce our methodology for evaluating the BM25L model in Subsection 2.3.2, and the evaluation results comparing with BM25 baseline is presented in Subsections 2.3.3. The impact of parameters b and β is investigated in Subsection 2.3.4.

2.3.1 The TREC Test Collections

We examine the impact of document length on relevance using four standard TREC test collections. These four test collections are the most recent TREC datasets, and provide a good coverage on the a variety of commonly used datasets in IR evaluation, and are used for different test purposes and vary in size in term of the document length. Basic information about the test collections and topics are given in Table 2.1.

Table 2.1: Information about the test collections

Coll.	TREC Task	Topics	# Docs
disk1&2	1-3, Ad-hoc	51-200	741,856
WT10G	9, 10 Web	451-550	1,692,096
.GOV2	2004-2006 Terabyte Ad-hoc	701-850	25,178,548
ClueWeb B	2009 Relevance Feedback	rf.01-rf.50	49,375,681

The disk1&2 collection contains newswire articles from various sources, such as

Association Press (AP), Wall Street Journal (WSJ), Financial Times (FT), etc., which are usually considered as high-quality text data with little noise. It usually used for ad hoc test. The WT10G collection is a medium size crawl of Web documents, which was used in the TREC 9 and 10 Web tracks. It contains 10 Gigabytes of uncompressed data. The .GOV2 collection, which has 426 Gigabytes of uncompressed data, is a crawl from the .gov domain. This collection has been employed in the TREC 14 (2004), 15 (2005) and 16 (2006) Terabyte tracks. The ClueWeb collection is a very large crawl of the Web, and is currently the largest TREC test collection. We use the category B of ClueWeb, which contains about 50 million English Web pages, and its associated topics used in the TREC 2009 Relevance Feedback track. We index all documents in the above four collections. For all four test collections used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed.

2.3.2 Evaluation Methodology

We evaluate our proposed BM25L model over the 4 test collections used, namely disk1&2, WT10G, .GOV2, and ClueWeb B. Each topic contains three topic fields, namely title, description and narrative. We only use the title topic field that contains very few keywords related to the topic. The title-only queries are usually short which is a realistic snapshot of real user queries in practice.

On each collection, the associated topics are divided into the odd-numbered and even-numbered topics. Over those two topic subsets, our proposed model is evaluated by a 2-fold cross-validation. In each fold, one of the topic subsets is used for training, and the other subset is used for testing purposes. More specifically, the half of the training topics with lower topic numbers are used to train the length distribution estimation parameters, and the other half of the training topics are used to train the score combination parameter β in Equation (2.8). Finally, our proposed BM25L model is evaluated by its retrieval performance on average over the two subsets of test topics. We use the TREC official evaluation measures in our experiments, namely the statMAP on ClueWeb B in Voorhees and Buckland (2009), and the Mean Average Precision (MAP) on the other three collections in Voorhees et al. (2005).

The baseline of our evaluation is the classical BM25 model with different settings of its parameter b. By varying the b value, we investigate to which extent BM25L improves the retrieval performance. In particular, we compare the retrieval performance of BM25L to BM25 with $b = 0$, that is, BM25 without tf normalization, and BM25 with its parameter b optimized. All statistical tests are based on Wilcoxon Matched-pairs Signed-rank test.

2.3.3 Comparison with BM25

Tables 2.2 and 2.3 compare the retrieval performance of BM25L using data transform technique to the original BM25 without tf normalization (i.e. when $b = 0$), and with the tf normalization with its parameter b optimized, respectively.

Table 2.2: Evaluation results over the BM25 baseline with $b = 0$. A star indicates a statistically significant improvement over the baseline.

Coll.	BM25	Normal	Log-Gamma	Log-IGD	Box-Cox-IGD	Box-Cox-GEV	Lognormal	Weibull
disk1&2	0.1698	0.1700	0.2195*	0.1821	0.1856*	0.2336	0.1685	0.2339*
WT10G	0.1571	0.1570	0.1663	0.1772*	0.1769	0.1647	0.1576	0.1604
.GOV2	0.1782	0.1812	0.2051*	0.2079*	0.2058*	0.1995	0.2058*	0.2564*
ClueWeb B	0.1930	0.2035	0.3192*	0.3084*	0.3265*	0.3117*	0.1931	0.2973*

Table 2.3: Evaluation results over the BM25 baseline with optimized setting of b. A star indicates a statistically significant improvement over this baseline.

Coll.	BM25	Normal	Log-Gamma	Log-IGD	Box-Cox-IGD	Box-Cox-GEV	Lognormal	Weibull
disk1&2	0.2324	0.2326	0.2421	0.2504*	0.2579*	0.2501*	0.2491*	0.2432*
WT10G	0.2090	0.2090	0.2115	0.2143	0.2101	0.2125	0.2111	0.2109
.GOV2	0.3044	0.3051	0.3121	0.3056	0.3321*	0.3227*	0.3134	0.3039
ClueWeb B	0.2322	0.2401	0.2722*	0.3350*	0.3561*	0.3612*	0.1586	0.3963*

From Tables 2.2 and 2.3, we see that modeling document length distribution using GEV distribution leads to the most stable retrieval performance of our proposed length-based BM25L model. This is not of a surprise as we have shown that the GEV density fits the best to the actual document length distribution. Using the GEV density fitting of the document length, BM25L appears to outperform the BM25 baseline, and the improvement is statistically significant in most cases on all four test collections except WT10G.

The use of other distribution functions, in particular Gamma distribution, also leads to retrieval performance over the BM25 baseline on some of the test collections. However, their retrieval performance does not appear to be as robust as that obtained by GEV distribution. An extreme case is Normal distribution, which does not improve the BM25 baseline significantly on disk1&2, WT10G, and .GOV2 wether or not the parameter b is optimized. Out of the four test collections used, ClueWeb B has the most incomplete relevance assessments, for which only the top-10 documents returned by the TREC participating runs are judged by human assessors, see Buckley and Robertson. (2008). As the top ranked documents are mostly overlong, the biase towards long documents in the document ranking could be so evident that the length distribution of relevant and non-relevant documents fits very well with the distribution functions on both training and testing topics. As a consequence, BM25L leads to extremely high retrieval performance on ClueWeb B at most cases.

Tables 2.4 also show that retrieval performance of BM25L using mixture model to the original BM25 improved.

Table 2.4: Evaluation results over the BM25 baseline on disak1&2.

Coll.	BM25	Mixture
BM25 with b=0	0.1698	0.1973
BM25 with optimal b	0.2324	0.2426

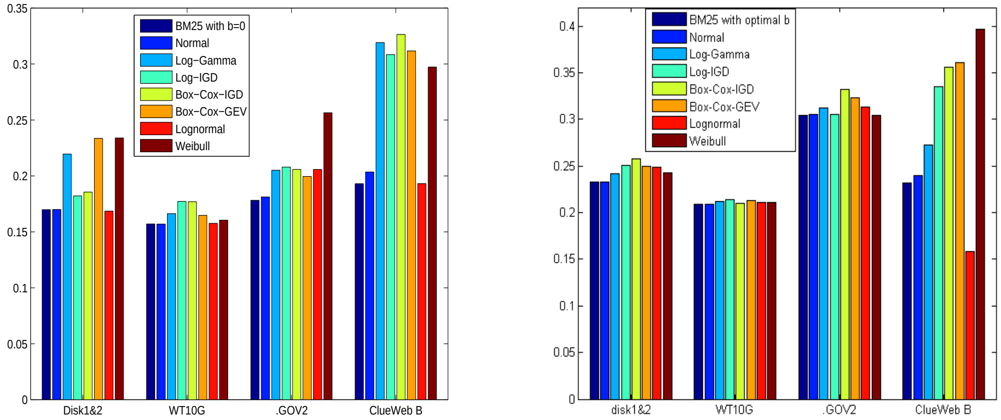


Figure 2.4: Performance of BM25L over BM25

To visualize the improvement brought the proposed length-based BM25L model, we plot the results in Figures 2.4 for the comparison to BM25 with $b = 0$ and with b optimized, respectively. As we can see on the WT10G collection, although the improvement is not as much as that obtained on other three test collections using all six distributions, the increase in retrieval performance is the evidence of length effect in information retrieval. Using Weibull distribution and normalization transformation has the best results, this may due to that Weibull distribution does retain the skewness of data between zero and one scale on all collections very well as we can see from Figure 2.3.

By comparing the performance improvement by BM25L over BM25 with $b = 0$ with the improvement over BM25 with the optimized b, we can see that the improvement over the optimized b is overall of a less scale than that over BM25 without tf normalization. This is because optimizing the parameter b in BM25 has exaggerated the length impact on the relevance weighting of term frequency tf, and in return, it reduces the impact of length relevance weighting itself on improving the document ranking.

When comparing BM25L with the best known results, for the WT10G, BM25L's best MAP is 0.2143, and the best published MAP is 0.2085. A possible explanation of the relatively minor improvement is as follows: the data transformation on WT10G does not show much difference between the length distribution of relevance and non-relevance documents. Compared to large-scale collections such as .GOV2 and ClueWeb B, it leaves little room for the BM25L model to further improve the retrieval performance by utilizing such difference (in document length distribution). In other words, the TREC pools are biased by the length distribution. Such bias is minor on WT10G, and becomes evident on heterogeneous collections like .GOV2 and ClueWeb B, which is captured by BM25L to boost the ranking effectiveness. For ClueWeb B, we believe the best statMAP in the TREC 2009 Relevance Feedback track, i.e. 0.2638, is achieved by combining BM25 with relevance feedback reported by Ye et al. (2009), although the overview paper is not available. Our model BM25L gives an MAP of 0.3963. For .GOV2, on top of the retrieval baselines, e.g. BM25 and language model, the best run in TREC 2006 further improved the effectiveness using pseudo relevance feedback and term dependency by Li and Yan (2006), Metzler et al. (2006). Since our model only considers document length, the best MAP presented in this paper, i.e. 0.3321, is not directly comparable to the best known MAP of 0.3737. For disk1&2, there hasn't been known best result for all 150 topics used in the TREC 1-3 ad-hoc tasks. According to evaluatir.org, the best known MAP on each task is 0.2062, 0.2475 and 0.3231, respectively, with an average of 0.2589. Note that the above best known results are achieved by stacking additional techniques such as relevance feedback over the retrieval baseline. Therefore, the results are not directly comparable. Even though, BM25L provides an MAP of 0.2579.

2.3.4 Impact of Parameters

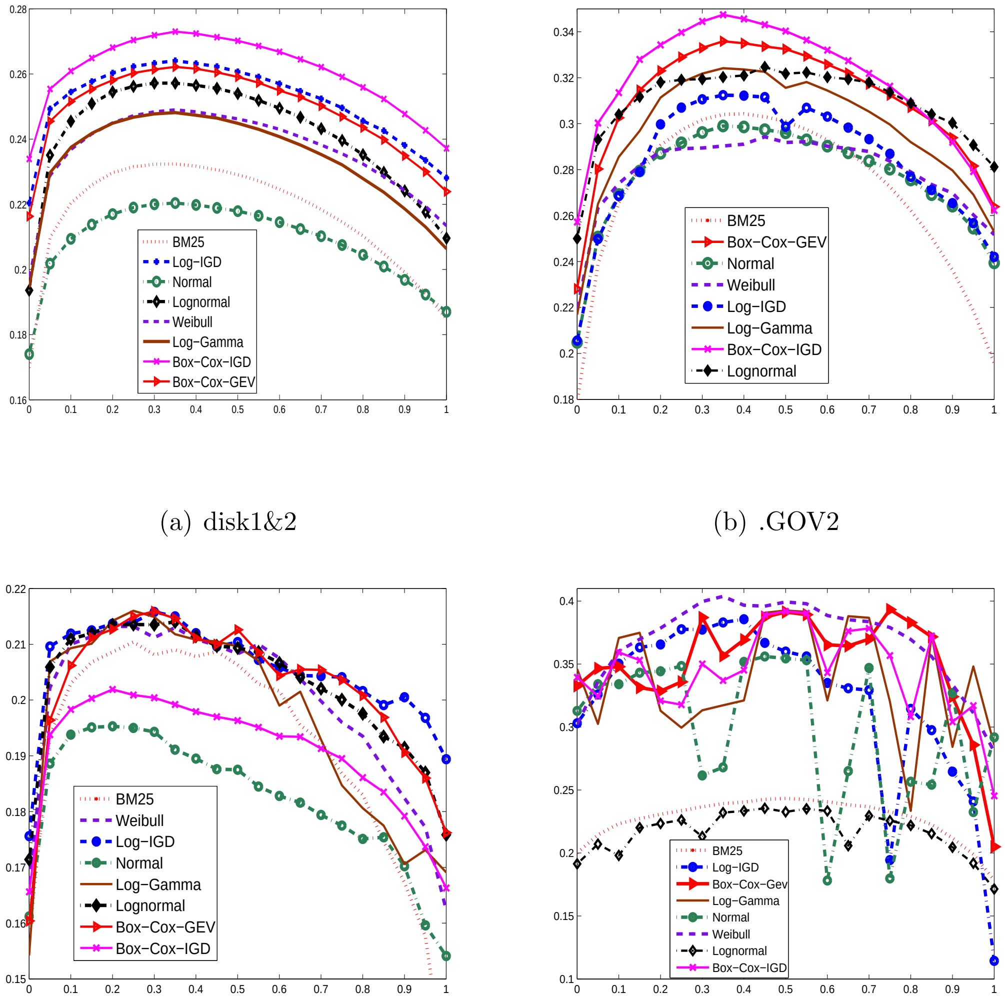
Experimental results in the previous section shows that, on one hand, BM25L leads to more improvement over BM25 when tf normalization is disabled. This is expected since there is no length information added to BM25 with $b = 0$ when compare to BM25L. On the other hand, BM25L provides higher MAP/statMAP values when the tf normalization parameter b is optimized. From this observation, a question arises: what is the impact of the setting of b on BM25L's effectiveness? To answer this question, Figure 2.5 plots the MAP/statMAP obtained by BM25L using the

6 different statistics of the document length distribution against different b values, from 0 to 1. BM25L and the original BM25's retrieval performance is seen to be correlated. A better setting of BM25's b leads to a better retrieval performance of BM25L. The document length itself do have a power as a stand-alone factor on the document relevance weighting other than normalization adjustment. The results for the full range of b are illustrated in Figure 2.5 for all 4 test collections.

Another important factor that could heavily affect BM25L's retrieval performance is the parameter β in Equation (2.8). Figure 2.6 plots the MAP/statMAP obtained by BM25L against β on the four collections used. As we can see that length impact on the relevance weighting increase first as β increase, then either decrease or remain flat as β increase. This is no coincidence because with only one factor, *i.e.* length, among other many important factors that can affect document relevance weighting, the improvement would be limited.

Another parameter in Equation (1.5) is k_1. BM25L outperforms BM25 with different settings of k_1, although the latter's retrieval performance is fairly sensitive to k_1's setting. The related experiments are not included in this paper for brevity, since changing k_1's setting does not affect the conclusions.

In summary, we have shown that the length information can be used for leveraging the bias towards long documents in the document ranking. The retrieval per-



(a) disk1&2

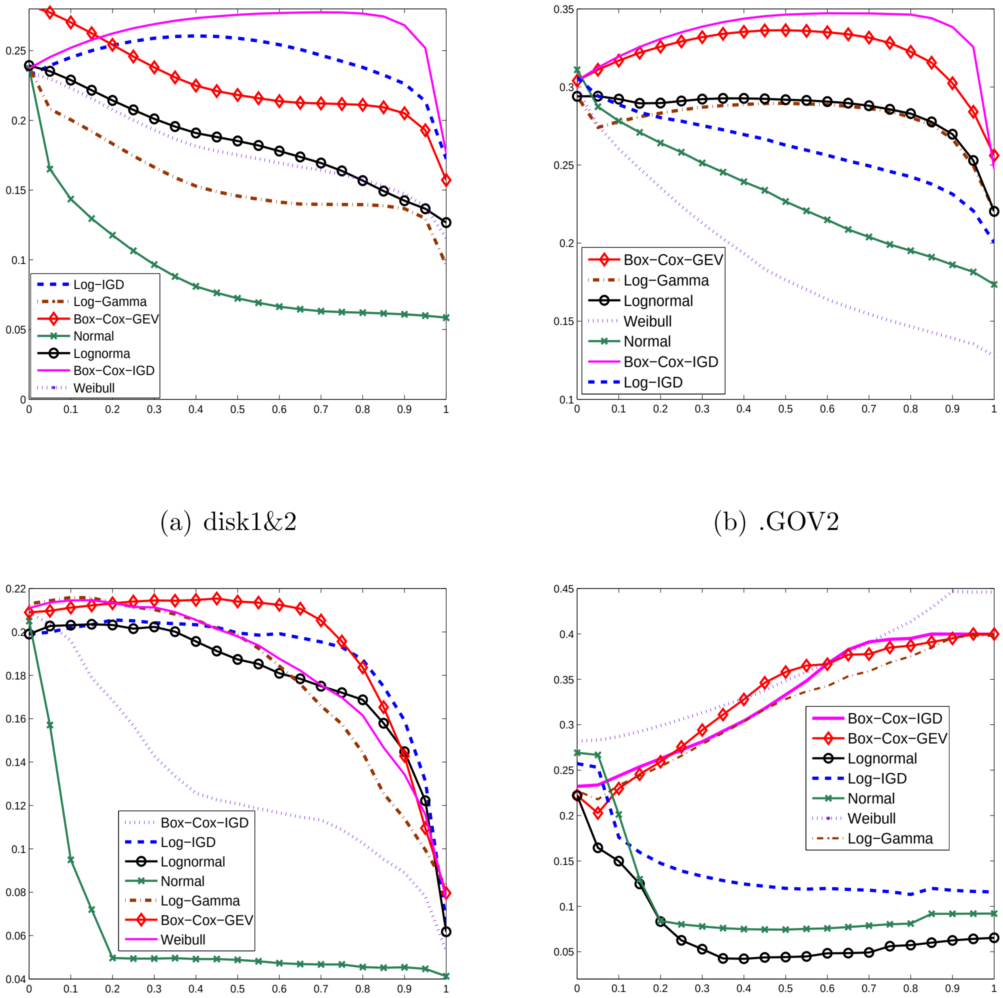
(b) .GOV2

(c) WT10G

(d) ClueWeb B

Figure 2.5: The MAP/statMAP values obtained against the parameter b

formance of the classical well-established BM25 model can be marked improved by incorporating a length-based weighting component with different settings of BM25's



(a) disk1&2

(b) .GOV2

(c) WT10G

(d) ClueWeb B

Figure 2.6: The MAP/statMAP values obtained against the parameter β

tf normalization parameter, including the optimal setting. Finally, we recommend applying GEV distribution for modeling the document length distribution as it has demonstrated effective and robust retrieval performance in our experiments. In our experiments, all parameters are learned from the training data in the two-fold cross-validation and our proposed model is trained and tested with different queries.

3 Survival Approach to Diversity and Proximity

3.1 Survival Analysis

Survival analysis is a statistical methodology used for modeling and evaluating survival data, also called time-to-event data, where one is interested in the occurrence of events (Cox and Oakes 1984) . Survival time refers to a variable which measures the time from a particular starting time to a particular endpoint of interest. Events are usually referred as birth, death and failure that happen to an individual in the context of study. For example, in clinical trial, one may be interested in the number of days that patient can survive in the study of the effectiveness of a new treatment for a disease. Formally, the survival function is defined as:

$$\begin{aligned} S(t) &= P(surviving\ longer\ than\ time\ t) \\ &= P(T > t) \end{aligned} \tag{3.1}$$

where t is a specific time, T is a random variable denoting the time of death, and "P" stands for probability. That is, the survival function gives the probability that the time of death is later than a specified time t. The survival function must be non-increasing: $S(u) \leq S(t) \quad \text{if} \quad u > t$ and $S(0) = 1$, that is, at the start of the study, the probability of surviving past time zero is one. The survival function is also assumed to approach zero as t goes to infinity.

We assume that the occurrence of event follows Poisson distribution with a parameter of rate λ, the probability mass function is

$$P(X = x) = \frac{\lambda^x \, e^{-\lambda}}{x!} \tag{3.2}$$

where X is a random variable denoting the number of occurrences of an event, λ ($\lambda > 0$) is the rate parameter denoting the expected number of occurrences. The reason of choosing Poisson distribution will be explained in the next two sections. Therefore, the survival function derived from Equation (3.1) can be formally written as:

$$S(x) = P(X > x) = 1 - e^{-\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!} \tag{3.3}$$

3.2 Survival Approach to Diversity

In the context of information retrieval, aspects covered by a document can be considered as treatments, a document can be considered as a patient in the clinical trial case. The number of times that an aspect has been observed can be considered as the survival time. The new information that can be provided by an aspect corresponds to the effectiveness of a treatment.

In clinical trial, as the number of times that a treatment has been given to a patient, the effectiveness of the treatment to the patient decreases. While in an IR ranked list, one can expect that, as the number of times that an aspect has been observed increases, the new information provided by this aspect decreases. This means that the effect of an aspect to a document's novelty decreases as the number of its occurrences increases. For example, in a ranked document list, when aspect "stroke treatment" is observed in the jth document at the first time, the information provided by this aspect should be counted as completely new. We presume that "stroke treatment" in the jth document covers the topic of "medications taken by mouth for long-term stroke treatment". Then, when aspect "stroke treatment" is observed again in the kth $(k > j)$ document of the ranked list, it may provide new information about "injection for short-term stroke treatment", but it is also possible that it only provides redundant information about "medications taken by mouth for long-term stroke treatment". As we can see, this situation satisfies the properties of the survival function described above:

- When an aspect is observed at the first time, this aspect can provide completely new information. This means that, at time point zero, the probability of surviving is one.

- The probability of obtaining new information from an aspect decreases as the number of the aspect's occurrences increases. This means that the probability of surviving decreases as time increases.

- When the number of occurrences of an aspect is approaching infinity, this aspect can not provide any new information. This means that the probability of surviving is approaching zero as time increases without bound.

The occurrences of a term in a document have a stochastic element was studied by Robertson and Walker (1994). In this chapter, we consider that the occurrences of an aspect in retrieved documents have the same stochastic property. So, it is reasonable to assume that the occurrences of an aspect follow Poisson distribution. When the aspect a_j observed in the retrieved documents, the probability of new information provided by the aspect a_j can be written as

$$S_{a_j}(x_j) = P(X_j > x_j) = 1 - e^{-\lambda_j} \sum_{i=0}^{x_j} \frac{\lambda_j^i}{i!} \tag{3.4}$$

where X_j is a random variable denoting the number of the aspect a_j observed in the retrieved documents; λ_j is the rate parameter denoting the expected number of occurrences of aspect a_j. When all aspects detected at the same rates of $\lambda_j = \lambda,\ j = 1, \ldots, n$, we name the above as single survival model. In real world, the distributions of different aspects' occurrences may be different from each other. Therefore, instead of assuming that all aspects' frequencies follow the same distribution, we assume that each observed aspect's frequency follows a specific Poisson distribution with rate λ_j, and we referred it to multiple survival model.

3.2.1 The Process of Re-ranking

In order to promote ranking diversity, we propose a document re-ranking method which combines the novelty and the relevance of retrieved documents at the aspect level. We ranked the first document according to

$$d_1 = \arg\max_{d_i}\{\sum_{a_j \in A_{d_i}} P(a_j|Q)\} \tag{3.5}$$

where d_i is a retrieved document and A_{d_i} is the set of aspects that can be detected from d_i.

For other retrieved documents, the document rankings should depend on which documents the user has already seen. Suppose that we have ranked top $i-1$ $(i>1)$ documents, and now we need to decide which document should be ranked at the ith position in the ranking list. The document which can deliver the newest and most relevant aspects should be considered as the ith document in the ranking list. Therefore, given previous ranked $i-1$ documents, we rank the ith document using the following scoring function:

$$score(d_i; d_1, ..., d_{i-1}) = P(New\ and\ Rel|d_i) \tag{3.6}$$

We assume that aspect novelty and aspect query-relevance are independent of each other. Moreover, since document d_i can be presented by the aspects detected from d_i, Equation (3.6) thus can be written as:

$$\begin{aligned} score(d_i; d_1, ..., d_{i-1}) &= \sum_{a_j \in A_{d_i}} P(New\ and\ Rel|a_j) \\ &= \sum_{a_j \in A_{d_i}} P(New|a_j)P(Rel|a_j) \\ &\propto \sum_{a_j \in A_{d_i}} P(New|a_j)\frac{P(a_j|Rel)}{P(a_j)} \end{aligned} \tag{3.7}$$

where a_j is an aspect detected from document d_i, which follows Poisson distribution with an estimated rate parameter. $P(New\ and\ Rel|a_j)$ denotes the probability that a_j is query-relevant and can provide new information as well.

$P(New|a_j)$ in Equation (3.7) states the probability of obtaining new information from aspect a_j, which can be calculated using the survival models proposed in Section 3.1. Since we do not usually have relevance information, $P(a_j|Rel)$ is unavailable. One possible solution, as introduced in Lavrenko and Croft (2001), is to consider that the best bet by relating the probability of aspect a_j to the conditional probability of observing a_j given the query: $P(a_j|Rel) \approx P(a_j|Q)$. Thus we can use the ranking scores of aspects from Cao et al. (2005), Zhu et al. (2010). This two-stage model that combines a relevance model and a co-occurrence model is used for ranking detected aspects. More formally, the two-stage model is defined as:

$$P(a_i|Q) = \sum_j P(a_i, d_j|Q) = \sum_j P(d_j|Q)P(a_i|d_j, Q) \tag{3.8}$$

where $P(d_j|Q)$ is the relevance model presenting whether a retrieved document $d_j(j = 1, 2, ..., N;$ where N is the number of retrieved documents) is relevant to the query Q; $P(a_i|d_j, Q)$ is the co-occurrence model presenting whether an aspect a_i is associated with the query. More details can be found at (Yin et al. 2010).

3.2.2 Experiment Settings and Evaluation Measures

In order to evaluate the proposed approach for promoting ranking diversity in biomedical information retrieval, we use the TREC 2006 and 2007 Genomics track full-text collection as the test corpus. It is a full-text biomedical corpus consisting of 162,259 documents from 49 genomics-related journals indexed by MEDLINE introduced by Hersh et al. (2007, 2006b). 28 official topics from the 2006 Genomics track and 36 official topics from the 2007 Genomics track are used as queries. Topics are in the form of questions asking for lists of specific entities that cover different portions of full answers to the topicsGenomics collections only present a fraction of millions of biomedical literatures indexed by MEDLINE. However, to the best of our knowledge, they are the largest and the only biomedical text collections with both manual relevance assessments and diversity evaluation available for biomedical text retrieval research so far.

There were three levels of retrieval performance that were measured in the TREC 2006 and 2007 Genomics tracks: passage retrieval, aspect retrieval and document retrieval. Each was measured by some variants of mean average precision (MAP). Passage MAP, Passage2 MAP[2], Aspect MAP and Document MAP were four evaluation measures corresponding to the three levels of retrieval performance introduced in (Hersh et al. 2007, 2006b). In this paper, we mainly focus on aspect level and passage level retrieval performance, since our objective is to promote diversity in the ranked list of retrieved passages. Moreover, aspect retrieval and passage retrieval were also the major tasks of these two Genomics tracks.

3.2.3 Information Retrieval Baseline Runs

For the 2007's topics, three IR baseline runs are used. NLMinter (Demner-Fushman et al. 2007) and MuMshFd (Stokes et al. 2007) were two of the most competitive IR runs submitted to the TREC 2007 Genomics track. NLMinter developed by the U.S. National Library of Medicine achieved the best performance in the TREC

[2] Passage2 MAP was defined in the TREC 2007 Genomics track, which is an alternative measure to the Passage MAP defined in the TREC 2006 Genomics track.

2007 Genomics track in terms of Aspect MAP, Passage2 MAP and Document MAP reported in (Demner-Fushman et al. 2007, Hersh et al. 2007). It merged the retrieval results obtained by Essie (Ide et al. 2007), Indri[3], Terrier (Ounis et al. 2006a), Theme (Demner-Fushman and Lin 2007), and EasyIR Gobeill et al. (2007) and employed a human-involved relevance feedback method. MuMshFd was developed by the National ICT Australia, Victoria Research Laboratory and also achieved top ranked performance in the 2007 Genomics track. Stokes et al. (2007) employed ontology-based (MeSH and Entrez Gene) query expansion and entity-based relevance feedback for genomics search. Another IR baseline run is an Okapi run, which is solely based on the probabilistic weighting model BM25 proposed by Hancock-Beaulieu et al. (1996). The performance of the Okapi run is also above average among all results reported in the TREC 2007 Genomics track by Hersh et al. (2007).

For 2006's topics, we test our approach on three Okapi runs since other retrieval results submitted to the TREC 2006 Genomics track are not available. In order to find out wether the proposed methods can work well on strong baselines as well as on average and weak baselines, we (Yin et al. 2011) set different values to BM25 parameters to obtain different baselines. The performance of the baseline run Okapi06b is also among the top performances reported in the TREC 2006 Genomics track Hersh

[3] http://www.lemurproject.org/indri/

et al. (2006b).

Since relevance scores from different baselines differ in range and relevance scores and novelty scores also differ in range,in our experiments we use the 0-1 normalization method to normalize the scores.

The best and mean results reported in the 2006 and 2007 Genomics tracks are shown in Table 3.1. The performance of baseline runs are shown in Table 3.2.

