Delft University of Technology

Master's Thesis

Exploring Copula-Based Models for the Stochastic Simulation of Information Retrieval Evaluation Data

by

Dimitris Theodorakopoulos

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

in

Computer Science

December, 2022

TU Delft Delft University of Technology

Challenge the future

To be defended publicly on Monday December 19, 2022 at 12:00pm.

Thesis committee:	Prof. dr. A. Hanjalic,	TU Delft
	Dr. J. Urbano,	TU Delft, supervisor
	Dr. J. Yang,	TU Delft
Student :	Dimitris Theodorakopoulos (4620534)	
Programme:	MSc Computer Science	
Track:	Software Technology	
University:	Delft University of Technology	
Faculty:	Electrical Engineering, Mathematics & Computer Science	
Department:	Intelligent Systems	
Research Group:	Multimedia Computing	
Project:	IN5000 Final Project (45EC)	

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Abstract

In the field of Information Retrieval (IR), the reliable evaluation of systems is a key component in order to progress the state-of-the-art. Much of IR research focuses on optimizing the various aspects of evaluation. Stochastic simulation is one technique that can be used to assist this kind of research. It allows researchers to overcome certain limitations associated with IR data, such as limited size, and lack of control. Recently, there have been two parallel lines of work that use stochastic simulation to study the question of ”which statistical significance test is optimal for IR evaluation data?”. Surprisingly, the authors reach different conclusions, despite the fact that both use stochastic simulation. One line of work, lead by Urbano et al., simulates scores for a fixed set of systems on new random topics, and concluded that the t-test is optimal. Another line of work, lead by Parapar et al., simulates new random retrieval runs for a fixed set of topics, and concluded that the Wilcoxon test is optimal. Interestingly these two tests are the most popular in IR literature. In an attempt to shed some light on this disagreement between the two conclusions, we made a first attempt at providing some empirical evidence regarding the quality of the simulation approach that was used by Urbano et al. Our main findings is that the quality of the simulation is moderately good, and also discovered some opportunities to refine it. In addition, we proposed a new model selection criterion, that showed some promising results, and in many cases managed to select models more optimally than other, more established criteria, such as AIC.

Contents

1 Introduction **1**
- 1.1 Stochastic Simulation for IR Research . . . 1
- 1.2 Motivation . . . 2
- 1.3 Research Goals . . . 3
- 1.4 Main Findings . . . 3
- 1.5 Thesis Outline . . . 4

2 Background **5**
- 2.1 Description of the Simulation . . . 5
  - 2.1.1 Marginal and Copula Families . . . 6
  - 2.1.2 Model Selection Criteria . . . 8
- 2.2 Related Work . . . 9

3 Margins **11**
- 3.1 Defining Goodness-of-Fit . . . 11
- 3.2 Experiments . . . 13
  - 3.2.1 Comparing Model Selection Criteria . . . 21
  - 3.2.2 Extrapolating Results to Larger Topic Set Sizes . . . 23
  - 3.2.3 Summary of Results . . . 25

4 Copulas **27**
- 4.1 Experiments . . . 27
  - 4.1.1 Comparing Model Selection Criteria . . . 30
  - 4.1.2 Extrapolating Results to Larger Topic Set Sizes . . . 31
  - 4.1.3 Summary of Results . . . 32

5 Conclusion **35**

1 Introduction

1.1. Stochastic Simulation for IR Research

In the field of Information Retrieval (IR), the reliable evaluation of systems is a key component in order to progress the state-of-the-art. One method of evaluation, is through the use of test-collections, which are composed of three main components: a set of documents, a set of topics and a set of relevance judgments (the ground truth). The Text REtrieval Conference (TREC[1]) provides such collections to the IR research community, as well as evaluation software which allows researchers to evaluate their (text retrieval) systems, in an offline setting. The evaluation results of a system are given in the form of per-topic scores and by then averaging these scores, a single system-score can be obtained.

Much of IR research focuses on optimizing the various aspects of such offline experiments. Stochastic simulation is one technique that can be used to assist this kind of research, and it allows us to overcome some limitations associated with IR data. Firstly, through simulation it is possible to generate *endless* amounts of data. This is generally helpful because IR data tends to be limited in size, for example due to the high costs associated with judging the relevance of documents on topics, which is typically done manually by human assessors. Secondly, simulation relies on statistical models that are fitted on data and describe some underlying population (i.e., the scores of systems on topics, or the clicks of users on hyperlinks). The advantage of drawing samples from statistical models, is that not only do we generate N observations, but now we also know some *true* characteristics about the underlying population from which these observations came from, i.e., the true mean value or the variance. Knowledge of such *true* characteristics would have otherwise been impossible, given the fact that the set of documents and queries can potentially be infinite. Thirdly, if we use models that are flexible enough, it is possible to generate data under some given condition, so that the resulting data have specific, desired characteristics. For example, we might want to study two systems that have a target difference in overall performance. In such cases, simulation can be useful, especially when our data contains a few (or no) observations with the desired characteristics we are looking for.

Simulation in IR is not a new concept. In the early days of IR, prior to the development and availability of large test-collections, simulation was used in an attempt to generate the entire test-collections themselves [11, 28]. In more recent years, simulation has been used to model various aspects of human interaction with the IR systems [36, 35, 4, 5, 16]. The term *human interaction* refers to how the user interacts with the system, for example formulating queries, clicking on results, re-formulating queries and so on. In [3], simulation was used to generate queries by selecting a document, the known-item, and producing a query for that known-item. This approach has the advantage that no additional relevance judgments are required, since the relevant document is simply the specified known-item. In [21], document scores are simulated in order to study the inherent noise that the per-topic evaluation scores carry due to the fact that test-collections contain a mere *sample* of documents rather than the entire *population* of documents.

In this thesis, we focus on a specific kind of simulation, which is the simulation of evaluation scores of systems on topics. More specifically, we explore the method proposed by Urbano and Nagler in [33]. This particular simulation approach, uses existing collections of system scores, to build a model for the

[1] `https://trec.nist.gov/overview.html`

joint distribution of system scores on topics, which can then be used to endlessly simulate scores by the same systems, but on random new topics. The simulation has two separate components: one *marginal* model for each system, which models the individual distribution of scores of the system (regardless of the other systems) and a *copula* that models the dependence among systems, meaning, how they tend to perform on the same topic. The main advantage of a copula approach, over classical multivariate models (i.e., the multivariate Gaussian distribution), is flexibility. For example, the marginal distribution of scores of each system, can be modeled with an entirely different distribution family. This is particularly useful for IR data, since not all systems follow the same distribution. Furthermore, it allows flexibility on the modeling of the dependence as well. For example, in one case the dependence could be stronger for small values of the system scores, and weaker for large values, and in another case it could be the other way around. Or, in different case, the dependence could be highly symmetric.

