Score Distributions in Information Retrieval

Avi Arampatzis[1], Stephen Robertson[2], and Jaap Kamps[1]

[1] University of Amsterdam, the Netherlands
[2] Microsoft Research, Cambridge UK

Abstract. We review the history of modeling score distributions, focusing on the mixture of normal-exponential by investigating the theoretical as well as the empirical evidence supporting its use. We discuss previously suggested conditions which valid binary mixture models should satisfy, such as the Recall-Fallout Convexity Hypothesis, and formulate two new hypotheses considering the component distributions under some limiting conditions of parameter values. From all the mixtures suggested in the past, the current theoretical argument points to the two gamma as the most-likely universal model, with the normal-exponential being a usable approximation. Beyond the theoretical contribution, we provide new experimental evidence showing vector space or geometric models, and BM25, as being "friendly" to the normal-exponential, and that the non-convexity problem that the mixture possesses is practically not severe.

1 Introduction

Current best-match retrieval models calculate some kind of score per collection item which serves as a measure of the degree of relevance to an input request. Scores are used in ranking retrieved items. Their range and distribution varies wildly across different models making them incomparable across different engines [1], even across different requests on the same engine if they are influenced by the length of requests. Even most probabilistic models do not calculate the probability of relevance of items directly, but some order-preserving (monotone or isotone) function of it [2].

For single-collection ad-hoc retrieval, the variety of score types is not an issue; scores do not have to be comparable across models and requests, since they are only used to rank items per request per system. However, in advanced applications, such as distributed retrieval, fusion, or applications requiring thresholding such as filtering or recall-oriented search, some form of score normalization is imperative. In the first two applications, several rankings (with non-overlapping and overlapping sets of items respectively) have to be merged or fused to a single ranking. Here, score normalization is an important step [3]. In practice, while many users never use meta-search engines directly, most conventional search engines have the problem of combining results from many discrete sub-engines. For example, blending images, text, inline answers, stock quotes, and so on, has become common.

In filtering, bare scores give no indication on whether to retrieve an incoming document or not. Usually a user model is captured into some evaluation measure. Some of these measures can be optimized by thresholding the probability of relevance at some specific level [4], thus a method of normalizing scores into probabilities is needed.

Moreover, thresholding has turned out to be important in recall-oriented retrieval setups, such as legal or patent search, where ranked retrieval has a particular disadvantage in comparison with traditional Boolean retrieval: there is no clear cut-off point where to stop consulting results [5]. Again, normalizing scores to expected values of a given effectiveness measure allows for optimal rank thresholding. In any case, the optimal threshold depends on the effectiveness measure being used—there is no single threshold suitable for all purposes.

Simple approaches, e.g. range normalization based on minimum and maximum scores, are rather naive, considering the wild variety of score outputs across search engines, because they do not take into account the *shape* of score distributions (SDs). Although these approaches have worked reasonably well for merging or fusing results [6], advanced approaches have been seen which try to improve normalization by investigating SDs. Such methods have been found to work at least as well (or in some cases better than) the simple ones in the context of fusion [7, 8]. They have also been found effective for thresholding in filtering [9–11] or thresholding ranked lists [12]. We are not aware of any empirical evidence in the context of distributed retrieval.

We review the history of modeling SDs in Information Retrieval, focusing on the currently most popular model, namely, the mixture of normal-exponential, by investigating the theoretical as well as the empirical evidence supporting its use. We discuss conditions which any valid—from an IR perspective—binary mixture model should satisfy, such as the Recall-Fallout Convexity Hypothesis, and formulate new hypotheses considering the component distributions individually as well as in pairs. Although our contribution is primarily theoretical, we provide new experimental evidence concerning the range of retrieval models that the normal-exponential gives a good fit, and try to quantify the impact of non-convexity that the mixture possesses. We formulate yet unanswered questions which should serve as directions for further research.

