Inf Retrieval (2011) 14:47–67
DOI 10.1007/s10791-010-9156-2

THE SECOND INTERNATIONAL CONFERENCE ON THE THEORY OF INFORMATION RETRIEVAL (ICTIR2009)

Variational bayes for modeling score distributions

Keshi Dai · Evangelos Kanoulas · Virgil Pavlu · Javed A. Aslam

Received: 2 August 2010 / Accepted: 16 August 2010 / Published online: 3 December 2010

Abstract Empirical modeling of the score distributions associated with retrieved documents is an essential task for many retrieval applications. In this work, we propose modeling the relevant documents' scores by a mixture of Gaussians and the non-relevant scores by a Gamma distribution. Applying Variational Bayes we automatically trade-off the goodness-of-fit with the complexity of the model. We test our model on traditional retrieval functions and actual search engines submitted to TREC. We demonstrate the utility of our model in inferring precision-recall curves. In all experiments our model outperforms the dominant exponential-Gaussian model.

Keywords Score distributions · Gaussian mixtures · Variational inference · Recall-precision curves

1 Introduction

Information retrieval systems assign scores to documents according to their relevance to a user's request and return documents in a descending order of their scores. In reality, however, a ranked list of documents is a mixture of both relevant and non-relevant documents. For a wide range of retrieval applications (e.g. information filtering, topic

K. Dai · V. Pavlu · J. A. Aslam
College of Computer and Information Science, Northeastern University,
360 Huntington Ave, #202 WVH, Boston, MA 02115, USA
e-mail: daikeshi@ccs.neu.edu

V. Pavlu
e-mail: vip@ccs.neu.edu

J. A. Aslam
e-mail: jaa@ccs.neu.edu

E. Kanoulas (✉)
Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street,
Sheffield S1 4DP, UK
e-mail: e.kanoulas@sheffield.ac.uk

detection, meta-search, distributed IR), *modeling* and *inferring* the distribution of relevant and non-relevant documents over scores in a reasonable way could be highly beneficial. For instance, in information filtering and topic detection score distributions can be utilized to find an appropriate threshold between relevant and non-relevant documents (Swets 1963, 1969; Arampatzis and van Hameran 2001; Zhang and Callan 2001; Collins-Thompson et al. 2003; Spitters and Kraaij 2000). Identifying a cut-off threshold between relevant and non-relevant documents is also important in recall-oriented applications such as legal and patent retrieval (Oard et al. 2009). In distributed IR inferring score distributions can be used for collection fusion (Baumgarten 1999), and in meta-search to combine the outputs of several search engines (Manmatha et al. 2001).

Inferring the score distribution for relevant and non-relevant documents in the absence of any relevance information is an extremely difficult task, if at all possible. *Modeling* score distributions in the right way is the basis of any possible inference. Due to this, numerous combinations of statistical distributions have been proposed in the literature to model score distributions of relevant and non-relevant documents. In the 60's and 70's Swets attempted to model the score distributions of non-relevant and relevant documents with two Gaussians of equal variance (Swets 1963), two Gaussians of unequal variance and two exponentials (Swets 1969). Bookstein instead proposed a two-Poisson model (Bookstein 1977) and Baumgarten a two-Gamma model (Baumgarten 1999). A negative exponential and a Gamma distribution (Manmatha et al. 2001) has also been proposed in the literature. The dominant model, however, has been a negative exponential for the non-relevant documents and a Gaussian for the relevant ones (Arampatzis and van Hameran 2001; Manmatha et al. 2001; Zhang and Callan 2001).

As mentioned earlier the right choice of distributions (that is distributions that reflect the underlying process that produces the scores of relevant and non-relevant documents) can enhance the ability to infer these distributions, while a bad choice may make this task practically impossible. Clearly a strong argument for choosing any particular combination of distributions is the goodness-of-fit to a set of empirical data. However, the complexity of the underlying process that generates document scores makes the selection of the appropriate distributions a hard problem. Hence, even though the exponential - Gaussian model is the dominant one, there is no real consensus on the choice of the distributions. For instance, Bennett (2003) observed that when using a two-Gaussians model for text classification document scores outside the modes of the two Gaussians (corresponding to "extremely irrelevant" and "obviously relevant" documents) demonstrated different empirical behavior than the scores between the two modes (corresponding to "hard to discriminate" documents). This motivated him to introduce several asymmetric distributions to capture these differences.

Even though the goodness-of-fit can be a reasonable indicator of whether a choice of statistical distributions is the right one, from an IR perspective, these distributions should also possess a number of IR theoretical properties. Robertson considered various combinations of distributions and examined whether these combinations exhibit anomalous behavior with respect to theoretical properties of precision and recall (Robertson 2007). In a similar line of work Arampatzis et al. (2009) also developed theoretical hypotheses that score distribution models should satisfy.

In order to study what is the appropriate choice of distributions for relevant and non-relevant documents, we assume that the relevance information for all documents is available. We revisit the choice of distributions used to model document scores. Similarly to Bennett (2003) we observe that the scores of relevant documents demonstrate different behaviors in different score ranges and the exponential—Gaussian model cannot capture

these behaviors. We propose a more flexible model that empirically fit a Gamma distribution in the scores of the non-relevant documents and a mixture of Gaussians in the scores of the relevant documents. In order to balance between the flexibility of this richer class of density function and the generalization power of the model we take a Bayesian treatment on the model that automatically trades-off the goodness-of-fit with the complexity of the model. Further, we examine the IR utility of our model by testing how well one can infer precision-recall curves from the fit probability distributions. We show that our model outperforms the dominant exponential—Gaussian model.

2 Motivation for the choice of score distributions

In the remaining of this work we will essentially let the data itself dictate how to model score distributions. In this section however we give an intuitive explanation for our choice of a richer set of density functions for this purpose. A theoretical analysis on the subject can be found in (Kanoulas et al. 2010). In their work, Kanoulas et al. (2010) derive the score distributions of non-relevant and relevant documents based on rudimentary only assumptions about the distribution of terms in documents and the behavior of good retrieval systems. The Gamma distribution appears to be a good approximation of the theoretically derived distribution for the non-relevant documents. A Gaussian-like distribution with heavy right tail, which could essentially be modeled by a mixture of two Gaussians, seems to be the appropriate model under some reasonably assumptions. Some more intuitive arguments follow.

2.1 Non-relevant documents score distribution

Previous work (Manmatha et al. 2001; Robertson 2007) argues that the score distribution of non-relevant documents can be well approximated by a negative exponential density function. Often, however, a more flexible distribution is necessary. The Gamma distribution, which can range (in skewness) from an exponential to a Gaussian distribution is flexible enough. In order to explain why a Gamma distribution is a better choice, several factors should be considered.