Table 3.1: The best and mean results in the Genomics tracks

	Best MAP		Mean MAP	
MAP	2006	2007	2006	2007
Aspect	0.4411	0.2631	0.1643	0.1326
Passage	0.1486	0.0976	0.0392	0.0560
Passage2		0.1148		0.0398
Document	0.5439	0.3286	0.2887	0.1862

3.2.4 Experimental Results on Genomics collections

Evaluation results of the proposed approach for document re-ranking are shown in Table 3.2, where "Single-SM" and "Multiple-SM" denote the single survival model and the multiple survival model respectively. The values in the parentheses are the relative rates of improvement over the original results and * denotes the improvement over the baseline is statistically significant (Wilcoxon test at the 5% significance level). As we can see, our approach achieves promising performance improvements

Table 3.2: Re-ranking Performance based on Aspect Detection Using Wikipedia

on 2007's topics				
MAP	Aspect	Passage	Passage2	Document
NLMinter	0.2631	0.0968	0.1148	0.3286
Single-SM	**0.3117***	0.1007	**0.1270***	**0.3440***
	(+18.5%)	(+4.0%)	(+10.6%)	(+4.7%)
Multiple-SM	**0.3128***	0.1009	**0.1274***	**0.3447***
	(+18.9%)	(+4.2%)	(+11.0%)	(+4.9%)
MuMshFd	0.2068	0.0840	0.0895	0.2906
Single-SM	**0.2432***	0.0877	0.0926	0.3030
	(+17.6%)	(+4.4%)	(+3.5%)	(+4.3%)
Multiple-SM	**0.2448***	**0.0883***	0.0931	**0.3056***
	(+18.4%)	(+5.1%)	(+4.0%)	(+5.2%)
Okapi07	0.1428	0.0633	0.0641	0.2025
Single-SM	**0.1660***	**0.0662***	0.0669	0.2086
	(+16.2%)	(+4.6%)	(+4.4%)	(+3.0%)
Multiple-SM	**0.1686***	**0.0671***	**0.0677***	**0.2124***
	(+18.1%)	(+6.0%)	(+5.6%)	(+4.9%)

on 2006's topics				
MAP	Aspect	Passage	Passage2	Document
Okapi06a	0.2176	0.0362	0.0450	0.3476
Single-SM	**0.2379***	0.0381	0.0472	0.3557
	(+9.3%)	(+5.2%)	(+4.9%)	(+2.3%)
Multiple-SM	**0.2383***	**0.0388***	**0.0487***	0.3604
	(+9.5%)	(+7.2%)	(+8.2%)	(+3.7%)
Okapi06b	0.3147	0.1559	0.0968	0.4705
Single-SM	0.3236	0.1606	0.1009	0.4885
	(+2.8%)	(+3.0%)	(+4.2%)	(+3.8%)
Multiple-SM	**0.3299***	0.1627	**0.1030***	0.4934
	(+4.8%)	(+4.4%)	(+6.4%)	(+4.8%)
Okapi06c	0.2596	0.0759	0.0601	0.4388
Single-SM	0.2697	0.0796	0.0624	0.4564
	(+3.9%)	(+4.9%)	(+3.8%)	(+4.0%)
Multiple-SM	**0.2709***	**0.0803***	**0.0637***	0.4619
	(+4.4%)	(+5.8%)	(+6.0%)	(+5.3%)

in most cases. It is worth mentioning that our approach can further improve the best result (NLMinter) reported in the TREC 2007 Genomics track by achieving 18.9% improvement on Aspect MAP and 11% improvement on Passage2 MAP.

Experimental results also show that the multiple survival model slightly outperforms the single survival model. In our experiments, we use the Maximum Likelihood Estimation (MLE) to estimate the Poisson parameter λ, then calculate the $P(a_j)$ according to Poisson distribution. We do not manually set the rate parameters. As described in Section 3.1, the multiple survival model estimates the distribution rate for each detected aspect, while the single survival model only estimates one distribution rate for all aspects as it assumes that all aspects follow the same distribution. Thus, it is not surprised that the multiple survival model outperforms the single survival model.

We also note that, in terms of Aspect MAP, the improvements on the 2007's topics are more significant than the improvements on the 2006's topics. This might be due to that the average number of distinct aspects of each 2007's topic (72.3 aspects per topic) is much larger than that of each 2006's topic (27.9 aspects per topic) descried by Hersh et al. (2007, 2006b). A topic with more distinct aspects indicates the information need of this topic could be more diverse. In this case, our approach performs better.

In biomedical IR, the use of domain-specific thesauri is still the most commonly used method of integrating external knowledge. Therefore, it is worthwhile to compare the re-ranking performance based on aspect detection using Wikipedia and using domain-specific thesauri. Table 3.3 presents re-ranking results based on aspect detection using the largest thesaurus UMLS[4] in the biomedical domain as the knowledge resource. When the UMLS is used for aspect detection, performance improvements can be obtained in terms of Aspect MAP and Passage2 MAP. However, Passage MAP and Document MAP may decrease on some baselines. From experimental results shown in Table 3.2 and Table 3.3, we can find that, compared with aspect detection using the UMLS, aspect detection based on Wikipedia can achieve more evident and more stable performance improvements. This is because the enriched entity pages in Wikipedia could result in a better mapping between terms in biomedical text and concepts. Moreover, instead of only providing hierarchical relationships (synonyms, hypernyms, hyponyms) among biomedical concepts like the UMLS, plenty of Wikipedia links and anchor texts can also provide more natural relationships among Wikipedia concepts.

[4] http://www.nlm.nih.gov/pubs/factsheets/umls.html

Table 3.3: Re-ranking Performance based on Aspect Detection Using UMLS

on 2007's topics				
MAP	Aspect	Passage	Passage2	Document
NLMinter	0.2631	0.0968	0.1148	0.3286
Single-SM	0.2688	0.0962	0.1173	0.3240
	(+2.2%)	(-0.6%)	(+2.1%)	(-1.4%)
Multiple-SM	0.2695	0.0969	0.1183	0.3243
	(+2.4%)	(+0.1%)	(+2.9%)	(-1.3%)
MuMshFd	0.2068	0.0840	0.0895	0.2906
Single-SM	**0.2233***	0.0836	0.0907	0.2829
	(+8.0%)	(-0.5%)	(+1.3%)	(-2.6%)
Multiple-SM	**0.2256***	0.0844	0.0918	0.2844
	(+9.1%)	(+0.5%)	(+2.6%)	(-2.1%)
Okapi07	0.1428	0.0633	0.0641	0.2025
Single-SM	**0.1564***	0.0638	0.0654	0.2070
	(+9.5%)	(+0.8%)	(+2.0%)	(+2.2%)
Multiple-SM	**0.1576***	0.0647	0.0655	0.2080
	(+10.4%)	(+2.2%)	(+2.1%)	(+2.7%)

on 2006's topics				
MAP	Aspect	Passage	Passage2	Document
Okapi06a	0.2176	0.0362	0.0450	0.3476
Single-SM	0.2202	0.0359	0.0460	0.3420
	(+1.1%)	(-0.8%)	(+2.2%)	(-1.6%)
Multiple-SM	0.2219	0.0363	0.0466	0.3431
	(+2.0%)	(+0.3%)	(+3.6%)	(-1.3%)
Okapi06b	0.3147	0.1559	0.0968	0.4705
Single-SM	0.3184	0.1512	0.0966	0.4738
	(+1.2%)	(-3.0%)	(-0.2%)	(+0.7%)
Multiple-SM	0.3195	0.1538	0.0973	0.4774
	(+1.5%)	(-1.3%)	(+0.5%)	(+1.5%)
Okapi06c	0.2596	0.0759	0.0601	0.4388
Single-SM	0.2702	0.0763	0.0603	0.4200
	(+4.1%)	(+0.5%)	(+0.3%)	(-4.3%)
Multiple-SM	**0.2732***	0.0769	0.0624	0.4255
	(+5.2%)	(+1.3%)	(+3.8%)	(-3.0%)

3.2.5 Impact of the Survival Models

In order to investigate the effect of the survival models on promoting ranking diversity, we substitute the survival model with a binary novelty measuring method. The binary method measures the novelty of an aspect as follows: if an aspect is observed at the first time, the novelty score of the aspect is 1, otherwise, 0. The re-ranking results of using these two methods are shown in Figure 3.1 and Figure 3.2, where "Multi-SM" denotes the system using the multiple survival model with aspect filtering for re-ranking, "Binary" denotes the system using the binary novelty measuring method for re-ranking and "No_Filter" denotes the system using the multiple survival model without aspect filtering for re-ranking. The binary method only involves filtered aspects. It re-ranks retrieved passages using Equation (3.7), where $P(New|a_j) = 1$ when a_j appears the first time, otherwise, $P(New|a_j) = 0$. We can see that the IR runs using the survival model outperform those using the binary novelty measure. Statistically significant improvement on Aspect MAP can be observed using the survival model for re-ranking, while the binary re-ranking method could only achieve significant improvements on Aspect MAP over the baseline NLMinter(Wilcoxon test at the 5% significance level). The performance differences between the use of two methods are more evident on Aspect MAP, while the differences on Passage MAP, Passage2 MAP and Document MAP are minor. This observation indicates that using the survival model has a substantial impact on promoting diversity of the ranked list. The survival model successfully measures the probability of obtaining novel information from an aspect and makes a positive contribution to retrieval result diversification.

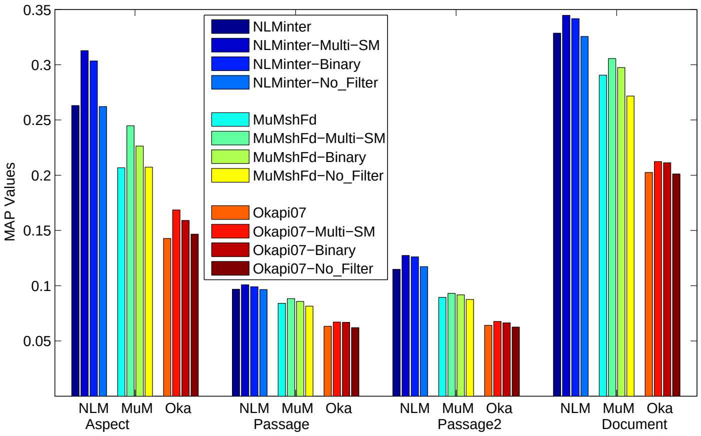


Figure 3.1: Effects of Survival Model and Aspect Filtering on 2007's topics.(The x-axis presents the evaluation measures, where "NLM", "MuM" and "Oka" in the left figure stand for three baselines corresponding to NLMinter, MuMshFd and Okapi07.)

3.2.6 Experiment Results on Web Track

Experimental results on the TREC Genomics collections demonstrate that the proposed approach is effective in promoting ranking diversity for biomedical text retrieval. However, promoting ranking diversity is not only a research topic in biomed-

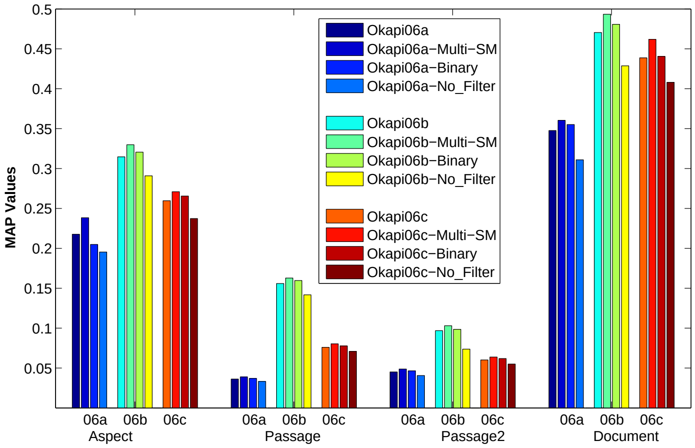


Figure 3.2: Effects of Survival Model and Aspect Filtering on 2006's topics.(The x-axis presents the evaluation measures, where "06a", "06b" and "06c" in the right figure stand for three baselines corresponding to Okapi06a, Okapi06b and Okapi06c.)

ical IR, but also a research topic in many other retrieval environments (e.g. ad-hoc retrieval and Web retrieval). In this section, we further conduct a series of experiments on the ClueWeb09-T09B collection (also known as the "Category B" collection in the TREC 2009) to evaluate our approach in the Web environment. ClueWeb09-T09B consists of about 50 million English-language Web pages and was used as the test collection in the TREC 2009 Web track (Clarke et al. 2009b).

We mainly focus on the diversity task of the TREC 2009 Web track, since the goal of the diversity task is to return a ranked list of Web pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list in (Clarke et al. 2009b). The evaluation measures of the diversity task are α-nDCG (with $\alpha = 0.5$ in the (Clarke et al. 2009b, 2008b) and an intent-aware version of precision (IA-P) in (Agrawal et al. 2009b). Both evaluation measures reflect the diversity of a ranked list as well as the query-relevance of retrieved documents. 50 topics (we only use the "<query>" field of the topics) from the 2009 Web track are used as queries in our experiments.

Table 3.4: Re-ranking Performance on the Diversity Task of TREC 2009 Web Track

	α-nDCG@5	α-nDCG@10	α-nDCG@20	IA-P@5	IA-P@10	IA-P@20
BM25	0.125	0.170	0.205	0.061	0.077	0.085
Single-SM	**0.169***	**0.194***	**0.231***	**0.081***	**0.081***	0.086
	(+35.2%)	(+14.1%)	(+12.7%)	(+32.8%)	(+5.2%)	(+1.2%)
Multiple-SM	**0.195***	**0.226***	**0.253***	**0.090***	**0.088***	**0.090***
	(+56.0%)	(+32.9%)	(+23.4%)	(+47.5%)	(+14.3%)	(+5.9%)
BM25+RF	0.209	0.224	0.238	0.108	0.106	0.089
Single-SM	**0.218***	0.225	0.239	0.110	0.106	0.089
	(+4.3%)	(+0.4%)	(+0.4%)	(+1.9%)	(+0%)	(+0%)
Multiple-SM	**0.225***	0.232	0.243	**0.115***	0.108	0.089
	(+7.7%)	(+3.6%)	(+2.1%)	(+6.5%)	(+1.9%)	(+0%)

We compare our approach with two baseline runs. The first baseline is solely based on the probabilistic weighting model BM25. The second baseline is much stronger, which employs BM25 for document ranking and Kullback-Leibler Divergence for relevance feedback (RF) based on 5 human labeled documents reported by Carpineto et al. (2001), Ye et al. (2009). Experimental results are shown in Table 3.4. As we can see, the proposed approach achieves significant performance improvements over the BM25 baseline in terms of α-nDCG@5, α-nDCG@10, α-nDCG@20, IA-P@5 and IA-P@10. The multiple survival model still consistently outperforms the single survival model. Moreover, our approach can further improve the performance of the baseline with relevance feedback in terms of α-nDCG@5. This indicates that, based on this strong baseline, our approach can still rank more diverse and relevant documents to the top of the ranked list. Overall, the evaluation results on the ClueWeb09-T09B collection demonstrate the effectiveness of our approach in promoting diversity of Web search result.

3.3 Survival Approach to Proximity

In literature, term proximity is interpreted as query term co-occurrences, or phrases. Follow the co-occurrence interpretation of the term proximity information in Croft et al. (1991), the proximity among query terms is interpreted as n-gram frequencies. In the context of modeling term proximity information, the importance of the n-gram in the document is measured by the co-occurrence probability. The main focus of this section is modeling co-occurrence probability of n-gram among query terms.

3.3.1 Re-ranking Model Based on Term Proximity

The probabilistic document ranking (PDR) for a given query using the term proximity evidence p_i is given by van Rijsbergen (1977) as follows:

$$PDR \overset{rank}{=} \sum_{i=1,2,\ldots,n} \alpha_i \cdot Score(d, p_i) \tag{3.9}$$

where $Score(d, p_i)$ is the ith n-gram relevance score. Different n-gram models are combined through a linear function with a parameter α_i to adjust the relative importance of each n-gram model.Thus, a major difficulty is how to assign the relevance score using term proximity between more than one terms, i.e. when i is larger than or equal to 2.

There could be many ways to assign the proximity-based relevance scores, such as those mentioned in Section 1.3. Here, we adopt the co-occurrence interpretation of term proximity introduced by van Rijsbergen (1977), where the proximity among query terms is represented by the n-gram frequencies, then the n-gram relevance score of a document d for a query Q is given as follows:

$$Score(d, Q, p_n) = \sum_{t_1,t_2,\ldots,t_n \in Q} w^{(1)} \frac{(k_{1n}+1)tf_n}{K + tf_n} \cdot qtw_n \tag{3.10}$$

where an n-gram consists of n query terms t_1, $t_2, \ldots$, and t_n. In the context of the n-gram model, the above variables can be interpreted as follows:

- tf_n is the n-gram frequency. A simple method for computing the n-gram frequency is to count the number of co-occurrences of query terms. A more sophisticated solution is to consider information provided by n-gram frequency a decreasing function of the distance between query terms. We propose various novel methods for computing the n-gram frequency, which is described in details in the next section.

- $w^{(1)}$ is the raw term weight, following the re-written point-5 formula proposed by Robertson et al. (1995), is given as

$$w^{(1)} = \log \frac{N - N_n + 0.5}{N_n + 0.5} \tag{3.11}$$

  where N_n is the n-gram document frequency , *i.e.* the number of documents in which the n-gram occurs for at least once. This Equation is corresponding to the $w^{(1)}$ in Equation (1.5).

- qtw_n is the n-gram weight, it is given by the average query term weight of the n-gram terms t_1, t_2, ..., and t_n: $qtw_n = \frac{\sum_{i \in 1,2,...,n} qtw_{t_i}}{n}$.

- K is given by: $K = k_{1n}\left((1 - b_n) + b_n \frac{l_n}{avg\_l_n}\right)$, where the variables are defined in the context of the n-gram model. The length l_n refers to the number of windows in a document, which is given by $(l - n + 1)$. avg_l_n is the average number of windows in a document, given by $\left(\sum_{d \in Coll} l_d - N \cdot n - 1\right)$, where

$\sum_{d \in Coll} l_d$ is the number of tokens in the whole collection.

- Each individual b_i of n-gram model is optimized by Simulated Annealing proposed by Kirkpatrick et al. (1983) on training queries, while only the n-gram model is enabled. This is equivalent to setting the linear combination weight α_i (see Equation (3.9)) of other n-gram models to 0.

- The other parameters k_{1n} and k_{3n} are the same parameters for each n-gram model, as those in the content-based BM25 in Equation (1.5). In experiment, we default them to $k_1 = 1.2$ and $k_3 = 1000$ for every n-gram model to reduce the number of parameters to be optimized.

3.3.2 Term Proximity Modeling

The simple way of modeling term proximity is the window-based n-gram frequency counting (NC) method, which has been popular in previous studies on using term proximity for IR (e.g. Metzler and Croft (2005b), Plachouras and Ounis (2007b)). The basic idea of the window-based n-gram counting method is to segment the document into a list of sliding windows, with each window having a fixed window size *wSize*. If a document has a length of *l*, and the window size is set to *wSize*, the document is then segmented into *1-wSize* sliding windows, where each window contains *wSize* consecutive tokens. For example, if a document has four tokens A, B, C, and D, and the window size is 3, there are two windows in this document, namely A, B, C and B, C, D. The n-gram frequency is then defined as the number of windows in which all n-gram terms co-occur.

However, the n-gram counting method does not take the actual distance between query terms into account, and any n-gram terms appear together within a window is counted as one occurrence of the n-gram. Another possible downside is that it is unable to differentiate between n-grams in which the n-gram terms appear loosely and tightly in the text. An alternate solution is to approximate a proximity-based co-occurrence probability of the n-gram terms in the document d. The distance between the n-gram terms is characterized by the shortest distance between any two n-gram terms in the n-gram, and other n-gram terms appears in between. For example, if a window consists of a sequence of the following 6 tokens: Q_1, A, Q_2, Q_1, B, Q_3, where Q_1, Q_2, and Q_3 are the tri-gram terms, and A, B are tokens in the document, the distance between the tri-gram terms is 2 in this case, which is the number of tokens between Q_2 and Q_3.

Therefore, the n-gram frequency becomes:

$$tf_n = \sum_{i=1,\ldots,x} P_i(Q_n, \tau_i) \tag{3.12}$$

where x is the number of windows in which the n-gram terms co-occur. τ_i is the distance between the n-gram terms in the ith window in which the n-gram terms co-occur. To simplify the approximation, we assume that only the two n-gram terms that have the shortest distance in the window (and the other n-gram terms appear in between) is taken into account. In the context of modeling the term proximity information, the importance of the n-gram in the document is measured by the probability of co-occurrence and its proximity, denoted as $S(\tau)$. The co-occurrence probability corresponds to the effectiveness of a treatment, or a mean of modeling of the distance, where τ is the distance between Q_i and Q_j. More details can be founded in He et al. (2011). Remind that Q_i and Q_j are the two n-gram terms with the shortest distance in the text window. A text window can be interpreted as a patient in the clinical trial case, and the distance between Q_i and Q_j, composed of a sequence of words, can be interpreted as the survival time, that is, the time that it takes for the treatment to cure the patient. In other words, in a text window, as the distance between the n-gram terms increases, the importance of the n-gram in the document decreases. The n-gram has the highest importance when the n-gram terms are adjacent to each other, that is, when the distance or survival time is zero. So the probability of the n-gram co-occurrence and proximity can be approximated by utilizing survival function on the Homogeneous Poisson Process (SurvHPP), exponential distribution (SurvExp), and empirical distribution (SurvEmp) as follows:

1. T follows Poisson distribution:

$$S(\tau) = 1 - F(\tau) = 1 - e^{-\lambda} \cdot \sum_{i=0}^{\tau} \frac{\lambda^i}{i!} \tag{3.13}$$

2. T follows exponential distribution:

$$S(\tau) = P(T > \tau) = e^{-\lambda\tau} \tag{3.14}$$

3. The empirical survivor function is

$$S(\tau) = 1 - \frac{1}{n}\sum_{i=1}^{n} I(\tau) \tag{3.15}$$

   where $I(\tau)$ is an indicator function.

3.3.3 Experimental Setup

We evaluate our proposed methods on ad-hoc topics over four large-scale TREC Web collections. These four collections are the most recent Web collections with associated ad-hoc test topics, which are currently the best snapshot of the real Web environment[5]. The WT10G, .GOV2 and ClueWebB collections are introduced in the Subsection 2.3.1. The Blog06 collection includes 100,649 blog feeds collected over an 11 week period from December 2005 to February 2006. Following the official TREC setting in (Ounis et al. 2006b), we index only the permalinks, which are the

[5] Other TREC Web collections are not used in that they do not have corresponding ad-hoc test topics.

blog posts and their associated comments. For all four test collections used, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed.

Each topic contains three topic fields, namely title, description and narrative. We only use the title topic field that contains few keywords related to the topic. The title-only queries are usually short which is a realistic snapshot of real user queries in practice. On each collection, we evaluate our proposed model by a two-fold cross-validation. The test topics associated to each collection are split into two equal subsets, namely the odd-numbered and even-numbered topics. In each fold, we use one subset of topics for training, and use the remaining subset for testing. The overall retrieval performance is averaged over the two test subsets of topics. We use the TREC official evaluation measures in our experiments, namely the Mean Average Precision (MAP) on WT10G and .GOV2 in Voorhees et al. (2005), the topical MAP on Blog06 in Ounis et al. (2006b), and the statMAP on ClueWeb B[6].

3.3.4 Experimental Results

In this section, we evaluate our proposed proximity-based model in a bi-gram setting first. The evaluation baseline is the original unigram BM25 model in Equation 1.5

[6] Information about the ClueWeb dataset can be found at the following URL: http://boston.lti.cs.cmu.edu/Data/clueweb09/

proposed by (Robertson et al. 1995) . Compared to this baseline, we evaluate our proposed model at two different levels. First, we compare the performance of the bi-gram BM25 model with the unigram BM25 baseline. Second, we compare the unigram BM25 model alone with the combination of the unigram and bi-gram BM25 models.

Table 3.5: The MAP/statMAP values obtained by the unigram BM25 baseline and the bi-gram model.

	Unigram	Bi-gram, unordered			
Coll.	BM25	NC	SurvHPP	SurvExp	SurvEmp
WT10G	0.2217	0.1605	0.1359	0.1623	0.1468
Blog06	0.3403	0.2205	0.2111	0.2163	0.2082
.GOV2	0.3055	0.2445	0.1977	0.2434	0.2023
ClueWeb B	0.2085	0.1437	0.1405	0.1423	0.1379
	Unigram	Bi-gram, ordered			
WT10G	0.2217	0.1575	0.1302	0.1600	0.1206
Blog06	0.3403	0.2189	0.1945	0.2082	0.1955
.GOV2	0.3055	0.2423	0.1907	0.2412	0.2002
ClueWeb B	0.2085	0.1376	0.1045	0.1451	0.1310

Table 3.5 compares the retrieval performance of the bi-gram models with the unigram BM25 model. The two variants, namely the ordered and the unordered bi-gram models, are also tested. Results in Table 3.5 show that the bi-gram BM25 model is not as good as the unigram model. This observation is in line with the findings in previous studies (e.g. Metzler and Croft (2005b)). We suggest that the

relatively low performance of the bi-gram model could be due to the fact that the bi-gram model is unable to properly measure the informativeness of the individual query terms that do not form a concept with other query terms. Even though, the bi-gram model still provides a descent retrieval performance, which has the potential to lead to improved retrieval performance by combining the unigram model with the bi-gram model.

Table 3.6: The MAP values obtained by the unigram BM25 baseline and the bi-gram model on WT10G.

Model	Unigram	Unordered	Ordered
NC	0.2104	0.2240, 6.46*%	0.2201, 4.61*%
SurvHPP	0.2104	0.2168, 3.04%	0.2158, 2.57%
SurvExp	0.2104	0.2229, 5.94*%	0.2199, 4.52*%
SurvEmp	0.2104	0.2167, 2.99%	0.2146, 2.00%

Table 3.7: The MAP values obtained by the unigram BM25 baseline and the bi-gram model on Blog06.

Model	Unigram	Unordered	Ordered
NC	0.3190	0.3521, 10.38*%	0.3384, 6.08*%
SurvHPP	0.3190	0.3449, 8.12*%	0.3417, 7.12*%
SurvExp	0.3190	0.3551, 11.32*%	0.3397, 6.49*%
SurvEmp	0.3190	0.3417, 7.12*%	0.3376, 5.83*%

Next, we examine the retrieval performance of the linear combination of the unigram and bi-gram models. The related experimental results on the four collections

Table 3.8: The MAP values obtained by the unigram BM25 baseline and the bi-gram model on .GOV2.

Model	Unigram	Unordered	Ordered
NC	0.3045	0.3211, 5.45*%	0.3146, 3.32%
SurvHPP	0.3045	0.3118, 2.40%	0.2842, -6.67%
SurvExp	0.3045	0.3196, 4.96*%	0.3119, 2.43%
SurvEmp	0.3045	0.3117, 2.36%	0.2797, -8.15%

Table 3.9: The statMAP values obtained by the unigram BM25 baseline and the bi-gram model on ClueWeb B.

Model	Unigram	Unordered	Ordered
NC	0.2107	0.2136, 1.38%	0.2043, -3.04%
SurvHPP	0.2107	0.2107, 0%	0.1854, -12.01%
SurvExp	0.2107	0.2131, 1.39%	0.2057, -2.37%
SurvEmp	0.2107	0.2113, 0.28%	0.1877, -10.92%

used are listed in Tables 3.6 - 3.9, respectively. From these tables, we observe that the combined model outperforms the unigram BM25 model in most cases. To be more specific, the N-Gram Counting (NC) method and the Survival Analysis over exponential distribution (SurvExp) result in statistically significant improvement over the unigram BM25 baseline on three out of four test collections used. We also observe insignificant improvement brought by BM25P on ClueWeb B. Furthermore, the unordered combined model is consistently better than the ordered one. This indicates that the order of appearances of query terms is not important for ad-hoc retrieval on the collections used. In addition to the bi-gram models tested in the previous section, we evaluate the effectiveness of our proposed model in tri-gram setting in this section. We do not experiment beyond the tri-gram model since only few title-only TREC test queries contain more than four non-stopword unique keywords.

Table 3.10: The MAP/statMAP values obtained for evaluating the tri-gram model. A star indicates a significant improvement over the unigram BM25 baseline.

Coll.	BM25	NC	SurvHPP	SurvExp	SurvEmp
	Unigram	Unigram+Bi-gram, unordered			
WT10G	0.2104	0.2240*	0.2168	0.2229*	0.2167
Blog06	0.3190	0.3521*	0.3449*	0.3494*	0.3417*
.GOV2	0.3045	0.3211*	0.3118*	0.3196*	0.3117
ClueWeb B	0.2107	0.2136	0.2107	0.2131	0.2113
	Unigram	Unigram+Trigram, unordered			
WT10G	0.2104	0.2198*	0.2166	0.2203*	0.2143
Blog06	0.3190	0.3254*	0.3267*	0.3301*	0.3271*
.GOV2	0.3045	0.3102	0.3056	0.3074	0.2956
ClueWeb B	0.2107	0.2187	0.2132	0.2215	0.2197
	Unigram	Unigram+Bi-gram+Trigram, unordered			
WT10G	0.2104	0.2283*	0.2194*	0.2275*	0.2184
Blog06	0.3190	0.3535*	0.3491*	0.3584*	0.3468*
.GOV2	0.3045	0.3245*	0.3187*	0.3213*	0.3117
ClueWeb B	0.2107	0.2177	0.2187	0.2208	0.2173

Table 3.10 presents the results for evaluating the tri-gram model. The combination of the unigram model with either the bi-gram or the tri-gram model can

provide a statistically significant improvement over the unigram baseline in many cases. However, the improvement brought by the tri-gram model is not as high as those brought by the bi-gram model. This might be due to the fact that the bi-gram model better utilizes the latent association between query words than the tri-gram model. Moreover, a combination of the unigram, bi-gram and tri-gram models performs better than the combination of the unigram and bi-gram models, while the improvements are rather marginal on all four collections. We only report the results with the unordered models in that the ordered models lead to similar conclusions while their retrieval performance is slightly lower. In summary, the tri-gram model is shown to be a useful addition to the bi-gram model, although the combination of the unigram and bi-gram models are sufficient to provide reliable retrieval performance on the collections used.