2 Modeling Score Distributions

Under the assumption of a binary relevance, classic attempts model SDs, on a per-request basis, as a mixture of two distributions: one for relevant and the other for non-relevant documents [13–17, 7]. Given the two component distributions and their mix weight, the probability of relevance of a document given its score can be calculated straightforwardly [17, 7], essentially allowing the normalization of scores into probabilities of relevance. Furthermore, the expected numbers of relevant and non-relevant documents above and below any rank or score can be estimated, allowing the calculation of precision, recall, or any other traditional measure at any given threshold enabling its optimization [12]. Assuming the right component choices, such methods are theoretically "clean" and non-parametric.

A more recent attempt models aggregate SDs of many requests, on per-engine basis, with single distributions [18, 8]; this enables normalization of scores to probabilities—albeit not of relevance—comparable across different engines. The approach was found to perform better than the simple methods in the context of fusion [8]. Nevertheless, it is not clear—if it is even possible—how using a single distribution can be applied to thresholding, where for optimizing most common measures a reference to relevance is

needed. For this reason, we will next concentrate on binary mixture models; moreover, we are not aware of any approach using SDs in beyond binary relevance setups.

Various combinations of distributions have been proposed since the early years of IR—two normal of equal variance [13], two normal of unequal variance or two exponential [14], two Poisson [15], two gamma [16]—with currently the most popular model being that of using a normal for relevant and an exponential for non-relevant, introduced in [9] and followed up by [17, 7, 10, 11] and others. For a recent extended review and theoretical analysis of the above choices, we refer the reader to [1]. The latest improvements of the normal-exponential model use truncated versions of the component densities, trying to deal with some of its shortcomings [12]. Next we focus on the original normal-exponential model.

3 The Normal-Exponential Model

In this section, we review the normal-exponential model. We investigate the theoretical as well as the empirical evidence and whether these support its use.

3.1 Normal for Relevant

A theorem by Arampatzis and van Hameren [17] claims that the distribution of relevant document scores converges to a *Gaussian central limit* (GCL) quickly, with "corrections" diminishing as $O(1/k)$ where k is the query length. Roughly, three explicit assumptions were made:

1. Terms occur independently.
2. Scores are calculated via some linear combination of document term weights.
3. Relevant documents cluster around some point in the document space, with some hyper-ellipsoidal density (e.g. a hyper-Gaussian) with tails falling fast enough.

Next, we re-examine the validity and applicability of these assumptions in order to determine the range of retrieval models for which the theorem applies.

Assumption 1 is generally untrue, but see the further discussion below. Assumption 2 may hold for many retrieval models; e.g. it holds for dot-products in vector space models, or sums of partially contributing log-probabilities (log-odds) in probabilistic models. Assumption 3 is rather geometric and better fit to vector space models; whether it holds or not, or it applies to other retrieval models, is difficult to say. Intuitively, it means that the indexing/weighting scheme does its job: it brings similar documents close together in the document space. This assumption is reasonable and similar to the Cluster Hypothesis of K. van Rijsbergen [19, Chapter 3]. Putting it all together, the proof is more likely to hold for setups combining the following three characteristics:

- Vector space model, or some other geometric representation.
- Scoring function in the form of linear combination of document term weights, such as the dot-product or cosine similarity of geometric models or the sum of partially contributing log-probabilities of probabilistic models.
- Long queries, due to the convergence to a GCL depending on query length.

This does not mean that there exists no other theoretical proof applicable to more retrieval setups, but we have not found any in the literature.

A Note on Term Independence. Term independence assumptions are common in the context of probabilistic models and elsewhere, but are clearly not generally valid. This has elicited much discussion. The following points have some bearing on the present argument:

- Ranking algorithms derived from independence models have proved remarkably robust, and unresponsive to attempts to improve them by including dependencies.
- Making the independence assumption conditional on relevance makes it a little more plausible than a blanket independence assumption for the whole collection.
- Cooper [20] has shown that for the simple probabilistic models, one can replace the independence assumptions with linked dependence (that is, linked between the relevant and non-relevant sets), and end up with the same ranking algorithms. This may be a partial explanation for the robustness of the independence models.
- This linked dependence unfortunately does not help us with the present problem.
- Cooper et al. [21] show that if we want to estimate an explicit, well-calibrated probability of relevance for each document (to show to the user), then corrections need to be made to allow for the inaccuracies of the (in)dependence assumptions.

What these points emphasise is the very strong distinction between on the one hand having a scoring system which ranks well and on the other hand placing any stronger interpretation on the scores themselves.