- Truncation at a cut-off: If a list is arbitrarily truncated very early (say at rank 1,000) the distribution of the top scores may indeed look as an exponential. However looking deep down in the list (say up to rank 200,000), the shape of score distribution changes (Fig. 1, bottom).
- Query complexity: Arguments for the score distribution for a single term queries have been given in the literature (Manmatha et al. 2001). For a query with two or more terms, most non-trivial documents (i.e. the ones that contain at least two query terms) will have the following property: the contribution of the two or more terms to the final score of a document would often times be very different for the two or more terms, with some terms having a low contribution while others having a higher contribution. Averaging such effects is likely to produce a "hill" of score frequencies, perhaps with different slopes at the left and the right side of the mean; the Gamma distribution is known to be an average of exponential distributions.
- Inversion of term-frequency (TF): many scoring functions contain fractions with TF in the denominator (for example BM25, Robertson-TF etc). Given that the TF values are most of the time distributed zipfian-like, such inversion will likely produce "hill"

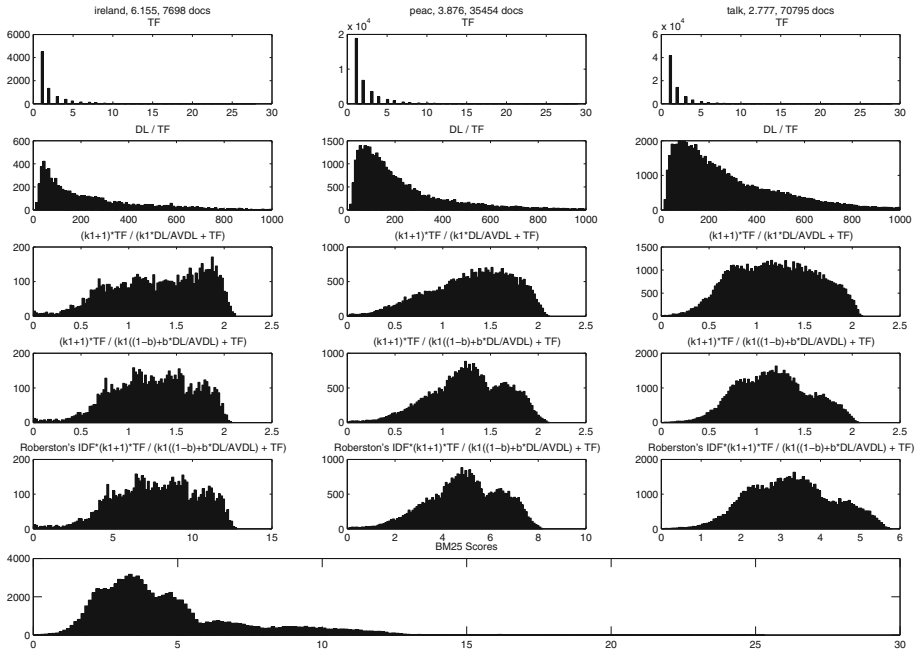

Fig. 1 The distribution of BM25 scores for all 113,947 documents (containing at least one query term) on query "Ireland peace talks". Note the different slopes at the left and at the right of the mean. Truncating the list at rank 1,000 would cause the scores' distribution to look like an exponential one. Histograms of BM25 computation are shown step-by-step starting with TF for individual query terms(top) and ending the BM25 final score (bottom)

histograms. Figure 1 shows the histograms of BM25 computation step by step for each query term followed by the final score; it can be observed that the "hill" appears when TF inversion takes place.

- Retrieval function: We mostly look at scoring functions that are decomposable into a sum of scores per query terms, like TF-IDF or Language Models (after taking logs); such scores also induce averaging effects(Fig. 1).

2.2 Relevant documents score distribution

The Gaussian density function has been the most widely used one to model the score distribution of relevant documents. However, due to its simplicity it often places unreasonable limitations over the fitting process. An example of a single Gaussian density function failing to capture the distribution of relevant documents can be viewed in the upper plot of Fig. 2. The figure shows the histogram over the scaled scores of relevant (thin red bars) and non-relevant (wide yellow bars) document for the TREC 8 query "Estonia economy". In the top plot a negative exponential and a single Gaussian density functions are separatelly fit into the scores, while in the bottom plot shows a Gamma density function and a mixture of two Gaussians are fit into the scores. As one can observe there are two clusters of relevant documents, one centered around score 0.3 and another centered around

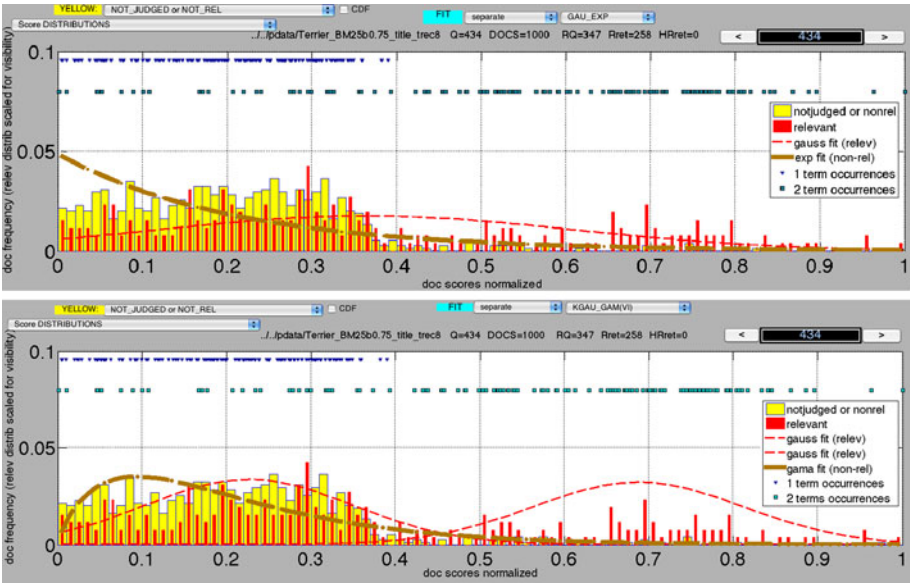

Fig. 2 The histogram over the scores of non-relevant and relevant documents along with the best fit exponential and Gaussian distributions (top plot) the Gamma and *k*-Gaussians distribution (bottom plot)

score 0.7. A single Gaussian fails to capture these two separated masses. On the contrary, it underestimates documents from two masses and overestimates documents with scores in the middle range, which in fact are less likely to be relevant. This leads to an incorrect prediction of relevance probability for a given score, especially for the low score mass that has the largest support of relevant documents. This very same phenomenon led Bennett (2003) to skew the single Gaussian distribution towards the low score relevant documents so that it does not underestimated the probability of being relevant.

An intuition behind the shape of the distribution that models the scores of relevant documents is given by Manmatha et al. (2001). Assuming that a query consists of a single term, Manmatha shows that the scores of relevant documents can be modeled as a Poisson distribution with a large λ parameter, which approaches a Gaussian distribution. Now, let's consider queries that consist of multiple terms and let's revisit Fig. 2. The query used in the example is: "Estonia economy". Each relevant document in the plot corresponds either to a triangular or to a rectangular marker at the top of the plot. The triangular markers denote the relevant documents for which only one out of the two query terms occur in the document, while the rectangular ones denote the relevant documents for which both terms occur in the document. By visual inspection, the relevant documents containing a single term clearly correspond to the low-scores' Gaussian, while the relevant documents containing both terms clearly correspond to the high-scores' Gaussian (bottom plot). Essentially, the former documents get a low score due to the fact that only one terms appear in them but they happen to be relevant to the query, while the latter correspond to documents that are obviously relevant. We observed the same phenomenon for many different queries independently of the IR model used for retrieval and independent of the query formulation. In the case of queries with multiple terms (e.g. queries that consists of both the title and the description), even though the possible number of query terms that may co-occur in a

document is greater than 2 (e.g. for a query with 3 terms, all terms may occur in a document or only two of them or only a single one of them), we observed that there is a threshold on the number of terms occurring in the document; relevant documents containing a number of terms that is less than this threshold are clustered towards low scores (first Gaussian), while relevant documents containing a number of terms that is greater than the threshold are clustered towards high scores (second Gaussian).