One application of this stochastic simulation, is to help researchers answer questions such as *"how many topics do we need to achieve a certain level of confidence in our evaluation results?"* [9, 29]. Previous work has largely relied on data-oriented approaches that repeatedly split the topic set in two halves and treat one of the halves as the ground truth and the other half as the actual test-collection. Then, a statistic of agreement between the evaluation results on each of the two splits is computed, for example Kendall's τ [38, 34, 32]. By extrapolating these observations, it is possible to obtain empirical estimates regarding the reliability of a test-collection, at a target number of topics. This so called *split-half* approach is limited due to the small amounts of available data, the fact that the ground truth is not *actual* ground truth, and that extrapolation is required. Beyond this approach, there is another one that relies on statistical theory, for example, *test theory* [6]. This kind of work is limited as well, due to the fact that it relies on various assumptions, that are typically not satisfied by IR data.

A second application of this stochastic simulation is to help researchers answer questions such as *"which statistical significance test is optimal for IR evaluation data?"*. Statistical significance tests can be used to determine if a small observed difference in mean system performance has occurred due to chance, or not. This is because random errors can occur due to the fact that systems are evaluated on a mere *sample* of topics, rather than the entire (possibly infinite) *population* of topics. Even though these tests are heavily used in IR related research papers [22, 8], it is still quite unclear which test is optimal for IR, and a lot of previous work contradicts each other. Statistical significance tests rely on various assumptions that are not actually satisfied by IR evaluation data. Earlier research argued on theoretical grounds about the robustness of statistical significance tests to having their assumptions violated [20, 13], but did not provide any empirical evidence. In later years, works such as [38, 23, 32], provided empirical results, by splitting the topic set in half, running a test on each split, and computing the agreement rate of the tests between splits. However, their results are limited by the fact that the tests can be consistently wrong on both splits. In [27], the authors compared various statistical significance tests, by measuring their agreement with the permutation test. This work is also limited since its conclusions are based on the assumption that the permutation test is optimal.

Stochastic simulation is a useful tool that can be used to overcome many of these aforementioned limitations that are found in previous work. With regards to the second example application that we mentioned (about finding the optimal statistical significance test for IR evaluation data), stochastic simulation has been employed in earlier works such as [37, 10, 7], however the models were simplistic and likely unrealistic. More recently, there have been two parallel lines of work that use more sophisticated simulation approaches. On the one hand, Urbano et al. [31] use a simplified version of the simulation model proposed in [33, 29], to generate paired scores for a given pair of systems, on *new random topics*. The reason why they focus on only two systems is because they only study paired statistical significance tests, which is the most common use case. On the other hand, Parapar et al. [19, 18] study the same problem, but instead simulate new random system *runs*[2] for the *same topics*.

1.2. Motivation

Surprisingly, the two aforementioned parallel lines of work of Urbano et al. and Parapar et al. reach different conclusions, despite the fact that both perform stochastic simulation.

In short, the works of Urbano et al. point in the direction of *t-test* being optimal and advocate for

[2]The run of system *s* on topic *t* refers to the resulting ranked list of documents that is produced when *s* is queried about *t*. However, the authors actually generate *relevance profiles*, not runs. The difference is that instead of providing ranked lists of documents, they only provide ranked lists of (binary) relevance values.

the discontinuation of the *Wilcoxon* test. Whereas the works of Parapar et al. make almost the exact opposite recommendations, advocating in favor of the *Wilcoxon* test. Interestingly, those two tests appear to be the most popular tests [8], with the t-test being used about 65% of time and the Wilcoxon about 25%. It is therefore important to investigate the reason why these two studies reach opposite conclusions and determine which of the tests is actually optimal.

On the one side, the main criticism of Urbano et al. aimed at Parapar et al. is that in order to study the tests ”one needs to simulate new topics for the same systems” [31]. This is because the significance tests are essentially trying to deal with the errors due to the sampling of topics. Another concern is that the simulation of Parapar et al. simulates retrieval scores[3] (as opposed to directly simulating effectiveness scores), which are then used to compute the effectiveness scores. As a consequence, the error of the simulation of retrieval scores, propagates to the effectiveness scores, to an extent that we do not know. Moreover, in the experiments of Parapar et al., the performance of a retrieval system is shifted across all topics, to obtain a better or worse system. However, in reality, systems may improve on some topics, but may not on others. Lastly, the coverage and scale of the study of Parapar et al. was small, and because we are interested in the behavioral trends of the systems (rather than specific cases), a wide range of factors needs to be studied.

On the other side, the main criticism of Parapar et al. aimed at Urbano et al. is that the quality of the simulation is unknown. It is argued that the simulated data could be biased towards specific tests, due to the fact that ”models are fitted from pre-selected classes of distributions” [18]. For example, if the simulated data come from a statistical model whose assumptions align with the assumptions of a certain test, that would favor the said test. So the results could be an artifact of the simulation. Furthermore, it is argued that, ”the best fit for each combination of measure and retrieval system might be still a poor fit”. In other words, they argue that the statistical models used to perform the simulation, despite their flexibility, may still be not good enough to describe IR evaluation data.

To some extend the concerns of Parapar et al. have been addressed in [30], arguing that a variety of both parametric and non-parametric families were used, including distributions based on Kernel Smoothing, which are as free of assumptions as they can be. Furthermore, it was shown that if some of the marginal distribution or copula families were in fact biased towards some tests; that bias would actually favor the *Wilcoxon* test. However, the authors did not provide any empirical results regarding the quality of the simulation, which is precisely the objective of this thesis.

1.3. Research Goals

In this thesis, we try to shed some light on this disagreement between the conclusions of the lines of work of Urbano et al. and Parapar et al. Our main goal is to provide empirical evidence regarding the quality of the simulation proposed in [33]. The idea is that if the simulation is of high quality, then this should further validate the recommendations made in [31], advocating for the use of the *t-test* and discontinuation of the *Wilcoxon* test.

We essentially try to answer the question: *”how good is the stochastic simulation proposed in [33]?”*. To this end, we evaluate the statistical models used to perform the simulation, in terms of how well they fit (or describe) the data; we call this: *goodness-of-fit*.

We explore the following:

- How well can copula-based models capture the joint distribution of IR system scores on topics?
- How can we improve the quality of a copula-based simulation approach, for the purposes of stochastically simulating scores of IR systems on new random topics?