4 Conclusion and Future Work in IR

We have demonstrated the effectiveness of incorporating document length, aspect and term proximity information into the classical probabilistic IR models. In Chapter 2, we study the impact of document length on its relevance based on the assumption that a document may exhibit both Verbosity and Scope hypotheses and show the evidence to its retrieval. In Chapter 3, we apply the survival function to estimate the novel information provided by aspects to promoting the diversity in the ranked document lists, and term proximity to improve retrieval performance.

Based on the current work, we plan to extend our work to the other IR models, such as language model and PL2, possibly by incorporating with machine learning methods.

- Document length has been empirically proved as an useful factor in document retrieval. There is a need of finding a simple way to model the impact of document length on document relevance when document style and structure change dynamically nowadays.

- In fact, whether a repeated aspect could provide novel information may depends on the context it appears. We thus plan to take aspects' contextual information into account in our future work. We also plan to adapt our approach to other domains, such as Web retrieval.

- Term proximity that can be employed not only by BM25 in text retrieval, we also consider to extend the application of our proposed model to retrieval from structured documents, such as the XML retrieval. An XML document is an ordered and labeled tree, where the nodes are usually composed of short text, and tokens of the same word appear at different levels in the tree structure. This poses a challenge for the application of our proposed model to the XML documents, since it is difficult to define the notions of n-grams and window size in such a tree structure.

From next Chapter, we will start on the topic of asymptotic methods and its applications.

5 Literature Survey on Asymptotic Methods

Fisher (1921) defined the likelihood function as a function of the parameters of a statistical model, which is proportional to the joint density of the model. In 1922, Fisher extended the studies and proposed the method of maximum likelihood estimate (MLE) as a mean of a systematic way of estimating the parameters based on the observed sample. Neyman and Pearson (1933) proved that the likelihood ratio test is the most powerful test for any given size γ test. Under regularity conditions as stated in Wilks (1938), he proved that the limiting distribution of the likelihood ratio is a χ^2 distribution. Wald (1943) extended Wilks' work on the distribution of the likelihood ratio statistic to a more general situation, and derived the limiting distribution of the maximum likelihood estimate. Additionally, Rao (1948) introduced score test which is also based on the likelihood function. These likelihood inference methods have rates of convergence $O(n^{-1/2})$ and is referred to as the first-order approximation methods. Even though the first-order methods are widely used, they generally do not give accurate approximation especially when the sample size is small. In recent literature, Fraser and Reid (1995) derived a more accurate likelihood-based asymptotic method with rate of convergence $O(n^{-3/2})$.

This section reviews some key concepts and definitions in asymptotic approximation inference. In Section 5.1, we review the likelihood function and maximum likelihood estimate. The standard first-order asymptotic techniques are also examined. In Section 5.2, the saddlepoint approximation to the density of the mean of n independent identically distributed random variables is introduced. We give the brief presentation about the third-order asymptotic method in Section 5.3.

5.1 Standard First-order Likelihood-based Asymptotic Methods

Assume a statistical model has probability density function (or probability function) $f(\cdot\,;\theta)$, where $\theta = (\theta_1, \theta_2, \cdots, \theta_k)' \in \Theta$ is a k-dimensional parameter. The likelihood function defined by Fisher (1921) for any random sample $y = (y_1, y_2, \cdots, y_n)'$ obtained from the above model is

$$L(\theta) = L(\theta; y_1, \cdots, y_n) = cf(y_1, \cdots, y_n; \theta),$$

for values of θ within a given domain, where $c > 0$ is a multiplicative constant, and $f(y_1, \cdots, y_n; \theta)$ is the value of the joint probability distribution or the joint probability density function of the random variables $Y_1, \cdots, Y_n$ evaluated at $Y_1 = y_1, \cdots, Y_n = y_n$. The corresponding log-likelihood function is

$$\ell(\theta) = \ell(\theta; y_1, \ldots, y_n) = a + \sum_{i=1}^{n} \log(f(y_i; \theta)), \tag{5.1}$$

where $a \in \mathbb{R}$ is an additive constant. Without lost of generality, a is set to be zero hereafter in this dissertation.

Let $\hat{\theta}$ be the overall MLE, which can be obtained by solving

$$\ell_\theta(\hat{\theta}) = \left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0.$$

Let $\psi = \psi(\theta)$ be a scalar parameter of interest and λ be the nuisance parameter with size $k-1$, denoted as $\theta = (\psi, \lambda')'$. Now, consider the null hypothesis

$$H_0 : \psi(\theta) = \psi_0,$$

$\hat{\theta}_\psi$ is the constrained MLE of θ which maximize (5.1) for a given $\psi(\theta) = \psi_0$. Generally, when λ is explicitly known, $\hat{\theta}_\psi$ is obtained by solving

$$\ell_\lambda(\hat{\theta}_\psi) = \left.\frac{\partial \ell(\theta)}{\partial \lambda}\right|_{\theta=\hat{\theta}_\psi} = 0.$$

However, when λ is not explicitly available, $\hat{\theta}_\psi$ can be obtained by applying the Lagrange multiplier technique for maximizing (5.1) subject to the constraint $\psi(\theta) = \psi_0$. Since the closed form of $\hat{\theta}$ and $\hat{\theta}_\psi$ are not always available, numerical methods are often required. The tilted log-likelihood function is defined as

$$\tilde{\ell}(\theta) = \ell(\theta) + \hat{\kappa}[\psi(\theta) - \psi_0] \tag{5.2}$$

with maximum value at $\hat{\theta}_\psi$ and $\hat{\kappa}$ is the Lagrange multiplier which maximized the (5.2). Note that this tilted likelihood has the property $\tilde{\ell}(\hat{\theta}_\psi) = \ell(\hat{\theta}_\psi)$ for a given constraint $\psi(\theta) = \psi_0$.

Moreover, throughout this dissertation with $\theta \in \Theta$, the following regularity conditions are assumed to hold:

- $f(y;\theta) > 0$ is twice continuously differentiable in a neighborhood of θ;
- $\displaystyle\int \sup_{\theta \in N} |f_\theta(y;\theta)| dy < \infty$ and $\displaystyle\int \sup_{\theta \in N} |f_{\theta\theta'}(Y;\theta)| dy < \infty$, where $f_\theta(y;\theta) = \dfrac{\partial f(y;\theta)}{\partial \theta}$ and $f_{\theta\theta'}(y;\theta) = \dfrac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta'}$;
- $E[\ell_\theta(y;\theta)\ell'_\theta(y;\theta)]$ exists and is nonsingular, where $\ell_\theta(\theta) = \dfrac{\partial \ell(\theta)}{\partial \theta}$;
- $\displaystyle\int \sup_{\theta \in N} |\ell_{\theta\theta'}(y;\theta)| dy < \infty$, where $\ell_{\theta\theta'}(\theta) = \dfrac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}$.

Under these regularity conditions, applying Central Limit Theorem and Taylor expansion to (5.1), the following can be obtained:

- $\ell_\theta(\theta) \xrightarrow{d} N_k(0, \mathrm{var}[\ell_\theta(\theta)])$, or equivalently $\ell'_\theta(\theta) \{\mathrm{var}[\ell_\theta(\theta)]\}^{-1} \ell_\theta(\theta) \xrightarrow{d} \chi^2_k$, where $\mathrm{var}(\ell_\theta(\theta)) = i_{\theta\theta'}(\theta) = \mathrm{E}[j_{\theta\theta'}(\theta)]$ is the Fisher expected full information matrix, and $j_{\theta\theta'}(\theta) = -\ell_{\theta\theta'}(\theta)$ is the observed information matrix.
- $\hat{\theta} \xrightarrow{d} N_k(\theta, \mathrm{var}(\hat{\theta}))$, or equivalently $(\hat{\theta} - \theta)'[\mathrm{var}(\hat{\theta})]^{-1}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_k$, where $\mathrm{var}(\hat{\theta}) \approx i^{-1}_{\theta\theta'}(\theta)$.

- $2\left[\ell(\hat{\theta})-\ell(\theta)\right] \xrightarrow{d} \chi^2_k$.

Note that although $i_{\theta\theta'}(\theta)$ could be difficult to obtain, it can be approximated by $j_{\theta\theta'}(\hat{\theta})$. When we are interested in the inference for ψ, the following three test statistics are usually used:

- Signed log-likelihood ratio statistic

$$r = r(\psi) = \operatorname{sgn}(\hat{\psi}-\psi)\{2[\ell(\hat{\theta})-\ell(\hat{\theta}_\psi)]\}^{\frac{1}{2}}. \tag{5.3}$$

- Wald statistic

$$q = q(\psi) = \frac{\hat{\psi}-\psi}{\sqrt{\operatorname{var}(\hat{\psi})}}\ . \tag{5.4}$$

- Score statistic

$$S = S(\psi) = \frac{\ell_\psi(\hat{\theta}_\psi)}{\sqrt{\operatorname{var}(\hat{\theta}_\psi)}}\ , \tag{5.5}$$

  where

$$\operatorname{var}(\hat{\psi}) \approx \psi_\theta(\hat{\theta}) j^{-1}_{\theta\theta'}(\hat{\theta}) \psi_\theta(\hat{\theta})',$$

$$\psi_\theta(\hat{\theta}) = \left.\frac{\partial \psi(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}},$$

$$\ell_\psi(\hat{\theta}_\psi) = \left.\frac{\partial \ell(\psi,\lambda)}{\partial \psi}\right|_{\lambda=\hat{\lambda}_\psi}.$$

Engle (1984) showed that the three tests are asymptotically equivalent.

The p-value functions based on (5.3), (5.4) and (5.5) are defined as

$$p(\theta) = \begin{cases} \Phi(r) \\ \Phi(q) \\ \Phi(S) \end{cases} \tag{5.6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of N(0,1). These methods all have rate of convergence $O(n^{-1/2})$ and are generally referred to as the first-order methods. Hence a central $(1-\gamma) \times 100\%$ confidence interval for ψ is

$$\left(\min\{p^{-1}(\gamma/2), p^{-1}(1-\gamma/2)\}, \max\{p^{-1}(\gamma/2), p^{-1}(1-\gamma/2)\}\right). \tag{5.7}$$

The first two statistics are more popular in terms of application. For finite sample, Doganaksoy and Schmee (1993) found that generally (5.3) has better coverage properties than (5.4). As we have discussed before, Neyman and Pearson (1933) proved that (5.3) gives the most powerful test. However, they also pointed out the (5.4) is more popular in applied analysis than (5.3) because of its simplicity in application. For example, a $(1-\gamma)100\%$ confidence interval for ψ is approximately $\hat{\psi} \pm z_{\gamma/2}\sqrt{\text{var}(\hat{\psi})}$.

Although the three first-order methods are widely used in hypothesis testing, they do not perform well when the sample size is small or when the underlying distribution is far away from the normal distribution. In the next two subsections, higher-order asymptotic methods will be introduced.

5.2 Saddlepoint Approximation

Daniels introduced saddlepoint approximation to statistics in 1954. This method approximate the density of the mean of n independent and identically distributed (*i.i.d.*) random variables, and it is very accurate but could be very complicated in terms of computation.

Assume $Y_1, \ldots, Y_n$ are *i.i.d.* random vectors with size k from a model with density $f_Y(\cdot;\ \theta)$. The moment generating function is defined as

$$M(t) = E[e^{t^{'}Y}]$$

and cumulant generating function is

$$K(t) = \log(M(t)).$$

The saddlepoint approximation for the density of the mean of n independent, identically distributed random variables, *i.e.* $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, given by Daniels (1954), is

$$f_{\bar{Y}}(\bar{y}) = (2\pi)^{-k/2}\left\{\frac{n}{\left|\mathrm{K}_{tt'}(\hat{t})\right|}\right\}^{1/2} \exp\left[n\{\mathrm{K}(\hat{t}) - \hat{t}'\bar{y}\}\right]\{1 + O(n^{-1})\} \qquad (5.8)$$

where

$$\mathrm{K}_t(\hat{t}) = \left.\frac{\partial \mathrm{K}(t)}{\partial t}\right|_{t=\hat{t}} = \bar{y} \tag{5.9}$$

is the saddlepoint and

$$\mathrm{K}_{tt'}(\hat{t}) = \left.\frac{\partial \mathrm{K}^2(t)}{\partial t \partial t'}\right|_{t=\hat{t}}$$

$\hat{t}$ is known as the saddlepoint. Note that (5.8) has a relative error of $O(n^{-1})$.

A statistical version of the derivation of the saddlepoint approximation can be obtained by combing exponential tilting and Edgeworth expansion techniques and then collect the first four terms. For detailed review of the derivation, please refer to Barndorff-Nielsen (1978), Barndorff-Nielsen and Cox (1979, 1989, 1994) and Reid (1988). The resulting density is

$$f_{\bar{Y}}(\bar{y}) = a \left\{ \frac{n}{\left|\mathrm{K}_{tt'}(\hat{t})\right|} \right\}^{1/2} \exp\left[n\{\mathrm{K}(\hat{t}) - \hat{t}'\bar{y}\}\right] \{1 + O(n^{-3/2})\} \tag{5.10}$$

where a is the normalizing constant. Durbin (1980) showed that error term of the saddlepoint approximation (5.8) is reduced to $O(n^{-3/2})$ when $(2\pi)^{-k/2}$ is replaced by a normalizing constant a in (5.10).

For a canonical exponential family model with density

$$f_Y(y;\theta) = \exp\{\theta' y - c(\theta) + h(y)\}, \tag{5.11}$$

and the log-likelihood function can be written as

$$\ell(\theta) = n\theta'\bar{y} - nc(\theta),$$

with the cumulant generating function

$$K(t) = \log(M(t)) = c(\theta + t) - c(\theta),$$

there exists a one-one transformation from $\bar{Y}$ to $\hat{\theta}$

$$K_t(\hat{\theta}) = \bar{Y},$$

where $\bar{Y}$ is a minimal sufficient statistic for θ. Barndorff-Nielsen(1980, 1983) showed that the saddlepoint approximation of the density function of $\hat{\theta}$ is

$$f(\hat{\theta}; \theta) = a|j_{\theta\theta'}(\hat{\theta})|^{1/2} \exp\left\{\ell(\theta) - \ell(\hat{\theta})\right\} \left\{1 + O(n^{-3/2})\right\}, \tag{5.12}$$

where a is a normalizing constant. For models outside the exponential family setting, if an ancillary statistic is available, there is one-to-one correspondence between the minimum sufficient statistic and an ancillary statistic. An example of the construction of the ancillary statistic can be found in Barndorff-Nielsen (1980) or Barndorff-Nielsen and Chamberlin (1991). For a general model, if the MLE is not a one-one transformation of the minimal sufficient statistic, whereas (5.12) continues to provide an approximation to a conditional density of $\hat{\theta}$ by conditioning on an appropriate ancillary statistic A. The same result derived by Barndorff-Nielsen and Cox (1984) is

$$f(\hat{\theta}|A; \theta) = a(\theta, A)|j_{\theta\theta'}(\hat{\theta})|^{1/2} \exp\left\{\ell(\theta) - \ell(\hat{\theta})\right\} \left\{1 + O(n^{-3/2})\right\}. \tag{5.13}$$

The above approximation is accurate to $O(n^{-3/2})$ for exponential families and transformation models. Detailed discussion of the saddlepoint method and its application in statistics can be found in Barndorff-Nielsen (1983, 1986, 1991, 1991), McCullagh (1987), Fraser (1988), Reid (1988, 1996) and Barndorff-Nielsen and Cox (1989). However, there is no systematic method to construct ancillary statistic for a general model. Moreover, ancillary statistic may not exist and even if it exists, it may not be unique.

5.3 The Third-order Asymptotic Methods

Intensive research on higher-order asymptotics was intrigued by Barndorff-Nielsen and Cox (1979), which highlighted the usefulness and accuracy of the saddlepoint approximation to the density function of $\bar{Y}$, the mean of n *i.i.d.* random variables, as well as the cumulative distribution function for $\overline{Y}$, $F_{\bar{Y}}(\bar{y}) = \int_{-\infty}^{\bar{y}} f_{\bar{Y}}(s)ds$. However, this approximation may not have closed form and numerically integration is required, which will introduce significant error when analytical solution does not exist.

Following the idea of saddlepoint approximation, Lugannani and Rice (1980) calculated the cumulative distribution function by inverting the characteristic function (Fourier transformation):

$$F_{\bar{Y}}(\bar{y}) = P(\bar{Y} \leq \bar{y}) = \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q}\right) + O(n^{-3/2}), \tag{5.14}$$

where $r = \text{sgn}(\hat{t})\sqrt{2n[\hat{t}\bar{y} - \text{K}(\hat{t})]}$ and $q = \hat{t}\sqrt{n\text{K}_{tt}(\hat{t})}$. Note that $\hat{t}$ is the saddlepoint defined in Section 5.2 which satisfied $K_t(\hat{t}) = \bar{y}$.

For the canonical exponential family defined by (5.11), the likelihood formulation of the statistics r and q is

$$\begin{aligned} r &= \text{sgn}(q)\left[2\{\ell(\hat{\theta}) - \ell(\theta)\}\right]^{\frac{1}{2}}, \qquad (5.15) \\ q &= (\hat{\theta} - \theta)\{j_{\theta\theta}(\hat{\theta})\}^{\frac{1}{2}}. \qquad (5.16) \end{aligned}$$

which coincide with the signed likelihood ratio statistic and the maximum likelihood departure as defined by (5.3) and (5.4) for a scalar parameter of interest situation. Hence the p-value function of θ approximated by Lugannani and Rice (1980) formula is

$$p(\theta) = \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q}\right). \qquad (5.17)$$

Barndorff-Nielsen (1986) derived alternative approximation that incorporates the correction term into the quantile of the normal cumulative distribution:

$$F_{\bar{Y}}(\bar{y}) = P(\bar{Y} \leq \bar{y}) = \Phi(r^*)\left(1 + O(n^{-3/2})\right) \qquad (5.18)$$

where

$$r^* = r + \frac{1}{r}\log\frac{q}{r} \qquad (5.19)$$

and showed that it is asymptotically distributed as a standard normal distribution with a relative error of $O(n^{-3/2})$. r^* is referred as modified signed log-likelihood ratio statistic. Hence the approximated p-value function of θ

$$p(\theta) = \Phi(r^*). \tag{5.20}$$

It is interesting to note that the Barndorff-Nielsen's method adjusts the signed log-likelihood ratio statistic such that the p-value function obtained from r^* is close to the true p-value function; whereas the Lugannani & Rice method adjusted the p-value function obtained from the signed log-likelihood ratio statistic such that it is close to the true p-value function. Fraser (1990) and Jensen (1992) showed that these two adjustments are equivalent up to third-order accuracy.

Equations (5.17) and (5.20) are generally referred as the third-order methods. For exponential family models and transformation models, (5.17) and (5.20) can be obtained easily. However for general models, q could be difficult or impossible to derived since it is based on the existence of ancillary statistic. In practice, there is no accessible procedure available for the construction of an ancillary in a general context. In Chapter 6, a route to the third-order likelihood inference for any general statistical model will be discussed.

6 Third-Order Likelihood Inference for a General Statistical Model

This chapter details the mechanics of the likelihood-based third-order methods for a general statistical model. The advantage of these two proposed third-order methods, Lugannani and Rice, and Barndorff-Nielsen, are their prominent accuracy for even when sample size is small and applicability on obtaining inference for any scalar parameter of interest. It only depends on the likelihood function and its first sample space derivative at the data points. However, calculating the p-values may encountered singularity at the maximum likelihood value, and be numerically instable in the neighborhood of the maximum likelihood value. Fraser et al. (2003) proposed a bridging method to deal with the instability and singularity problem when the MLE is close to the hypothesized value, therefore it will not be discussed in detail in this dissertation. When the parameter of interest, ψ cannot be expressed explicitly, computation problems arise for any likelihood-based methods. For example, the ψ

considered in this dissertation is an integral which has no known closed-form. The numerical difficulties will be discussed in the next chapter and a penalized likelihood method is proposed to deal with the constrained maximization problem for the model we considered will also be presented in the next chapter.

The organization of the rest chapter is as follows. Section 6.1 studies the canonical exponential family model with the presence of the explicitly known nuisance parameter λ and then extended to a general exponential model in Section 6.2. An illustrated example is also given in this section via Behrens-Fisher problem. A general algorithm to the third-order likelihood inference for any statistical model will be introduced in Section 6.3.

6.1 Canonical Exponential Model

Consider a canonical exponential family model

$$f(y;\theta) = \exp\left\{\theta' y - c(\theta) + h(y)\right\} \tag{6.1}$$

where $\theta = (\psi, \lambda')'$ and our parameter of interest is ψ. It is easy to see that y is a sufficient statistic and θ is the canonical parameter for the model. For any random sample from above model, the signed log-likelihood ratio statistic r is remain unchanged as in (5.3) and it is

$$r = r(\psi) = \operatorname{sgn}(\hat{\psi} - \psi)\left\{2\left[\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)\right]\right\}^{1/2} \tag{6.2}$$

where $\ell(\hat{\theta}_\psi)$ is the log likelihood evaluated at constraint MLE.

Applying saddlepoint procedure (Fraser et al. (1991)) on the joint and on the marginal likelihood functions, the standardized maximum likelihood departure in the canonical parameter space becomes

$$q = q(\psi) = (\hat{\psi} - \psi)\left\{\frac{\left|j_{\theta\theta'}(\hat{\theta})\right|}{\left|j_{\lambda\lambda'}(\hat{\theta}_\psi)\right|}\right\}^{1/2} \tag{6.3}$$

where $j_{\theta\theta'}(\hat{\theta})$ is the observed overall information matrix evaluated at the overall MLE $\hat{\theta}$ and $j_{\lambda\lambda'}(\hat{\theta}_\psi)$ is the observed nuisance information matrix evaluated at the constrained MLE $\hat{\theta}_\psi$. Detailed derivation of (6.3) is discussed in Fraser et al. (1991). Therefore we can apply either Lugannani-Rice formula (5.17) or Barndorff-Nielsen formula (5.20) to obtain $p(\psi)$ with r and q defined above. These two approximations both have third-order accuracy.

To illustrate the application of third-order method for canonical exponential family model, the exponential distribution, the simplest case, is considered. Suppose $(y_1, \cdots, y_n)'$ is a sample from exponential distribution with rate $\theta > 0$, and its probability density function can be written in canonical form as

$$f(y;\theta) = \exp\{-\theta y + \log\theta\},\ y > 0,\ \theta > 0,$$

This is also known as the Gamma(1, θ) distribution. Then the log-likelihood is

$$l(\theta) = n\log(\theta) - \theta\sum_{i=1}^{n} y_i. \tag{6.4}$$

The overall MLE can be obtained easily *i.e.* $\hat{\theta} = \dfrac{n}{\sum_{i=1}^{n} y_i}$, and $j_{\theta\theta}(\theta) = -l_{\theta\theta}(\theta) = \dfrac{n}{\theta^2}$.

The signed log-likelihood ratio statistic and the standardized maximum likelihood estimate departure given in (6.2) and (6.3) can be easily obtained as

$$\begin{aligned} r &= sgn(\hat{\theta} - \theta)\{2[\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi)]\}^{1/2} \\ q &= (\hat{\theta} - \theta)\left(\frac{n}{\hat{\theta}^2}\right)^{1/2}. \end{aligned}$$

Hence, we can approximate p-value function for θ by (5.17) or (5.20).

Moreover, let $T = \sum_{i=1}^{n} Y_i$, where Y_i is *i.i.d.* from exponential model above, from distribution theory, $T \sim Gamma(n, \theta)$. Hence, for a given θ,

$$P(T \leq t; \theta) = \int_0^t f_T(s;\theta)ds.$$

Then the exact p-value function for θ is

$$p(\theta) = 1 - P(T \leq t; \theta).$$

To compare the accuracy of third-order methods, we generated three data set from above exponential distribution with $\theta = 4$ and sample size to be 5, 10 and 20. The simulated data are recorded in Table 6.1.

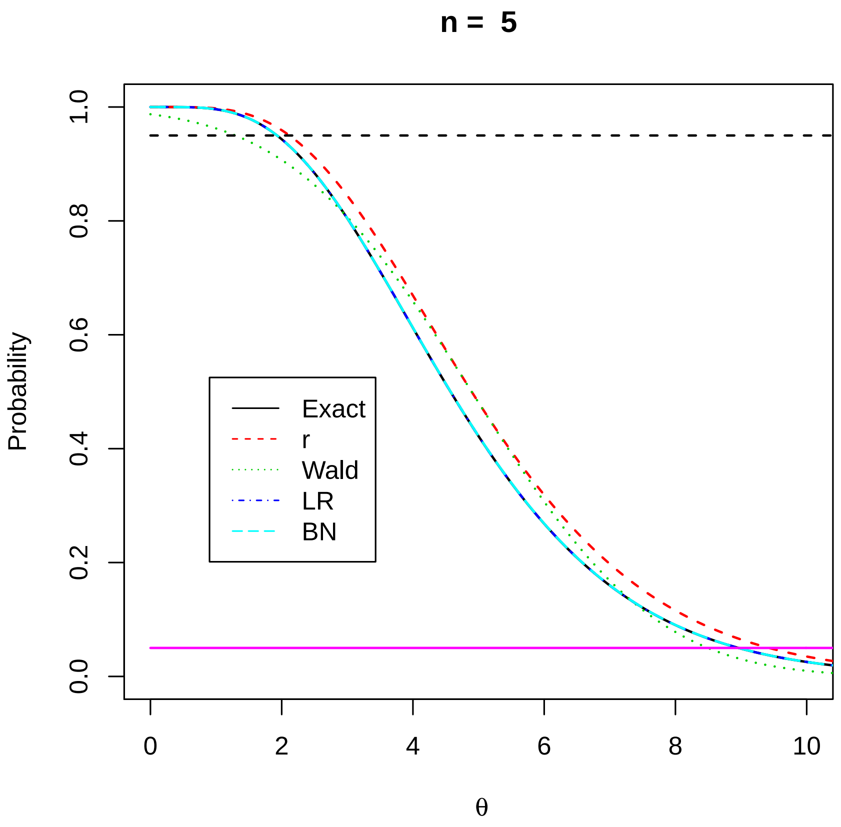


Figure 6.1: $p(\theta)$ for Data Set 1.

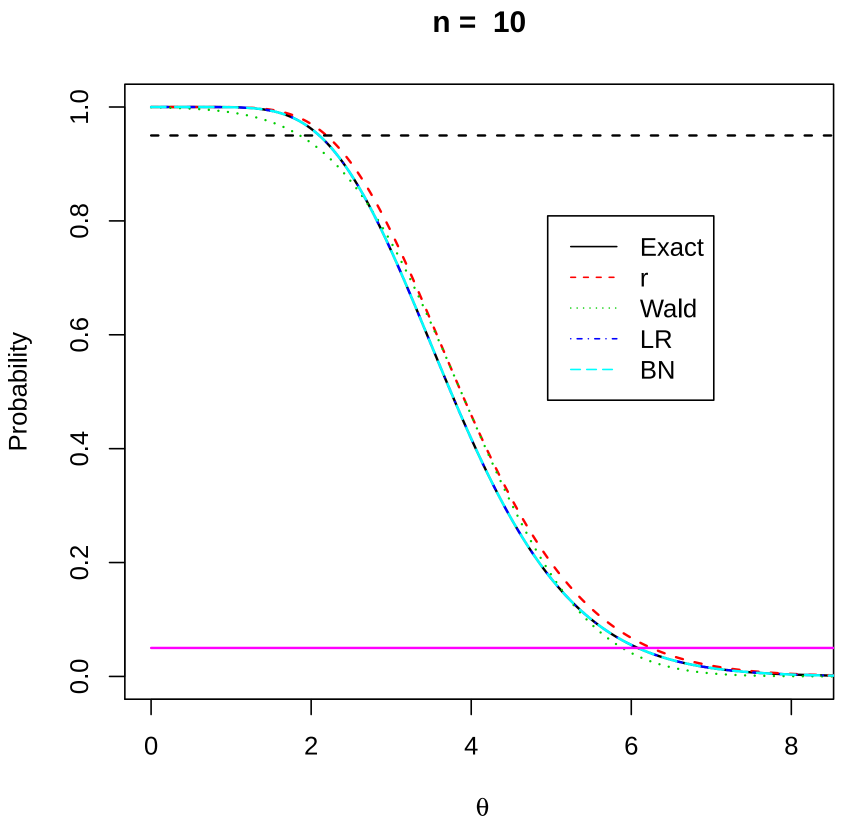


Figure 6.2: $p(\theta)$ for Data Set 2.