3 Methodology

The Gamma distribution is used to model the scores of the non-relevant documents. The Gamma density function with scale θ and shape M is given by,

$$P(x|M,\theta) = x^{M-1}\frac{\exp^{-x/\theta}}{\theta^M\Gamma(M)} \quad \text{for } x > 0 \text{ and } M,\theta > 0$$

where, $\Gamma(M)$ is an extension of the factorial function to real numbers, while for M positive integer $\Gamma(M) = (M-1)!$. The mean of the distribution is $M\theta$, while the variance is $M\theta^2$. The maximum likelihood (ML) estimation is used to estimate the Gamma parameters. When $M = 1$, the Gamma distribution degrades to an exponential distribution with rate parameter $1/\theta$. The scores of relevant documents are modeled by a mixture of K Gaussians

$$P(x|\pi,\mu,\Lambda) = \sum_{i=1}^{K}\pi_i\mathcal{N}\left(x|\mu_i,\Lambda_i^{-1}\right)$$

where $\boldsymbol{\pi} = \{\pi_i\}$ are the mixing coefficients, and satisfy $0 \le \pi_i \le 1$ and $\sum_{i=1}^{K}\pi_i = 1$. $\mathcal{N}\left(x|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right)$ is Gaussian probability density function with parameters $\boldsymbol{\mu} = \{\mu_i\}$ and $\boldsymbol{\Lambda} = \{\Lambda_i\}$, the mean and the precision of the Gaussian components, respectively. The mixture coefficients $\{\pi_i\}$ essentially express the contribution of each Gaussian to the mixture.

Fitting the mixture of Gaussians into scores could be easily done by employing the Expectation Maximization (EM) algorithm if the number of Gaussian components K was known. However, we assume that we only know an upper bound on K. Given the fact that the larger the number of components the better the fit and that EM finds the maximum likelihood mixture of Gaussians regardless of the model complexity, the EM algorithm is not appropriate for our problem. Instead, to avoid over-fitting, we employ a Bayesian treatment on the model by utilizing the Variational Bayes (VB) framework (Bishop, 2006; Attias 1999, 2000).

The VB framework takes a fully Bayesian treatment of the mixture modeling problem by introducing prior distributions over all the parameters of the model, i.e. $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ and thus accounting for the uncertainty of the value of these parameters. Given a fixed number of potential components (an upper bound on K) the variational inference approach causes the mixing coefficients of unwanted components to go to zero and essentially leads to an automatic trade-off between the goodness-of-fit and the complexity of the model.

To give some insight into how VB trades the goodness-of-fit with the complexity of the model, let's consider the function that VB optimizes, in the general case. Given a set of variables, X, and a set of parameters, Θ , Variational Bayes aims at optimizing the log of the *marginal likelihood* or *evidence*, $p(X)$, where the hidden variables along with the parameters have been integrated out. That is,

$$\log p(X) = \log \int p(X, \Theta) d\Theta = \log \int q(\Theta|X) \frac{p(X, \Theta)}{q(\Theta|X)} d\Theta \quad (1)$$

$$\geq \mathcal{L} \equiv \int q(\Theta|X) \log \frac{p(X, \Theta)}{q(\Theta|X)} d\Theta \text{ by Jensen's Inequality} \quad (2)$$

Inequality 2 holds for any arbitrary conditional distribution q. The difference between the l.h.s and the r.h.s of the inequality is in fact the KL-divergence between the arbitrary conditional distribution q and the posterior distribution $p(\Theta|X)$, and thus the optimal q is obtained by letting $q = p(\Theta|X)$. Given the above inequality, VB actually optimizes the lower bound $\mathcal{L}$ instead of the log marginal likelihood. If we further expand the lower bound $\mathcal{L}$ we obtain,

$$\mathcal{L} = \int q(\Theta|X) \log \frac{p(X, \Theta)}{q(\Theta|X)} d\Theta \quad (3)$$

$$= \int q(\Theta|X) \log \frac{p(X|\Theta)p(\Theta)}{q(\Theta|X)} d\Theta, \quad (4)$$

$$= \int q(\Theta|X) \log p(X|\Theta) d\Theta - \int q(\Theta|X) \log \frac{q(\Theta|X)}{p(\Theta)} d\Theta \quad (5)$$

$$= \int q(\Theta|X) \log p(X|\Theta) d\Theta - KL[q(\Theta|X)||p(\Theta)] \quad (6)$$

where $p(\Theta)$ is the prior distribution over the model parameters and $KL[q(\Theta|X)||p(\Theta)]$ is the KL-divergence between the posterior distribution of the model parameters and their prior distribution. The left term of the r.h.s. of Eq. 6 expresses the goodness-of-fit of the model to the data and increases with the complexity of the model, while the right term of the r.h.s. of Eq. 6 is the Occam factor which penalizes over-complex models. Essentially, VB penalizes the departure of the parameters from their prior distribution. Finally, note that the Bayesian information criterion (BIC) (Schwarz 1978) and the minimum description length criterion (MDL) (Rissanen 1987) both emerge as a special case of a large sample expression of Eq. 6 (Attias 1999).

Moving back to the mixture of Gaussians, we introduced priors over the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$. To simplify the mathematics of VB and achieve an analytic solution, we only consider conjugate prior distributions (as in Bishop (2006)), such that the posterior distribution of the model parameters given the data is in the same family of distributions with the prior. Thus, we chose a Dirichlet distribution over the mixing coefficients $\boldsymbol{\pi}$, i.e. $p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\alpha_0)$, and an independent Gaussian-Wishart distribution over the mean and the precision of each Gaussian component, i.e. $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}|m_0, \left(\beta_0 \Lambda\right)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}|W_0, \nu_0)$. Regarding the Dirichlet distribution, by symmetry we chose the same parameter α_0 for all the mixture components. Given that α_0 can be interpreted as the effective prior number of observations associated with each component of the mixture we set $\alpha_0 = 10^{-3}$, such that the posterior distribution will be influenced primary by the data. Regarding the Gaussian distribution, m_0 corresponds to the mean value of the distribution of the Gaussian means, thus we assigned the same value m_0 to all the mixture components and set it to the mean of the data, i.e. the mean score of the relevant documents. Regarding the Wishart distribution, ν_0 corresponds to the dimensionality of the data which in our case is 1 and thus we set ν_0 equal to 1. The hyperparameter W_0, in the general case, is a positive definite matrix,

however in the case of one-dimensional data W_0 is simply a number, which we initialized by the precision of the data. The parameter β_0 is a scalar corresponding to the ratio between the precision of the distribution of the Gaussian mixture means and the precision of the distribution of the data. We initialized it by clustering the data into K clusters and setting β_0 equal to *variance of data/variance of cluster means*.

To initialize the VB process we first run the k-means algorithm and obtain 10 initial clusters and then use these clusters to initialize the expectation-maximization (EM) algorithm that results in a mixture of 10 Gaussians. This mixture is then used to initialize the VB process that finds the optimal model in terms of goodness-of-fit and model complexity. The same process is run 10 times with 10 different initializations of the k-means algorithm. Finally, we select the model that leads to the highest lower bound $\mathcal{L}$.

Remark In this work we only consider Variational Bayes as a technique to automatically select the number of components of the Gaussian mixture. However, several criteria that can lead to a trade-off between the goodness-of-fit and the model complexity have been proposed in the literature. Schwarzs BIC Schwarz (1978), Akaikes AIC Akaike (1974), Rissanens minimum description length, the Information Complexity Criterion Bozdogan (1993), the Normalised Entropy Criterion Celeux and Soromenho (1996) are some of them that have been used along with an EM algorithm to fit a mixture of Gaussians with unknown number of components. Markov Chain Monte Carlo (MCMC) methods have also been used for model selection (e.g. see Richardson and Green (1997).

4 Experimental setup

We use data from TREC 6, 7 and 8 ad-hoc tracks, TREC 9 and 10 Web tracks (ad-hoc tasks) and TREC 12 Robust track. TREC 6, 7, 8 and 12 collections consist of documents contained in the TREC Disk 4 and 5, excluding the *Congressional Record* sub-collection, while TREC 9 and 10 collections use the WT10g document collection. The topics used are the TREC topics 301–550 and 601–650 (Voorhees and Harman 2005). The Robust track topic set in TREC 12 consists of two subsets of topics, the topics 601–650 and 50 old topics selected based on topic hardness from past collections. In all results regarding TREC 12 only the topics 601–650 are used.