1.4. Main Findings

- Overall, both the marginal models and the copulas (to a lesser extent) fit the data moderately well.
- All marginal and copula families perform fairly consistently when they are selected (by AIC), with the exception of Beta Kernel Smoothing.

[3] The term retrieval score refers to the score that the system itself gives to each document, in order to rank the documents from best to worst, during the retrieval process.

- The high appearance of zero scores in the data, is a special case where none of our marginal candidate families appear to be flexible enough to describe well.
- We proposed a new model selection criterion, inspired by the split-half approach, that is able to select copula models more optimally than other criteria such as AIC, BIC and LL. This is consistent across all effectiveness measures.

1.5. Thesis Outline

The remainder of this thesis is structured as follows. In Chapter 2, we provide a description of the particular stochastic simulation approach which we explore in this thesis, as well as how it was used in previous work. In Chapter 3, we study the marginal models (separately from the copulas). We describe the approach we devised for measuring goodness-of-fit, and then employ the said approach. We present and discuss our results, and proceed to explore outliers. We explore ways of improving the quality of the margins, and propose a new selection criterion. Lastly, we provide some empirical evidence, that addresses a limitation of our methodology. In Chapter 4, we study the copula models (separately from the margins), by employing approaches analogous with the ones used in the case of the margins. In Chapter 5, we conclude this thesis by discussing our main findings and their implications, and provide future work directions.

2

Background

2.1. Description of the Simulation

In this section we summarize the simulation proposed in [33], which as we mentioned is the one we explore in this thesis. This particular simulation builds on previous work done in [29].

Given an existing collection of evaluation scores of systems on topics, the objective is to build a model that captures the joint distribution of scores of those systems on the underlying *population* of topics. Using this model, it is then possible to endlessly simulate scores that come from the same systems, on new random topics. It is worth noting that the model built does not represent the *exact* systems of the existing collection, but rather systems *similar* to those.

In order to capture this joint distribution of system scores on topics, *copulas* are used, because they have an important advantage over classical multivariate models such as the multivariate Gaussian distribution, which is that they are generally more flexible. This is because copulas allow us to separate the modeling of the: *i)* marginal distribution of each individual system; meaning the univariate distribution of topic-scores of a system regardless of all other systems and *ii)* the dependence among systems; meaning how they tend to behave on the same topic.

The theoretical foundation behind copulas is the theorem of Sklar [25], which states that every multivariate cumulative distribution function can be expressed in terms of its marginals and a copula. For example, in the bi-variate case where we have two continuous random variables X and Y, with joint cumulative distribution function $F(x, y) = Pr[X <= x, Y <= y]$, according to Sklar's theorem the function F can be expressed in term of its marginals $F_X(x) = Pr[X <= x]$, $F_y(y) = Pr[Y <= y]$ and a copula C, like so:

$$F(x, y) = C(F_X(x), F_Y(y)) \tag{2.1}$$

It further holds that if F has a density f, then the density can be expressed as:

$$f(x, y) = c(F_X(x), F_Y(y)) * f_X(x) * f_Y(y), \tag{2.2}$$

where c, f_X and f_Y are the densities corresponding to C, F_X and F_Y respectively.

The simulation procedure is schematically shown in Figure 2.1. For simplicity, we illustrate and discuss the case of only two systems. The cases of three or more systems are completely analogous. However there is one important difference. In the case of only two systems, the dependencies are modeled using bi-variate copulas. In the case of three or more systems, *vine* copulas are used instead. Vine copulas are a generalization of bi-variate copulas, because they combine several bi-variate copulas in a tree structure, in order to build a dependence structure for arbitrarily high dimensions (i.e. N systems instead of 2).

For any two given systems A and B, three separate models need to be fitted, in order to be able to simulate. Two marginal models that model the marginal distribution of system A and B respectively, and one bi-variate copula that models the dependence between A and B. These three models can be fitted in any order, since they are separate.

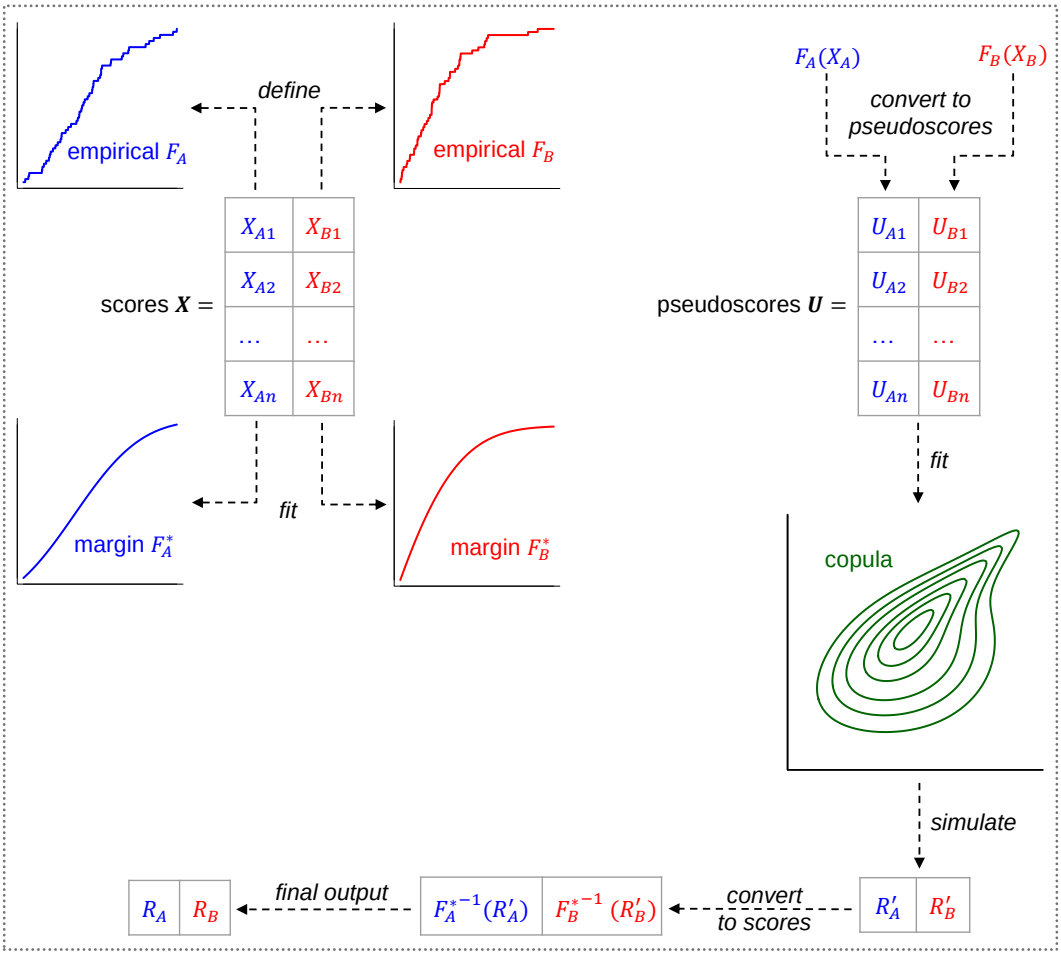


Figure 2.1: Schematic outline of the simulation.