3.2 Exponential for Non-Relevant

Under a similar set of assumptions and approximations, Arampatzis and van Hameren [17] investigate also the distribution of non-relevant document scores and conclude that a GCL is unlikely and if it appears it does only at a very slow rate with k (practically never seen even for massive query expansion). Although such a theorem does not help much in determining a usable distribution, under its assumptions it contradicts Swets' use of a normal distribution for non-relevant [13, 14].

The distribution in question does not necessarily have to be a known one. [17] provides a model for calculating numerically the SD of any class of documents (thus also non-relevant) using Monte-Carlo simulation. In absence of a related theory or a simpler method, the use of the exponential distribution has been so far justified empirically: it generally fits well to the *high-end* of non-relevant item scores, but not to all.

3.3 Normal-Exponential in Practice

The normal-exponential mixture model presents some practical difficulties in its application. Although the GCL is approached theoretically quickly as query length increases, practically, queries of length above a dozen terms are only possible through relevance feedback and other learning methods. For short queries, the Gaussian may simply not be there to be estimated. Empirically, using a vector space model with scores which were unbounded above on TREC data, [17] found usable Gaussian shapes to form at around $k = 250$. k also seemed to depend on the quality of a query; the better the query, the fewer the terms necessary for a normal approximation of the observed distribution.

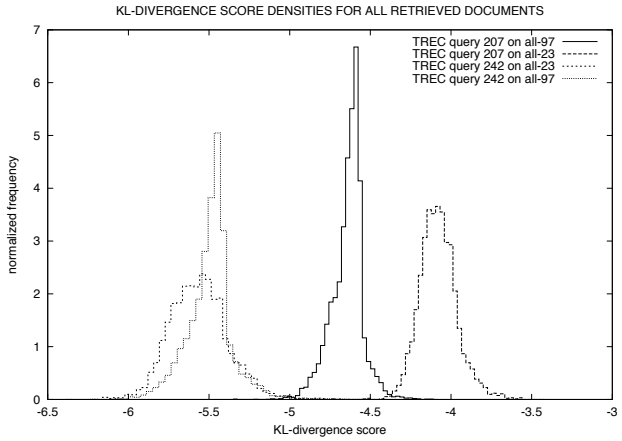

Fig. 1. KL-divergence score densities; two queries on two collections.

Along similar lines, [7] noticed that better systems (in terms of average precision) produce better Gaussian shapes.

It was also shown in previous research that the right tail of the distribution of non-relevant document scores can be very well approximated with an exponential: [17, 11] fit on the top 50–100, [7] fit on almost the top-1,000 (1,000 minus the number of relevant documents). [22] even fits on a non-uniform sample of the whole score range, but the approach seems system/task-specific. In general, it is difficult to fit an exponential on the whole score range. Figure 1 shows the total score densities produced by a combination of two queries and two sub-collections using KL-DIVERGENCE as a retrieval model. Obviously, none of these SDs can be fitted *in totality* with the mixture. Candidate ranges are, in general, $[s_{\text{peak}}, +\infty)$ where s_{peak} is set at the most frequent score or above.

Despite the above-mentioned practical problems, [7] used the model with success, with much shorter queries and even with a scoring system which produces scores between 0 and 1 without worrying about the implied truncation at both ends for the normal and at the right end for the exponential. In the context of thresholding for document filtering [11], with the generally unbounded scoring function BM25 and a maximum of 60 query terms per profile, the method performed well (2nd best, after Maximum Likelihood Estimation) on 3 out of 4 TREC data sets.

To further determine the retrieval models whose observed SDs can be captured well with a normal-exponential mixture, we investigated all 110 submissions to the TREC 2004 Robust track. This track used 250 topics combining the ad-hoc track topics in TRECs 6–8, with the robust track topics in TRECs 2003–2004. Table 1 shows the 20 submissions where the mixture obtained the best fit as measured by χ^2 goodness-of-fit test. The table shows the run names; the used topic fields; the median χ^2 upper probability indicating the goodness-of-fit; and the correlation between the optimal $F_1@K$ (with K a rank) based on the qrels and on the fitted distributions. The two remaining columns will be discussed in Section 4. Not surprisingly, over all runs, the 20 runs with the best fit also tend to have better predictions of $F_1@K$.