To avoid the effects of arbitrary query manipulations and score transformations that systems submitted to TREC (Text REtrieval Conference) often apply, in the sections that follow we instead use scores produced by traditional IR models. Later, in Sect. 8, we validate our model on TREC systems.

Indexing and search was performed using the Terrier search engine (Ounis et al. 2007). Porter stemming and stop-wording was applied. The document scores obtained are the outputs of (a) Robertson's and Spärck Jones' TF-IDF (Robertson and Jones 1976), (b) BM25 (Robertson and Walker 1994), (c) Hiemstra's Language Model (LM) (Hiemstra 2001), and (d) PL2 divergence from randomness (Amati and Van Rijsbergen 2002) (with Poisson estimation for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization). Further, three different topic formulations were used, (a) topic titles only, (b) topic titles and descriptions, and (c) topic titles, descriptions and narratives.

Finally, note that, by convention, documents not judged by TREC assessors are considered non-relevant, since there were not retrieved by any of the submitted to TREC runs in the top-k ranks, where k is usually 100. When fitting the Gamma distribution we

consider these documents as non-relevant and thus we essentially fit the Gamma distribution in both non-relevant and unjudged documents. For the rest of the article by non-relevant documents we refer to both judged non-relevant and unjudged documents.

5 Results

We separately fit the Gamma distribution and the mixture of Gaussians into the scores of the non-relevant and relevant documents, respectively, for each topic-system pair. There are 50 topics available per TREC data set and 3 query formulations (title, title and description and title, description and narrative), along with the relevance information for the top 1000 documents returned by 4 IR systems (TF-IDF, BM25, LM and PL2). Thus, there are in total 600 ranked lists of documents per TREC data set. The scores of the documents were first normalized into a 0 to 1 range by shifting and scaling to preserve the score distribution.

To summarize our results we report the parameter M of the Gamma distribution, which as mentioned earlier corresponds to the number of independent exponential density functions averaged, and the number K of Gaussian components in the mixture, for all four systems, all 150 topics (50 topics and 3 query formulations) for each TREC data set. Figure 3 shows the histograms over M and K. Each row corresponds to each one of the TREC 6, 7, 8, 9, 10 and 12 data sets in this order. As it can be observed, K is most of the times different than one, especially in the early TREC collections. Further, M is spread both above and below one. This illustrates that often times, taken into account the complexity of the model, the data suggests that a Gamma distribution and a mixture of Gaussians is a better fit to relevant and non-relevant scores than a negative exponential and a single Gaussian. In particular, the mean number of Gaussian components over all TREC data sets is 1.52 while the mean number of component for each TREC data set separately is 1.75, 1.74, 1.67, 1.11, 1.37, 1.51. The mean value of the parameter M over all TREC data sets is 0.98 while the mean value of the parameter M for each TREC data set separately is 0.92, 0.96, 0.93, 1.10, 1.00, 0.96. Even though the mean M is close to 1, as it can be viewed in Fig. 3, M varies in a wide range below and above 1.

6 Analysis

In this section, we attempt to analyze the parameters K (number of Gaussian mixture components) and M (number of independent exponential distributions averaged) to obtain a better understanding of the underlying process that generates this distribution of relevant and non-relevant documents. First, we examine whether different factors such as the IR model and the number of query terms affect the distribution of the parameters K and M. Then, we focus on the relevant document score distribution and examine whether the total number of relevant documents retrieved affect our ability to recover complex distributions, such as a mixture of Gaussians. Further, we examine whether the different components of the mixture correspond to relevant documents of different characteristics. In particular, we explore whether (a) different Gaussians explain relevant documents uniquely identified by manual runs and relevant documents retrieved by automatic runs, and (b) different Gaussians correspond to relevant versus highly relevant documents.

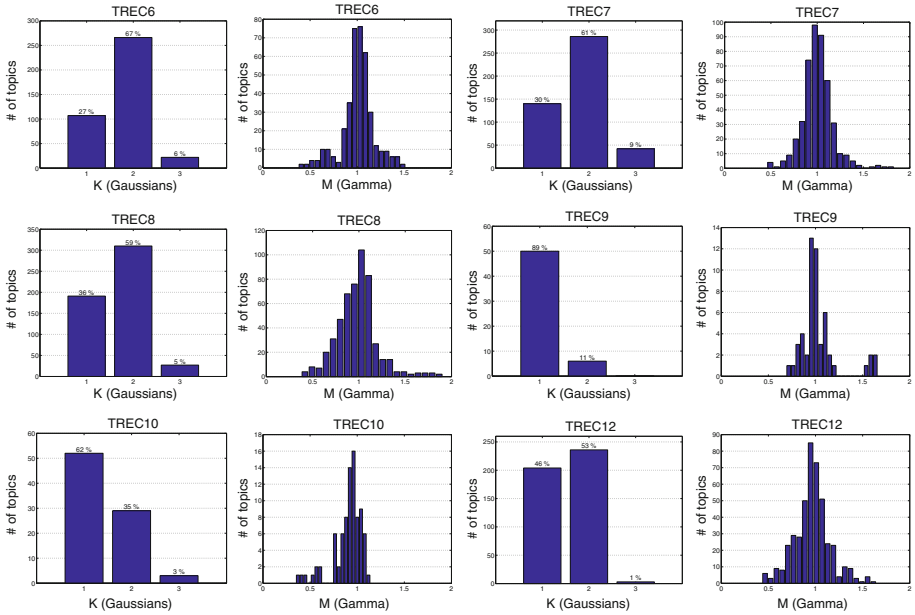

Fig. 3 The histograms over the number K of Gaussian components and the parameter M of the Gamma distribution, over all IR models, topics and topic formulations for TREC 6 and 7 (top row), 8 and 9 (middle row), and 10 and 12 (bottom row)

6.1 IR systems

First we test whether and how different IR systems affect the parameters M and K. In Figs. 4, 5 and 6 we report the histograms over K and M for each system separately (50 topics with 3 topic formulations) for TREC 6, 7 and 8 data sets (the plots for TREC 9, 10 and 12 resemble the ones presented here and thus omitted). As it can be observed, both the distribution over K and the distribution over M appear to be independent with respect to the IR model utilized. To validate our observations we run an n-way ANOVA testing whether the mean values of K per IR model are equal and we could not reject the hypothesis.

6.2 Number of query terms

Then, we tested how the number of query terms affect the parameters M and K. In Figs. 7, 8 and 9 we report the histograms over K and M for each different query formulation separately (50 topics with 3 topic formulations) for TREC 6, 7 and 8 data sets. As it can be observed, the distribution over K and M appear to be independent of the number of query terms. The distribution over M appears to be slightly flatter in the case of title-only queries than the rest of the formulations. However, the mean M appears to remain unaffected by the number of query terms. To further analyze the effect of the number of query terms on the values of K and M, we performed query expansion on the top of the title query formulation, by varying the number of terms to expand a query with from 4 to 512 terms (increasing powers of 2). These terms were extracted from the actual relevant documents and the term weighting model used for expanding the queries with the most informative terms was the Bose-Einstein 1 Method (Amati 2003) provided by the Terrier Toolkit