The simulation works as follows: Let **X** denote the $n \times 2$ matrix of evaluation scores of two given systems A and B on n topics of some test collection. Let X_A and X_B denote the column vectors with the scores of system A and B respectively. To model the marginal distribution of system A, fit a marginal model on X_A. Similarly for system B. Let F_A^* and F_B^* denote the cumulative distribution function CDF of the newly fitted marginal models of system A and B respectively. Let F_A and F_B denote the empirical cumulative distribution function ECDF (Equation 2.3) of system A and B respectively.

$$ECDF(x) = \frac{\text{number of elements in the sample} <= x}{\text{sample size}} \tag{2.3}$$

To model the dependence between systems A and B a copula needs to be fitted. Copulas are actually fitted on so-called *pseudo-scores*. Let **U** denote the $n \times 2$ matrix of pseudo-scores, and U_A and U_B the column vectors with the pseudo-scores of system A and B respectively. Scores are converted to pseudo-scores like so: $U_A = F_A(X_A)$ and $U_B = F_B(X_B)$. A copula model is then fitted on **U**.

The copula model can be used to generate *paired* pseudo-observations $\{R'_A, R'_B\}$ of system A and B on a new random topic. Those two pseudo-observations are then converted to actual observations using the inverse CDF of the (previously) fitted margins, like so: $R_A = F_A^{*-1}(R'_A)$ and $R_B = F_B^{*-1}(R'_B)$.

$\{R_A, R_B\}$ is the generated pair of scores of system A and B on a new random topic. This procedure can be endlessly repeated to generate scores for an arbitrarily large number of new random topics.

2.1.1. Marginal and Copula Families

In order to fit a margin or a copula, several distribution families are considered, which are listed in Table 2.1.

Margins	
for *continuous* measures (AP, nDCG@20, ERR@20)	Truncated[1] Normal
	Truncated Normal Kernel Smoothing
	Beta
	Beta Kernel Smoothing
for *discrete* measures (P@10, RR)	Beta-Binomial
	Discrete Kernel Smoothing[2]
Copulas	Gaussian
	Student t
	Frank
including their 90, 180 and 270 degree rotations	Clayton
	Gumbel
	Joe
	BB1
	BB6
	BB7
	BB8
	Tawn 1
	Tawn 2

Table 2.1: List of the family distribution candidates that are considered when fitting the marginal and copula models.

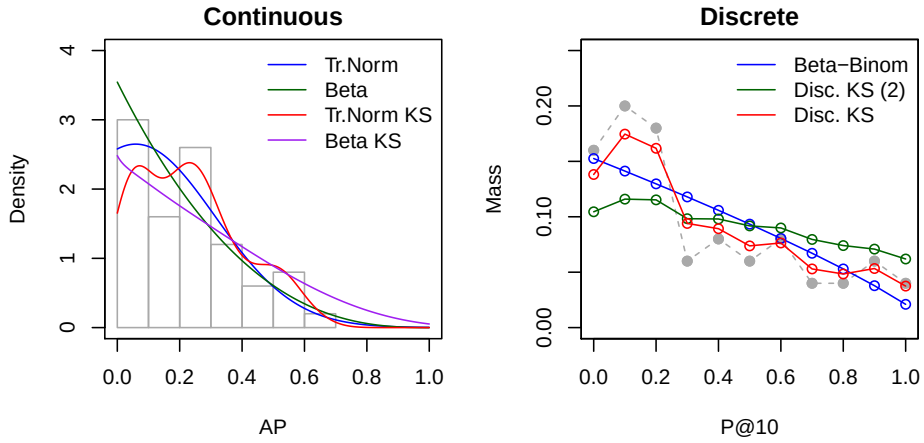
For modeling the marginal distribution of a system, a simple approach that does not require model fitting would have been to use the empirical cumulative distribution function (ECDF) of the given data. However, the problem with simply using the ECDF is that the scores which were not present in the data, would never come up in the simulation. For this reason, a model is required. Furthermore, a distinction is made depending on whether or not the evaluation scores are *continuous* or *discrete*. This is because there are multiple effectiveness measures (i.e., AP, nDCG@20, ERR@20, P@10 and RR) and even though all of them are technically discrete, some of them have a much larger set of possible values, which makes it reasonable to treat them as continuous.

Figure 2.2 shows a visual comparison of all candidate models for modeling the marginal distribution, in two examples. The left-hand side plot, shows all the candidate models (according to Table 2.1) that could be selected for modeling the marginal distribution of AP (Average Precision) scores of a system on the population of topics. Similarly, the right-hand side plot, all the candidate models that could be selected for modeling the marginal distribution of P@10 scores of another system. The original data are shown in gray color. Note that sometimes models may fail to fit, in cases where the fit is particularly bad. For example, on the right-hand side plot, two models are missing because they failed to fit. Only three models were fitted successfully, even though five models were considered (namely, the Beta-Binomial distribution plus four variations of the Discrete Kernel Smoothing distribution).

For modeling the dependence between the two systems, 12 copula families are considered, including some rotations. The Tawn 1 and 2 copulas are asymmetric copulas, whereas all others are symmetric. The main difference between these two categories is that when a symmetric copula is used, the distribution of the generated per-topic score differences tend to be symmetric. In contrast, asymmetric copulas may yield distributions with non-zero skewness.

Figure 2.3 shows a visual comparison of three candidate copulas for modeling the dependence between two systems, in one example. The left-hand side plot, shows the visual difference between a symmetric (Gaussian) and an asymmetric (Tawn) copula. The right-hand side plot, shows the Independence copula for comparison, which assumes no dependency between the two systems, and the contours are perfect circles. These copulas were selected for illustration purposes, even though many more are considered (as per Table 2.1). The contour plots show the joint probability density function $f(x, y)$ (Equation 2.2) that is modeled with a Tawn, Gaussian and Independence copula. The reason why the joint density f is plotted, as opposed to the copula density c, is because copula densities usually explode at some corners, which makes it difficult to visualize. A common approach is to combine

[1]The distributions are truncated, so that the simulated scores are within the $[0, 1]$ range, as all IR evaluation scores are.
[2]Four different variations of this distribution are considered by manipulating the smoothing parameter.