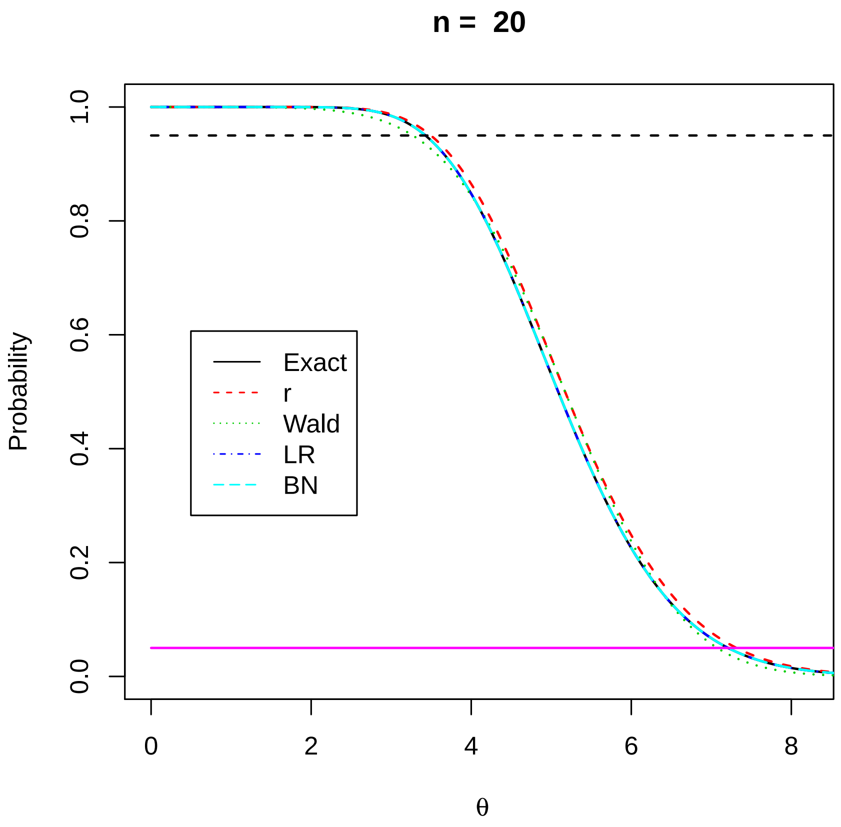


Figure 6.3: $p(\theta)$ for Data Set 3.

Table 6.1: Three simulated data sets

Data Set	Observations					Sample Size n
1	0.0413	0.1784	0.3138	0.2645	0.2235	5
2	0.1533	0.0529	0.2861	0.1332	0.8038	10
	0.0461	0.4998	0.0664	0.3269	0.2138	
3	0.0702	0.0680	0.0841	0.1256	0.0216	20
	0.5176	0.1051	0.0879	0.0760	1.0258	
	0.0862	0.0205	0.1602	0.1994	0.3186	
	0.2660	0.3802	0.0306	0.0189	0.2039	

Figures 6.1 to 6.3 display the p-value functions obtained from Exact, $\Phi(r)$, $\Phi(q)$(Wald), Equation (5.17)(LR) and Equation (5.20)(BN) for the three data sets. The horizontal lines indicate the two nominal levels, 0.95 and 0.05, for the 90% central confidence interval, respectively. Table 6.2 gives the corresponding 90% confidence intervals.

Both the plots and Table 6.2 show the outstanding performance of the third-order methods over the first-order methods, and third-order methods have remarkable accuracy when sample size is small.

Table 6.2: 90% central confidence intervals for θ

Method	$n = 5$	$n = 10$	$n = 20$
Exact	$(1.9303, 8.9603)$	(2.1003, 6.0803)	(3.4303, 7.2103)
r	$(2.1203, 9.4303)$	(2.1903, 6.2503)	(3.5003, 7.3203)
Wald	$(1.2903, 8.5003)$	(1.8603, 5.8903)	(3.2703, 7.0703)
LR	$(1.9303, 8.9603)$	(2.1003, 6.0803)	(3.4303, 7.2103)
BN	$(1.9303, 8.9603)$	(2.1003, 6.0803)	(3.4303, 7.2103)

6.2 General Exponential Family

Now, consider a general full rank exponential family model,

$$f(y;\theta) = \exp\{\varphi'(\theta)t(y) - c(\theta) + h(t(y))\} \tag{6.5}$$

with the scalar parameter of interest being $\psi = \psi(\theta)$ and $\theta = (\psi, \lambda')'$ with λ be nuisance parameters vector which is explicitly known. Under the set up of model (6.5), the natural parameter and natural variable are θ and y, and $\varphi(\theta)$ and $t(y)$ are the canonical parameter and canonical variable which are functions of the natural parameter and natural variable, respectively. We restrict our attention to the full-rank exponential family model where the canonical parameter and the natural parameter have the same dimensionality since most commonly used distributions in the exponential family belong to this category. For the curved exponential family model, that is, the dimension k of the parameter vector $\theta = (\theta_1, \cdots, \theta_k)'$ is less than the dimension s of the vector $\varphi(\theta) = (\varphi_1(\theta), \cdots, \varphi_s(\theta))'$, detailed discussion can be found in Barndorff-Nielsen (1978).

The signed log-likelihood ratio statistic, $r = r(\psi)$, is invariant to reparameterization, it remains unchanged and is same as in (6.2). However, the quantity $q = q(\psi)$ has to be reexpressed in the canonical parameter, $\varphi(\theta)$, scale.

For the general exponential model above, we have

$$\ell(\theta; y) = \varphi'(\theta)t(y) - c(\theta).$$

Fraser (1990) derived

$$\varphi'(\theta) = \dot{\ell}(\theta; y_0)\ \dot{S}^{-1}(\hat{\theta}; y_0)$$

for the observation y_0 with $\hat{\theta}$ being the corresponding MLE, where

$$\dot{\ell}(\theta; y) = \frac{\partial \ell(\theta; y)}{\partial y'}, \quad \dot{S}^{-1}(\hat{\theta}) = \frac{\partial S(\theta; y)}{\partial y'}, \quad S(\theta; y) = \frac{\partial \ell(\theta; y)}{\partial \theta}.$$

Let $\chi(\theta)$ be a rotated coordinate of $\varphi(\theta)$ that agrees with $\psi(\theta)$ at $\hat{\theta}_\psi$, is define as

$$\chi(\theta) = \frac{\varphi^{\psi}(\hat{\theta}_\psi)}{||\varphi^{\psi}(\hat{\theta}_\psi)||}\varphi(\theta), \tag{6.6}$$

which can be viewed operationally as the canonical parameter of in $\varphi(\theta)$ scale. The $\varphi^{\psi}(\theta)$ is the row of $\varphi_\theta^{-1}(\theta)$ that corresponds to ψ, $||\varphi^{\psi}(\theta)||^2$ is the square length of the vector $\varphi^{\psi}(\theta)$, and $\varphi_{\theta}(\theta)$ is the derivatives of $\varphi(\theta)$ with respect to θ. The calibrated version, $\chi(\theta)$ of $\varphi(\theta)$, is basically a vector from the space spanned by the columns of the $\varphi(\theta)$ and its direction depends on the constrained MLE for given $\varphi(\theta)$. Hence $|\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})|$ is a measure of departure of $\hat{\psi}$ from ψ in $\varphi(\theta)$ scale.

Since $\ell(\theta) = \ell(\varphi)$, by the chain rule in differentiation, we have the full and nuisance information determinants recalibrated on the $\varphi(\theta)$ scale:

$$|j_{\varphi\varphi'}(\hat{\theta})| = |j_{\theta\theta'}(\hat{\theta})||\varphi_{\theta}(\hat{\theta})|^{-2} \quad \text{and} \quad |j_{(\lambda\lambda')}(\hat{\theta}_{\psi})| = |j_{\lambda\lambda'}(\hat{\theta}_{\psi})||\varphi'_{\lambda}(\hat{\theta}_{\psi})\varphi_{\lambda}(\hat{\theta}_{\psi})|^{-1},$$

where $\varphi_{\lambda}(\theta)$ is the derivatives of $\varphi(\theta)$ with respect to λ.

An estimated variance for $\Big(\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})\Big)$ obtained by Fraser et al. (1999) in $\varphi(\theta)$ scale is:

$$\widehat{\text{var}}(\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})) = \frac{|j_{(\lambda\lambda')}(\hat{\theta}_{\psi})|}{|j_{\varphi\varphi'}(\hat{\theta})|}.$$

Thus the standardized maximum likelihood departure of ψ in $\varphi(\theta)$ scale is

$$q = q(\psi) = \text{sgn}(\hat{\psi} - \psi)|\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})| \left\{ \frac{|j_{\varphi\varphi'}(\hat{\theta})|}{|j_{(\lambda\lambda')}(\hat{\theta}_{\psi})|} \right\}^{1/2}. \tag{6.7}$$

Therefore $p(\psi)$ can be obtained by Lugannani & Rice method (5.17) and Barndorff-Nielsen method (5.20) with r and q being defined in (6.2) and (6.7) respectively. Thus, a $(1-\gamma)100\%$ confidence interval for ψ is

$$\left(\min\left\{p^{-1}(\gamma/2), p^{-1}(1-\gamma/2)\right\}, \max\left\{p^{-1}(\gamma/2), p^{-1}(1-\gamma/2)\right\}\right). \tag{6.8}$$

As an example, we consider inference for the difference of two independent normal means which has been widely studied in statistical literature. Typically, the variances are assumed to be unknown and must be estimated. When we assume equal variances, then a pooled estimate of the common variance is used and the test statistic is exactly distributed as a Student t distribution. However, without making the equality of variances assumption, the problem is then the well-known Behrens-Fisher problem, where no exact distribution of the test statistic is available. There exists many approximate solutions. In this section, we apply the third-order methods discussed in the previous chapters to the Behrens-Fisher Problem with an additional assumption that the ratio of the two variances is known, which is examined by Schechtman and Sherman (2007). More discussion can be found on the third-order methods to Behrens-Fisher Problem proposed by She et al. (2011) when the ratio of the two variances is unknown. The comparison with first-order approximation methods and the methods proposed by Schechtman and Sherman (2007) is presented.

Let $x = (x_1, ..., x_n)'$ and $y = (y_1, ..., y_m)'$ be random samples from two independent normal distribution with mean and variance (μ_x, σ_x^2) and (μ_y, σ_y^2), respectively. Assume σ_x^2 and σ_y^2 are unknown but with the ratio $\sigma_y^2/\sigma_x^2 = c$ known. Our parameter of interest is $\psi = \mu_x - \mu_y$.

The log-likelihood function can be written as

$$l(\theta) = -\frac{m+n}{2}\log(\sigma_x^2) - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i - \psi - \mu_y)^2 - \frac{1}{2c\sigma_x^2}\sum_{j=1}^{m}(y_j - \mu_y)^2$$

where $\theta = (\psi, \mu_y, \sigma_x^2)'$.

The overall MLE, $\hat{\theta} = (\hat{\psi}, \hat{\mu}_y, \hat{\sigma^2}_x)'$, are:

$$\hat{\psi} = \bar{x} - \bar{y}, \quad \hat{\mu}_y = \bar{y}, \quad \hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{c}\sum_{j=1}^{m}(y_j - \bar{y})^2}{m+n}.$$

Then the observed full information matrix $j_{\theta\theta'}(\theta)$ can be derived as

$$\begin{pmatrix} \frac{n}{\sigma_x^2} & \frac{n}{\sigma_x^2} & \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} \\ \frac{n}{\sigma_x^2} & \frac{n}{\sigma_x^2} + \frac{m}{c\sigma_x^2} & \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y} - \mu_y)}{c(\sigma_x^2)^2} \\ \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} & \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y} - \mu_y)}{c(\sigma_x^2)^2} & -\frac{m+n}{2(\sigma_x^2)^2} + \frac{1}{(\sigma_x^2)^3}(\sum (x_i - \psi - \mu_y)^2 + \frac{1}{c}\sum (y_j - \mu_y)^2) \end{pmatrix}.$$

Hence, the determinant of the observed information matrix evaluated at the overall MLE is $|j_{\theta\theta'}(\hat{\theta})| = \dfrac{mn(m+n)}{2c(\hat{\sigma}_x^2)^4}$.

The constrained MLE, $\hat{\theta}_\psi = (\psi, \tilde{\mu}_y, \tilde{\sigma}_x^2)'$, can also be obtained directly, where

$$\begin{aligned} \tilde{\mu}_y &= \frac{n}{n + m/c}\bar{x} + \frac{m/c}{n + m/c}\bar{y} - \frac{n}{n + m/c}\psi, \\ \tilde{\sigma}_x^2 &= \frac{\sum(x_i - \psi - \tilde{\mu}_y)^2 + \frac{1}{c}\sum(y_j - \tilde{\mu}_y)^2}{m+n}, \\ \text{and} \quad \tilde{\mu}_y &= \psi + \tilde{\mu}_y \end{aligned}$$

Similarly, the observed nuisance information matrix $j_{\lambda\lambda'}(\theta)$ is

$$\begin{pmatrix} \frac{n}{\sigma_x^2} + \frac{m}{c\sigma_x^2} & \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y} - \mu_y)}{c(\sigma_x^2)^2} \\ \frac{n(\bar{x} - \psi - \mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y} - \mu_y)}{c(\sigma_x^2)^2} & -\frac{m+n}{2(\sigma_x^2)^2} + \frac{1}{(\sigma_x^2)^3}(\sum (x_i - \psi - \mu_y)^2 + \frac{1}{c}\sum (y_j - \mu_y)^2) \end{pmatrix}.$$

Therefore, the determinant of constrained nuisance information matrix is

$$|j_{\lambda\lambda'}(\hat{\theta}_\psi)| \quad = \quad \frac{(m+n)(m+cn)}{2c(\tilde{\sigma}_x^2)^3}.$$

The signed log-likelihood ratio statistic, (6.2), can be written as

$$r = sgn(\hat{\psi}-\psi)\sqrt{2(m+n)log\left(\left(\frac{\tilde{\sigma}_x^2}{\hat{\sigma}_x^2}\right)^{1/2}\right)}. \tag{6.9}$$

This model is a canonical exponential family model with canonical parameter

$$\varphi(\theta) = \left(\frac{\psi+\mu_y}{\sigma_x^2},\ \frac{\mu_y}{\sigma_x^2},\ \frac{1}{\sigma_x^2}\right)'.$$

Then we have

$$\begin{aligned}
\varphi_\theta(\theta) &= \begin{pmatrix} \sigma_x^{-2} & \sigma_x^{-2} & -(\psi+\mu_y)\sigma_x^{-4} \\ 0 & \sigma_x^{-2} & -\mu_y\sigma_x^{-4} \\ 0 & 0 & -\sigma_x^{-4} \end{pmatrix}, \\
|\varphi_\theta(\theta)| &= -\sigma_x^{-8}, \\
\varphi_\lambda(\theta) &= \begin{pmatrix} \sigma_x^{-2} & -(\psi+\mu_y)\sigma_x^{-4} \\ \sigma_x^{-2} & -\mu_y\sigma_x^{-4} \\ 0 & -\sigma_x^{-4} \end{pmatrix}, \\
\varphi^\psi(\theta) &= (\sigma_x^2, -\sigma_x^2, -\psi\sigma_x^2).
\end{aligned}$$

Thus, from (6.7), we have

$$q = q(\psi) = \sqrt{\frac{mn}{m+cn}}\frac{\hat{\sigma}_x^2}{(\tilde{\sigma}_x^2)^{3/2}}(\hat{\psi}-\psi). \tag{6.10}$$

The p-value function for ψ can then be approximated from either by the Lugannani & Rice method (5.17) or by the Barndorff-Nielsen method (5.20) with third-order accuracy, where r and q are defined in (6.9) and (6.10) respectively.

Schechtman and Sherman (2007) showed that a $(1-\gamma)100\%$ confidence interval for $\psi = \mu_x - \mu_y$ under set up in the previous Section **??** is

$$\left((\bar{x}-\bar{y}) - t\left[s_p\sqrt{\frac{1}{n}+\frac{c}{m}}\right], (\bar{x}-\bar{y}) + t\left[s_p\sqrt{\frac{1}{n}+\frac{c}{m}}\right]\right)$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{j=1}^{m} y_i}{m}$, $s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, $s_y^2 = \frac{\sum_{j=1}^{m}(y_j-\bar{y})^2}{m-1}$, $s_p^2 = \frac{(n-1)s_x^2+(m-1)s_y^2/c}{n+m-2}$, and t is the $(1-\gamma/2)100^{th}$ percentile of the t-distribution with $(n+m-2)$ degrees of freedom. They claimed that their method is good in terms of size and power. To compare the accuracy of the third-order methods with the Wald method, the signed log likelihood ratio method, and the Schechtman and Sherman (2007) method, Monte Carlo simulation studies were conducted.

We generated 10,000 simulated samples for some combinations of the parameters. For each simulated sample, we calculate the 95% confidence intervals for ψ obtained by the third-order methods in Section 5.3, *i.e.* (5.17)(LR) and (5.20)(BN) with the Wald method (Wald), the signed log-likelihood ratio method (r), and the Schechtman and Sherman (2007) method (SS). For each simulated setting, we report the proportion of ψ that falls outside the lower bound of the confidence interval (lower error), the proportion of ψ that falls outside the upper bound of the confidence interval (upper error), the proportion of ψ that falls within the confidence interval (central coverage), and the average bias (Average Bias), which is defined as

$$\text{Average Bias} = \frac{|\text{lower error} - 0.025| + |\text{upper error} - 0.025|}{2}.$$

The nominal values for the lower and the upper errors, the central coverage and the average bias are 0.025, 0.025, 0.95 and 0 respectively. These values reflect the desired properties of the accuracy and symmetry of the interval estimates of ψ.

The simulation standard errors for these three quantities are 0.0022, 0.0016 and 0.0016 respectively. Results are recorded in Tables 6.3 to 6.5. It is clear that the results from Wald method and signed log-likelihood method are not satisfactory especially when the sample sizes are small. Results from the Schechtman and Sherman (2007) method, LR and BN are identical for large sample size, and almost indistinguishable for small sample sizes (they are all within 3 simulated standard errors).

Table 6.3: $\mu_x = 0, \psi = \mu_x - \mu_y = 1, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 100$ and $m = 50$.

c(known)	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.5	$Wald$	0.0268	0.0282	0.9450	0.00250
	r	0.0262	0.0275	0.9463	0.00185
	SS	0.0250	0.0264	0.9486	0.00070
	LR	0.0250	0.0264	0.9486	0.00070
	BN	0.0250	0.0264	0.9486	0.00070
1	$Wald$	0.0277	0.0273	0.9450	0.00250
	r	0.0272	0.0265	0.9463	0.00185
	SS	0.0258	0.0251	0.9491	0.00045
	LR	0.0258	0.0251	0.9491	0.00045
	BN	0.0258	0.0251	0.9491	0.00045
5	$Wald$	0.0278	0.0272	0.9450	0.00250
	r	0.0272	0.0257	0.9471	0.00145
	SS	0.0265	0.0249	0.9486	0.00080
	LR	0.0265	0.0249	0.9486	0.00080
	BN	0.0265	0.0249	0.9486	0.00080
10	$Wald$	0.0285	0.0266	0.9449	0.00255
	r	0.0280	0.0260	0.9460	0.00200
	SS	0.0268	0.0247	0.9485	0.00105
	LR	0.0268	0.0247	0.9485	0.00105
	BN	0.0268	0.0247	0.9485	0.00105

Table 6.4: $\mu_x = 0, \psi = \mu_x - \mu_y = 3, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 10$ and $m = 20$.

c(known)	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.5	$Wald$	0.0353	0.0336	0.9311	0.00945
	r	0.0322	0.0300	0.9378	0.00610
	SS	0.0273	0.0246	0.9481	0.00135
	LR	0.0276	0.0246	0.9478	0.00150
	BN	0.0276	0.0246	0.9478	0.00150
1	$Wald$	0.0363	0.0346	0.9291	0.01045
	r	0.0323	0.0308	0.9369	0.00655
	SS	0.0266	0.0258	0.9476	0.00120
	LR	0.0266	0.0258	0.9476	0.00120
	BN	0.0266	0.0258	0.9476	0.00120
5	$Wald$	0.0359	0.0321	0.9320	0.00900
	r	0.0316	0.0275	0.9409	0.00455
	SS	0.0257	0.0227	0.9516	0.00150
	LR	0.0258	0.0227	0.9515	0.00155
	BN	0.0258	0.0227	0.9515	0.00155
10	$Wald$	0.0358	0.0314	0.9328	0.00860
	r	0.0318	0.0274	0.9408	0.00460
	SS	0.0263	0.0219	0.9518	0.00220
	LR	0.0263	0.0219	0.9518	0.00220
	BN	0.0263	0.0219	0.9518	0.00220

Table 6.5: $\mu_x = 0, \psi = \mu_x - \mu_y = 5, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 5$ and $m = 5$.

c(known)	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.5	$Wald$	0.0586	0.0586	0.8828	0.03360
	r	0.0443	0.0445	0.9112	0.01940
	SS	0.0236	0.0237	0.9527	0.00135
	LR	0.0243	0.0244	0.9513	0.00065
	BN	0.0242	0.0243	0.9515	0.00075
1	$Wald$	0.0593	0.0603	0.8804	0.03480
	r	0.0456	0.0443	0.9101	0.01995
	SS	0.0244	0.0245	0.9511	0.00055
	LR	0.0254	0.0259	0.9487	0.00065
	BN	0.0252	0.0257	0.9491	0.00045
5	$Wald$	0.0590	0.0577	0.8833	0.03335
	r	0.0465	0.0440	0.9095	0.02025
	SS	0.0270	0.0240	0.9490	0.00150
	LR	0.0283	0.0251	0.9466	0.00170
	BN	0.0283	0.0249	0.9468	0.00170
10	$Wald$	0.0597	0.0569	0.8834	0.03330
	r	0.0467	0.0427	0.9106	0.01970
	SS	0.0286	0.0240	0.9474	0.00230
	LR	0.0298	0.0253	0.9449	0.00255
	BN	0.0297	0.0251	0.9452	0.00240

6.3 General Statistical Model

The key step of the third-order method is to find the canonical parameter $\varphi = \varphi(\theta)$. For a general statistical model, when sufficiency and ancillarity do not reduce the dimension of the variable to that of the parameter, an approximation of ancillary statistics could be obtained. Barndorff-Nielsen (1980) and McCullagh (1987) suggested a different ways of constructing approximate ancillary statistics but the methodologies are problem dependent. However, ancillary statistics may not exist and even if it exits, it may not be unique. Moreover, the feasible methods are lacking for tail probability approximation. Fraser (1988) developed the tangent exponential model to approximate an asymptotic model with third-order accuracy for a model with variable and parameter of the same dimension, and Fraser and Reid (1995) extended to more general case of variable of larger dimension than parameter by constructing an ancillary direction and using observed likelihood. The tangent exponential model can be fully characterized by its likelihood function and sample space derivative of the likelihood function at the data point showed by Fraser and Reid (1993). For the one dimension case, the tangent exponential model is determined by the first row and the first column of the coefficient matrix of the Taylor expansion of the log-likelihood function at both the MLE and observations.

The term ancillary direction is refer to the tangent direction to the ancillary statistic surface at the observed data point. It can be retrieved from an appropriate pivotal quantity in the applications without construction of approximate ancillary statistic (Fraser and Reid (1995)).

Consider real-valued parameter θ and response y. Assume the cumulative distribution function of Y, $F(y;\theta)$ has nonzero derivative with respect to θ and the variable y is monotone increasing in θ. Note that $F(y;\theta) \sim U(0,1)$, hence it is a pivotal quantity. The total differential for F at $\hat{\theta}$ is

$$dF(y;\theta) = F_y(y;\theta)|_{\hat{\theta},\, y^0} dy + F_\theta(y;\theta)|_{\hat{\theta},\, y^0} d\theta \tag{6.11}$$

with $F_y(y;\theta)$ and $F_\theta(y;\theta)$ being the partial derivatives of $F(y;\theta)$ with respect to both y and θ respectively. y^0 is the observed value for y. With the probability level F held constant, the ancillary direction is given by Fraser (1990) as

$$V = \frac{\partial y}{\partial \theta} = -F_y^{-1}(y;\theta)F_\theta(y;\theta)\big|_{\hat{\theta},y^0}. \tag{6.12}$$

which depicts the change in y that corresponds to the change in θ. Denote the ancillary direction as $V = (v_1, v_2, \cdots, v_d)$. V records how y changes when θ changes.

For a pivotal quantity, $z(y,\theta)$, other than the cumulative distribution function, the ancillary direction is given by

$$V = \frac{\partial y}{\partial \theta} = -z_y^{-1}(y,\theta)z_\theta(y,\theta)\bigg|_{\hat{\theta},y^0}. \tag{6.13}$$

Note that vectors $(v_1, v_2, \cdots, v_d)$ are tangent to first derivative ancillaries for parameter changes in linearly independent direction at $\theta = \hat{\theta}$. For more details on the ancillary directions tangent to the level surface of an approximate ancillary statistic see Fraser (1990), Fraser and Reid (1995,1996).

If the model is conditional model which is conditioning on some ancillary statistic, then the conditional likelihood gradient becomes the full likelihood gradient tangent to the ancillary surface. Thus Fraser et al. (1999) derive that the reparametrization $\psi(\theta)$ can be obtained by differentiating the full likelihood in the directions V

$$\begin{aligned} \varphi(\theta) &= \frac{\partial}{\partial V} l(\theta; y)\bigg|_{y^0} = l_y(\theta'; y^0)V \\ &= \left(\sum_{i=1}^{n} \frac{\partial}{\partial y_i} l(\theta; y^0) v_{i1}, \cdots, \sum_{i=1}^{n} \frac{\partial}{\partial y_i} l(\theta; y^0) v_{ik} \right). \end{aligned} \tag{6.14}$$

This quantity is the locally defined canonical parameter for the tangent exponential model at the data point y^0. So the observed log-likelihood $l(\theta; y^0)$ and the ancillary direction V together will produce a locally defined canonical parameter $\varphi(\theta)$. The tangent exponential model provides the full third-order p-values for the original model (Fraser et al. (1999)).

For a special case, when the dimension of variable y is same as the dimension of parameter θ, the new parameter $\psi(\theta)$ would be

$$\varphi = \varphi(\theta) = \ell_y(y^0; \theta) = \frac{\partial \ell(y; \theta)}{\partial y}\bigg|_{y^0}$$

Fraser (1990), and Fraser and Reid (1995) proved that to secure third-order accuracy in (5.17) and (5.20), an ancillary statistic of second-order is sufficient.

Once we have the tangent exponential family model and its locally defined canonical parameter, the methodology in Section 6.2 can be directly applied to approximating the tail probability by using either the Barndorff-Nielsen method (5.17) or the Lugannani-Rice method (5.20). Thus a $(1-\alpha) \times 100\%$ confidence interval for ψ can be obtained accordingly. The results always apply for the cases when the nuisance parameter is explicit or implicit. Computational issues encountered in obtaining constrained MLE for likelihood-based third-order method will be addressed in Chapter 7. Illustration of third-order method for general statistical model will be presented in Chapter 8. For the rest of this dissertation, we only consider the Barndorff-Nielsen method (5.17) since we can see the results from Barndorff-Nielsen method (5.17) and Lugannani-Rice method (5.20) in Section 6.2 are almost identical regardless the sample size.

7 Computational Issues

Calculating constrained MLE is one of the major steps of the likelihood-based third-order method. Computation problems always arise for the likelihood-based method because of the optimization problem, especially when the constrained MLE involves a constraint which cannot be expressed in closed-form. In our case, the constraint is equating the stress-strength reliability in a form of an integral. This problem will be discussed in Section 7.1, a penalized likelihood method is proposed in Section 7.2 to deal with this numerical complication of maximizing the constrained likelihood model, illustration of proposed penalized likelihood method using a real life example will also be presented.

7.1 Computation Problem

The constrained MLE $\hat{\theta}_\psi$ for a given ψ_0, is obtained by maximizing the log-likelihood function (5.1) subject to the constraint $R = \psi(\theta) = \psi_0$. Formally, this optimization problem can be approached using the Lagrange multiplier method. But when the constraint has no closed form representation, the standard Lagrange multiplier method will encounter numerical difficulties. In the rest of this chapter, Burr type X distribution will be used as an illustration.

The Burr type X distribution was introduced by Burr (1942). Using the notation in Raqab and Kundu (2006), let Y be distributed as a Burr type X distribution, $BurrX(\alpha, \sigma)$, the distribution function can be written as

$$F(y; \alpha, \sigma) = \left(1 - e^{-(\sigma y)^2}\right)^{\alpha} \qquad \alpha > 0,\ \sigma > 0,\ y > 0. \tag{7.1}$$

where α is the shape parameter and σ is the scale parameter. Let $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_m)'$ be the random samples from $BurrX(\alpha_1, \sigma_1)$ and $BurrX(\alpha_2, \sigma_2)$ respectively. Then the log-likelihood function is

$$\begin{aligned} & l(\alpha_1, \alpha_2, \sigma_1, \sigma_2; x, y) \\ = \;& n \log \alpha_1 + 2n \log \sigma_1 + \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} (\sigma_1 x_i)^2 + (\alpha_1 - 1) \sum_{i=1}^{n} \log \left[1 - e^{-(\sigma_1 x_i)^2}\right] \\ & + m \log \alpha_2 + 2m \log \sigma_2 + \sum_{j=1}^{m} \log y_j - \sum_{j=1}^{m} (\sigma_2 y_j)^2 + (\alpha_2 - 1) \sum_{j=1}^{m} \log \left[1 - e^{-(\sigma_2 y_j)^2}\right] \end{aligned} \tag{7.2}$$

where $\theta = (\alpha_1, \alpha_2, \sigma_1, \sigma_2)'$ and $\alpha_1, \alpha_2, \sigma_1, \sigma_2$ are positive. The parameter of interest,

ψ, is the stress-strength reliability, R, which is defined as:

$$\begin{aligned} R = R(\theta) &= P(Y < X) = \int_0^\infty \int_0^x f(x;\ \alpha_1, \sigma_1)\ f(y;\ \alpha_2, \sigma_2)\ dx\ dy \qquad (7.3) \\ &= \int_0^\infty 2\alpha_1 \sigma_1^2\ x\ e^{-(\sigma_1 x)^2}\ \left(1 - e^{-(\sigma_1 x)^2}\right)^{\alpha_1 - 1}\ \left[1 - e^{-(\sigma_2 x)^2}\right]^{\alpha_2}\ dx. \end{aligned}$$

The constrained MLE $\hat{\theta}_\psi = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\sigma}_1, \tilde{\sigma}_2)'$ for a given ψ_0, is obtained by maximizing the log-likelihood function in (7.2) subject to the constraint $R = \psi(\theta) = \psi_0$ in (7.3). In this case, the constraint on θ values is expressed by equating an integral, $R(\theta)$ (which has no closed form representation) to ψ_0. The highly nonlinear form of the constraint suggests that the standard Lagrange multiplier method will encounter numerical difficulties. Indeed, this was the case. Moreover, there is no obvious way to study convexity and other properties of the constraint set corresponding to the ψ_o^{th} contour of $R(\theta)$.