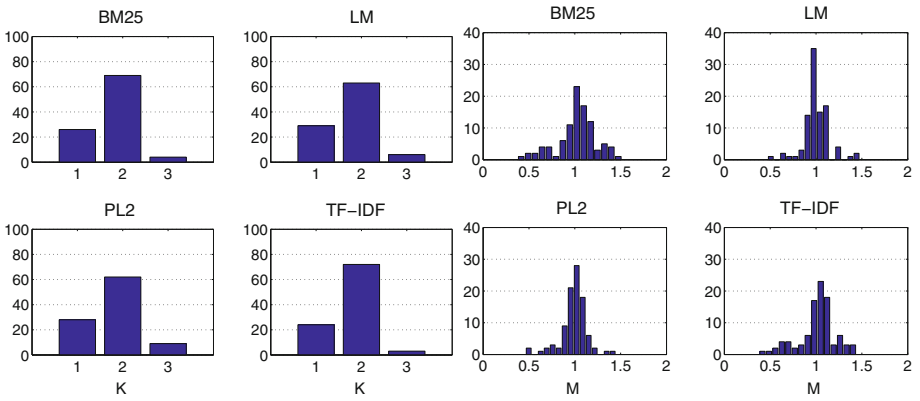

Fig. 4 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and topic formulations for each IR model for TREC 6

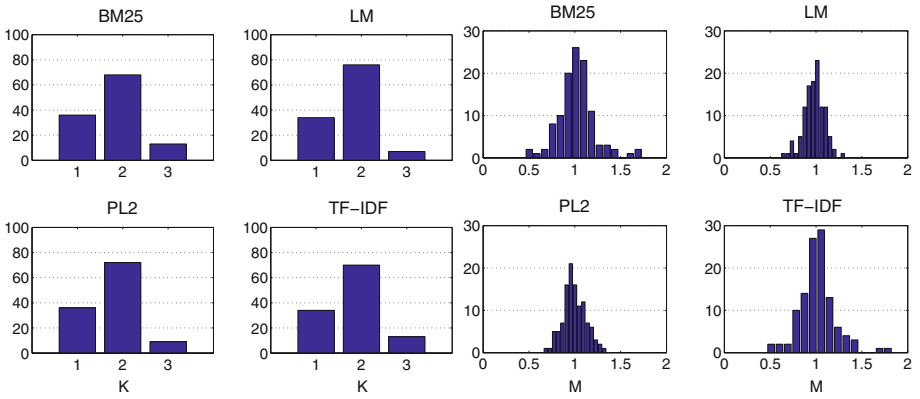

Fig. 5 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and topic formulations for each IR model for TREC 7

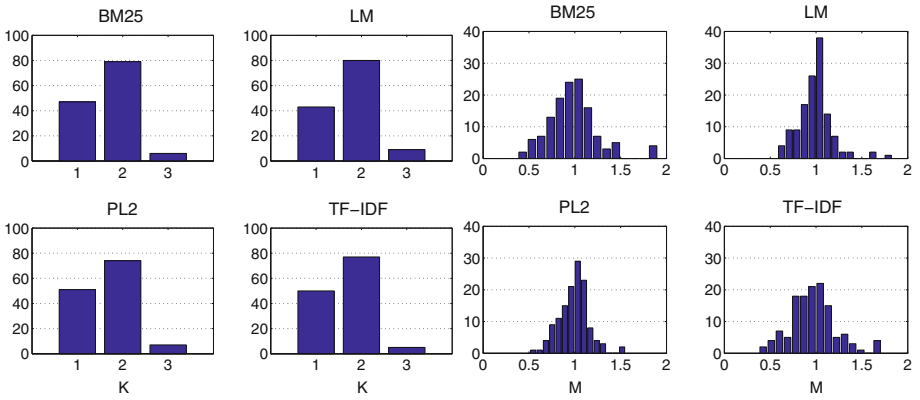

Fig. 6 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and topic formulations for each IR model for TREC 8

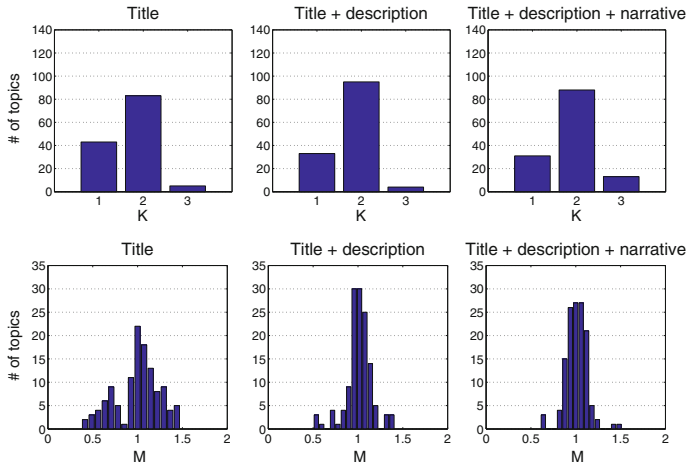

Fig. 7 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and IR models for each topic formulation, for TREC 6

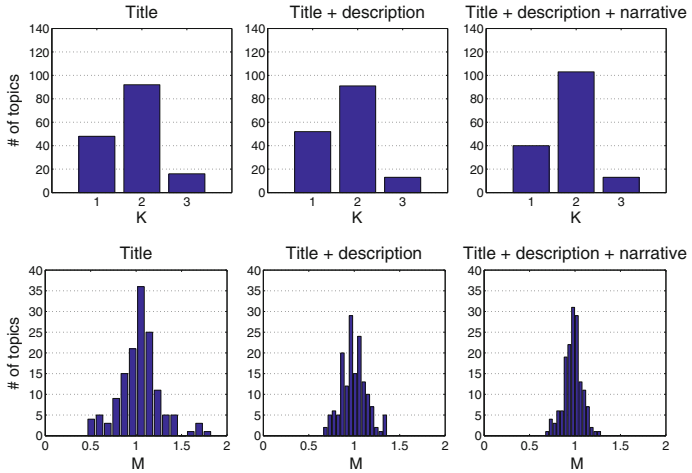

Fig. 8 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and IR models for each topic formulation, for TREC 7

(Ounis et al. 2007). The average values and their standard deviation (shaded area) for the parameters K and M against the number of expanding terms are illustrated in Fig. 10. As it can be observed, both K and M are independent on the number of query terms.

6.3 Number of relevant documents retrieved

As one can observe in Fig. 3, the number of components of the Gaussian mixture varies across different TREC data sets. In particular, in TREC 9 and 10 the dominant model to describe the distribution of relevant documents scores is the single Gaussian. The opposite is true in TREC 6, 7, 8, and 12. One particular characteristic of TREC 9 and 10 data sets is

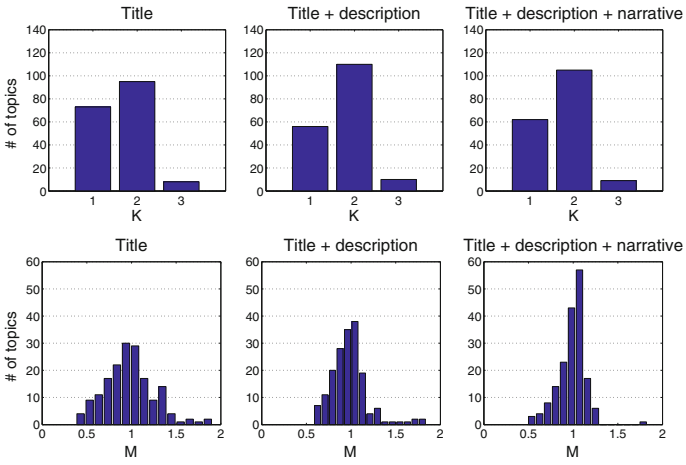

Fig. 9 The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and IR models for each topic formulation, for TREC 8