(a) System 52 from the Ad-hoc 2007 Track.

(b) System 29 from the Ad-hoc 2007 Track.

Figure 2.2: Visual comparison of the candidate marginal models. Original data in gray.

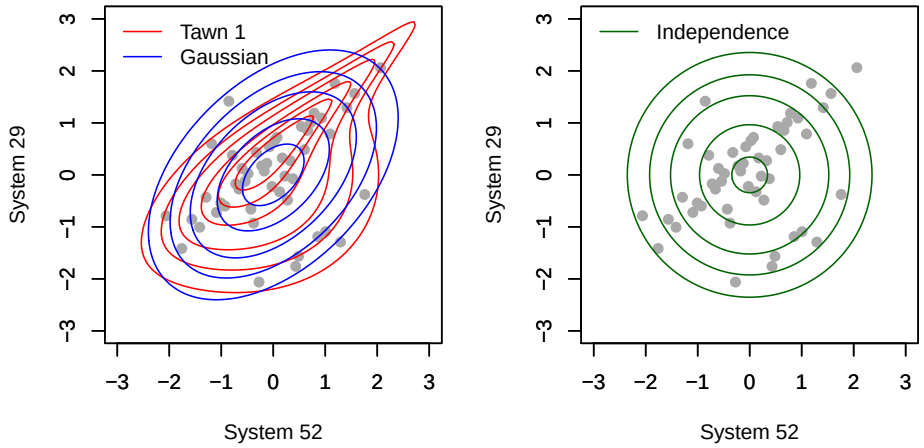


Figure 2.3: Visual comparison of three candidate copulas, fitted from the (paired) AP scores of systems 52 and 29 from the Ad-hoc 2007 Track. These contour plots show the joint densities, by combining the copulas with standard normal margins. Original data in gray.

the copula with standard normal margins, and plot the joint density instead, as we have done.

2.1.2. Model Selection Criteria

In order to fit a model, *all* candidate families of Table 2.1 are fitted first and then the best model has to be selected based on some model selection criterion. To this end, several options are available:

Log-likelihood (LL) is a basic criterion that can be used, and it is defined as the natural logarithm of the likelihood that the fitted model would have generated the observed data. Based on LL, two more criteria are defined, namely the Akaike Information Criterion (AIC) [1] and the Bayesian Information Criterion (BIC) [24]:

$$LL = \log\left(p\left(X|\theta\right)\right) \tag{2.4}$$

$$AIC = -2LL + 2\kappa \tag{2.5}$$

$$BIC = -2LL + \kappa \log\left(n\right) \tag{2.6}$$

where X is the observed data, θ is the vector of parameters of the model, κ is the number of parameters of the model (or the *effective degrees of freedom* in case the model is non-parametric) and n is the sample size (the number of topics).

For the case of LL, the model with the highest value is selected. The opposite is true for the case of AIC and BIC, where the model with the lowest value is selected. All three criteria are probabilistic measures that estimate a model's performance on the same data that was used to fit that model. This means that their computation does not require a hold-out set. The main difference between the three criteria is that AIC and BIC penalize complex models, and therefore favor simple models. LL can be problematic because it might favor models that are overly complex, and those models are often the result of overfitting. BIC penalizes models harsher than AIC, if $n \geq 8$.

2.2. Related Work

One application of the aforementioned simulation is that it can be used as a tool to study which statistical significance test is optimal for IR evaluation data. Statistical significance tests are used to assess if an observed difference in mean system performance is real, as opposed to an error due to the sampling of topics. This is because systems are evaluated on a mere sample of topics rather than the entire population of topics, and for this reason there is some random noise associated with evaluation results. There are various statistical significance tests, for example the Student's t-test, Wilcoxon test, Sign test, Bootstrap test and Permutation test. Every test relies on certain assumptions that are typically not satisfied by IR evaluation data. For example, the t-test assumes that the data are normally distributed, which is not true.

Most commonly, researchers are interested in comparing only two systems; an experimental system against a baseline system. To achieve this, the two systems are typically evaluated on the same test collection and as a consequence on the same set of topics. For this reason, the evaluation results are in the form of *paired* per-topic scores. After the results are obtained, if it is observed that the experimental system outperformed the baseline system, then a so-called *paired* statistical significance test is run, to determine if this observed difference in mean system performance is statistically significant or not. In other, rarer cases where a researcher wants to compare N systems as opposed to only two, ANOVA models are used instead.

The simulation approach we discussed in Section 2.1, was applied in [31], in order to study how well statistical significance tests really behave on IR evaluation data. The authors studied an extensive range of different factors, while focusing on the specific (but popular) case of *paired* tests. In such tests, only two systems (as opposed to n systems) are compared. In order to compare the various statistical significance tests, the main requirement is to estimate their Type I and Type II error rates. One way of accomplishing this, is by employing stochastic simulation.

In order to study Type I error rates, the following process is repeated. For a target effectiveness measure and topic set size, two systems B and E are randomly selected from the same collection. Then, two marginal models (F_B and F_E) and a copula are fitted on the data. The model selection is done using AIC. For computing Type I error rates, the data are generated under the null hypothesis $H_0 : \mu_E = \mu_B$. This is achieved by assigning $F_E \leftarrow F_B$ after the margins have been fitted. After the simulation, the significance tests are then run on the newly generated data. Any statistically significant result counts as a Type I error, due to the fact that the data were generated under the null hypothesis.

In order to study Type II error rates, the process is similar. However, this time the data need to be generated under the alternative hypothesis $H_A : \mu_E = \mu_B + \delta$. This requirement is achieved in two steps. Firstly, system B is selected from the bottom 75% performing systems, and system E is selected at random from the set of 10 systems whose mean is closest to the target $\mu_B + \delta$. Secondly, after the margins have been fitted, a small transformation is performed on F_E, such that the condition $\mu_E = \mu_B + \delta$ holds true. After the simulation, the significance tests are then run on the newly generated data. Any result that does not come up as statistically significant, counts as a Type II error, due to the fact that the data were generated under a false null hypothesis.

Their findings suggest that the t-test and Permutation tests are the most optimal, whereas the Wilcoxon, Sign and Bootstrap-Shift test are the least optimal. The authors' top recommendation is the t-test.

Beyond the line of work of Urbano et al., there is another recent line of work by Parapar et al. [19, 18], that also relies on simulation. In [19], for every system-topic pair, a model is built that models the *retrieval score* distribution (SD) [15]. The term retrieval score refers to the score that the system itself gives to each document, in order to rank the documents from best to worst, during the retrieval process. The model used is a mixture of two log-normal distributions: one for relevant and the other

for non-relevant documents.