Our first attempt is to solve the integral constraint iteratively for one of the parameters (say α_1) in terms of the others and then choose the others to maximize the unconstrained likelihood but with α_1 fixed. That is, first choose any positive number as a start value for α_2, σ_1 and σ_2, denoted as $\theta_c^0 = (\alpha_2^0, \sigma_1^0, \sigma_2^0)'$, to find the optimal solution for the following function for a fixed ψ_0,

$$R(\alpha_1^0) = (R(\theta) - \psi_0)^2, \qquad (7.4)$$

where $R(\theta)$ is given by (7.3). In other words, α_1 can be expressed as $\alpha_1 = f(\alpha_2, \sigma_1, \sigma_2)$, a function of other three parameters. Then we maximize the following log-likelihood function $l(\alpha_1^0, \alpha_2, \sigma_1, \sigma_2; x, y)$ for a given α_1^0

$$\begin{aligned}
& l(\alpha_1^0, \alpha_2, \sigma_1, \sigma_2; x, y) \\
= \;& n \log \alpha_1^0 + 2n \log \sigma_1 + \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} (\sigma_1 x_i)^2 + (\alpha_1^0 - 1) \sum_{i=1}^{n} \log \left[1 - e^{-(\sigma_1 x_i)^2} \right] \\
& + m \log \alpha_2 + 2m \log \sigma_2 + \sum_{j=1}^{m} \log y_j - \sum_{j=1}^{m} (\sigma_2 y_j)^2 + (\alpha_2 - 1) \sum_{j=1}^{m} \log \left[1 - e^{-(\sigma_2 y_j)^2} \right].
\end{aligned}$$

We repeat the above procedure until the log-likelihood function $l(\alpha_1, \alpha_2, \sigma_1, \sigma_2)$ given by (7.2) is maximized. We found out that this iterative method gives relative satisfactory results when R is great than 0.42 and smaller than 0.9 on the strength data for single carbon fibre at 20-mm, 50-mm reported by Badar and Priest (1982) despite the computing time due to the integration. The strength data are recorded in Table 7.1. This is not surprising, considering the sample sizes of 65 and 69 are relatively large. Out of this range, even this method claim there is a optimal solution at some points of R, the convergence condition is not satisfied. In statistical software R, it gives the error code 52 and error message of 'ERROR: ABNORMAL_TERMINATION_IN_LNSRCH', which indicates 'NaNs' is produced in optimization. The mathematical software Matlab also delivers a warning message of 'Matrix is singular, close to singular or badly scaled. Results may be inaccurate'. We will show some of these results in Table 7.2 comparing with the proposed penalized likelihood method (Penalized) which will be introduced in the next Section 7.2

Table 7.1: The strength data for single carbon fibre at 20-mm, 50-mm.

Data Set	Obs								Sample Size
fibre 20-mm	1.312	1.314	1.479	1.552	1.700	1.803	1.861	1.865	69
	1.944	1.958	1.966	1.997	2.006	2.021	2.027	2.055	
	2.063	2.098	2.140	2.179	2.224	2.240	2.253	2.270	
	2.272	2.274	2.301	2.301	2.359	2.382	2.382	2.426	
	2.434	2.435	2.478	2.490	2.511	2.514	2.535	2.554	
	2.566	2.570	2.586	2.629	2.633	2.642	2.648	2.684	
	2.697	2.726	2.770	2.773	2.800	2.809	2.818	2.821	
	2.880	2.957	3.012	3.067	3.084	3.090	3.096	3.128	
	2.848	3.233	3.433	3.585	3.585				
fibre 50-mm	1.339	1.434	1.549	1.574	1.589	1.613	1.746	1.753	65
	1.764	1.807	1.812	1.840	1.852	1.852	1.862	1.864	
	1.931	1.952	1.974	2.019	2.051	2.055	2.058	2.088	
	2.125	2.162	2.171	2.172	2.180	2.194	2.211	2.270	
	2.272	2.280	2.299	2.308	2.335	2.349	2.356	2.386	
	2.390	2.410	2.430	2.431	2.458	2.471	2.497	2.514	
	2.558	2.577	2.593	2.601	2.604	2.620	2.633	2.670	
	2.682	2.699	2.705	2.735	2.785	3.020	3.042	3.116	
	3.174								

with optimal settings. Table 7.2 shows that the estimated parameters using penalized likelihood method reach higher likelihood than the iterative method. Moreover, when R approaches the boundary and the sample size is getting smaller, this method runs into the singularity problem. For example, when $n = 10$ or $m = 10$ and R is approach 0.1 or greater than 0.8 in the simulations.

Table 7.2: Parameter Estimates for Example using iterative and penalized method

ψ_0	Method	α_1	α_2	σ_1	σ_2	Loglikelihood
0.22	Iterative	8.9836	0.7772	10.0005	0.7772	-216.9874
	Penalized	6.6900	0.7470	10.0570	0.6433	-214.1980
0.34	Iterative	9.1430	0.7476	10.0005	0.6733	-196.0151
	Penalized	7.3164	0.7184	10.6230	0.6838	-194.9954
0.40	Iterative	8.6891	0.7215	10.0005	0.6872	-189.2763
	Penalized	7.6284	0.7063	10.9243	0.7027	-188.7765
0.97	Iterative	9.3193	0.5231	10.0000	0.9003	- 254.8785
	Penalized	12.8284	0.5842	19.3404	0.9875	-250.8684

7.2 A Penalized Likelihood Method

The penalized likelihood function, PL, is defined as:

$$PL(\theta, \psi_0; x, y) = l(\theta; x, y) - K(\psi_0)[R(\theta) - \psi_0]^N \tag{7.5}$$

where N is an even natural number and $K(\psi_0)$ is a positive constant. If $K(\psi_0)$ is sufficiently small then it is approximately true that $PL(\theta, \psi_0; x, y) = l(\theta; x, y)$. Therefore, the unconstrained likelihood model is embedded in (7.5). Even when $K(\psi_0)$ is positive, we still consider the unconstrained maximization of $PL(x, y; \theta, \psi_0)$. Given that $[R(\theta) - \psi_0]^N$ is always positive, as $K(\psi_0)$ is set to successively larger values, the term $-K(\psi_0)[R(\theta) - \psi_0]^N$ will dominate unless the model parameters are chosen so that $-K(\psi_0)[R(\theta) - \psi_0]^N$ is very close to 0 and this only occurs if the integral constraint is satisfied.

In practice, for the problems we considered, we found that N could be set to 2 and that $K(\psi_0)$ could be set to 10,000 for $0.1 < \psi_0 < 0.9$ and that a value of $K(\psi_0) = 80,000$ worked well for the remaining more extreme values of ψ_0. It is sometimes noted in the optimization literature that a sequence of $K(\psi_0)$ values may help convergence. Table 7.3 presents the parameter estimates with $K(\psi_0) = 10,000$ and ψ_0 varying from 0.1 to 0.9 when N is chosen to be 2, 4 and 6. Note that when $N = 0$, it corresponds to the normal MLE. Table 7.4 presents the parameter estimates with $N = 2$ and ψ_0 varying from 0.1 to 0.9 when $K(\psi_0)$ is chosen to be

100, 1000 and 10000.

The final consideration is that the Hessian matrix of $PL(\theta, \psi_0; x, y)$ with respect to the parameters θ must be negative definite at the optimum. In all cases we found that this condition was satisfied. In particular, all of the eigenvalues of the Hessian matrix were negative at all optima.

In practice, we obtained parameter estimates through unconstrained optimization of PL using the *optim* function in the statistical software package, R. Within *optim*, we adopted the *L-BFGS-B* algorithm. *L-BFGS-B* was developed by Byrd et al. (1995) and is a quasi-Newton optimization variant of the Broydon-Fletcher-Goldfarb-Shanno (*BFGS*) algorithm. *L-BFGS-B* conveniently allows for simple upper and lower search bounds on the parameters.

Table 7.3: Parameter Estimates for Example when $N = 2, 4$ and 6.

ψ_0	N	α_1	α_2	σ_1	σ_2	Loglikelihood
	N=0	8.7900	0.6667	12.1310	0.7706	-179.2783
$\psi_0 = 0.1$	N=2	6.0017	0.7846	9.4026	0.5929	-248.7447
	N=4	6.4828	0.7579	9.8569	0.6285	-223.1123
	N=6	6.9610	0.7258	9.6508	0.6570	-201.4173
$\psi_0 = 0.2$	N=2	6.5816	0.7525	9.9657	0.6360	-218.5981
	N=4	7.3181	0.7187	10.6288	0.6837	-195.0653
	N=6	7.9000	0.6960	11.1848	0.6960	-184.6655
$\psi_0 = 0.3$	N=2	7.1196	0.7274	10.4446	0.6710	-200.2460
	N=4	7.7399	0.7019	11.0294	0.7095	-186.9096
	N=6	8.2829	0.6827	11.5748	0.7415	-180.9538
$\psi_0 = 0.4$	N=2	7.6284	0.7063	10.9243	0.7027	-188.7765
	N=4	8.1570	0.6869	11.4459	0.7342	-181.9196
	N=6	8.6082	0.6723	11.9255	0.7602	-179.4868
$\psi_0 = 0.5$	N=2	8.1440	0.6875	11.4361	0.7334	-182.0525
	N=4	8.5513	0.6741	11.8610	0.7697	-179.6424
	N=6	8.7743	0.6672	12.1129	0.7697	-179.2799
$\psi_0 = 0.6$	N=2	8.6889	0.6697	12.0260	0.7650	-179.3391
	N=4	8.7877	0.6668	12.1277	0.7704	-179.2784
	N=6	8.7900	0.6667	12.1310	0.7706	-179.2783
$\psi_0 = 0.7$	N=2	9.3100	0.6521	12.7628	0.7999	-180.8716
	N=4	8.9172	0.6630	12.2808	0.7778	-179.3780
	N=6	8.7930	0.6666	12.1347	0.7707	-179.2784
$\psi_0 = 0.8$	N=2	10.0654	0.6335	13.8002	0.8415	-188.1891
	N=4	9.3147	0.6519	12.7653	0.8001	-180.8990
	N=6	8.8916	0.6637	12.2494	0.7763	-179.3418
$\psi_0 = 0.9$	N=2	11.2616	0.6083	15.9035	0.9072	-209.9848
	N=4	9.8395	0.6388	13.4648	0.8289	-185.4150
	N=6	9.1884	0.6554	12.6062	0.7930	-180.2209

Table 7.4: Parameter Estimates for Example when $K(\psi_0) = 100, 1000$ and 10000 with $N = 2$

ψ_0	$K(\psi_0)$	α_1	α_2	σ_1	σ_2	Loglikelihood
$\psi_0 = 0.1$	$K(\psi_0) = 100$	7.8615	0.6973	11.1511	0.7169	-185.1367
	$K(\psi_0) = 1000$	6.6187	0.7504	9.9958	0.6386	-216.9586
	$K(\psi_0) = 10000$	6.0017	0.7846	9.4026	0.5929	-248.7447
$\psi_0 = 0.2$	$K(\psi_0) = 100$	8.0299	0.6912	11.3182	0.7268	-183.1303
	$K(\psi_0) = 1000$	6.9853	0.7330	10.3297	0.6629	-203.9641
	$K(\psi_0) = 10000$	6.5816	0.7525	9.9657	0.6360	-218.5981
$\psi_0 = 0.3$	$K(\psi_0) = 100$	8.2055	0.6852	11.4971	0.6852	-181.5117
	$K(\psi_0) = 1000$	7.3807	0.7160	10.6925	0.6878	-193.5799
	$K(\psi_0) = 10000$	7.1196	0.7274	10.4446	0.6710	-200.2460
$\psi_0 = 0.4$	$K(\psi_0) = 100$	8.3869	0.6792	11.6871	0.7476	-180.3202
	$K(\psi_0) = 1000$	7.7980	0.6997	11.0889	0.7131	-186.0212
	$K(\psi_0) = 10000$	7.6284	0.7063	10.9243	0.7027	-188.7765
$\psi_0 = 0.5$	$K(\psi_0) = 100$	8.5719	0.6734	11.8869	0.7582	-179.5773
	$K(\psi_0) = 1000$	8.2378	0.6842	11.5306	0.7389	-181.2657
	$K(\psi_0) = 10000$	8.1440	0.6875	11.4361	0.7334	-182.0525
$\psi_0 = 0.6$	$K(\psi_0) = 100$	8.5847	0.6677	12.0952	0.7688	-179.2845
	$K(\psi_0) = 1000$	8.7070	0.6692	12.0378	0.7659	-179.3210
	$K(\psi_0) = 10000$	8.6889	0.6697	12.0260	0.7650	-179.3391
$\psi_0 = 0.7$	$K(\psi_0) = 100$	8.9441	0.6622	12.3098	0.7792	-179.4221
	$K(\psi_0) = 1000$	9.2134	0.6547	12.6358	0.7943	-180.3362
	$K(\psi_0) = 10000$	9.3100	0.6521	12.7628	0.7999	-180.8716
$\psi_0 = 0.8$	$K(\psi_0) = 100$	9.1268	0.6571	12.5283	0.7895	-179.9527
	$K(\psi_0) = 1000$	9.7611	0.6407	13.3519	0.8246	-184.5582
	$K(\psi_0) = 10000$	10.0654	0.6335	13.8002	0.8415	-188.1891
$\psi_0 = 0.9$	$K(\psi_0) = 100$	9.3045	0.6523	12.7483	0.7994	-180.8252
	$K(\psi_0) = 1000$	10.3388	0.6276	14.1984	0.8559	-192.0126
	$K(\psi_0) = 10000$	11.2616	0.6083	15.9035	0.9072	-209.9848

8 Application to Stress-Strength Reliability

Inferences for the stress-strength reliability, $R = P(Y < X)$, where X and Y are independently distributed, are a subject of interest in statistical literature and have been widely studied in many areas. Wolfe and Hogg (1971) considered the quantity R as a measure of difference between distributions. Hanley (1989) discussed the importance of R in medical applications. In addition, R has been examined extensively in many other areas, such as radiology, reliability and material science. This chapter will discuss the inference for $R = P(Y < X)$ in two cases, that is, when X and Y are independent Burr type X distribution and when X and Y are independent Exponentiated Exponential distribution (EE). The reason of choosing these two distributions will be reviewed in Section 8.1.1 and 8.2. A penalized likelihood method proposed in Chapter 7 is adopted to deal with the numerical complications of maximizing the constrained likelihood model with differing scale parameters for the two distributions mentioned above. Then following the general procedure of third-order method discussed in previous Section 6.3, the detailed calculation of obtaining the inference for

$R = P(Y < X)$ with Burr type X is presented Section 8.1 and with Exponentiated Exponential distribution is presented in Section 8.2. To illustrate the accuracy of third-order method, a real life example and simulation study are presented.

8.1 Inference for Stress-Strength Reliability with Burr Type X Distributions

In this section, we will briefly introduce the Burr type X distribution and present how third-order method applied to Burr type X distribution. We will first consider the case that the two independent Burr type X distribution have equal scale parameter. Then we discuss the different scale parameters case, which is not commonly discussed in literature. Finally, numerical results for both cases are presented as well.

8.1.1 Burr type X Distribution

Burr introduced Burr type X distribution in 1942, and, following the notation in Section 7.1, let Y be distributed as a Burr type X distribution, $BurrX(\alpha, \sigma)$, where α is the shape parameter and σ is the scale parameter. For convenience, the distribution function is rewritten here as

$$F(y; \alpha, \sigma) = \left(1 - e^{-(\sigma y)^2}\right)^{\alpha} \qquad \alpha > 0,\ \sigma > 0,\ y > 0.$$

$BurrX(\alpha, \sigma)$ distribution is a generalized Rayleigh distribution with no shift parameter. It is also equivalent to the exponentiated Weibull ($\kappa = 2, \alpha, \sigma$) distribution as introduced in Mudholkar and Srivastava (1993), where κ is the first shape parameter, α is the second shape parameter and σ is the scale parameter. Moreover, if $\alpha = 1$, it reduced to the Weibull $(2, \sigma)$ model. Furthermore, $BurrX(\alpha, \sigma)$ is a useful model for lifetime data because when $\alpha \geq 1/2$, the model has a monotone increasing hazard function and when $\alpha < 1/2$, then it has bathtub hazard function. Its cumulative distribution function and survival function can be explicitly written in closed form.

Various aspects of $BurrX(\alpha, 1)$ model were studied by Sartawi and Abu-Salih (1991), Jaheen (1995, 1996), and Ahmad et al. (1997). Several interesting properties of the $BurrX(\alpha, \sigma)$ model were discussed by Surles and Padgett (2005) and Raqab and Kundu (2006). Raqab and Kundu (2006) also studied the relations of $BurrX(\alpha, \sigma)$ model with gamma, Weibull, generalized exponential and exponentiated Weibull distributions. Raqab and Kundu (2005) compared several methods of estimating for R when X and Y followed independent Burr type X distributions with same scale parameters. Surles and Padgett (1998, 2001) derived inference procedures for R when X and Y are independently distributed as $BurrX(\alpha_1, \sigma_1)$ and $BurrX(\alpha_2, \sigma_2)$.

The Burr type X distribution is generally difficult to work with since it is neither a transformation family model, nor an exponential family model. Moreover, it does not have closed-form for moments or moment generating function although some approximated moments can be obtained using results of Mudholkar et al. (1995). The case of equal scale parameters has been extensively study in literature (see, Surles and Padgett 2001, 2005). And, it is often stated in literature that the methodology can be extended to the non equal scale parameters case. However, when the scale parameters are not equal, computation problems arised for any likelihood-based method because of the optimization problem for the constrained MLE with the constraint being an integral without closed-form. The penalized likelihood method discussed in chapter 7 is applied to obtain the constrained MLE. The proposed likelihood-based third-order method is then applied to obtain inference for $R = P(Y < X)$.

8.1.2 Stress-Strength Reliability with Equal Scale Parameters

Let X and Y be independently distributed as $BurrX(\alpha_1, \sigma_1)$ and $BurrX(\alpha_2, \sigma_2)$ respectively. When the scale parameters are unequal, $\sigma_1 \neq \sigma_2$, the stress-strength reliability R is a form of integral in (7.3). An exact inference method for R is not available but Surles and Padgett (2001) obtained asymptotic inference results for R based on the expected Fisher information matrix which will be discussed in next subsection 8.1.3. When scale parameters are equal, $\sigma_1 = \sigma_2 = \sigma$, $R = R(\theta)$

can be simplified to $R = \dfrac{\alpha_1}{\alpha_1 + \alpha_2}$. Let $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_m)'$ be the random samples from $BurrX(\alpha_1, \sigma)$ and $BurrX(\alpha_2, \sigma)$ respectively. The log-likelihood function of the above model can be written as

$$l(\alpha_1, \alpha_2, \sigma; x, y) = n \log \alpha_1 + m \log \alpha_2 + 2(n+m) \log \sigma + \sum_{i=1}^{n} \log x_i + \sum_{j=1}^{m} \log y_j$$

$$- \ \sigma^2 \left(\sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{m} y_j^2 \right) + (\alpha_1 - 1) \sum_{i=1}^{n} \log(1 - e^{-(\sigma x_i)^2}) + (\alpha_2 - 1) \sum_{j=1}^{m} \log(1 - e^{-(\sigma y_j)^2}).$$

Denote the overall maximum likelihood estimate as $\hat{\theta} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma})'$, then observed information matrix $j_{\theta\theta}(\hat{\theta})$ can be obtained as

$$j_{\theta\theta}(\hat{\theta}) = \begin{pmatrix} \dfrac{n}{\hat{\alpha}_1^2} & 0 & -2\displaystyle\sum_{i=1}^{n} \dfrac{\hat{\sigma} \ x_i^2 \ e^{-(\hat{\sigma} x_i)^2}}{1 - e^{-(\hat{\sigma} x_i)^2}} \\ 0 & \dfrac{m}{\hat{\alpha}_2^2} & -2\displaystyle\sum_{j=1}^{m} \dfrac{\hat{\sigma} \ y_j^2 \ e^{-(\hat{\sigma} y_j)^2}}{1 - e^{-(\hat{\sigma} y_i)^2}} \\ -2\displaystyle\sum_{i=1}^{n} \dfrac{\hat{\sigma} \ x_i^2 \ e^{-(\hat{\sigma} x_i)^2}}{1 - e^{-(\hat{\sigma} x_i)^2}} & -2\displaystyle\sum_{j=1}^{m} \dfrac{\hat{\sigma} \ y_j^2 \ e^{-(\hat{\sigma} y_j)^2}}{1 - e^{-(\hat{\sigma} y_i)^2}} & j_{\sigma\sigma}(\hat{\theta}) \end{pmatrix},$$

where $j_{\sigma\sigma}(\hat{\theta}) = \dfrac{2(n+m)}{\hat{\sigma}^2} - 2(\hat{\alpha}_1 - 1) \displaystyle\sum_{i=1}^{n} \dfrac{x_i^2 \ e^{-(\hat{\sigma} x_i)^2} \ (1 - e^{-(\hat{\sigma} x_i)^2} - 2\hat{\sigma}^2 \ x_i^2)}{\left(1 - e^{-(\hat{\sigma} x_i)^2}\right)^2}$

$$+ \, 2 \left(\sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{m} y_j^2 \right) - 2(\hat{\alpha_2} - 1) \sum_{j=1}^{m} \frac{y_j^2 \ e^{-(\hat{\sigma} y_j)^2} \ (1 - e^{-(\hat{\sigma} y_j)^2} - 2\hat{\sigma}^2 \ y_j^2)}{\left(1 - e^{-(\hat{\sigma} y_j)^2}\right)^2}.$$

The parameter of interest is $\psi(\theta) = \dfrac{\alpha_1}{\alpha_1 + \alpha_2}$. Now we define the tilted log-likelihood function $\tilde{l}(\theta)$ as

$$\tilde{l}(\theta) = n \log \alpha_1 + m \log \alpha_2 + 2(n+m) \log \sigma + \sum_{i=1}^{n} \log x_i + \sum_{j=1}^{m} \log y_j - \sigma^2 \left(\sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{m} y_j^2 \right)$$

$$+ \, (\alpha_1 - 1) \sum_{i=1}^{n} \log(1 - e^{-(\sigma x_i)^2}) + (\alpha_2 - 1) \sum_{j=1}^{m} \log(1 - e^{-(\sigma y_j)^2}) + \hat{\kappa}[\psi(\theta) - \psi]. \tag{8.1}$$

Then penalized likelihood method is adopted to obtain the constrained MLE $\hat{\theta}_\psi = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\sigma})'$. Once $\hat{\theta}_\psi$ is obtained, we can calculate $\hat{\kappa}$, the Lagrange multiplier, by solving

$$\frac{\partial \ell(\hat{\theta}_\psi)}{\partial \alpha_1} - \hat{\kappa}\frac{\partial R(\hat{\theta}_\psi)}{\partial \alpha_1} = 0 \ .$$

Hence, the tilted log-likelihood function can be obtained by (8.1). Similarly, the constrained observed information matrix $\tilde{j}_{\theta\theta}(\hat{\theta}_\psi)$ is obtained as

$$\tilde{j}_{\theta\theta}(\hat{\theta}_\psi) = \begin{pmatrix} \dfrac{n}{\tilde{\alpha}_1^2} + \dfrac{2\hat{\kappa}\tilde{\alpha}_2}{(\tilde{\alpha}_1+\tilde{\alpha}_2)^3} & -\dfrac{\hat{\kappa}(\tilde{\alpha}_1-\tilde{\alpha}_2)}{(\tilde{\alpha}_1+\tilde{\alpha}_2)^3} & -2\displaystyle\sum_{i=1}^{n}\dfrac{\tilde{\sigma}\ x_i^2\ e^{-(\tilde{\sigma}x_i)^2}}{1-e^{-(\tilde{\sigma}x_i)^2}} \\ -\dfrac{\hat{\kappa}(\tilde{\alpha}_1-\tilde{\alpha}_2)}{(\tilde{\alpha}_1+\tilde{\alpha}_2)^3} & \dfrac{m}{\tilde{\alpha}_2^2} - \dfrac{2\hat{\kappa}\tilde{\alpha}_1}{(\tilde{\alpha}_1+\tilde{\alpha}_2)^3} & -2\displaystyle\sum_{j=1}^{m}\dfrac{\tilde{\sigma}\ y_j^2\ e^{-(\tilde{\sigma}y_j)^2}}{1-e^{-(\tilde{\sigma}y_i)^2}} \\ -2\displaystyle\sum_{i=1}^{n}\dfrac{\tilde{\sigma}\ x_i^2\ e^{-(\tilde{\sigma}x_i)^2}}{1-e^{-(\tilde{\sigma}x_i)^2}} & -2\displaystyle\sum_{j=1}^{m}\dfrac{\tilde{\sigma}\ y_j^2\ e^{-(\tilde{\sigma}y_j)^2}}{1-e^{-(\tilde{\sigma}y_i)^2}} & \tilde{j}_{\sigma\sigma}(\hat{\theta}_\psi) \end{pmatrix},$$

where $\tilde{j}_{\sigma\sigma}(\hat{\theta}_\psi) = \dfrac{2(n+m)}{\tilde{\sigma}^2} - 2(\tilde{\alpha}_1 - 1)\displaystyle\sum_{i=1}^{n}\dfrac{x_i^2\ e^{-(\tilde{\sigma}x_i)^2}\ (1-e^{-(\tilde{\sigma}x_i)^2} - 2\tilde{\sigma}^2\ x_i^2)}{\left(1-e^{-(\tilde{\sigma}x_i)^2}\right)^2}$

$$+ 2\left(\sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{m} y_j^2\right) - 2(\tilde{\alpha}_2 - 1)\sum_{j=1}^{m}\frac{y_j^2\ e^{-(\tilde{\sigma}y_j)^2}\ (1-e^{-(\tilde{\sigma}y_j)^2} - 2\tilde{\sigma}^2\ y_j^2)}{\left(1-e^{-(\tilde{\sigma}y_j)^2}\right)^2}.$$

Now, we have everything to calculate $r(\psi)$. The next step is to find the ancillary direction V. First, let $w = (x_1, \ldots, x_n,\ y_1, \ldots, y_m)'$ be the observed data and $z = (\alpha_1 \log(1-e^{-(\sigma x_1)^2}),\ \ldots,\ \alpha_1 \log(1-e^{-(\sigma x_n)^2}),\ \alpha_2 \log(1-e^{-(\sigma y_1)^2}),\ \ldots,\ \alpha_2 \log(1-e^{-(\sigma y_m)^2}))'$ be the vector-pivotal quantity. Note that z is continuously differentiable, and has one-one mappings between z_k, y_k for each k. Then,

$$\frac{\partial z_i}{\partial w_i} = \begin{cases} \dfrac{2\alpha_1\ \sigma^2\ x_i\ e^{-(\sigma x_i)^2}}{1-e^{-(\sigma x_i)^2}}\ , & \text{if} \quad 1 \le i \le n \\ \dfrac{2\alpha_2\ \sigma^2\ y_{i-n}\ e^{-(\sigma y_{i-n})^2}}{1-e^{-(\sigma y_{i-n})^2}}\ , & \text{if} \quad (n+1) \le i \le (n+m) \end{cases} \qquad (8.2)$$

and $\dfrac{\partial z_i}{\partial w_j} = 0$ for all $i \neq j$. Hence, we have the ancillary direction V

$$
\begin{aligned}
V &= (V_1, V_2, V_3) \\
&= \begin{pmatrix}
-\dfrac{(1-e^{-(\hat{\sigma}x_1)^2})\log(1-e^{-(\hat{\sigma}x_1)^2})}{2\hat{\alpha}_1\ \hat{\sigma}^2\ x_1\ e^{-(\hat{\sigma}x_1)^2}} & 0 & -\dfrac{x_1}{\hat{\sigma}} \\
\vdots & \vdots & \vdots \\
-\dfrac{(1-e^{-(\hat{\sigma}x_n)^2})\log(1-e^{-(\hat{\sigma}x_n)^2})}{2\hat{\alpha}_1\ \hat{\sigma}^2\ x_n\ e^{-(\hat{\sigma}x_n)^2}} & 0 & -\dfrac{x_n}{\hat{\sigma}} \\
0 & -\dfrac{(1-e^{-(\hat{\sigma}y_1)^2})\log(1-e^{-(\hat{\sigma}y_1)^2})}{2\hat{\alpha}_2\ \hat{\sigma}^2\ y_1\ e^{-(\hat{\sigma}y_1)^2}} & -\dfrac{y_1}{\hat{\sigma}} \\
\vdots & \vdots & \vdots \\
0 & -\dfrac{(1-e^{-(\hat{\sigma}y_m)^2})\log(1-e^{-(\hat{\sigma}y_m)^2})}{2\hat{\alpha}_2\ \hat{\sigma}^2\ y_m\ e^{-(\hat{\sigma}y_m)^2}} & -\dfrac{y_m}{\hat{\sigma}}
\end{pmatrix}.
\end{aligned}
$$

Thus, the locally defined canonical parameter from (6.14) is

$$\varphi(\theta) = \left(\sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\ V_{1i},\quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\ V_{2i},\quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i} V_{3i} \right)',$$

where

$$
\begin{cases}
\dfrac{\partial l(\theta)}{\partial w_i} = \dfrac{1}{x_i} - 2\sigma^2 x_i + (\alpha_1 - 1)\dfrac{2\sigma^2\ x_i\ e^{-(\sigma x_i)^2}}{1 - e^{-(\sigma x_i)^2}}\ , & \text{if} \quad 1 \leq i \leq n \\
\dfrac{\partial l(\theta)}{\partial w_i} = \dfrac{1}{y_{i-n}} - 2\sigma^2 y_{i-n} + (\alpha_2 - 1)\dfrac{2\sigma^2\ y_{i-n}\ e^{-(\sigma y_{i-n})^2}}{1 - e^{-(\sigma y_{i-n})^2}}\ , & \text{if} \quad (n+1) \leq i \leq (n+m)
\end{cases}
$$

and its derivative is

$$
\begin{aligned}
\varphi_\theta(\theta) &= \frac{\partial \varphi(\theta)}{\partial \theta} \\
&= \begin{pmatrix}
\displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_1} V_{1i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_2} V_{1i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma} V_{1i} \\
\displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_1}\ V_{2i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_2}\ V_{2i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma} V_{2i} \\
\displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_1} V_{3i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_2} V_{3i} & \displaystyle\sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma} V_{3i}
\end{pmatrix}
\end{aligned}
\tag{8.3}
$$

Finally,

$$\psi_\theta(\theta) = \left(\frac{\partial R(\theta)}{\partial \alpha_1} \quad \frac{\partial R(\theta)}{\partial \alpha_2} \quad \frac{\partial R(\theta)}{\partial \sigma} \right).$$

Therefore, $\chi(\theta)$ can be calculated from (6.6), $q(\psi)$ can be calculated from (6.7), $r^*(\psi)$ can be obtained from (5.19), and the $(1-\gamma)100\%$ confidence interval for R can be obtained. Since the results from BN (5.17) and LN (5.20) are almost identical, we will only consider BN hereafter.