Table 1. Twenty submissions with the best normal-exponential goodness-of-fit.

Run	Qry	χ^2	F_1 c.	NC	Inv.	Run	Qry	χ^2	F_1 c.	NC	Inv.
icl04pos2d	d	0.228	0.742	1.0	95.76	icl04pos2t	t	0.163	0.752	2.5	93.05
SABIR04FA	tdn	0.214	0.650	1.0	87.57	uogRobDWR10	d	0.158	0.642	1.0	89.35
icl04pos7f	tdn	0.197	0.663	2.0	93.64	wdo25qla1	tdn	0.157	0.579	4.0	83.12
icl04pos2f	tdn	0.190	0.629	1.0	93.66	icl04pos2td	td	0.154	0.718	1.0	95.87
SABIR04BA	tdn	0.185	0.658	1.0	90.25	uogRobLWR5	tdn	0.152	0.593	1.0	90.19
NLPR04OKapi	d	0.184	0.708	3.0	90.29	icl04pos7td	td	0.152	0.744	1.0	95.40
SABIR04FT	t	0.182	0.723	2.0	90.31	SABIR04BT	t	0.149	0.712	1.0	91.08
SABIR04FD	d	0.180	0.668	2.0	88.23	wdoqla1	tdn	0.149	0.637	2.0	85.66
SABIR04BD	d	0.174	0.647	2.0	88.05	uogRobDBase	d	0.148	0.646	1.0	88.31
icl04pos48f	tdn	0.166	0.694	1.0	95.78	fub04Dg	d	0.145	0.511	2.5	86.82

Looking at the retrieval models resulting in the best fits, we see seven runs of Peking University (icl) using a vector space model and the cosine measure. We also see 6 runs of Sabir Research, Inc. (SABIR) using the SMART vector space model. There are 3 runs of the University of Glasgow (uog) using various sums of document term weights in the DRF-framework. Two runs from Indiana University (wdo) using Okapi BM25. Finally, a single run from the Chinese Academy of Science (NLPR) using Okapi BM25, and one from Fundazione Ugo Bordoni (fub) also using sums of document term weights in the DRF-framework. Overall, we see support for vector space or geometrical models as being amenable to the normal-exponential mixture, as well as BM25.

Looking at query length, we see only 3 systems using the short title statement, and 8 systems using all topic fields. Many of the systems used query expansion, either using the TREC corpus or using the Web, leading to even longer queries. While longer queries tend to lead to smoother SDs and improved fits, the resulting $F_1@K$ prediction seems better for the short title queries with high quality keywords. The "pos2" runs of Peking University (icl) only index verbs and nouns, and considering only the most informative words seems to help distinguish the two components in the mixture.

4 The Recall-Fallout Convexity Hypothesis

From the point of view of how scores or rankings of IR systems should be, Robertson [1] formulates the Recall-Fallout Convexity Hypothesis:

> *For all good systems, the recall-fallout curve (as seen from the ideal point of recall=1, fallout=0) is convex.*

Similar hypotheses can be formulated as conditions on other measures, e.g., the probability of relevance should be monotonically increasing with the score; the same should hold for *smoothed* precision. Although, in reality, these conditions may not always be satisfied, they are expected to hold for good systems, i.e. those producing rankings satisfying the *probability ranking principle* (PRP), because their failure implies that systems can be easily improved.

As an example, let us consider smoothed precision. If it declines as score increases for a part of the score range, that part of the ranking can be improved by a simple random

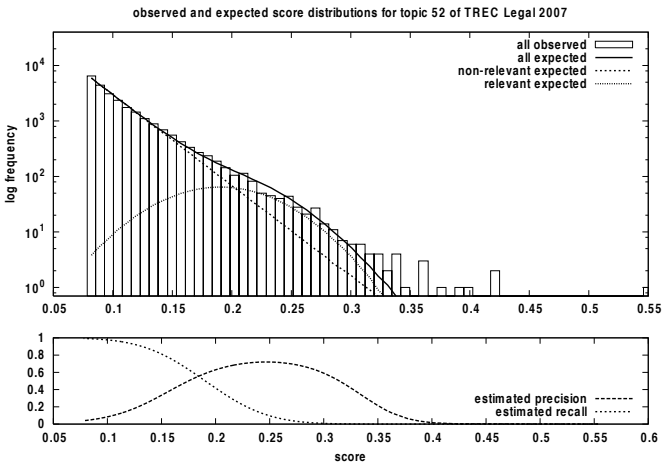

Fig. 2. Non-convexity inside the observed score range of a normal-exponential fit.