In order to study Type I error rates, the following process is repeated. A system is randomly selected, and all (50) of its mixture models are used to generate two outputs each. The outputs are *synthetic* lists of 1000 retrieval scores and corresponding relevance values, sorted from best to worst retrieval score. For each output, Average Precision (AP) is computed. The statistical significance tests are then run on the two resulting sequences of 50 AP scores. Any statistically significant result is a Type I error, because the data were generated under the null hypothesis, since they come from the same system.

In order to study Type II error rates, the approach is similar. The difference is that instead of generating two outputs per mixture model, one output is generated instead. Then, the parameter of the model is altered, and the second output is generated. The model is altered by increasing the true mean of the (log-normal) distribution for the relevant documents. This way, the data are generated under some alternative hypothesis.

In a more recent paper [18], the work done in [19] was improved by the same authors. The approaches used in the two papers are very similar. The main difference is in the simulation methodology. Given a system-topic pair, a model is build that captures the relationship between document ranks and relevance. More specifically, for a system-topic pair, a logistic regression model is fitted, where the target variable is relevance, and the only predictor is the position of the document in the ranking. In order to simulate a new ranking, for every position p in $\{1, 2, ..., 1000\}$, the value $h_\theta(p)$ is obtained (from the fitted Logistic model h_θ). The relevance value of the new ranking at position p is determined by drawing a sample from a Bernoulli distribution with parameter $h_\theta(p)$. This way a sequence of 0s and 1s is generated, which is sufficient for computing Average Precision. Studying Type I and Type II errors is done in the same manner. One difference is in regard to studying Type II errors, where the parameters θ_0 and θ_1 of the logistic regression model are manipulated in order to simulate under some alternative hypothesis.

Surprisingly, the authors reach opposite conclusions. The biggest disagreement in terms of recommendations is regarding the Wilcoxon test and the t-test, where the conclusions are opposite.

Parapar et al., in [18], provide some empirical results regarding the quality of the simulation used in that paper. This was done by computing the average correlation between original ranking and simulated rankings of the system-topic pairs. Their results showed that the simulated rankings only differ slightly compared to the original, and it is argued that this is a good sign of quality, because the generated rankings should represent the original. However, this is arguably not a very comprehensive analysis, because according to this comparison, adding a small noise to the rankings would be considered a simulation of high quality.

Thus far, no empirical evidence have been produced regarding the quality of the simulation used in the line of work of Urbano et al., which is the main objective of this thesis.

3

Margins

As already mentioned, the main goal of this thesis is to provide empirical evidence regarding the quality of the simulation proposed in [33]. To this end, we can evaluate the statistical models that are used to perform the simulation, in terms of how well they fit (or describe) the data. We refer to this concept as *goodness-of-fit*. If the models describe the data well, then consequently the simulation should be fairly realistic.

As we discussed, the simulation relies on three models that are fitted separately: two margins and a copula. Conveniently, this also allows us to evaluate the margins separately from the copula. In this chapter, we explore how well the marginal distribution of system scores is modeled, independent of the copulas.

3.1. Defining Goodness-of-Fit

One simple approach for measuring the goodness-of-fit of a model would be to compute its Log-Likelihood (Equation 2.4). However this statistical measure is meaningful only for comparing two or more models. The absolute Log-Likelihood value in isolation is mostly meaningless. Furthermore, Log-Likelihood only measures how well the model fits the observed data; not how well it describes the underlying population. In other words, it does not measure the model's predictive power on unseen data.

Ideally, the best way to measure the goodness-of-fit of a fitted model F^* is to compute its similarity to the *true* distribution F (the distribution of scores on the entire population of topics), also known as the *ground truth*. Assuming that we have knowledge of the true distribution, there are several options for computing this similarity. We have considered three metrics that actually measure dissimilarity, which means that low values equate to high similarity. Namely, the *i)* Kolmogorov–Smirnov (KS) statistic, *ii)* Cramér–von Mises (CvM) criterion and *iii)* Anderson–Darling (AD) statistic. All three of these metrics essentially define a distance Δ between the two cumulative distribution functions: F^* and F.

The Kolmogorov–Smirnov (KS) statistic [14, 26] is a simple metric which is defined as the largest absolute difference between two cumulative distribution functions across all x-values. The left plot of Figure 3.1 shows the computed value of the KS statistic in one example. The value is the length of the red line.

$$\text{KS}\,(F^*, F) = \sup_x \left|F^*(x) - F(x)\right| \tag{3.1}$$

The Cramér–von Mises (CvM) criterion [17] is defined as the squared area between the two curves. The right plot of Figure 3.1 shows the computed value of the CvM statistic in one example. The value is the square of the red area.

$$\text{CvM}\,(F^*, F) = \int_{-\infty}^{\infty} \left[F^*(x) - F(x)\right]^2 dF^*(x) \tag{3.2}$$

Finally, the Anderson–Darling (AD) statistic [2] is defined similar to the CvM criterion. The main difference is that it places more weight on observations at the tails of the distribution, due to the $w(x)$

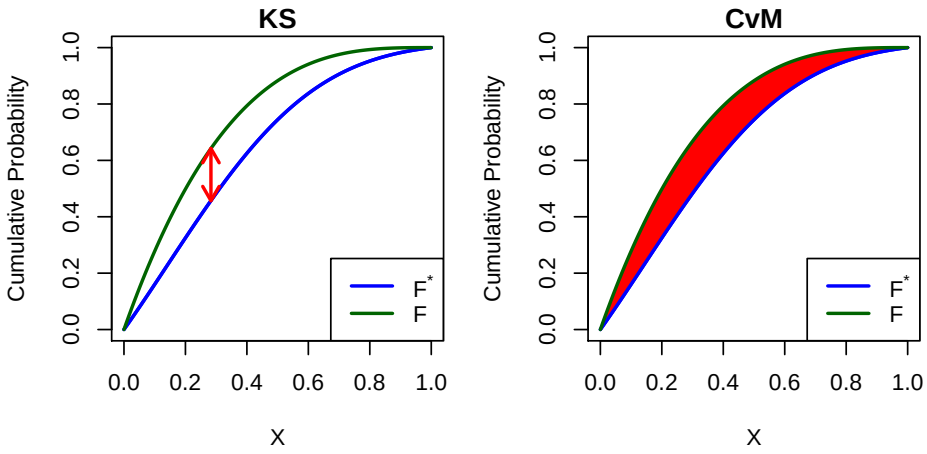


Figure 3.1: Visualization of Kolmogorov–Smirnov statistic (left) and Cramér–von Mises criterion (right).

weighting function. When the weighting function is $w(x) = 1$, the statistic is identical to the CvM criterion.

$$\text{AD}\left(F^*, F\right) = \int_{-\infty}^{\infty} \left[F^*(x) - F(x)\right]^2 w(x) dF^*(x),$$
$$\textit{where } w(x) = \frac{1}{F^*(x)\left(1 - F^*(x)\right)} \tag{3.3}$$

In practice, having knowledge of the true distribution of a system's scores is not feasible, because it would imply that the system has been evaluated on the entire (possibly infinite) population of topics. This is either impractical due to the enormous amount of relevance judgments required, or even impossible if the potential set of topics is infinite or not well-defined. For this reason, some estimate of this true distribution is required.