Here, we analyze the strength data for single carbon fibre reported by Badar and Priest (1982). Badar and Priest reported the strength data for single carbon fibre at 20-mm, 50-mm, 150-mm and 300-mm gauge length which is given in Table 7.1. Surles and Padgett (1998, 2001) drew inferences for $R = P(Y < X)$ on 20-mm and 50-mm gauge lengths, where X represent the strength of 20-mm fiber and Y represents the strength of 50-mm fiber, with sample sizes of 69 and 65, respectively. Assuming X and Y follow Burr type X distributions with equal scale parameters, Surles and Padgett (1998) reported the MLE of R is $\hat{R} = 0.57284$, and concluded that R is greater than 0.5 using the approximate inference procedures they suggested. This is consistent with the fact of that short fibers tends to be stronger than long ones.

Assume that X and Y are independently distributed as $BurrX(\alpha_1, \sigma)$ and $BurrX(\alpha_2, \sigma)$ respectively. We are interested in testing $H_0 : R = 0.5$ vs. $H_1 : R > 0.5$ where

$R = P(Y < X)$ is given in (7.3). Using the techniques discussed in Chapter 6, the p-values are 0.0426, 0.0493 and 0.0503 obtained by the Wald method, the signed log-likelihood ratio method and the proposed method, respectively. At the 5% significance level, the proposed third-order method gives different conclusion than the other two first-order methods. That is, proposed third-order method fail to reject H_0 at 5% level whereas the first-order methods reject H_0 at 5% level. Moreover, the 90% and 95% confidence intervals (CI) for R based on Wald, r and proposed third-order (Proposed) intervals are presented in Table 8.1. In examining Table 8.1, we observe that all three methods give similar interval estimations. The result is not surprising because both samples are relatively large.

Table 8.1: Interval Estimates of ψ for Example assuming equal scale parameter

	90% Confidence Interval	95% Confidence Interval
$Wald$	(0.5032, 0.6425)	(0.4899, 0.6558)
r	(0.5003, 0.6429)	(0.4863, 0.6559)
$Proposed$	(0.4999, 0.6427)	(0.4858, 0.6557)

To examine the accuracy of the methods discussed in this dissertation, simulation studies are performed. To compare the accuracy of the proposed method with the Wald method and the signed log-likelihood ratio method, Monte Carlo simulation studies were conducted.

For each parameter configuration and for each sample size, we generate 10,000 random samples from Burr type X distributions by using the following transformation:

$$T = \frac{1}{\sigma}\left[-\log(1 - U^{\frac{1}{\alpha}})\right]^{\frac{1}{2}} \tag{8.4}$$

where U is a uniform variate between 0 and 1. For each simulated sample, we calculated the 95% confidence interval for ψ obtained by the proposed method (Proposed) with the Wald method (Wald), and the signed log-likelihood ratio method (r). For each simulated setting, we report the same comparison criterions used in Chapter 6: lower error, upper error, central coverage and the average bias. The nominal values for the lower and the upper errors, the central coverage and the average bias are 0.025, 0.025, 0.95 and 0 respectively. These values reflect the desired properties of the accuracy and symmetry of the interval estimates of R.

Tables 8.2, 8.3 and 8.4 present the simulation results for the cases that $\sigma_1 = \sigma_2 = 2$. More specifically, we present results using the parameters setting: $\sigma_1 = \sigma_2 = 2, \alpha_1 = 5$ and $R = 0.1(0.1)0.9$ with $(n, m) = (10, 10), (10, 50)$ and $(50, 10)$. Note that α_2 is uniquely determined from $R = \alpha_1/(\alpha_1 + \alpha_2)$.

Table 8.2: $\sigma_1 = \sigma_2 = 2, \alpha_1 = 5, (n, m) = (10, 10)$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
	Wald	0.1605	0.0024	0.8371	0.07905
0.1	r	0.0506	0.0176	0.9318	0.01650
	Proposed	0.0217	0.0249	0.9534	0.00170
	Wald	0.1223	0.0122	0.8655	0.05505
0.2	r	0.0455	0.0232	0.9313	0.01115
	Proposed	0.0246	0.0255	0.9499	0.00045
	Wald	0.0971	0.0234	0.8795	0.03685
0.3	r	0.0432	0.0261	0.9307	0.00965
	Proposed	0.0252	0.0246	0.9502	0.00030
	Wald	0.0740	0.0369	0.8891	0.03045
0.4	r	0.0393	0.0298	0.9309	0.00955
	Proposed	0.0261	0.0250	0.9489	0.00055
	Wald	0.0560	0.0540	0.8900	0.03000
0.5	r	0.0356	0.0337	0.9307	0.00965
	Proposed	0.0255	0.0257	0.9488	0.00060
	Wald	0.0391	0.0716	0.8893	0.03035
0.6	r	0.0310	0.0382	0.9308	0.00960
	Proposed	0.0264	0.0241	0.9495	0.00115
	Wald	0.0246	0.0925	0.8829	0.03395
0.7	r	0.0274	0.0414	0.9312	0.00940
	Proposed	0.0257	0.0245	0.9498	0.00060
	Wald	0.0127	0.1164	0.8709	0.05185
0.8	r	0.0233	0.0456	0.9311	0.01115
	Proposed	0.0250	0.0246	0.9504	0.00020
	Wald	0.0031	0.1459	0.8510	0.07140
0.9	r	0.0185	0.0473	0.9342	0.01440
	Proposed	0.0238	0.0241	0.9521	0.00105

Table 8.3: $\sigma_1 = \sigma_2 = 2, \alpha_1 = 5, (n, m) = (10, 50)$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.0730	0.0093	0.9177	0.03185
	r	0.0302	0.0269	0.9429	0.00355
	Proposed	0.0241	0.0258	0.9501	0.00085
0.2	Wald	0.0565	0.0215	0.9220	0.01750
	r	0.0277	0.0289	0.9434	0.00330
	Proposed	0.0247	0.0253	0.9500	0.00030
0.3	Wald	0.0447	0.0324	0.9229	0.01355
	r	0.0259	0.0296	0.9445	0.00275
	Proposed	0.0239	0.0249	0.9512	0.00060
0.4	Wald	0.0326	0.0443	0.9231	0.01345
	r	0.0249	0.0317	0.9434	0.00340
	Proposed	0.0252	0.0234	0.9514	0.00090
0.5	Wald	0.0235	0.0547	0.9218	0.01560
	r	0.0233	0.0326	0.9441	0.00465
	Proposed	0.0244	0.0233	0.9523	0.00115
0.6	Wald	0.0165	0.0657	0.9178	0.02460
	r	0.0227	0.0334	0.9439	0.00535
	Proposed	0.0261	0.0228	0.9511	0.00165
0.7	Wald	0.0098	0.0782	0.912	0.03420
	r	0.0211	0.0337	0.9452	0.00630
	Proposed	0.0259	0.0219	0.9522	0.00200
0.8	Wald	0.0054	0.0947	0.8999	0.04465
	r	0.0201	0.0354	0.9445	0.00765
	Proposed	0.0263	0.0221	0.9516	0.00210
0.9	Wald	0.0018	0.1144	0.8838	0.05630
	r	0.0171	0.0401	0.9428	0.01150
	Proposed	0.0241	0.0223	0.9536	0.00180

Table 8.4: $\sigma_1 = \sigma_2 = 2, \alpha_1 = 5, (n, m) = (50, 10)$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1114	0.0009	0.8877	0.05525
	r	0.0402	0.0158	0.9440	0.01220
	Proposed	0.0228	0.0241	0.9531	0.00155
0.2	Wald	0.0909	0.0042	0.9049	0.04335
	r	0.0379	0.0191	0.9430	0.00940
	Proposed	0.0226	0.0249	0.9525	0.00125
0.3	Wald	0.0781	0.0098	0.9121	0.03415
	r	0.0347	0.0219	0.9434	0.00640
	Proposed	0.0228	0.0259	0.9513	0.00155
0.4	Wald	0.0650	0.0182	0.9168	0.02340
	r	0.0340	0.0228	0.9432	0.00560
	Proposed	0.0232	0.0258	0.9510	0.00130
0.5	Wald	0.0542	0.0250	0.9208	0.01460
	r	0.0324	0.0245	0.9431	0.00395
	Proposed	0.0235	0.0258	0.9507	0.00115
0.6	Wald	0.0427	0.0342	0.9231	0.01345
	r	0.0309	0.0255	0.9436	0.00320
	Proposed	0.0245	0.0256	0.9499	0.00055
0.7	Wald	0.0319	0.0425	0.9256	0.01220
	r	0.0287	0.0275	0.9438	0.00310
	Proposed	0.0232	0.0263	0.9505	0.00155
0.8	Wald	0.0201	0.0531	0.9268	0.01650
	r	0.0282	0.0280	0.9438	0.00310
	Proposed	0.0242	0.0250	0.9508	0.00040
0.9	Wald	0.0087	0.0667	0.9246	0.02900
	r	0.0273	0.0287	0.9440	0.00300
	Proposed	0.0253	0.0249	0.9498	0.00020

It is clear that the coverage probabilities for R are poor and the two-tail error probabilities are extremely asymmetric from the Wald method in all three settings. The results from signed log-likelihood method are not satisfactory. The proposed method gives not only an almost exact coverage probability but also it has symmetric two-tail error probabilities even for small sample sizes.

8.1.3 Stress-Strength Reliability with Unequal Scale Parameters

When scale parameters are unequal, $\sigma_1 \neq \sigma_2$, R, in (7.3) becomes difficult to calculate since integral has no known closed-form. Let $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_m)'$ be the random samples from $BurrX(\alpha_1, \sigma_1)$ and $BurrX(\alpha_2, \sigma_2)$ respectively. By maximizing the log-likelihood function in (7.2), the overall MLE can be obtained. Denote the overall MLE as $\hat{\theta} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}_1, \hat{\sigma}_2)'$. The observed information matrix evaluated at $\hat{\theta}$ is

$$j_{\theta\theta}(\hat{\theta}) =$$

$$\begin{pmatrix} \dfrac{n}{\hat{\alpha}_1^2} & 0 & -2\displaystyle\sum_{i=1}^{n} \frac{\hat{\sigma_1}\ x_i^2\ e^{-(\hat{\sigma_1} x_i)^2}}{1 - e^{-(\hat{\sigma_1} x_i)^2}} & 0 \\ 0 & \dfrac{m}{\hat{\alpha}_2^2} & 0 & -2\displaystyle\sum_{j=1}^{m} \frac{\hat{\sigma_2}\ y_j^2\ e^{-(\hat{\sigma_2} y_j)^2}}{1 - e^{-(\hat{\sigma_2} y_i)^2}} \\ -2\displaystyle\sum_{i=1}^{n} \frac{\hat{\sigma_1}\ x_i^2\ e^{-(\hat{\sigma_1} x_i)^2}}{1 - e^{-(\hat{\sigma_1} x_i)^2}} & 0 & j_{\sigma_1\sigma_1}(\hat{\theta}) & 0 \\ 0 & -2\displaystyle\sum_{j=1}^{m} \frac{\hat{\sigma_2}\ y_j^2\ e^{-(\hat{\sigma_2} y_j)^2}}{1 - e^{-(\hat{\sigma_2} y_i)^2}} & 0 & j_{\sigma_2\sigma_2}(\hat{\theta}) \end{pmatrix},$$

where $j_{\sigma_1\sigma_1}(\hat{\theta}) = \dfrac{2n}{\hat{\sigma_1}^2} + 2\sum_{i=1}^{n} x_i^2 - 2(\hat{\alpha}_1 - 1)\sum_{i=1}^{n} \dfrac{x_i^2\ e^{-(\hat{\sigma_1}x_i)^2}\ (1 - e^{-(\hat{\sigma_1}x_i)^2} - 2\hat{\sigma_1}^2\ x_i^2)}{(1 - e^{-(\hat{\sigma_1}x_i)^2})^2}$

and $j_{\sigma_2\sigma_2}(\hat{\theta}) = \dfrac{2m}{\hat{\sigma_2}^2} + 2\sum_{j=1}^{m} y_j^2 - 2(\hat{\alpha_2} - 1)\sum_{j=1}^{m} \dfrac{y_j^2\ e^{-(\hat{\sigma_2}y_j)^2}\ (1 - e^{-(\hat{\sigma_2}y_j)^2} - 2\hat{\sigma_2}^2\ y_j^2)}{\left(1 - e^{-(\hat{\sigma_2}y_j)^2}\right)^2}$.

The constrained MLE $\hat{\theta}_\psi = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\sigma}_1, \tilde{\sigma}_2)'$ for a given ψ_0, is obtained by penalized likelihood method discussed in Chapter 7 subject to the constraint $R = \psi(\theta) = \psi_0$. Once $\hat{\theta}_\psi$ is obtained, we can calculate $\hat{\kappa}$, the Lagrange multiplier, by solving

$$\frac{\partial \ell(\hat{\theta}_\psi)}{\partial \alpha_1} - \hat{\kappa}\frac{\partial R(\hat{\theta}_\psi)}{\partial \alpha_1} = 0\ .$$

Hence, the tilted log-likelihood function can be obtained by (5.2). Thus, the observed information matrix for the tilted log-likelihood function evaluated at constrained MLE $\hat{\theta}_\psi$ is $\tilde{j}_{\theta\theta}(\hat{\theta}_\psi) = -\tilde{l}_{\theta\theta}(\hat{\theta}_\psi)$ and can be written as

$$\tilde{j}_{\theta\theta}(\hat{\theta}_\psi) = \begin{pmatrix} \tilde{j}_{\alpha_1\alpha_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\sigma_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\sigma_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\alpha_1\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\sigma_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\sigma_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\alpha_1\sigma_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\sigma_2}(\hat{\theta}_\psi) & \tilde{j}_{\sigma_1\sigma_1}(\hat{\theta}_\psi) & \tilde{j}_{\sigma_1\sigma_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\alpha_1\sigma_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\sigma_2}(\hat{\theta}_\psi) & \tilde{j}_{\sigma_1\sigma_2}(\hat{\theta}_\psi) & \tilde{j}_{\sigma_2\sigma_2}(\hat{\theta}_\psi) \end{pmatrix}$$

where

- $\tilde{j}_{\alpha_1\alpha_1}(\hat{\theta}_\psi) = \dfrac{n}{\tilde{\alpha}_1^2} - \hat{\kappa}\ R_{\alpha_1\alpha_1}(\hat{\theta}_\psi)$, where $R_{\alpha_1\alpha_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_1\alpha_1}\right|_{\hat{\theta}_\psi}$
- $\tilde{j}_{\alpha_1\alpha_2}(\hat{\theta}_\psi) = -\hat{\kappa}\ R_{\alpha_1\alpha_2}(\hat{\theta}_\psi)$, where $R_{\alpha_1\alpha_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_1\alpha_2}\right|_{\hat{\theta}_\psi}$
- $\tilde{j}_{\alpha_1\sigma_1}(\hat{\theta}_\psi) = -2\sum_{i=1}^{n} \dfrac{\tilde{\sigma_1}\ x_i^2\ e^{-(\tilde{\sigma_1}x_i)^2}}{1 - e^{-(\tilde{\sigma_1}x_i)^2}} - \hat{\kappa}\ R_{\alpha_1\sigma_1}(\hat{\theta}_\psi)$, where $R_{\alpha_1\sigma_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_1\sigma_1}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\alpha_1\sigma_2}(\hat{\theta}_\psi) = -\hat{\kappa}\; R_{\alpha_1\sigma_2}(\hat{\theta}_\psi)$, where $R_{\alpha_1\sigma_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_1 \sigma_2}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\alpha_2\alpha_2}(\hat{\theta}_\psi) = \dfrac{m}{\tilde{\alpha}_2^2} - \hat{\kappa}\; R_{\alpha_2\alpha_2}(\hat{\theta}_\psi)$, where $R_{\alpha_2\alpha_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_2 \alpha_2}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\alpha_2\sigma_1}(\hat{\theta}_\psi) = -\hat{\kappa}\; R_{\alpha_2\sigma_1}(\hat{\theta}_\psi)$, where $R_{\alpha_2\sigma_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_2 \sigma_1}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\alpha_2\sigma_2}(\hat{\theta}_\psi) = -2\displaystyle\sum_{j=1}^{m} \dfrac{\tilde{\sigma}\; y_j^2\; e^{-(\tilde{\sigma} y_j)^2}}{1 - e^{-(\tilde{\sigma} y_i)^2}} - \hat{\kappa}\; R_{\alpha_2\sigma_2}(\hat{\theta}_\psi)$, where $R_{\alpha_2\sigma_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \alpha_2 \sigma_2}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\sigma_1\sigma_1}(\hat{\theta}_\psi) = \dfrac{2n}{\tilde{\sigma_1}^2} + 2\displaystyle\sum_{i=1}^{n} x_i^2 - 2(\tilde{\alpha}_1 - 1)\sum_{i=1}^{n} \dfrac{x_i^2\; e^{-(\tilde{\sigma_1} x_i)^2}\; (1 - e^{-(\tilde{\sigma_1} x_i)^2} - 2\tilde{\sigma_1}^2\; x_i^2)}{(1 - e^{-(\tilde{\sigma_1} x_i)^2})^2} -$ $\hat{\kappa}\; R_{\sigma_1\sigma_1}(\hat{\theta}_\psi)$, where $R_{\sigma_1\sigma_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \sigma_1 \sigma_1}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\sigma_1\sigma_2}(\hat{\theta}_\psi) = -\hat{\kappa}\; R_{\sigma_1\sigma_2}(\hat{\theta}_\psi)$, where $R_{\sigma_1\sigma_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \sigma_1 \sigma_2}\right|_{\hat{\theta}_\psi}$

- $\tilde{j}_{\sigma_2\sigma_2}(\hat{\theta}_\psi) = \dfrac{2m}{\tilde{\sigma_2}^2} + 2\displaystyle\sum_{j=1}^{m} y_j^2 - 2(\tilde{\alpha}_2 - 1)\sum_{j=1}^{m} \dfrac{y_j^2\; e^{-(\tilde{\sigma_2} y_j)^2}\; (1 - e^{-(\tilde{\sigma_2} y_j)^2} - 2\tilde{\sigma_2}^2\; y_j^2)}{\left(1 - e^{-(\tilde{\sigma_2} y_j)^2}\right)^2} -$ $\hat{\kappa}\; R_{\sigma_2\sigma_2}(\hat{\theta}_\psi)$, where $R_{\sigma_2\sigma_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial R(\theta)}{\partial \sigma_2 \sigma_2}\right|_{\hat{\theta}_\psi}$

Now, let $w = (x_1, \ldots, x_n,\; y_1, \ldots, y_m)'$ be the observed data and $z = (z_1, \ldots, z_n, z_{n+1}, \ldots, z_{n+m})'$ denoted the vector-pivotal quantity, in this case, $z = (\log F(x_1; \alpha_1, \sigma_1),\; \ldots,\; \log F(x_n; \alpha_1, \sigma_1),$ $\log F(y_1; \alpha_2, \sigma_2),\; \ldots,\; \log F(y_m; \alpha_2, \sigma_2))'$ and

$$\frac{\partial z_i}{\partial w_i} = \begin{cases} \dfrac{2\alpha_1\; \sigma_1^2\; x_i\; e^{-(\sigma_1 x_i)^2}}{1 - e^{-(\sigma_1 x_i)^2}}\;, & \text{if} \quad 1 \leq i \leq n \\ \dfrac{2\alpha_2\; \sigma_2^2\; y_{i-n}\; e^{-(\sigma_2 y_{i-n})^2}}{1 - e^{-(\sigma_2 y_{i-n})^2}}\;, & \text{if} \quad (n+1) \leq i \leq (n+m) \end{cases} \tag{8.5}$$

and $\dfrac{\partial z_i}{\partial w_j} = 0$ for all $i \neq j$. Hence, from (6.13), we have

$$
\begin{aligned}
V \;&= (V_1, V_2, V_3, V_4) \qquad\qquad (8.6)\\
&= \begin{pmatrix}
-\dfrac{\log(1-e^{-(\hat{\sigma_1}x_1)^2})(1-e^{-(\hat{\sigma_1}x_1)^2})}{2\hat{\alpha}_1\ \hat{\sigma_1}^2\ x_1\ e^{-(\hat{\sigma_1}x_1)^2}} & 0 & -\dfrac{x_1}{\hat{\sigma_1}} & 0\\
\vdots & \vdots & \vdots & \vdots\\
-\dfrac{\log(1-e^{-(\hat{\sigma_1}x_n)^2})(1-e^{-(\hat{\sigma_1}x_n)^2})}{2\hat{\alpha}_1\ \hat{\sigma_1}^2\ x_n\ e^{-(\hat{\sigma_1}x_n)^2}} & 0 & -\dfrac{x_n}{\hat{\sigma_1}} & 0\\
0 & -\dfrac{\log(1-e^{-(\hat{\sigma_2}y_1)^2})(1-e^{-(\hat{\sigma_2}y_1)^2})}{2\hat{\alpha}_2\ \hat{\sigma_2}^2\ y_1\ e^{-(\hat{\sigma_2}y_1)^2}} & 0 & -\dfrac{y_1}{\hat{\sigma_2}}\\
\vdots & \vdots & \vdots & \vdots\\
0 & -\dfrac{\log(1-e^{-(\hat{\sigma_2}y_m)^2})(1-e^{-(\hat{\sigma_2}y_m)^2})}{2\hat{\alpha}_2\ \hat{\sigma_2}^2\ y_m\ e^{-(\hat{\sigma_2}y_m)^2}} & 0 & -\dfrac{y_m}{\hat{\sigma_2}}
\end{pmatrix}
\end{aligned}
$$

Thus, the locally defined canonical parameter from (6.14) is

$$
\varphi(\theta) = \left(\sum_{i=1}^{n+m}\frac{\partial l(\theta)}{\partial w_i}V_{1i},\quad \sum_{i=1}^{n+m}\frac{\partial l(\theta)}{\partial w_i}V_{2i},\quad \sum_{i=1}^{n+m}\frac{\partial l(\theta)}{\partial w_i}V_{3i},\quad \sum_{i=1}^{n+m}\frac{\partial l(\theta)}{\partial w_i}V_{4i}\right)' \qquad (8.7)
$$

where

$$
\begin{cases}
\dfrac{\partial l(\theta)}{\partial w_i} = \dfrac{1}{x_i} - 2\sigma_1^2 x_i + (\alpha_1 - 1)\dfrac{2\sigma_1^2\ x_i\ e^{-(\sigma_1 x_i)^2}}{1-e^{-(\sigma_1 x_i)^2}}\ , & \text{if} \quad 1 \leq i \leq n\\
\dfrac{\partial l(\theta)}{\partial w_i} = \dfrac{1}{y_{i-n}} - 2\sigma_2^2 y_{i-n} + (\alpha_2 - 1)\dfrac{2\sigma_2^2\ y_{i-n}\ e^{-(\sigma_2 y_{i-n})^2}}{1-e^{-(\sigma_2 y_{i-n})^2}}\ , & \text{if} \quad (n+1) \leq i \leq (n+m)
\end{cases}
$$

and its derivative is

$$\begin{aligned}\varphi_\theta(\theta) &= \frac{\partial \varphi(\theta)}{\partial \theta} \\ &= \begin{pmatrix} \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_1} V_{1i} & 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma_1} V_{1i} & 0 \\ 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_2} V_{2i} & 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma_2} V_{2i} \\ \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_1} V_{3i} & 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma_1} V_{3i} & 0 \\ 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \alpha_2} V_{4i} & 0 & \sum_{i=1}^{n+m} \frac{\partial^2 l(\theta)}{\partial w_i \partial \sigma_2} V_{4i} \end{pmatrix}\end{aligned} \tag{8.8}$$

Finally,

$$\psi_\theta(\theta) = \begin{pmatrix} \frac{\partial R(\theta)}{\partial \alpha_1} & \frac{\partial R(\theta)}{\partial \alpha_2} & \frac{\partial R(\theta)}{\partial \sigma_1} & \frac{\partial R(\theta)}{\partial \sigma_2} \end{pmatrix}$$

Therefore $q(\psi)$ can be calculated from (6.7), $r^*(\psi)$ can be obtained from (5.19), and the $(1-\gamma)100\%$ confidence interval for R can be obtained.

To illustrate the proposed method with unequal scale parameters, we will use the same data sets of carbon fibre reported by Badar and Priest (1982). Surles and Padgett (2001) reported the MLE of R is $\hat{R} = 0.616592$, and concluded that R is greater than 0.5 using the approximate inference procedures they suggested. This is consistent with the fact of that short fibers tends to be stronger than long ones.

Again, assuming that X and Y are independently distributed as $BurrX(\alpha_1, \sigma_1)$ and $BurrX(\alpha_2, \sigma_2)$ respectively. We are interested in testing $H_0 : R = 0.5$ vs.

$H_1 : R > 0.5$ where $R = P(Y < X)$ is given in (7.3). Using the techniques discussed in Section 8.1.3, the p-values are 0.0067, 0.0081 and 0.0086, obtained by the Wald method, the signed log-likelihood ratio method and the proposed method, respectively. They provide **strong** evidence to suggest that 20-mm fiber is typically stronger than 50-mm fiber.

Moreover, the 90% and 95% confidence intervals (CI) for R based on Wald, r and proposed third-order method (Proposed) intervals are presented in Table 8.5. In examining Table 8.5, we observe that all three methods give similar interval estimations. The result is not surprising because both samples have relatively large sample sizes.

Table 8.5: Interval Estimates of ψ for Example assuming unequal scale parameters

	90% Confidence Interval	95% Confidence Interval
Wald	(0.5392, 0.6946)	(0.5243, 0.7095)
r	(0.5357, 0.6937)	(0.5199, 0.7077)
Proposed	(0.5346, 0.6927)	(0.5188, 0.7067)

We randomly sampled from this data set using $n = 30, 50$ and $n = 30, 50$, for each combination of (n, m), we obtained the 90% and 95% CI for R and the results are reported in Table 8.6. As we can see in Table 8.6, the proposed method and the other methods give quite different results for small sample size situations.

We also conduct the Monte Carlo simulation studies when the scale parameters

Table 8.6: Interval Estimates of ψ for Example using resampling

Sample sizes	Methods	90% Confidence Interval	95% Confidence Interval
n=30, m=30	Wald	(0.4702, 0.7076)	(0.4476, 0.7293)
	r	(0.4681, 0.7011)	(0.4450, 0.7210)
	Proposed	(0.4660, 0.6994)	(0.4429, 0.7027)
n=50, m=50	Wald	(0.5279, 0.7080)	(0.5106, 0.7252)
	r	(0.5255, 0.7041)	(0.5076, 0.7196)
	Proposed	(0.5240, 0.7027)	(0.5044, 0.7182)

of two Burr type X distributions are differ with similar settings comparing when scale parameter is same. For each parameter configuration and for each sample size, we generate 10,000 random samples from Burr type X distribution by using the transformation from (8.4) For each simulated sample, we calculated the 95% confidence interval for ψ obtained by the proposed third-order method with the Wald method, and the signed log-likelihood ratio method. For each simulated setting, we report the same criterions: lower error, upper error, central coverage and the average bias.