re-ordering [23]. This is equivalent of "forcing" the two underlying distributions to be uniform in that score range. This will replace the offending part of the precision curve with a flat one—the least that can be done—improving the overall effectiveness of the system. In fact, rankings can be further improved by reversing the offending sub-rankings; this will force the precision to increase with an increasing score, leading to better effectiveness than randomly re-ordering the sub-ranking.

Such hypotheses put restrictions on the relative forms of the two underlying distributions. Robertson [1] investigated whether the following mixtures satisfy the convexity hypothesis: two normals, two exponentials, two Poisson, two gamma, and normal-exponential. From this list, the following satisfy the hypothesis: two normal (only when their variances are equal), two exponential, two Poisson, and two gamma (for a quite wide range of parameters but not all).

Let us consider the normal-exponential mixture which violates such conditions only (and always) at both ends of the score range. Although the low-end scores are of insignificant importance, the top of the ranking is very significant. The problem is a manifestation of the fact that a normal falls more rapidly than an exponential and hence the two density functions intersect twice. Figure 2 depicts a normal-exponential fit on score data, together with the estimated precision and recall. The problem can be seen here as a declining precision above score 0.25.

In adaptive filtering, [9, 22] deal with the problem by selecting as threshold the lower solution of the 2nd degree equation resulting from optimizing linear utility measures, while [10, 11] do not seem to notice or deal with it. In meta-search, [7] noted the problem and *forced* the probability to be monotonic by drawing a straight line from the point where the probability is maximum to the point $[1, 1]$. Both procedures, although they may have been suitable for the above tasks, are theoretically unjustified. In [12], the two component distributions were set to uniform within the offending score range; as noted above, this is equivalent to randomization.

The problem does not seem severe for thresholding tasks. For example, [12] tried to optimize the F_1 measure and found that the impact of randomization on thresholding is that the SD method turns "blind" inside the offending range. As one goes down the corresponding ranks, estimated precision would be flat, recall naturally rising, so the optimal F_1 threshold can only be below the range. On average, the optimal rank threshold was expected to be deeper than the affected ranks, so the impact of non-convexity on thresholding deemed to be insignificant. Sometimes the problem may even appear above the maximum observed score. Furthermore, the truncated normal-exponential model used in [12] also helped to alleviate non-convexity by sometimes out-truncating it; a modest and conservative theoretical improvement over the original model which always violates the hypothesis.

To further determine the effect of the non-convexity of the normal-exponential, we again investigate the 110 submissions to the TREC 2004 Robust track. Table 1 also shows the median rank at which the estimated precision peaks (hence there is a non-convexity problem before this rank). We also show the effect of inverting the initial non-convex ranks, in percentage of overall MAP. That is, if precision increases up to rank 3 then it should make sense to invert the ranking of the first 2 documents. Two main observations are made. First, the median rank down to which the problem exists is very low, in the range of 1 (i.e. no practical problem) to 4, suggesting a limited impact on at least half the topics. Although there are outlier topics where the problem occurs far down the ranking, some of these may be due to problematic fits [12]. Second, "fixing" the problematic initial ranks by inverting the order leads to a loss of MAP throughout. This signals that the problem is not inherent in the underlying retrieval model violating the PRP. Rather, the problem is introduced by the fitted normal-exponential; both practical and fundamental problems can cause a misfit given the limited information available.