To this end, the so-called *split-half* approach can be utilized, as schematically outlined in Figure 3.2. This approach is fairly common in the field of IR research [38, 34, 32, 23, 12, 27], and it is used in cases where obtaining ground truth data is too expensive or not feasible. Following this approach, the observations (in our case, the 50 topic-scores of a given system) are randomly split in two halves. The first half is treated as *'the sample'* (the actual observations) and the second half as *'the population'* (the ground truth). This means that for the purposes of measuring goodness-of-fit, the models are actually fitted on only the first half of the data. The empirical cumulative distribution ECDF of the second half of the data is used as an estimate of the true distribution (the ground truth). We use F_1^* to denote the distribution of the fitted model and F_2 to denote the estimated true distribution. In order to measure the goodness-of-fit of the model F_1^*, we can use any of the three aforementioned metrics (Equations 3.1-3.3).

For our purposes, there is likely no benefit in placing more weight on the tails of the distribution, or only considering the largest distance. Therefore, the CvM criterion (Equation 3.2) can be safely chosen as the measure of choice. We use Δ_{obs} to denote the observed distance (or dissimilarity) between a fitted model and the estimated ground truth, that corresponds to a given random split. Intuitively, it represents the area between the two curves.

$$\Delta_{\text{obs}} = \sqrt{\text{CvM}\left(F_1^*, F_2\right)} \tag{3.4}$$

Utilizing Δ_{obs}, the goodness-of-fit of a model F_1^* can simply be defined as its $-\Delta_{\text{obs}}$. The minus sign is due to the fact that the goodness-of-fit of a model and its deviation from the ground truth are inversely related.

As an example, in Figure 3.3a we show the computed Δ_{obs} values of two models in one particular split. One limitation regarding the interpretation of these measurements, is that it is not obvious which specific range of values constitute a good fit. In relative terms, we could say that the model with the

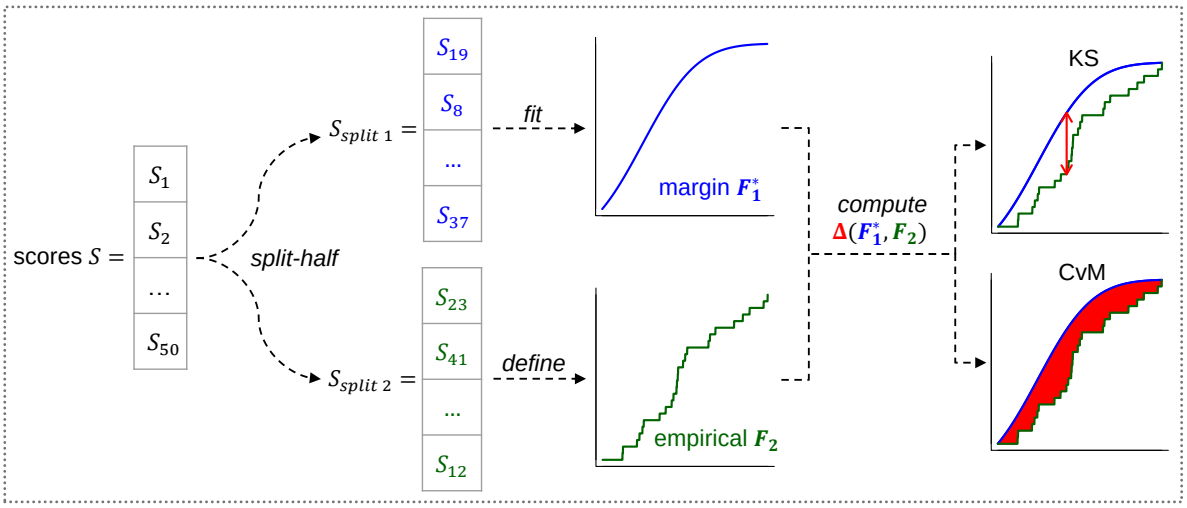


Figure 3.2: Diagrammatic representation of the Split-Half approach.

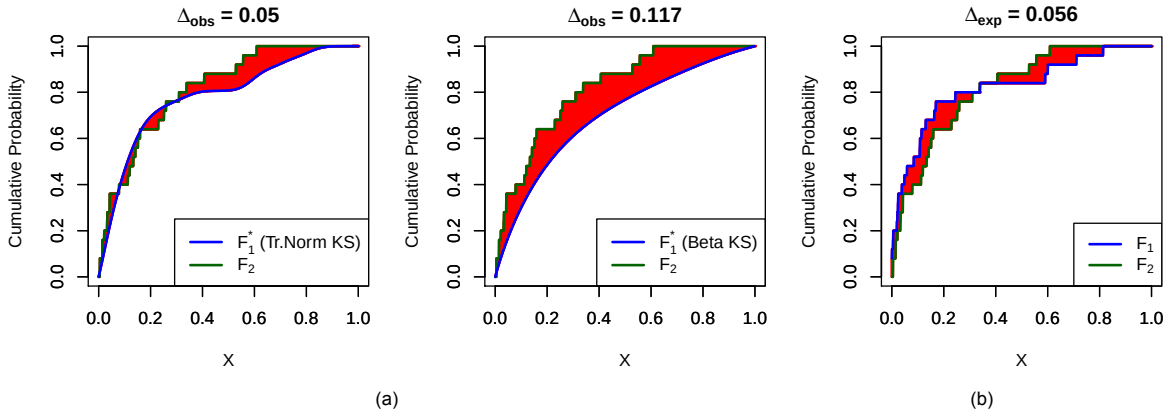


Figure 3.3: Visualization of: (a) the Δ_{obs} of two models on the same split, and (b) corresponding Δ_{exp}.

lowest Δ_{obs}, in this case 0.05, is better. However, in absolute terms, it is not obvious how to assess if the model that measured $\Delta_{obs} = 0.05$ constitutes a good fit or not. To overcome this problem, we need to determine what value we should approximately expect from a good fit. We use Δ_{exp} to denote this value.