Table 8.7 presents the simulation results of , lower errors, upper errors, coverage probabilities and average bias by choosing $\alpha_1 = 8,\ \sigma_1 = 0.7$ and $\alpha_2 = 12.5,\ \sigma_2 = 0.8$ which is closed to the overall MLE obtained from example in the previous section, when the same sample sizes are equal, $n = m = 10$, and when sample sizes are not equal, $n = 10,\ m = 50$ and $n = 50,\ m = 10$. In this case, $R = \psi = 0.5966$. It is clear

Table 8.7: Simulation results for R when σ is not equal

Sample Size	Method	Lower Error	Upper Error	Central Coverage	Average Bias
n=m=10	Wald	0.0394	0.0613	0.8993	0.02535
	r	0.0308	0.0364	0.9376	0.00620
	Proposed	0.0273	0.0257	0.9470	0.00150
n=10, m=50	Wald	0.0372	0.0567	0.9061	0.02195
	r	0.0289	0.0335	0.9376	0.00620
	Proposed	0.0236	0.0223	0.9541	0.00205
n=50, m=10	Wald	0.0338	0.0562	0.9100	0.02000
	r	0.0283	0.0362	0.9355	0.00725
	Proposed	0.0222	0.0256	0.9522	0.00170

that the coverage probabilities for R are poor and the two-tail error probabilities are extremely asymmetric from the Wald method in all three settings. The results from signed log-likelihood method are not satisfactory. The proposed method gives not only an almost exact coverage probability but also it has symmetric two-tail error probabilities even for small or uneven sample sizes.

Tables 8.8 to 8.10 present more simulation results for $\alpha_1 = 5, \alpha_2 = 10, \sigma_1 = 2$, and $R = 0.1(0.1)0.9$ with $(n, m) = (10, 10), (10, 50)$ and $(50, 10)$. Note that, we fixed R and σ_2 is determined uniquely by (7.3). Again, the proposed method outperformed the other two methods even when the sample sizes are small or uneven.

Table 8.8: $\sigma_1 \neq \sigma_2$, $\alpha_1 = 5, \alpha_2 = 10, \sigma_1 = 2$ and σ_2 is obtained by Equation (7.3), $(n, m) = (10, 10)$.

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1607	0.0036	0.8357	0.07855
	r	0.0406	0.0183	0.9411	0.01115
	Proposed	0.0204	0.0259	0.9537	0.00275
0.2	Wald	0.1140	0.0123	0.8737	0.05085
	r	0.0399	0.0234	0.9367	0.00825
	Proposed	0.0215	0.0259	0.9526	0.00220
0.3	Wald	0.0854	0.0236	0.8910	0.03090
	r	0.0372	0.0268	0.9360	0.00700
	Proposed	0.0236	0.0258	0.9506	0.00110
0.4	Wald	0.0652	0.0362	0.8986	0.02570
	r	0.0358	0.0304	0.9338	0.00810
	Proposed	0.0244	0.0259	0.9497	0.00075
0.5	Wald	0.0497	0.0508	0.8995	0.02525
	r	0.0326	0.0337	0.9337	0.00815
	Proposed	0.0266	0.0267	0.9467	0.00165
0.6	Wald	0.0357	0.0674	0.8969	0.02655
	r	0.0290	0.0371	0.9339	0.00805
	Proposed	0.0267	0.0251	0.9482	0.00090
0.7	Wald	0.0232	0.0889	0.8879	0.03285
	r	0.0262	0.0383	0.9335	0.00725
	Proposed	0.0255	0.0256	0.9489	0.00055
0.8	Wald	0.0145	0.1207	0.8648	0.05310
	r	0.0234	0.0396	0.9370	0.00810
	Proposed	0.0261	0.0240	0.9499	0.00105
0.9	Wald	0.0036	0.1693	0.8271	0.08285
	r	0.0195	0.0442	0.9363	0.01235
	Proposed	0.0264	0.0231	0.9505	0.00165

Table 8.9: $\sigma_1 \neq \sigma_2$, $\alpha_1 = 5, \alpha_2 = 10, \sigma_1 = 2$ and σ_2 is obtained by Equation (15), $(n, m) = (10, 50)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
	Wald	0.1322	0.0047	0.8631	0.06375
0.1	r	0.0395	0.0190	0.9415	0.01025
	Proposed	0.0227	0.0254	0.9519	0.00135
	Wald	0.0983	0.0131	0.8886	0.04260
0.2	r	0.0376	0.0245	0.9379	0.00655
	Proposed	0.0238	0.0254	0.9508	0.00080
	Wald	0.0772	0.0251	0.8977	0.02615
0.3	r	0.0352	0.0291	0.9357	0.00715
	Proposed	0.0228	0.0271	0.9501	0.00215
	Wald	0.0604	0.0352	0.9044	0.0280
0.4	r	0.0335	0.0312	0.9353	0.00735
	Proposed	0.0219	0.0264	0.9517	0.00225
	Wald	0.0456	0.0474	0.9070	0.02150
0.5	r	0.0306	0.0330	0.9364	0.00680
	Proposed	0.0221	0.0246	0.9533	0.00165
	Wald	0.0340	0.0661	0.8999	0.02505
0.6	r	0.0280	0.0348	0.9372	0.00640
	Proposed	0.0235	0.0218	0.9547	0.00235
	Wald	0.0224	0.0889	0.8877	0.03375
0.7	r	0.0250	0.0391	0.9359	0.00705
	Proposed	0.0233	0.0234	0.9533	0.00165
	Wald	0.0131	0.1179	0.8690	0.05240
0.8	r	0.0221	0.0420	0.9359	0.00995
	Proposed	0.0229	0.0237	0.9534	0.00170
	Wald	0.0045	0.1654	0.8301	0.08045
0.9	r	0.0175	0.0457	0.9368	0.01410
	Proposed	0.0222	0.0232	0.9546	0.00230

Table 8.10: $\sigma_1 \neq \sigma_2$, $\alpha_1 = 5, \alpha_2 = 10, \sigma_1 = 2$ and σ_2 is obtained by Equation (15), $(n, m) = (50, 10)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1348	0.0052	0.8600	0.06480
	r	0.0404	0.0208	0.9388	0.00980
	Proposed	0.0235	0.0249	0.9516	0.00080
0.2	Wald	0.0912	0.0153	0.8935	0.03795
	r	0.0355	0.0227	0.9418	0.00640
	Proposed	0.0221	0.0231	0.9548	0.00240
0.3	Wald	0.0679	0.0239	0.9082	0.02200
	r	0.0337	0.0260	0.9403	0.00485
	Proposed	0.0228	0.0236	0.9536	0.00180
0.4	Wald	0.0552	0.0350	0.9128	0.01860
	r	0.0321	0.0293	0.9386	0.00570
	Proposed	0.0237	0.0244	0.9519	0.00095
0.5	Wald	0.0405	0.0464	0.9131	0.018455
	r	0.0296	0.0320	0.9384	0.00580
	Proposed	0.0225	0.0245	0.9530	0.00150
0.6	Wald	0.0289	0.0557	0.9154	0.01730
	r	0.0263	0.0349	0.9388	0.00560
	Proposed	0.0225	0.0246	0.9529	0.00145
0.7	Wald	0.0184	0.0681	0.9135	0.02485
	r	0.0227	0.0371	0.9402	0.00720
	Proposed	0.0217	0.0247	0.9536	0.00180
0.8	Wald	0.0109	0.0818	0.9073	0.03545
	r	0.0185	0.0375	0.9440	0.00950
	Proposed	0.0191	0.0245	0.9564	0.00320
0.9	Wald	0.0042	0.1033	0.8925	0.04955
	r	0.0163	0.0341	0.9496	0.00890
	Proposed	0.0196	0.0216	0.9588	0.00440

8.2 Inference for Stress-Strength Reliability with Exponentiated Exponential Distribution

Gupta et al. (1998) and Gupta and Kundu (2001) introduced a two-parameter exponentiated exponential (EE) distribution as an alternative to the two-parameter gamma and Weibull distributions for analysing failure time data. For the two-parameter gamma distribution, its cumulative distribution function, or equivalently, the survival function cannot be expressed in closed form if the shape parameter is not an integer. This makes it less popular than the Weibull distribution, whose distribution function, survival function, and hazard function can all be expressed as a closed form. However, for the Weibull distribution, the asymptotic coverage to normality for the distribution of the maximum likelihood estimators (MLEs) is very slow (Bain 1976). Therefore, most of the asymptotic inferences are not very accurate, unless the sample size is very large. Compare to these two most commonly used distributions, the cumulative distribution function of EE distribution can be expressed in explicit form, share some common properties with gamma and Weibull distributions for failure time data, and has fast convergence of the MLEs.

The distribution function, $F(y;\alpha,\beta)$, of the EE distribution takes the form

$$F(y;\alpha,\beta) = (1-e^{-\beta y})^{\alpha} \qquad \alpha,\ \beta,\ y > 0 \tag{8.9}$$

where α is the shape parameter and β is the scale parameter.

Similar to the two-parameter gamma distribution, the EE distribution also has an increasing or decreasing failure rate that depends on the shape parameter. The density function of the EE distribution also varies significantly depending on the shape parameter. The advantage of the EE distribution is that the cumulative distribution function and the survival function can be explicitly written in closed form.

8.2.1 Stress-Strength Reliability with Equal Scale Parameters

Let X and Y be independently distributed as $EE(\alpha_1, \beta)$ and $EE(\alpha_2, \beta)$ respectively. Then the stress-strength reliability with same scale parameter, β, is

$$\begin{aligned} R = P(Y < X) &= \int_0^\infty \int_0^x f(x; \alpha_1, \beta)\ f(y; \alpha_2, \beta)\ dy\ dx \\ &= \frac{\alpha_1}{\alpha_1 + \alpha_2} \end{aligned} \tag{8.10}$$

Let $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_m)'$ be the random samples from $EE(\alpha_1, \beta)$ and $EE(\alpha_2, \beta)$ respectively. Then the log-likelihood function of above model can be written as

$$l(\alpha_1, \alpha_2, \beta; x, y) = n \log \alpha_1 + m \log \alpha_2 + (n+m) \log \beta - \beta \left(\sum_{i=1}^n x_i + \sum_{j=1}^m y_j \right)$$

$$+ (\alpha_1 - 1) \sum_{i=1}^n \log(1 - e^{-\beta x_i}) + (\alpha_2 - 1) \sum_{j=1}^m \log(1 - e^{-\beta y_j})$$

Denote the overall maximum likelihood estimate as $\hat{\theta} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta})'$, then the observed information matrix $j_{\theta\theta}(\hat{\theta})$ can be obtained as

$$
\begin{aligned}
j_{\theta\theta}(\hat{\theta}) &= -l_{\theta\theta}(\hat{\theta}) \\
&= \begin{pmatrix}
\dfrac{n}{\hat{\alpha_1^2}} & 0 & -\displaystyle\sum_{i=1}^{n} \frac{x_i}{e^{\hat{\beta} x_i} - 1} \\
0 & \dfrac{m}{\hat{\alpha_2^2}} & -\displaystyle\sum_{j=1}^{m} \frac{y_j}{e^{\hat{\beta} y_j} - 1} \\
-\displaystyle\sum_{i=1}^{n} \frac{x_i}{e^{\hat{\beta} x_i} - 1} & -\displaystyle\sum_{j=1}^{m} \frac{y_j}{e^{\hat{\beta} y_j} - 1} & \dfrac{n+m}{\hat{\beta}^2} + A
\end{pmatrix},
\end{aligned}
$$

where $A = (\hat{\alpha}_1 - 1) \displaystyle\sum_{i=1}^{n} \frac{x_i^2 \, e^{\hat{\beta} x_i}}{(e^{\hat{\beta} x_i} - 1)^2} + (\hat{\alpha}_2 - 1) \sum_{j=1}^{m} \frac{y_j^2 \, e^{\hat{\beta} y_j}}{(e^{\hat{\beta} y_j} - 1)^2}$.

The tilted log-likelihood function $\tilde{l}(\theta)$ is defined as

$$
\tilde{l}(\theta) = l(x, y; \alpha_1, \alpha_2, \beta) + \hat{\kappa}[\psi(\theta) - \psi]
$$

where $\psi(\theta) = R = \dfrac{\alpha_1}{\alpha_1 + \alpha_2}$. Then the constrained MLE can be obtained from penalized likelihood method and denote as $\hat{\theta}_\psi = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\beta})'$. $\hat{\kappa}$ can be calculated by Lagrange multiplier method, and constrained observed information matrix $\tilde{j}_{\theta\theta}(\hat{\theta}_\psi)$ can be written as

$$
\tilde{j}_{\theta\theta}(\hat{\theta}_\psi) = \begin{pmatrix}
\dfrac{n}{\tilde{\alpha}_1^2} + \dfrac{2\hat{\kappa}\tilde{\alpha}_2}{(\tilde{\alpha}_1 + \tilde{\alpha}_2)^3} & -\dfrac{\hat{\kappa}(\tilde{\alpha}_1 - \tilde{\alpha}_2)}{(\tilde{\alpha}_1 + \tilde{\alpha}_2)^3} & -\displaystyle\sum_{i=1}^{n} \frac{x_i}{e^{\tilde{\beta} x_i} - 1} \\
-\dfrac{\hat{\kappa}(\tilde{\alpha}_1 - \tilde{\alpha}_2)}{(\tilde{\alpha}_1 + \tilde{\alpha}_2)^3} & \dfrac{m}{\tilde{\alpha}_2^2} - \dfrac{2\hat{\kappa}\tilde{\alpha}_1}{(\tilde{\alpha}_1 + \tilde{\alpha}_2)^3} & -\displaystyle\sum_{j=1}^{m} \frac{y_j}{e^{\tilde{\beta} y_j} - 1} \\
-\displaystyle\sum_{i=1}^{n} \frac{x_i}{e^{\tilde{\beta} x_i} - 1} & -\displaystyle\sum_{j=1}^{m} \frac{y_j}{e^{\tilde{\beta} y_j} - 1} & \dfrac{n+m}{\tilde{\beta}^2} + B
\end{pmatrix},
$$

where $B = (\tilde{\alpha}_1 - 1)\sum_{i=1}^{n} \frac{x_i^2 \, e^{\tilde{\beta} x_i}}{(e^{\tilde{\beta} x_i} - 1)^2} + (\tilde{\alpha}_2 - 1)\sum_{j=1}^{m} \frac{y_j^2 \, e^{\tilde{\beta} y_j}}{(e^{\tilde{\beta} y_j} - 1)^2}$. Now, we have everything to calculate $r(\psi)$.

Let $z = (z_1, \dots, z_n, z_{n+1}, \dots, z_{n+m})'$ denotes the vector-pivotal quantity, in this case, $z = (\log F(x_1; \alpha_1, \beta),\ \dots,\ \log F(x_n; \alpha_1, \beta),\ \log F(y_1; \alpha_2, \beta),$ $\dots,\ \log F(y_m; \alpha_2, \beta))'$. Thus, the ancillary direction V is

$$V = (V_1, V_2, V_3) = \begin{pmatrix} -\log(1 - e^{-\hat{\beta} x_1})\,\dfrac{e^{\hat{\beta} x_1} - 1}{\hat{\alpha_1}\hat{\beta}} & 0 & -\dfrac{x_1}{\hat{\beta}} \\ \vdots & \vdots & \vdots \\ -\log(1 - e^{-\hat{\beta} x_n})\,\dfrac{e^{\hat{\beta} x_n} - 1}{\hat{\alpha_1}\hat{\beta}} & 0 & -\dfrac{x_n}{\hat{\beta}} \\ 0 & -\log(1 - e^{-\hat{\beta} y_1})\,\dfrac{e^{\hat{\beta} y_1} - 1}{\hat{\alpha_2}\hat{\beta}} & -\dfrac{y_1}{\hat{\beta}} \\ \vdots & \vdots & \vdots \\ 0 & -\log(1 - e^{-\hat{\beta} y_m})\,\dfrac{e^{\hat{\beta} y_m} - 1}{\hat{\alpha_2}\hat{\beta}} & -\dfrac{y_m}{\hat{\beta}} \end{pmatrix}$$

Then we can calculate the locally defined canonical parameter $\varphi(\theta)$

$$\varphi(\theta) = \left(\sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\, V_{1i}, \quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\, V_{2i}, \quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i} V_{3i} \right)'$$

where $w = (x_1, \dots, x_n,\ y_1, \dots, y_m)'$ be the observed data. Hence, we also have $\varphi_\theta(\theta)$. For this particular case, $\chi(\theta)$ can be obtained accordingly. Therefore, $v\hat{a}r\left(\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)\right)$, $Q(\psi)$ and $r^*(\psi)$ can be obtained. Hence $(1-\gamma)100\%$ confidence interval can be obtained from the modified signed log-likelihood ratio statistics.

To illustrate the proposed third-order method for interval estimation, the following two data sets with sample size of 11 and 9 are used: 2.1828, 0.5911, 1.0711, 0.9007, 1.7814, 1.3616, 0.8629, 0.2301, 1.5183, 0.8481, 1.0845 and 0.8874, 1.1482, 0.8227, 0.4086, 0.5596, 1.1978, 1.1324, 0.5625, 1.0679. Assume that X and Y are independently distributed as $EE(\alpha_1, \beta)$ and $EE(\alpha_2, \beta)$ respectively. We are interested in testing $H_0 : R = 0.5$ vs. $H_1 : R > 0.5$, where $R = P(Y < X)$ is given in (8.10). The 90% and 95% confidence intervals for R based on Wald, r and r^* methods are presented in Table 8.11. In examining Table 8.11, we observe that all three methods give different interval estimations.

Table 8.11: Interval Estimates of ψ Assuming EED with Same Scale Parameter β

	90% Confidence Interval	95% Confidence Interval
Wald	(0.3441, 0.7138)	(0.3086, 0.7492)
r	(0.3485, 0.7058)	(0.3165, 0.7363)
Proposed	(0.3501, 0.7110)	(0.3176, 0.7415)

To compare the accuracy of the proposed method with the MLE method, the signed log-likelihood ratio method, and the proposed method, Monte Carlo simulation studies were conducted. The cases of unequal and equal scale parameters are both examined. For each parameter configuration and for each sample size, we generate 10,000 random samples from EE distribution by using the following transformation:

$$T = -\frac{1}{\beta} \log(1 - U^{1/\alpha}),$$

where U is a uniform variate between 0 and 1.

The performance of a method is judged by using the same criteria as before: central coverage, upper error, lower error and average bias. These values reflect the desired properties of the accuracy and symmetry of the interval estimates of R.

Tables 8.12 to 8.14 present simulation results for the equal scale parameter case, *i.e.* $\alpha_1 = 4, \beta = 8$, and $R = 0.1(0.1)0.9$ with $(n, m) = (10, 10), (10, 50)$ and $(50, 10)$. Note that, we fixed R and α_2 is determined uniquely by $R = \alpha_1/(\alpha_1 + \alpha_2)$. Again, the proposed method outperformed the other two methods even when the sample sizes are small.

Table 8.12: $\alpha_1 = 4, \beta = 8$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$, $(n, m) = (10, 10)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1608	0.0024	0.8368	0.07920
	r	0.0526	0.0189	0.9285	0.01685
	Proposed	0.0236	0.0267	0.9497	0.00155
0.2	Wald	0.1195	0.0121	0.8684	0.05370
	r	0.0502	0.0236	0.9262	0.01330
	Proposed	0.0269	0.0264	0.9467	0.00165
0.3	Wald	0.0944	0.0224	0.8832	0.03600
	r	0.0428	0.0256	0.9316	0.00920
	Proposed	0.0225	0.0234	0.9541	0.00205
0.4	Wald	0.0724	0.0347	0.8929	0.02855
	r	0.0387	0.0277	0.9336	0.00820
	Proposed	0.0256	0.0224	0.9520	0.00160
0.5	Wald	0.0523	0.0517	0.8960	0.02700
	r	0.0335	0.0328	0.9337	0.00815
	Proposed	0.0244	0.0230	0.9526	0.00013
0.6	Wald	0.0378	0.0753	0.8869	0.03155
	r	0.0295	0.0400	0.9305	0.00975
	Proposed	0.0234	0.0260	0.9506	0.00130
0.7	Wald	0.0245	0.0944	0.8811	0.03495
	r	0.0282	0.0454	0.9264	0.01180
	Proposed	0.0261	0.0262	0.9477	0.00115
0.8	Wald	0.0109	0.1187	0.8704	0.05390
	r	0.0226	0.0467	0.9307	0.01205
	Proposed	0.0239	0.0262	0.9499	0.00115
0.9	Wald	0.0027	0.1464	0.8509	0.07185
	r	0.0211	0.0499	0.9290	0.01440
	Proposed	0.0268	0.0252	0.9480	0.00100

Table 8.13: $\alpha_1 = 4, \beta = 8$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$, $(n, m) = (10, 50)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.0722	0.0093	0.9185	0.03145
	r	0.0302	0.0273	0.9425	0.00375
	Proposed	0.0250	0.0261	0.9489	0.00055
0.2	Wald	0.0577	0.0206	0.9217	0.01855
	r	0.0287	0.0268	0.9445	0.00275
	Proposed	0.0260	0.0244	0.9496	0.00080
0.3	Wald	0.0431	0.0327	0.9242	0.01290
	r	0.0276	0.0306	0.9418	0.00410
	Proposed	0.0261	0.0257	0.9482	0.00090
0.4	Wald	0.0299	0.0412	0.9289	0.01055
	r	0.0216	0.0290	0.9494	0.00370
	Proposed	0.0218	0.0227	0.9555	0.00275
0.5	Wald	0.0475	0.0488	0.9037	0.02315
	r	0.0234	0.0358	0.9408	0.00620
	Proposed	0.0243	0.0269	0.9488	0.00130
0.6	Wald	0.0166	0.0688	0.9146	0.02610
	r	0.0221	0.0350	0.9429	0.00645
	Proposed	0.0249	0.0254	0.9497	0.00250
0.7	Wald	0.0105	0.0789	0.9106	0.03420
	r	0.0206	0.0366	0.9428	0.00800
	Proposed	0.0276	0.0229	0.9495	0.00235
0.8	Wald	0.0059	0.1021	0.8920	0.04810
	r	0.0195	0.0426	0.9379	0.01155
	Proposed	0.0259	0.0269	0.9472	0.00140
0.9	Wald	0.0016	0.1210	0.8774	0.05970
	r	0.0160	0.0456	0.9384	0.01480
	Proposed	0.0239	0.0249	0.9512	0.00060

Table 8.14: $\alpha_1 = 4, \beta = 8$ and α_2 satisfies $R = \alpha_1/(\alpha_1 + \alpha_2)$, $(n, m) = (50, 10)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1120	0.0008	0.8872	0.05560
	r	0.0412	0.0161	0.9427	0.01255
	Proposed	0.0230	0.0247	0.9523	0.00115
0.2	Wald	0.1011	0.0048	0.8941	0.04815
	r	0.0407	0.0202	0.9319	0.01025
	Proposed	0.0261	0.0260	0.9479	0.00105
0.3	Wald	0.0805	0.0108	0.9087	0.03485
	r	0.0366	0.0232	0.9402	0.00670
	Proposed	0.0248	0.0275	0.9477	0.00135
0.4	Wald	0.0648	0.0140	0.9212	0.02540
	r	0.0338	0.0215	0.9447	0.00615
	Proposed	0.0238	0.0247	0.9515	0.00075
0.5	Wald	0.0551	0.0258	0.9191	0.01545
	r	0.0317	0.0256	0.9427	0.00365
	Proposed	0.0242	0.0276	0.9482	0.00170
0.6	Wald	0.0437	0.0296	0.9267	0.01165
	r	0.0329	0.0228	0.9443	0.00505
	Proposed	0.0247	0.0229	0.9524	0.00120
0.7	Wald	0.0332	0.0400	0.9268	0.01160
	r	0.0307	0.0247	0.9446	0.00300
	Proposed	0.0257	0.0241	0.9502	0.00080
0.8	Wald	0.0228	0.0548	0.9224	0.01600
	r	0.0294	0.0266	0.9440	0.00300
	Proposed	0.0256	0.0242	0.9502	0.00070
0.9	Wald	0.0098	0.0670	0.9232	0.02860
	r	0.0247	0.0299	0.9454	0.00260
	Proposed	0.0234	0.0261	0.9505	0.00135

8.2.2 Stress-Strength Reliability with Unequal Scale Parameters

Let X and Y be independently distributed as $EE(\alpha_1, \beta_1)$ and $EE(\alpha_2, \beta_2)$ respectively. When scale parameters are unequal, $\beta_1 \neq \beta_2$, R is difficult to calculate since integral has no known closed-form, which is

$$R = P(Y < X) = \int_0^\infty \alpha_1 \beta_1 \, (1 - e^{-\beta_1 x})^{\alpha_1 - 1} \, e^{-\beta_1 x} \, (1 - e^{-\beta_2 x})^{\alpha_2} \, dx \tag{8.11}$$

Exact inference procedure for R is not available. Let $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_m)'$ be the random samples from $EE(\alpha_1, \beta_1)$ and $EE(\alpha_2, \beta_2)$ respectively. Then the log-likelihood function of above model can be written as

$$\begin{aligned} l(\alpha_1, \beta_1, \alpha_2, \beta_2; x, y) &= n \log \alpha_1 + n \log \beta_1 + (\alpha_1 - 1) \sum_{i=1}^{n} \log(1 - e^{-\beta_1 x_i}) \\ -\beta_1 \sum_{i=1}^{n} x_i + m \log \alpha_2 &+ m \log \beta_2 + (\alpha_2 - 1) \sum_{j=1}^{m} \log(1 - e^{-\beta_2 y_j}) - \beta_2 \sum_{j=1}^{m} y_j \end{aligned}$$

Denote the overall maximum likelihood estimate as $\hat{\theta} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2)'$, and observed information matrix $j_{\theta\theta}(\hat{\theta})$ can be obtained as

$$\begin{aligned} j_{\theta\theta}(\hat{\theta}) &= -\ell_{\theta\theta}(\hat{\theta}) \\ &= \begin{pmatrix} \dfrac{n}{\hat{\alpha}_1^2} & -\displaystyle\sum_{i=1}^{n} \frac{x_i \, e^{-\hat{\beta}_1 x_i}}{1 - e^{-\hat{\beta}_1 x_i}} & 0 & 0 \\ -\displaystyle\sum_{i=1}^{n} \frac{x_i \, e^{-\hat{\beta}_1 x_i}}{1 - e^{-\hat{\beta}_1 x_i}} & \dfrac{n}{\hat{\beta}_1^2} + A & 0 & 0 \\ 0 & 0 & \dfrac{m}{\hat{\alpha}_2^2} & -\displaystyle\sum_{j=1}^{m} \frac{y_j \, e^{-\hat{\beta}_2 y_j}}{1 - e^{-\hat{\beta}_2 y_j}} \\ 0 & 0 & -\displaystyle\sum_{j=1}^{m} \frac{y_j \, e^{-\hat{\beta}_2 y_j}}{1 - e^{-\hat{\beta}_2 y_j}} & \dfrac{m}{\hat{\beta}_2^2} + B \end{pmatrix} \end{aligned}$$

where $A = (\hat{\alpha}_1 - 1)\displaystyle\sum_{i=1}^{n} \frac{x_i^2 \; e^{-\hat{\beta}_1 x_i}}{(1 - e^{-\hat{\beta}_1 x_i})^2}$ and $B = (\hat{\alpha}_2 - 1)\displaystyle\sum_{j=1}^{m} \frac{y_j^2 \; e^{-\hat{\beta}_2 y_j}}{(1 - e^{-\hat{\beta}_2 y_j})^2}$.

The tilted log-likelihood function $\tilde{l}(\theta)$ is defined as

$$\tilde{l}(\theta) = l(x, y; \alpha_1, \alpha_2, \beta_1, \beta_2) + \hat{\kappa}[\psi(\theta) - \psi]$$

where $\psi(\theta) = R$ defined by (8.11). Similarly, we can obtain the constrained MLE $\hat{\theta}_\psi = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\beta}_1, \tilde{\beta}_2)'$ by penalized likelihood method and $\hat{\kappa}$ by Lagrange multiplier method, then constrained observed information matrix $\tilde{j}_{\theta\theta}(\hat{\theta}_\psi)$ can be written as

$$\tilde{j}_{\theta\theta}(\hat{\theta}_\psi) = -\tilde{l}_{\theta\theta}(\hat{\theta}_\psi) = \begin{pmatrix} \tilde{j}_{\alpha_1\alpha_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\beta_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_1\beta_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\beta_1\alpha_1}(\hat{\theta}_\psi) & \tilde{j}_{\beta_1\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\beta_1\beta_1}(\hat{\theta}_\psi) & \tilde{j}_{\beta_1\beta_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\alpha_2\alpha_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\beta_1}(\hat{\theta}_\psi) & \tilde{j}_{\alpha_2\beta_2}(\hat{\theta}_\psi) \\ \tilde{j}_{\beta_2\alpha_1}(\hat{\theta}_\psi) & \tilde{j}_{\beta_2\alpha_2}(\hat{\theta}_\psi) & \tilde{j}_{\beta_2\beta_1}(\hat{\theta}_\psi) & \tilde{j}_{\beta_2\beta_2}(\hat{\theta}_\psi) \end{pmatrix}$$

where

- $\tilde{j}_{\alpha_1\alpha_1}(\hat{\theta}_\psi) = \dfrac{n}{\tilde{\alpha}_1^2} - \hat{\kappa}\; R_{\alpha_1\alpha_1}(\hat{\theta}_\psi)$, where $R_{\alpha_1\alpha_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial \alpha_1^2}\right|_{\theta=\hat{\theta}_\psi}$.
- $\tilde{j}_{\alpha_1\alpha_2}(\hat{\theta}_\psi) = -\hat{\kappa}\; R_{\alpha_1\alpha_2}(\hat{\theta}_\psi)$, where $R_{\alpha_1\alpha_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial \alpha_1 \partial \alpha_2}\right|_{\theta=\hat{\theta}_\psi}$.
- $\tilde{j}_{\alpha_1\beta_1}(\hat{\theta}_\psi) = -\displaystyle\sum_{i=1}^{n} \frac{x_i \; e^{-\tilde{\beta}_1 x_i}}{1 - e^{-\tilde{\beta}_1 x_i}} - \hat{\kappa}\; R_{\alpha_1\beta_1}(\hat{\theta}_\psi)$, where $R_{\alpha_1\beta_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial \alpha_1 \partial \beta_1}\right|_{\theta=\hat{\theta}_\psi}$.
- $\tilde{j}_{\alpha_1\beta_2}(\hat{\theta}_\psi) = -\hat{\kappa}\; R_{\alpha_1\beta_2}(\hat{\theta}_\psi)$, where $R_{\alpha_1\beta_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial \alpha_1 \partial \beta_2}\right|_{\theta=\hat{\theta}_\psi}$.
- $\tilde{j}_{\alpha_2\alpha_2}(\hat{\theta}_\psi) = \dfrac{m}{\tilde{\alpha_2}^2} - \hat{\kappa}\; R_{\alpha_2\alpha_2}(\hat{\theta}_\psi)$, where $R_{\alpha_2\alpha_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial \alpha_2^2}\right|_{\theta=\hat{\theta}_\psi}$.