In the bottom line, the PRP dictates that any theoretically sound choice of component densities should satisfy the convexity condition; from all the mixtures suggested in the past, the normal-exponential as well as the normal-normal of unequal variances do not, for all parameter settings. In practice, the problem does not seem to be severe in the case of normal-exponential; the affected ranks are usually few. Given the theoretical and empirical evidence, we argue that the problem is introduced by the exponential, not by the normal. Moreover, many distributions—especially "peaky" ones—have a GCL. For example, assuming Poisson-distributed relevant document scores, for a system or query with a large mean score the Poisson would converge to a normal.

5 In-the-limit Hypotheses

The Recall-Fallout Convexity Hypothesis considers the validity of *pairs* of distributions under the PRP. There are some reasons for considering distributions in pairs, as follows:

- The PRP is about the relative ranking of relevant and non-relevant documents under conditions of uncertainty about the classification; it makes no statements about either class in isolation.
- Consideration of the pair makes it possible for the hypothesis to ignore absolute scores, and therefore to be expressed in a form which is not affected by any mono-

tonic transformation of the scores. Since ranking itself is not affected by such a transformation, this might be considered a desirable property.
– If we wish in the future to extend the analysis to multiple grades of relevance, a desirable general form would be a parametrised family of distributions, with different parameter values for each grade of relevance (including non-relevance), rather than a separately defined distribution for each grade.

However, the evidence of previous work suggests that the distributions of relevant and non-relevant look very different. This renders the third point above difficult to achieve, and further suggests that we might want to identify suitable hypotheses to apply to each distribution separately. Here we consider two hypotheses, the first of which achieves some degree of separation but may be difficult to support; the second is expressed in relative terms but may be more defensible.

Note that both hypothesis are "in the limit" conditions—they address what happens to the SDs under some limiting conditions of parameter values. They do not address the behaviour of distributions in other than these limiting conditions. Therefore they do not imply anything like the Recall-Fallout Convexity Hypothesis under actually observed parameter values.

5.1 The Strong Hypothesis

The ultimate goal of a retrieval system is not to produce some SD, but rather deliver the right items. In this light, the observed SD can be seen as an artifact of the inability of current systems to do a direct classification. Therefore, the ultimate SD all systems are trying to achieve is to the one with all relevant documents at the same high score $s_{\max}$, and all non-relevant documents at the same low score $s_{\min}$. The better the system, the better it should approximate the ultimate SD. This imposes restrictions on the two underlying components:

The Strong SD Hypothesis. *For good systems, the score densities of relevant and non-relevant documents should be capable of approaching Dirac's delta function, shifted to lie on the maximum score for the relevant and on the minimum score for the non-relevant, in some limiting condition.*

Let us now investigate which of the historically suggested distributions can approximate a delta and how.

The normal goes to delta via $\sigma \to 0$, and it can be positioned on demand via μ. The exponential approximates delta only via $\lambda \to +\infty$. The Poisson has one parameter λ, which incidentally equals both its mean and variance. For large λ, it approximates a normal with a mean and variance of λ. Consequently, as λ grows, the variance grows as well and it will never reach a delta. At the other side, for $\lambda = 0$ it becomes Kronecker's delta, i.e. the discrete analogue of Dirac's delta. The gamma has two parameters, $\Gamma(k, \theta)$. For large k it converges to a Gaussian with mean $k\theta$ and variance $k\theta^2$. The variance grows with k, but for $\theta \to 0$ it declines faster than the mean. So, the gamma can approximate a delta via an increasingly narrow Gaussian, and it can be positioned on demand via proper choices of k and θ.

Consequently, under the Strong SD Hypothesis, good candidates for relevant document scores are the normal or gamma, while for non-relevant are the normal, Poisson, exponential, or gamma. We only manage to reject the use of exponential and Poisson for relevant; although these could be simply shifted at $s_{\max}$ or vertically mirrored to end at $s_{\max}$, those setups would seem rather strange and unlikely.

Considering the historically suggested pairs of distributions, we can reject the mixture of two exponentials—at least as it was suggested in [14]: while the non-relevant exponential can approximate $\delta(s - s_{\min})$ for $\lambda \to +\infty$, the relevant exponential cannot approximate $\delta(s - s_{\max})$ for any λ. The two Poisson mixture of [15] is similarly rejected. The pairs remaining are the two normal, two gamma, or normal-exponential. Since a normal for non-relevant is unlikely according to [17] and Section 3.2, that leaves us with the two gamma or normal-exponential with only the former satisfying the convexity hypothesis for a range of parameter settings—not all. Note also that the two exponential or two Poisson constructions with the relevant component vertically mirrored would violate the Recall-Falout Convexity Hypothesis.