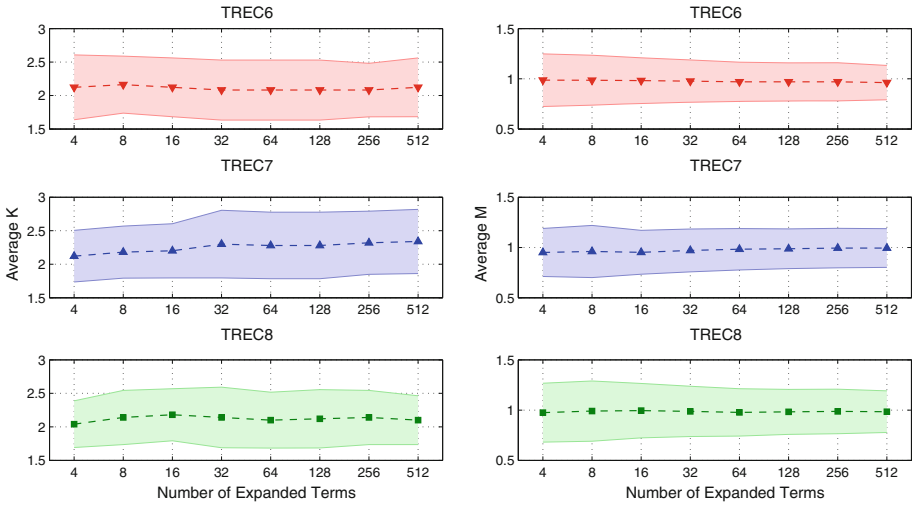

Fig. 10 The average values of K and M against the number of query terms for TREC 6, 7, 8

that the number of relevant documents retrieved is particularly smaller when compared with the traditional ad-hoc tracks of TREC 6, 7, 8 and 12. Thus, we analyze K with respect to the number of relevant documents in the ranked lists. The results are illustrated in Fig. 11. In the scatter plot, each point corresponds to a ranked list for a given query. As it can be observed, when the number of retrieved relevant documents is particularly small, usually less than 50, the data cannot support more than one Gaussian. This clearly explains the histogram over the parameter K in Fig. 3 for TREC 9 and 10, since as it can be viewed in Fig. 11, the number of relevant documents retrieved per ranked list in these two TREC data sets is always less than 35. However, when the number of retrieved relevant documents is above 50 there is no real indication that the more the number of relevant

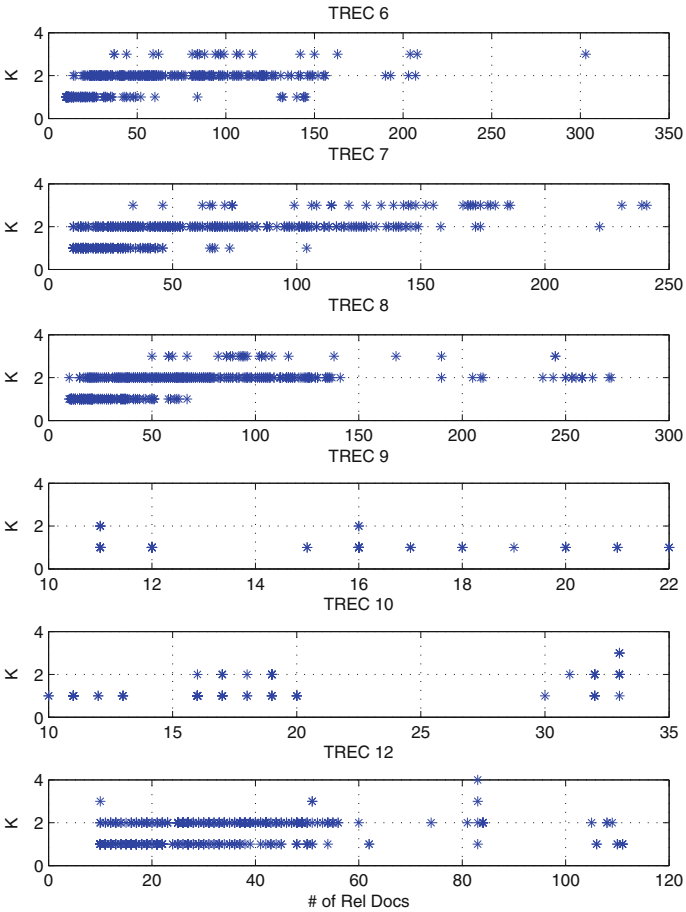

Fig. 11 The number of Gaussian components, K, against number of relevant documents retrieved over all ranked lists for each TREC data set

documents retrieved the more the Gaussian. Note that, if such a phenomenon was observed, it would have raised the question of whether $K > 1$ is simply artifact.

6.4 Relevant documents retrieved by manual vs. automatic runs

In TREC 6, 7 and 8, along with the ranked lists of documents returned by different retrieval systems, a number of ranked lists of documents was retrieved with some manual effort (manual runs[1]). In particular, 17 manual runs were submitted to TREC 6, 17 were submitted to TREC 7 and 13 were submitted to TREC 8. Often times, relevant documents identified with manual effort have very different characteristics from the ones that automatic system are built to find, and thus many manually identified relevant documents do not appear in the top-100 ranks of the lists returned by automatic systems. These

[1] A manual run may require, for instance, extensive human search in the collection to identify relevant documents

documents may be retrieved by automatic systems as well but they are usually assigned smaller scores. One of the hypothesis we test is whether the different Gaussian components separately capture the distributions of the automatically and manually retrieved relevant documents. The question we want to answer is whether the low scores Gaussian components solely model the distribution of relevant documents uniquely identified by manual runs while the high score Gaussian components solely model the distribution of relevant documents retrieved by automatic runs. If a document is identified by both manual runs and automatic runs, we consider it retrieved by automatic runs. To answer this question, we first construct a contingency table for each ranked list of documents (i.e. for each one of the 600 system-query pairs per TREC data set). The two rows of the table correspond to manual and automatic runs, while each column corresponds to each one of the Gaussian component in the best fit mixture. To fill in the table, we first compute the responsibility of each of the Gaussian components for each relevant documents as $\pi_i * P\left(x|\mu_i, \Lambda_i^{-1}\right)$, where x is the score of a relevant document; if the relevant document is uniquely identified by manual (automatic) runs then the responsibility values are appropriately added to the manual (automatic) runs row. In this manner a probabilistic table of counts is created. Each row is the normalized to give the distribution of manually (automatically) identified relevant documents over the Gaussian components, i.e. $P\left(i^{th}\text{Gaussian}|\text{type of run}\right)$. If different components correspond to manual and automatic runs the distance between these two distributions will be high. We measure this distance by the Jensen-Shannon divergence between the two distribution. The average Jensen-Shannon divergence values for TREC6, 7, 8 are 0.05, 0.06, and 0.10 respective, which do not reveal a correlation between manual and automatic relevant documents and the different Gaussian components. However, whether the Jensen-Shannon divergence is a good measure of the correlation between the manual and automatic with the different Gaussian components needs to be further investigated.

6.5 Relevant versus highly relevant documents

Further, we want to test the hypothesis of whether different Gaussian components correspond to different grades of relevance. In TREC 9, 10 and 12 documents were judged in a three-grades scale as non-relevant, relevant and highly relevant. We repeated the same analysis as the one above, by constructing a contingency table for relevant and highly relevant documents. The Jensen-Shannon divergence values computed here also do not demonstrated any correlation between the different Gaussian components and the relevance graded. However, as mentioned before, we intent to further investigate these explanatory factors by different means other than the Jensen-Shannon divergence.