- $\tilde{j}_{\alpha_2\beta_1}(\hat{\theta}_\psi) = -\hat{\kappa}\ R_{\alpha_2\beta_1}(\hat{\theta}_\psi)$, where $R_{\alpha_2\beta_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial\alpha_2\partial\beta_1}\right|_{\theta=\hat{\theta}_\psi}$.

- $\tilde{j}_{\alpha_2\beta_2}(\hat{\theta}_\psi) = -\displaystyle\sum_{j=1}^{m} \frac{y_j\ e^{-\tilde{\beta}_2 y_j}}{1 - e^{-\tilde{\beta}_2 y_j}} - \hat{\kappa}\ R_{\alpha_2\beta_2}(\hat{\theta}_\psi)$, where $R_{\alpha_2\beta_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial\alpha_2\partial\beta_2}\right|_{\theta=\hat{\theta}_\psi}$.

- $\tilde{j}_{\beta_1\beta_1}(\hat{\theta}_\psi) = \dfrac{m}{\tilde{\beta}_1^2} + (\tilde{\alpha}_1 - 1)\displaystyle\sum_{i=1}^{n} \frac{x_i^2\ e^{-\tilde{\beta}_1 x_i}}{(1 - e^{-\tilde{\beta}_1 x_i})^2} - \hat{\kappa}\ R_{\beta_1\beta_1}(\hat{\theta}_\psi)$, where $R_{\beta_1\beta_1}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial\beta_1^2}\right|_{\theta=\hat{\theta}_\psi}$.

- $\tilde{j}_{\beta_1\beta_2}(\hat{\theta}_\psi) = -\hat{\kappa}\ R_{\beta_1\beta_2}(\hat{\theta}_\psi)$, where $R_{\beta_1\beta_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial\beta_1\partial\beta_2}\right|_{\theta=\hat{\theta}_\psi}$.

- $\tilde{j}_{\beta_2\beta_2}(\hat{\theta}_\psi) = \dfrac{m}{\tilde{\beta}_2^{\ 2}} + (\tilde{\alpha}_2 - 1)\displaystyle\sum_{j=1}^{m} \frac{y_j^2\ e^{-\tilde{\beta}_2 y_j}}{(1 - e^{-\tilde{\beta}_2 y_j})^2} - \hat{\kappa}\ R_{\beta_2\beta_2}(\hat{\theta}_\psi)$, where $R_{\beta_2\beta_2}(\hat{\theta}_\psi) = \left.\dfrac{\partial^2 R(\theta)}{\partial\beta_2^2}\right|_{\theta=\hat{\theta}_\psi}$.

Thus $r(\psi)$ can be obtained accordingly.

Let $z = (z_1, \ldots, z_n, z_{n+1}, \ldots, z_{n+m})'$ denotes the vector-pivotal quantity, in this case, $z = (\log F(x_1; \alpha_1, \beta),\ \ldots,\ \log F(x_n; \alpha_1, \beta),\ \log F(y_1; \alpha_2, \beta),$

$\ldots,\ \log F(y_m;\alpha_2,\beta))'$. Hence the ancillary direction V is

$$V = (V_1, V_2, V_3, V_4)$$
$$= \begin{pmatrix} -\log(1-e^{-\hat{\beta}_1 x_1})\,\dfrac{1-e^{-\hat{\beta}_1 x_1}}{\hat{\alpha}_1\hat{\beta}_1 e^{-\hat{\beta}_1 x_1}} & -\dfrac{x_1}{\hat{\beta}_1} & 0 & 0 \\ \vdots & \vdots & \vdots & \\ -\log(1-e^{-\hat{\beta}_1 x_n})\,\dfrac{1-e^{-\hat{\beta}_1 x_n}}{\hat{\alpha}_1\hat{\beta}_1 e^{-\hat{\beta}_1 x_n}} & -\dfrac{x_n}{\hat{\beta}_1} & 0 & 0 \\ 0 & 0 & -\log(1-e^{-\hat{\beta}_2 y_1})\,\dfrac{1-e^{-\hat{\beta}_2 y_1}}{\hat{\alpha}_2\hat{\beta}_2 e^{-\hat{\beta}_2 y_1}} & -\dfrac{y_1}{\hat{\beta}_2} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\log(1-e^{-\hat{\beta}_2 y_m})\,\dfrac{1-e^{-\hat{\beta}_2 y_m}}{\hat{\alpha}_2\hat{\beta}_2 e^{-\hat{\beta}_2 y_m}} & -\dfrac{y_m}{\hat{\beta}_2} \end{pmatrix}$$

Then we can calculate the locally defined canonical parameter $\varphi(\theta)$ as

$$\varphi(\theta) = \left(\sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\, V_{1i},\quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i}\, V_{2i},\quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i} V_{3i},\quad \sum_{i=1}^{n+m} \frac{\partial l(\theta)}{\partial w_i} V_{4i} \right)'$$

where $w = (x_1, \ldots, x_n,\ y_1, \ldots, y_m)'$ be the observed data. Hence, we also have $\varphi_\theta(\theta)$. Therefor, for this unequal scale parameter case, $\chi(\theta)$ can be obtained accordingly. Therefore, $v\hat{a}r\left(\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)\right)$, $Q(\psi)$ and $r^*(\psi)$ can be obtained. Hence $(1-\gamma)100\%$ confidence interval can be obtained from the modified signed log-likelihood ratio statistics.

Table 8.15 present the 90% and 95% confidence intervals (CI) for R based on Wald, r and proposed third-order methods by using the same data set recorded in Section 8.2.1.

Table 8.15: Interval Estimates of ψ for Example

	90% Confidence Interval	95% Confidence Interval
Wald	(0.4223, 0.8179)	(0.3843, 0.8557)
r	(0.4151, 0.7966)	(0.3767, 0.8241)
Proposed	(0.4080, 0.7910)	(0.3698, 0.8188)

We also conduct Monte Carlo simulation studies. The performance of a method is judged by using the following same criteria: central coverage, upper error, lower error and average bias.

Tables 8.16 to 8.18 present simulation results for the unequal scale parameters case, *i.e.* $\alpha_1 = 2, \alpha_2 = 5, \beta_1 = 3$, and $R = 0.1(0.1)0.9$ with $(n, m) = (10, 10), (10, 50)$ and $(50, 10)$. Note that, we fixed R and β_2 is determined uniquely by Equation (8.11). It is clear that the coverage probabilities for R are poor and the two-tail error probabilities are extremely asymmetric from the Wald method. The results from signed log-likelihood method are not satisfactory especially when two sample sizes are small or sample sizes are unequal, and it also shows some evidence of asymmetry of two-tail error probabilities. However, the proposed method gives not only an almost exact coverage probability but also it has symmetric two-tail error probabilities even for small or uneven sample sizes.

Table 8.16: $\beta_1 \neq \beta_2$, $\alpha_1 = 2, \alpha_2 = 5, \beta_1 = 3$ and β_2 is obtained by Equation (8.11), $(n, m) = (10, 10)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
	Wald	0.1602	0.0033	0.8365	0.07845
0.1	r	0.0401	0.0177	0.9422	0.01120
	Proposed	0.0207	0.0252	0.9541	0.00255
	Wald	0.1138	0.0125	0.8737	0.05065
0.2	r	0.0388	0.0235	0.9377	0.00765
	Proposed	0.0218	0.0258	0.9524	0.00200
	Wald	0.0857	0.0225	0.8918	0.03160
0.3	r	0.0372	0.0262	0.9366	0.00670
	Proposed	0.0230	0.0259	0.9527	0.00135
	Wald	0.0656	0.0362	0.8982	0.02590
0.4	r	0.0352	0.0294	0.9354	0.00730
	Proposed	0.0244	0.0259	0.9497	0.00075
	Wald	0.0505	0.0506	0.8989	0.02555
0.5	r	0.0317	0.0328	0.9355	0.00725
	Proposed	0.0249	0.0255	0.9496	0.00030
	Wald	0.0353	0.0670	0.8977	0.02615
0.6	r	0.0290	0.0359	0.9351	0.00745
	Proposed	0.0244	0.0246	0.9510	0.00050
	Wald	0.0235	0.0900	0.8865	0.03325
0.7	r	0.0257	0.0394	0.9349	0.00755
	Proposed	0.0246	0.0238	0.9516	0.00080
	Wald	0.0142	0.1234	0.8624	0.05460
0.8	r	0.0239	0.0419	0.9342	0.00900
	Proposed	0.0261	0.0239	0.9500	0.00110
	Wald	0.0035	0.1763	0.8202	0.08640
0.9	r	0.0198	0.0465	0.9337	0.01335
	Proposed	0.0262	0.0240	0.9498	0.00110

Table 8.17: $\beta_1 \neq \beta_2$, $\alpha_1 = 2, \alpha_2 = 5, \beta_1 = 3$ and β_2 is obtained by Equation (8.11), $(n, m) = (10, 50)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
0.1	Wald	0.1341	0.0046	0.8613	0.06475
	r	0.0399	0.0186	0.9415	0.01065
	Proposed	0.0231	0.0246	0.9523	0.00115
0.2	Wald	0.1007	0.0131	0.8862	0.04380
	r	0.0370	0.0243	0.9387	0.00635
	Proposed	0.0239	0.0251	0.9510	0.00060
0.3	Wald	0.0774	0.0249	0.8977	0.02625
	r	0.0349	0.0289	0.9362	0.00690
	Proposed	0.0228	0.0270	0.9502	0.00210
0.4	Wald	0.0615	0.0353	0.9032	0.02340
	r	0.0327	0.0311	0.9362	0.00690
	Proposed	0.0220	0.0260	0.9520	0.00200
0.5	Wald	0.0475	0.0488	0.9037	0.02315
	r	0.0294	0.0323	0.9383	0.00585
	Proposed	0.0229	0.0240	0.9531	0.00155
0.6	Wald	0.0351	0.0682	0.8967	0.02665
	r	0.0278	0.0348	0.9374	0.00630
	Proposed	0.0222	0.0222	0.9556	0.00280
0.7	Wald	0.0225	0.0962	0.8813	0.03685
	r	0.0256	0.0388	0.9356	0.00720
	Proposed	0.0226	0.0227	0.9547	0.00235
0.8	Wald	0.0126	0.1249	0.8625	0.05615
	r	0.0217	0.0431	0.9352	0.01070
	Proposed	0.0224	0.0242	0.9534	0.00170
0.9	Wald	0.0063	0.1779	0.8158	0.08580
	r	0.0168	0.0473	0.9359	0.01525
	Proposed	0.0215	0.0238	0.9547	0.00235

Table 8.18: $\beta_1 \neq \beta_2$, $\alpha_1 = 2, \alpha_2 = 5, \beta_1 = 3$ and β_2 is obtained by Equation (8.11), $(n, m) = (50, 10)$

R	Method	Lower Error	Upper Error	Central Coverage	Average Bias
	Wald	0.1257	0.0046	0.8697	0.06055
0.1	r	0.0392	0.0167	0.944115	0.01125
	Proposed	0.0207	0.0203	0.9590	0.00450
	Wald	0.0885	0.0154	0.8961	0.03655
0.2	r	0.0347	0.0223	0.9430	0.00620
	Proposed	0.0226	0.0226	0.9548	0.00240
	Wald	0.0657	0.0234	0.9109	0.02115
0.3	r	0.0332	0.0252	0.9416	0.00420
	Proposed	0.0221	0.0233	0.9546	0.002130
	Wald	0.0497	0.0335	0.9168	0.01660
0.4	r	0.0317	0.0286	0.9397	0.00515
	Proposed	0.0228	0.0241	0.9531	0.00155
	Wald	0.0368	0.0428	0.9204	0.01480
0.5	r	0.0285	0.0309	0.9405	0.00475
	Proposed	0.0222	0.0236	0.9542	0.00210
	Wald	0.0264	0.0499	0.9237	0.01315
0.6	r	0.0248	0.034830	0.9422	0.00410
	Proposed	0.0215	0.0231	0.9554	0.00270
	Wald	0.0184	0.0595	0.9221	0.02055
0.7	r	0.0222	0.0332	0.9446	0.00550
	Proposed	0.0206	0.0235	0.9559	0.00295
	Wald	0.0112	0.0715	0.9173	0.03015
0.8	r	0.0186	0.0324	0.9490	0.00690
	Proposed	0.0196	0.0212	0.9592	0.00460
	Wald	0.0055	0.0927	0.9018	0.04360
0.9	r	0.0149	0.0291	0.9560	0.00710
	Proposed	0.0183	0.0209	0.9608	0.00540

8.3 Conclusion and Future Work on Third-order Asymptotic Methods

A likelihood-based higher order asymptotic method is proposed to apply inference in the stress-strength reliability model when the two populations are distributed independently. When the scale parameters are different, the constrained MLE cannot be obtained by standard Lagrange multiplier methods and a penalized likelihood method is proposed. The proposed third-order method exhibits high accuracy and the penalized likelihood method can be easily employed for relatively small data set.

Based on the current work, there are several possible directions that research could be extended to:

- Theoretically, the proposed method can be applied to obtain inference for stress-strength reliability from any distributions. In this dissertation, the parameter is from a continuous distribution. However, when the distribution is of a discrete nature, the ancillary direction V cannot be obtained by differentiation. It will be important if an appropriate approach can be developed.

- The third-order method results in extremely accurate p-values and confidence intervals. In applied work, researchers may deal with inference for hierarchical structure data, for example, in medical science. Extending third-order methods to those fields will be an interesting future project.

- The exponential approximation plays an important role of providing accurate approximations to general statistical models in third-order inference. The is due to that $p(\theta)$ does not dependent on the non-exponential term to the third order, and it depends only on the observed likelihood and the gradient of the likelihood at the data point. This dissertation illustrates that this property hold for univariate model. When working with multivariate or non independent situation data, does this property still hold? This topic remains challenging.

Bibliography

[1] Adèr, H. J., Mellenbergh, G. J., and Huizen, D. J. H. (2011). Advising on research methods: A consultant's companion. *Journal of Applied Statistics*, 38.

[2] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009a). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14.

[3] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009b). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14.

[4] Ahmad, K., Fakhry, M., and Jaheen, Z. (1997). Empirical bayes estimation of $p(y < x)$ and characterizations of the burr-type x model. *Journal of Statistical Planning and Inference*, 64:297–308.

[5] Badar, M. and Priest, A. (1982). Statistical aspects of fibre and bundle strength

in hybrid composites. *Progress in Science and Engineering Composites,(T. Hayashi, K.Kawata, and S. Umekawa, eds.) ICCM-IV,*, pages 1129–1136.

[6] Bain, L. (1976). *Statistical analysis of reliability and life testing model.* Marcel and Dekker Inc., New York.

[7] Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley Series in probability and mathematical statistics. John Wiley & Sons.

[8] Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, 67:293–310.

[9] Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimate. *Biometrika*, 70:343–365.

[10] Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters, based on the standardized signed log likelihood ratio. *Biometrika*, 73:307–322.

[11] Barndorff-Nielsen, O. E. (1991). Modified signed log-likelihood ratio statistic. *Biometrika*, 78:557–563.

[12] Barndorff-Nielsen, O. E. (1994). *Inference and Aspptotics.* London: Chapman and Hall.

[13] Barndorff-Nielsen, O. E. and Chamberlin, S. R. (1991). An ancillary invariant modi-

fication of the signed log likelihood ratio. *Scandinavian Journal of Statistics*, 18:341–52.

[14] Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications (with discussion). *Journal of the Royal Statistical Society B*, 41:279–312.

[15] Barndorff-Nielsen, O. E. and Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of the Royal Statistical Society B*, 46:483–495.

[16] Barndorff-Nielsen, O. E. and Cox, D. R. (1989). Asymptotic techniques for use in statistics. *Metrika*, 37:216–216.

[17] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252.

[18] Bttcher, S. and Clarke, C. L. A. (2005). Efficiency vs. effectiveness in terabyte-scale information retrieval. In *TREC*, volume Special Publication 500-266.

[19] Buckley, C. and Robertson., S. E., editors (2008). *Relevance feedback track overview: TREC 2008*.

[20] Burr, I. W. (1942). Cumulative frequency functions. *Annuals of Mathematical Statistics*, 13:215–232.

[21] Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16:1190–1208.

[22] Cao, Y., Liu, J., Bao, S., and Li, H. (2005). Research on expert search at enterprise track of trec 2005. volume Special Publication 500-266.

[23] Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *In Research and Development in Information Retrieval*, pages 335–336.

[24] Carpineto, C., de Mori, R., Romano, G., and Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19:1–27.

[25] Clarke, C. L., Craswell, N., and Soboroff, I. (2009a). Overview of the trec 2009 web track. Technical report.

[26] Clarke, C. L., Craswell, N., and Soboroff, I. (2009b). Preliminary report on the trec 2009 web track. In *Proc. of TREC-18*.

[27] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008a). Novelty and diversity in information retrieval evaluation.

In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666.

[28] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008b). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666.

[29] Cooper, W. (1976). *The Suboptimality of Retrieval Rankings Based on Probability of Usefulness*. School of Library and Information Studies.

[30] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*.

[31] Craswell, N., Fetterly, D., Najork, M., Robertson, S., and Yilmaz, E. (2009). Microsoft research at trec 2009: Web and relevance feedback track. In *TREC*, volume Special Publication 500-278.

[32] Croft, W. B., Turtle, H. R., and Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum)*, pages 32–45. ACM.

[33] Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, 25:631–650.

[34] Demner-Fushman, D., Humphrey, S. M., Ide, N. C., Loane, R. F., Mork, J. G., Ruch, P., Ruiz, M. E., Smith, L. H., Wilbur, W. J., and Aronson, A. R. (2007). Combining resources to find answers to biomedical questions. In *TREC*.

[35] Demner-Fushman, D. and Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Comput. Linguist.*, 33:63–103.

[36] Doganaksoy, N. and Schmee, J. (1993). Comparisons of approximate confidence intervals for distributions used in life-data analysis. *Technometrics*, 35:175–184.

[37] Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika*, 67:334–487.

[38] Engle, R. F. (1984). *Handbook of Econometrics*, volume Volume 2, chapter Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, pages 775–826. Elsevier.

[39] Fagan, J. L. (1987). Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *SIGIR*, pages 91–101. ACM.

[40] Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

[41] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A*, 222:309–368.

[42] Fraser, D. A. S. (1988). Normed likelihood as saddlepoint approximation. *Multivariate Analysis*, 26:181–193.

[43] Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, 77:65–76.

[44] Fraser, D. A. S. and Reid, N. (1993). Third order asymptotic models: likelihood functions leading to accurate approximations for distribution functions. *Statistica Sinica*, 3:67–82.

[45] Fraser, D. A. S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematics*, pages 33–53.

[46] Fraser, D. A. S. and Reid, N. (1996). Ancillary information for statistical inference. *Technical report*.

[47] Fraser, D. A. S., Reid, N., and Wong, A. (1991). Exponential linear model: a two-

pass procedure for saddlepoint approximation. *Journal of Royal Statistical Society B*, 53:483–492.

[48] Fraser, D. A. S., Reid, N., and Wong, A. (2003). p-value formulas from likelihood asymptotics bridging the singularities. *Journal of Statistical Research*, 37:1–15.

[49] Fraser, D. A. S., Reid, N., and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and bayesian inference. *Biometrika*, 86:249–264.

[50] Gao, J., Nie, J.-Y., Wu, G., and Cao, G. (2004). Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 170–177.

[51] Gentle, J. E. (2002). *Elements of computational statistics.* Springer.

[52] Gobeill, J., Tbahriti, I., Ehrler, F., and Ruch, P. (2007). Vocabulary-driven passage retrieval for question-answering in genomics. In *Proceedings of the 16th Text REtrieval Conference (TREC)*, Maryland, USA.

[53] Gumbel, E. J. (1958). Statistics of extremes. *Columbia University Press, New York.*

[54] Gupta, R., Gupta, P., and Gupta, R. (1998). Modeling failure time data by lehman alternatives. *Communications in Statistics - Theory and Methods*, 27(4):887–904.

[55] Gupta, R. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical Journal*, 43:117–130.

[56] Hancock-Beaulieu, M., Gatford, M., Huang, X., Robertson, S. E., Walker, S., and Williams, P. W. (1996). Okapi at trec-5.

[57] Hanley, J. (1989). Receiver operating characteristic(roc) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging*, 29:307–335.

[58] He, B., Huang, J. X., and Zhou, X. (2011). Modeling term proximity for probabilistic information retrieval models. *Information Sciences*, 181(14):3017 – 3031.

[59] Hersh, W., Cohen, A., Ruslen, L., and Roberts, P. (2007). TREC 2007 Genomics Track Overview. In *In Proceedings of the Text REtrieval Conference (TREC*.

[60] Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006a). TREC 2006 genomics track overview.

[61] Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006b). Trec 2006 genomics track overview. In *In Proceedings of the Text REtrieval Conference (TREC*.

[62] Hogg, R. V., Craig, A., and Mckean, J. W. (2004). *Introduction to Mathematical Statistics*. Prentice Hall, 6th edition.

[63] Ide, N. C., Loane, R. F., and Demner-Fushman, D. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of American Medical Informatics Association*, 14:253–263.

[64] Jaheen, Z. (1995). Bayesian approach to prediction with outliers from the burr type x model. *Microelectronics Reliability*, 35:703–705.

[65] Jaheen, Z. (1996). Empirical bayes estimation for the reliability and failure rate function of burr type x failure model. *Journal of Applied Statistical Science*, 3:281–288.

[66] Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika*, 79(4):693–703.

[67] Kaptein, R., Koolen, M., and Kamps, J. (2009). Result diversity and entity ranking experiments: Anchors, links, text and wikipedia. In *TREC*, volume Special Publication 500-278.

[68] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

[69] Kraaij, W., Westerveld, T., and Hiemstra, D. (2002). The importance of prior probabilities for entry page search. SIGIR '02, pages 27–34.

[70] Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127.

[71] Li, J. and Yan, H. (2006). Peking university at the trec 2006 terabyte track. volume Special Publication 500-272.

[72] Losada, D. E. and Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Inf. Retr.*, 11(2):109–138.

[73] Losee, R. M. and Jr. (1994). Term dependence: Truncating the bahadur lazarsfeld expansion. In *Information Processing and Management*, pages 293–303.

[74] Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Advanced Applied Probability*, 12:475–490.

[75] Lv, Y. and Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 299–306.

[76] McCreadie, R., Macdonald, C., Ounis, I., Peng, J., and Santos, R. L. T. (2009). University of glasgow at trec 2009: Experiments with terrier. In *TREC*.

[77] McCullagh, P. (1987). *Tensor methods in statistics.* Monographs on statistics and applied probability. Chapman and Hall, London [u.a.].

[78] Metzler, D. and Croft, W. B. (2005a). A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479.

[79] Metzler, D. and Croft, W. B. (2005b). A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479.

[80] Metzler, D., Strohman, T., and Croft, W. B. (2006). Indri trec notebook 2006: Lessons learned from three terabyte tracks. volume Special Publication 500-272.

[81] Mishne, G. and de Rijke, M. (2005). Boosting web retrieval through query operations. In *ECIR*, volume 3408, pages 502–516.

[82] Mudholkar, G. and Srivastava, D. (1993). Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability.*, 42:299–302.

[83] Mudholkar, G., Srivastava, D., and Freimer, M. (1995). The exponentiated weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37:436–445.

[84] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests

of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

[85] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006a). Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

[86] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., and Soboroff, I. (2006b). Overview of the trec-2006 blog track. In *Text Retrieval Conference*.

[87] Plachouras, V. and Ounis, I. (2007a). Multinomial randomness models for retrieval with document fields. In *ECIR*, volume 4425, pages 28–39.

[88] Plachouras, V. and Ounis, I. (2007b). Multinomial randomness models for retrieval with document fields. In *ECIR*, volume 4425, pages 28–39. Springer.

[89] Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44:50–57.

[90] Raqab, M. and Kundu, D. (2005). Comparison of different estimators of $p(y < x)$ for a scaled burr type x distribution. *Communications in Statistics - Simulation and Computation*, 34:465–482.

[91] Raqab, M. and Kundu, D. (2006). Burr type x distribution: Revisited. *Journal of Probability and Statistical Sciences*, 2:179–193.

[92] Rasolofo, Y. and Savoy, J. (2003). Term proximity scoring for keyword-based retrieval systems. In *In Proc. of the 25th European Conf. on IR Research*, pages 207–218.

[93] Reid, N. (1988). Saddlepoint Methods and Statistical Inference. *Statistical Science*, 3:213–227.

[94] Reid, N. (1996). Likelihood and higher-order approximation to tail areas: a review and annotated bibliography. *Canadian Journal of Statistics*, 24:141–166.

[95] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at trec-3. pages 109–126.

[96] Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of Documentation*, 33:294–304.

[97] Robertson, S. E. and Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

[98] Robertson, S. E., van Rijsbergen, C. J., and Porter, M. F. (1981). Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, SIGIR '80, pages 35–56.

[99] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

[100] Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gatford, M., and Payne, A. (1995). Okapi at trec-4. In *TREC*.

[101] Santos, R. L., Macdonald, C., and Ounis, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 881–890.

[102] Santos, R. L. T., Peng, J., Macdonald, C., and Ounis, I. (2010b). Explicit search result diversification through sub-queries. In *ECIR*, volume 5993, pages 87–99.

[103] Sartawi, H. and Abu-Salih, M. (1991). Bayesian prediction bounds for the burr type x model. *Communications in Statistics-Theory and Methods*, 20:2307–2330.

[104] Schechtman, E. and Sherman, M. (2007). The two-sample t-test with a known ratio of variances. *Statistical Methodology*, 4:508–514.

[105] Shao, J. (2003). *Mathematical Statistics.* Springer-Verlag New York Inc, 2nd edition.

[106] She, Y., Wong, A., and Zhou, X. (2011). Revisit the two sample t-test with a known ratio of variances. *Open Journal of Statistics*, 1:151–156.

[107] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

[108] Singhal, A., Buckley, C., and Mitra, M. (1996a). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 21–29.

[109] Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996b). Document length normalization. *Inf. Process. Manage.*, 32:619–633.

[110] Song, R., Wen, J.-R., Shi, S., Xin, G., Liu, T.-Y., Qin, T., Zheng, X., Zhang, J., Xue, G.-R., and Ma, W.-Y. (2004). Microsoft research asia at web track and terabyte track of trec 2004. In *TREC*, volume Special Publication 500-261.

[111] Stokes, N., Li, Y., Cavedon, L., Huang, E., Rong, J., and Zobel, J. (2007). Entity-based relevance feedback for genomic list answer retrieval. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007.*

[112] Surles, J. and Padgett, W. (1998). Inference for $p(y < x)$ in the burr type x model. *Journal of Applied Statistical Science*, 7:225–238.

[113] Surles, J. and Padgett, W. (2001). Inference for reliability and stress-strength for a scaled burr type x distribution. *Lifetime Data Analysis*, 7:187–200.

[114] Surles, J. and Padgett, W. (2005). Some properties of a scaled burr type x distribution. *Journal of Statistical Planning and Inference*, 128:271–280.

[115] van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33:106–199.

[116] Voorhees, E., Harman, D., of Standards, N. I., and (US), T. (2005). *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridgeˆ eMA MA.

[117] Voorhees, E. M. and Buckland, L. P., editors (2009). *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*.

[118] Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54:426–482.

[119] Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

[120] Wolfe, D. A. and Hogg, R. V. (1971). On constructing statistics and reporting data. *The American Statistician*, 25:27–30.

[121] Ye, Z., Huang, X., He, B., and Lin, H. (2009). York university at trec 2009: Relevance feedback track. In *TREC*, volume Special Publication 500-278.

[122] Yin, X., Huang, J. X., and Li, Z. (2011). Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information Processing and Management*, 47:53–67.

[123] Yin, X., Huang, J. X., Zhou, X., and Li, Z. (2010). A survival modeling approach to biomedical search result diversification using wikipedia. In *SIGIR*, pages 901–902.

[124] Yu, C. T., Buckley, C., Lam, K., and Salton, G. (1983). A Generalized Term Dependence Model in Information Retrieval. Technical report, Cornell University.

[125] Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 10–17.

[126] Zhang, Y., Callan, J., and Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR*

conference on Research and development in information retrieval, SIGIR '02, pages 81–88.

[127] Zhao, J. and Yun, Y. (2009). A proximity language model for information retrieval. In *SIGIR*, pages 291–298.

[128] Zhou, X., Huang, J. X., and He, B. (2011). Enhancing ad-hoc relevance weighting using probability density estimation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 175–184.

[129] Zhu, J., Huang, X., Song, D., and Rger, S. M. (2010). Integrating multiple document features in language models for expert finding. *Knowl. Inf. Syst.*, 23:29–54.