5.2 The Weak Hypothesis

The Strong SD Hypothesis would like to see all relevant documents at the same (high) score, and all non-relevant documents at the same (low) score. This requirement is not really compatible with any notion that there may actually be degrees of relevance (even if the user makes a binary decision), and is also not necessary for perfect ranking performance—either or both classes might cover a range of scores, provided only that they do not overlap. Thus we can formulate a weaker hypothesis:

The Weak SD Hypothesis. *For good systems, the score densities of relevant and non-relevant documents should be capable of approaching full separation in some limiting condition.*

Clearly, the Strong Hypothesis implies the Weak Hypothesis, because the Dirac delta function gives full separation.

The Weak Hypothesis, however, would not reject the mixture of two exponentials: as we push the mean of the non-relevant distribution down, non-relevant scores are increasingly concentrated around zero, while if we push the mean of the relevant distribution up, the relevant scores are more and more widely spread among high values. In the limit, perfect separation is achieved. The Weak Hypothesis also does not reject the Poisson mixture, if we achieve the limit by letting lambda go to zero for non-relevant and to infinity for relevant. This is similar to the mixture of two exponentials, except that the relevant scores are uniformly distributed over the positive integers only, instead of the positive real line.

The Weak Hypothesis is indeed weak, in that it does not reject any of the combinations previously discussed. However, it reveals significant differences in the notions of "perfect" retrieval effectiveness implicit in different combinations (and therefore what form improvements should take in SD terms). This "in the limit" behaviour is worth further exploration.

6 Conclusions and Directions for Future Research

The empirical evidence so far confirm that SD methods are effective for thresholding in filtering or ranked lists, as well as score normalization in meta-search. Specifically, the normal-exponential model seems to fit best vector space or geometric and BM25 retrieval models. Some mixtures have theoretical problems with an unclear practical impact. For example, using the normal-exponential model for thresholding the impact of non-convexity seems insignificant, however, elsewhere the effect may vary. Latest improvements of the model, namely, using truncated component densities alleviate the non-convexity problem—providing also better fits on data and better end-effectiveness in thresholding—without eliminating it [12].

The classic methods assume a binary relevance. A different approach would have to be taken, if degrees of relevance are assumed. For example, in TREC Legal 2008, there was a 3-way classification into non-relevant, relevant, and highly relevant. This complicates the analysis considerably, suggesting the need for three distributions. In this respect, it would fit more naturally with a model where both or all distributions came from the same family. It is difficult to see how one could adapt something like the normal-exponential combination to this situation. On the flip-side, approaches that analyze SDs without reference to relevance are just beginning to spring up [8]; nevertheless, these seem more suitable for score normalization for distributed IR or fusion rather than thresholding tasks.

An alternative approach would be to devise new scoring functions that have good distributional properties, or seek a calibration function by trying out different transformations on the scores of an existing system. Following the discussion on independence, we make a connection with the work of Cooper et al. [21], who argue that systems *should* give users explicit probability-of-relevance estimates, and use logistic regression techniques to achieve this. The idea of using logistic regression in this context dates back in [24], and re-iterated by others, e.g., [2]. The SD analysis indicates that in principle there should be such a calibration, which would take the form of a monotonic transformation of the score function, and therefore not affecting the ranking. Probability of relevance itself is sufficient for some of the thresholding tasks identified in the introduction but not for all—some require more complete distributional information. However, given probabilities of relevance we may find it easier to perform SD analysis and the chances of discovering a universal pair of distributions greater.

A universal pair should satisfy some conditions from an IR perspective. Although the two new hypotheses we introduced do not seem to align their demands with each other or with the older one, the pair that seems more "bullet-proof" is that of the two gamma suggested by [16]. The gamma can also become normal via a GCL or exponential via $k = 1$, thus allowing for the two exponential and normal-exponential combinations which are also likely depending on which conditions/hypotheses one considers. The increased degrees of freedom offered by the two gamma, however, is a two-edged sword: it may just allow too much. Parameter estimation methods introduce another layer of complexity, approximations, and new problems, as voiced by most previous experimental studies and more recently by [25]. At any rate, the distributions in question do not necessarily have to be known ones.