7 Precision-recall curves

As a utility of our model for IR purposes, we estimate the precision-recall (PR) curve separately from both the Exponential-Gaussian (*EG*) and Gamma-Gaussian mixture (*GkG*) model. Similarly to Robertson Robertson (2007), let f_r and f_n denote the model densities of relevant and non-relevant scores, respectively; $F_r(x) = \int_x^1 f_r(x)dx$ and $F_n(x) = \int_x^1 f_n(x)dx$ are the cumulative density functions *from the right*. While the density models might have support outside the range [0,1], we use integrals up to 1 because our scores are normalized. For each recall level r we estimate the retrieval score at which r happens, from the relevant

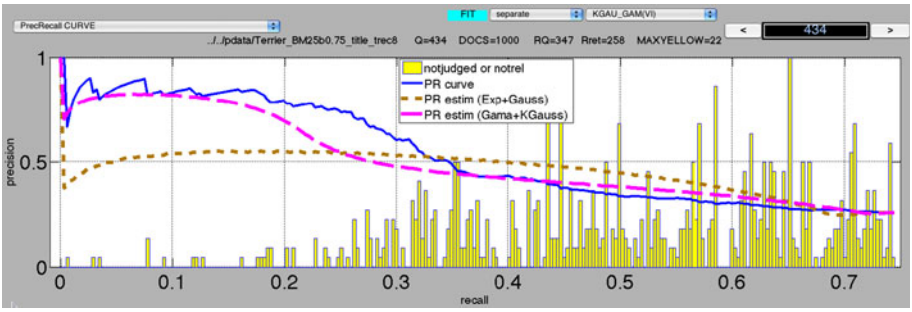

Fig. 12 Precision-Recall curve (blue) for query 434 and the BM25 retrieval function implemented by Terrier. It is easy to see that the PR curve estimated from the GkG model (magenta) is much better than the PR estimated from the EG model (brown). Yellow bars indicate the number of non-relevant documents in each recall interval

cumulative density: $score(r) = F_r^{-1}(r)$, which we compute numerically. Then we have $n(r) = F_n(score(r))$ as the percentage of non-relevant documents found up to recall r in the ranked list. Finally, the precision at recall r can be computed as in Robertson (2007), $prec(r) = \frac{r}{r+n(r)*G}$, where G is the ratio of non-relevant to relevant documents in the collection searched. Computing precision at all recall levels from the score distribution models f_r and f_n gives an estimated PR curve. In the remainder of this section we show that estimating PR curves from the *GkG* model clearly outperforms PR curves estimated from the dominant *EG* model.

To measure the quality of the estimated PR curves we report the RMS error between the actual and the predicted precisions at all recall levels for both models (see Fig. 12). The results are summarized in Table 1, separately for each model. Language model (LM) and Divergence from randomness (PL2) seem to produce slightly better PR estimates, independent of the query formulation. The over-all RMSE of *GkG* vs. *EG* is .155 vs .176, or about 12% improvement.

Further, we report the mean absolute error between the actual and predicted precisions at all recall levels. This is the area difference between the estimated and the actual curve, which immediately gives a bound for the difference in Average Precision of the two curves (because the AP metric is approximated by the area under the PR curve). The results are reported in Table 2. Note that the best fit with respect to MAE are given for the full query formulation (title, description and narrative); the overall MAE for *GkG* is .099 vs *EG* with .120, or an improvement of about 21%.

Table 1 RMS error between the actual and the inferred precision-recall curves, averaged over about 200 queries from TREC collections

	Title		Title+desc		Title+desc+narrative	
	EG	GkG	EG	GkG	EG	GkG
BM25	.196	.172	.182	.163	.180	.159
LM	.166	.153	.151	.134	.142	.125
PL2	.173	.149	.178	.157	.175	.151
TFIDF	.195	.170	.185	.165	.180	.158

Table 2 Mean Absolute Error between actual and inferred precision-recall curves, averaged over about 200 queries from TREC collections

	Title		Title+desc		Title+desc+narrative	
	EG	GkG	EG	GkG	EG	GkG
BM25	.141	.117	.124	.104	.118	.099
LM	.118	.102	.101	.085	.093	.076
PL2	.121	.098	.120	.098	.114	.092
TFIDF	.140	.115	.126	.105	.119	.098

8 TREC search engines

To avoid the effects of arbitrary query manipulations and score transformations that systems submitted to TREC (Text REtrieval Conference) often applied, we used in our experiments scores produced by traditional IR models. In this section we apply our methodology over the score distributions returned by search engines submitted to TREC 6, 7, 8, 9, 10 and 12. Out of all the runs submitted to TREC a number of them was excluded from our experiments since they are manual runs and scores are computed based on the retrieved rank list. No other quality control was performed. As earlier, we report the parameter M of the Gamma distribution, and the number K of Gaussian components in the mixture, for all systems and all queries as histograms in Fig. 13. As it can be observed, similarly to the case of the traditional IR models, K, in most cases, is different from 1,

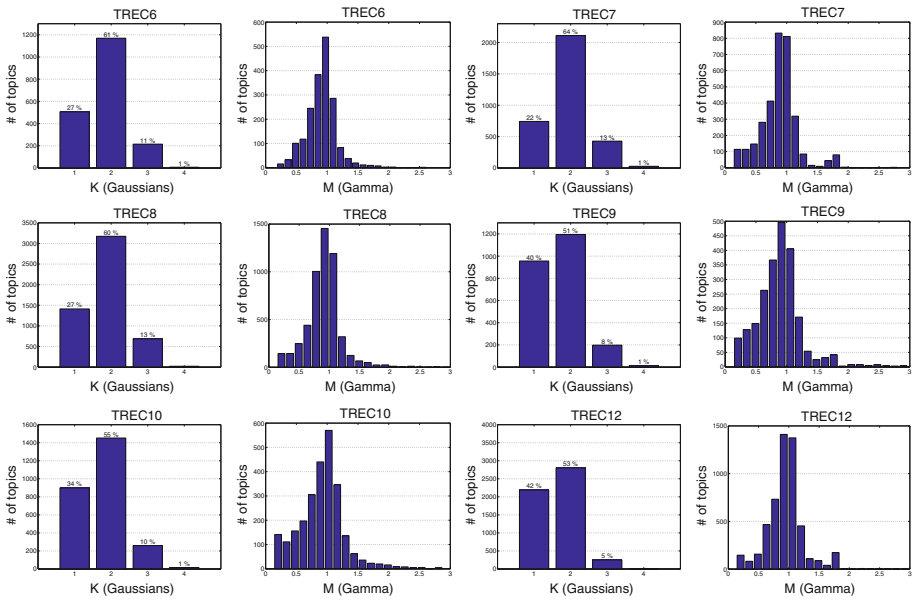

Fig. 13 The histograms over the number K of Gaussian components and the parameter M of the Gamma distribution for all automatic runs in TREC 6 and 7 (top row), 8 and 9 (middle row), and 10 and 12 (bottom row)

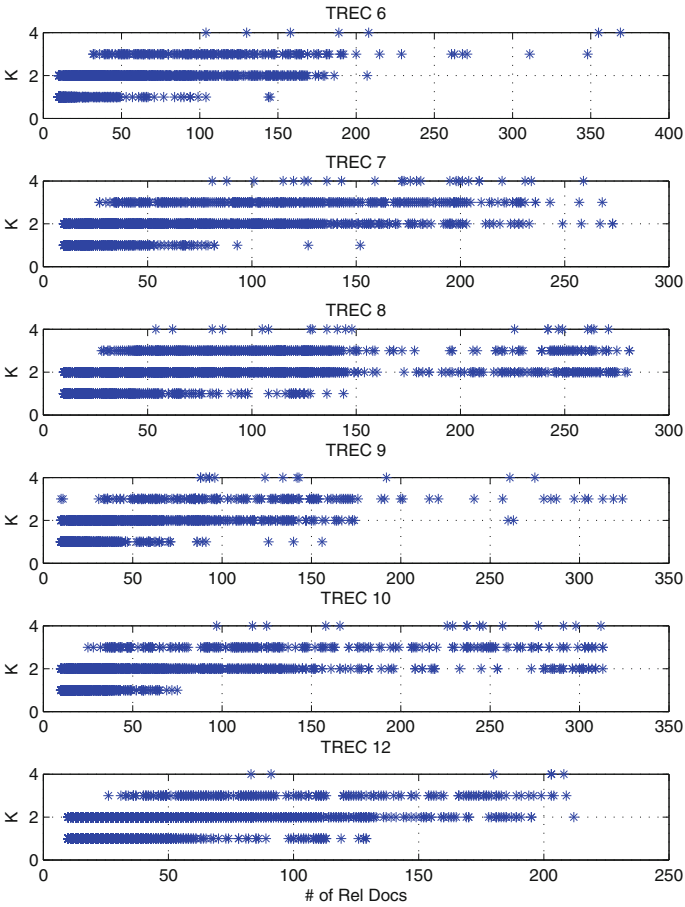

Fig. 14 The number of Gaussian components, K, against number of relevant documents retrieved over all ranked lists for each TREC data set

while M varies in a wide range confirming that a Gamma distribution and a mixture of Gaussians is a better fit than a negative exponential and a single Gaussian.