References

1. Robertson, S.: On score distributions and relevance. In: Proceedings ECIR'07, Springer (2007) 40–51
2. Nottelmann, H., Fuhr, N.: From uncertain inference to probability of relevance for advanced IR applications. In: Proceedings ECIR'03. (2003) 235–250
3. Callan, J.: Distributed information retrieval. In: Advances Information Retrieval: Recent Research from the CIIR. Kluwer Academic Publishers (2000) 127–150
4. Lewis, D.D.: Evaluating and optimizing autonomous text classification systems. In: Proceedings SIGIR'95, ACM Press (1995) 246–254
5. Oard, D.W., Hedin, B., Tomlinson, S., Baron, J.R.: Overview of the TREC 2008 legal track. In: Proceedings TREC 2008. (2009)
6. Lee, J.H.: Analyses of multiple evidence combination. In: Proceedings SIGIR'97, ACM Press (1997) 267–276
7. Manmatha, R., Rath, T.M., Feng, F.: Modeling score distributions for combining the outputs of search engines. In: Proceedings SIGIR'01, ACM Press (2001) 267–275
8. Fernández, M., Vallet, D., Castells, P.: Using historical data to enhance rank aggregation. In: Proceedings SIGIR'06, ACM Press (2006) 643–644
9. Arampatzis, A., Beney, J., Koster, C.H.A., van der Weide, T.P.: Incrementality, half-life, and threshold optimization for adaptive document filtering. In: Proceesing TREC 2000. (2000)
10. Zhang, Y., Callan, J.: Maximum likelihood estimation for filtering thresholds. In: Proceedings SIGIR'01, ACM Press (2001) 294–302
11. Collins-Thompson, K., Ogilvie, P., Zhang, Y., Callan, J.: Information filtering, novelty detection, and named-page finding. In: Proceedings TREC 2002. (2002)
12. Arampatzis, A., Robertson, S., Kamps, J.: Where to stop reading a ranked list? threshold optimization using truncated score distributions. In: Proceedings SIGIR'09, ACM Press (2009)
13. Swets, J.A.: Information retrieval systems. Science **141**(3577) (1963) 245–250
14. Swets, J.A.: Effectiveness of information retrieval methods. American Documentation **20** (1969) 72–89
15. Bookstein, A.: When the most "pertinent" document should not be retrieved – an analysis of the Swets model. Information Processing and Management **13**(6) (1977) 377–383
16. Baumgarten, C.: A probabilitstic solution to the selection and fusion problem in distributed information retrieval. In: Proceedings SIGIR'99, ACM Press (1999) 246–253
17. Arampatzis, A., van Hameren, A.: The score-distributional threshold optimization for adaptive binary classification tasks. In: Proceedings SIGIR'01, ACM Press (2001) 285–293
18. Fernández, M., Vallet, D., Castells, P.: Probabilistic score normalization for rank aggregation. In: Proceedings ECIR'06, Springer (2006) 553–556
19. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
20. Cooper, W.S.: Some inconsistencies and misnomers in probabilistic information retrieval. In: Proceedings SIGIR'91, ACM Press (1991) 57–61
21. Cooper, W.S., Gey, F.C., Dabney, D.P.: Probabilistic retrieval based on staged logistic regression. In: Proceedings SIGIR'92, ACM Press (1992) 198–210
22. Arampatzis, A.: Unbiased s-d threshold optimization, initial query degradation, decay, and incrementality, for adaptive document filtering. In: Proceedings TREC 2001. (2002)
23. Robertson, S.E.: The parametric description of retrieval tests. part 1: The basic parameters. Journal of Documentation **25**(1) (1969) 1–27
24. Robertson, S.E., Bovey, J.D.: Statistical problems in the application of probabilistic models to information retrieval. Technical Report Report No. 5739, BLR&DD (1982)
25. Arampatzis, A., Kamps, J.: Where to stop reading a ranked list? In: Proceedings TREC 2008. (2008)