Further, we tested the same hypotheses as in the case of the traditional IR models, that is whether the number of Gaussian components is correlated with the number of relevant documents retrieved in the ranked lists, the manual versus the automatic runs and the relevant versus the highly relevant documents. Figure 14 illustrates the fact that as in the case of traditional IR models, small number of relevant documents retrieved cannot support the reconstruction of complex models like the mixture of Gaussians. Further, the Jensen-Shannon divergence between the distributions of Gaussians for manual and automatic runs and relevant and highly relevant documents did not reveal any correlation between them and the different Gaussians.

The Precision-Recall curve estimate obtained for TREC runs is measured in Table 3. The proposed model (GkG) easily outperforms the EG model in terms of both RMSE and MAE. Overall RMSE improvement is about 18%, while overall MAE improvement is about 25%. Note that "harder" TRECs (latest collections) also mean harder estimates for both GkG ad EG models.

Table 3 RMS and MAE error between the actual and the inferred precision-recall curves reported separately on TREC6-12 ad-hoc runs

Valid q-runs	TREC6 1906	TREC7 3317	TREC8 5300	TREC9 2370	TREC10 2633	TREC12 5275
EG RMSE	.140	.181	.171	.214	.220	.216
GkG RMSE	.099	.127	.126	.184	.198	.199
EG MAE	.097	.130	.123	.147	.143	.150
GkG MAE	.060	.081	.081	.116	.121	.130

9 Conclusions and discussion

In this work, we proposed modeling the relevant document scores by a mixture of Gaussians and the non-relevant document scores by a Gamma distribution. An intuition about the choice of the particular model from an IR perspective was given but the main approach was data-driven. We extensively studied the correlation between the number of Gaussian components K in the mixture distribution and the value of the parameter M in the Gamma distribution with a number of explanatory factors. The results of our experiments demonstrated that the distributions of K and M are independent both of the IR model used and the number of query terms. Further, we demonstrated that a small number of retrieved relevant documents retrieved may prohibit the Variational Bayes framework to fit complex models such as the mixture of Gaussians to the relevant document scores. Finally, we demonstrated the utility of our model in inferring precision-recall curves.

The aim of this work was to revisit the problem of modeling score distributions under the observation that the negative exponential - Gaussian model fails to capture the underlying process that generates document scores. However, an important question that arises here is the practical consequences of this work. Applying the proposed model to traditional applications such as information filtering, distributed retrieval or data fusion is not straight forward. In all our experiments we intentionally made use of the actual relevance judgments, while all the afore-mentioned applications require blind fit of score distributions. The fact that we proposed the use of a richer model than the traditional ones makes a blind fit harder. Given that, the applications of our model or the practical consequencies of the results we obtained from the analyses we conducted is a topic of future research.

Acknowledgments We would like to thank Avi Arampatzis, Jaap Kamps and Stephen Robertson for many useful discussions. Further, we gratefully acknowledge the support provided by NSF grants IIS-0533625 and IIS-0534482 and by the European Commission who funded parts of this research within the Accurat project under contract number FP7-ICT-248347.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions onAutomatic Control, 19*, 716–723.

Amati, G. (2003). Probability models for information retrieval based on divergence from randomness. PhD thesis, University of Glasgow.

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems, 20*(4), 357–389.

Arampatzis, A., & van Hameran, A. (2001). The score-distributional threshold optimization for adaptive binary classification tasks. In: *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 285–293). New York, NY: ACM. doi:10.1145/383952.384009.

Arampatzis, A. T., Robertson, S., & Kamps, J. (2009). Score distributions in information retrieval. In: *ICTIR* (pp. 139–151).

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In: *Proceedings of the 15th conference on uncertainty in artificial intelligence* (pp. 21–30). San Francisco: Morgan Kaufmann Publishers.

Attias, H. (2000). A variational bayesian framework for graphical models. In: *In advances in neural information processing systems* (Vol. 12, pp. 209–215). Cambridge: MIT Press.

Baumgarten, C. (1999) A probabilistic solution to the selection and fusion problem in distributed information retrieval. In: *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 246–253). New York, NY: ACM. doi:10.1145/312624.312685.

Bennett, P. N. (2003). Using asymmetric distributions to improve text classifier probability estimates. In: *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 111–118), ACM, New York, NY, USA, doi:10.1145/860435.860457.

Bishop, C. M. (2006). *pattern recognition and machine learning (information science and statistics)*. New York: Springer.

Bookstein, A. (1977). When the most "pertinent" document should not be retrieved—an analysis of the swets model. *Information Processing & Management, 13*(6), 377–383.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new information complexity criterion "choosing the number of component clusters in the mixture model using a new information complexity criterion of the inverse-fisher information matrix. *Information and Classification*, 40–54.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(195–212).

Collins-Thompson, K., Ogilvie, P., Zhang, Y., & Callan, J. (2003). Information filtering, novelty detection, and named-page finding. In *Proceedings of the 11th text retrieval conference*.

Hiemstra, D. (2001). Using language models for information retrieval. PhD thesis, Centre for Telematics and Information Technology, University of Twente.

Kanoulas, E., Dai, K., Pavlu, V., & Aslam, J. A. (2010). Score distribution models: Assumptions, intuition, and robustness to score manipulation. In: *To appear in proceedings of the 33rd annual international ACM SIGIR conference on research and development in information retrieval*.

Manmatha, R., Rath, T., & Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In: *SIGIR '01: proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 267–275). New York, NY: ACM. doi:10.1145/383952.384005.

Oard, D. W., Hedin, B., Tomlinson, S., Baron, J. R. (2009). Overview of the trec 2008 legal track. In: *In Proceedings of the 17th text retrieval conference*.

Ounis, I., Lioma, C., Macdonald, C., & Plachouras, V. (2007). Research directions in terrier. In: R. Baeza-Yates. et al (Eds.), *Novatica/UPGRADE special issue on next generation web search. Invited Paper*, *8*(1), 49–56.

Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society B, 59*(4), 731–792.

Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society B, 49*, 223–239; 253–265.

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

Robertson, S. (2007). On score distributions and relevance. In: Amati, G., Carpineto, C., Romano, G. (Eds.), *Advances in information retrieval, 29th European conference on IR research, ECIR 2007. Lecture notes in computer science, vol 4425/2007* (pp. 40–51). Springer: New York.

Robertson, S. E., & Jones, S. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Spitters, M., & Kraaij, W. (2000). A language modeling approach to tracking news events. In: *Proceedings of TDT workshop 2000* (pp. 101–106).
Swets, J. A. (1963). Information retrieval systems. *Science, 141*(3577), 245–250.
Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation, 20*, 72–89.
Voorhees, E. M., & Harman, D. K. (2005). *TREC: experiment and evaluation in information retrieval*. Cambridge: Digital Libraries and Electronic Publishing, MIT Press.
Zhang, Y., & Callan, J. (2001). Maximum likelihood estimation for filtering thresholds. In: *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 294–302). New York, NY: ACM. doi:10.1145/383952.384012.