One way of calculating this Δ_{exp}, is to measure the distance that would have been observed if the empirical distribution of the first half of the data (F_1) was used instead of the fitted model, as shown in Figure 3.4.

$$\Delta_{exp} = \sqrt{\text{CvM}\,(F_1, F_2)} \tag{3.5}$$

The advantage of using the empirical distribution as a *reference*, is that it gives us an unbiased measurement, because it is not based on any model. Using this definition, in this particular example, Δ_{exp} was measured to be 0.56 (Figure 3.3b), which is actually higher than 0.05. This implies that the fit was slightly better than our expectation. In general, a model that fits the data well should measure a Δ_{obs} that is about the same as its corresponding expectation Δ_{exp}.

3.2. Experiments

For our experiments, we require existing collections of evaluation scores of systems on topics, so that models can be fitted on real data. The collections we use throughout this thesis come from actual results of systems that TREC participants submitted in previous years, in the Ad-hoc and Web track, as detailed in Table 3.1. This dataset contains a large number of systems, as well as a wide range

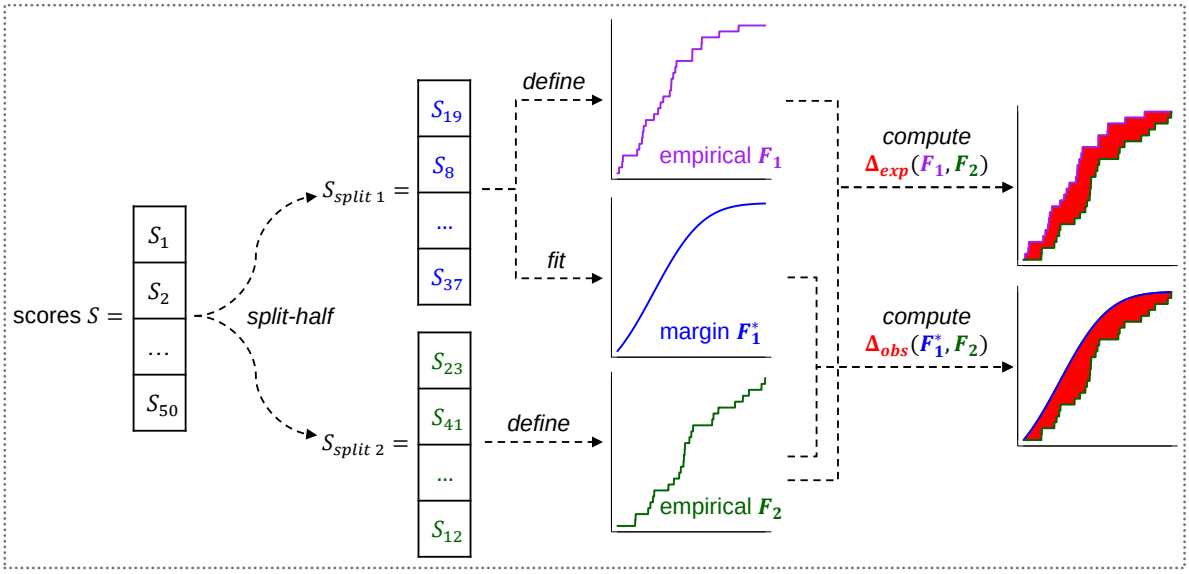


Figure 3.4: Diagrammatic representation of Δ_{obs} and Δ_{exp}.

TREC Track	Years	#Systems	#Topics	Effectiveness Measures
Ad-hoc	2005 to 2008	363	50	{AP, RR, P@10}
Web	2010 to 2013	216	50	{nDCG@20, ERR@20}

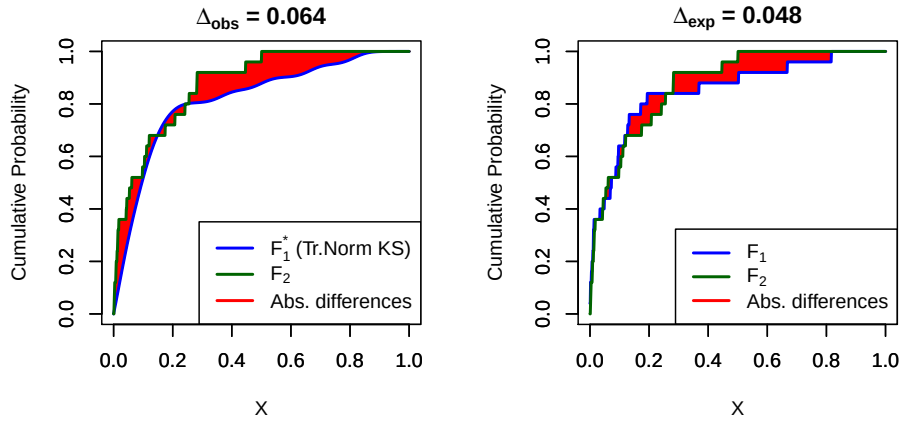
Table 3.1: Some descriptive statistics about the data collections.

of effectiveness measures, namely: AP, nDCG@20, ERR@20, P@10 and RR. This ensures that the quality of the stochastic simulation is explored on a variety of IR data. We pre-process the data by removing the bottom 10% performing systems, to avoid erroneous system implementations, so the final number of systems is slightly smaller.

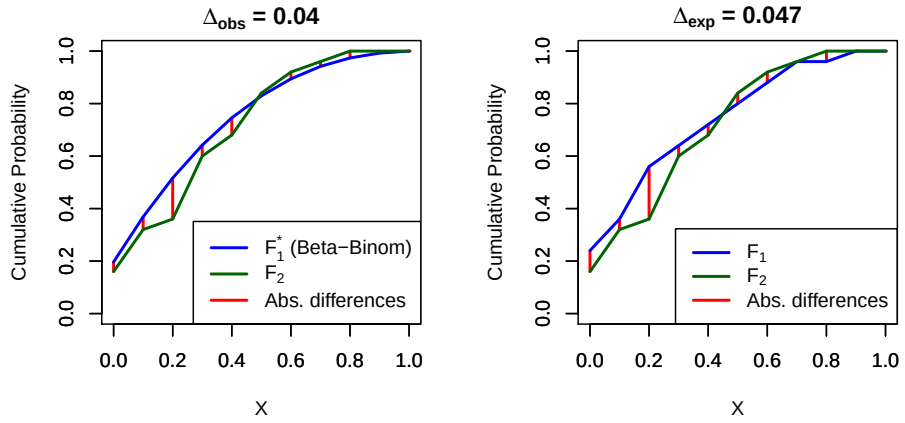
Our primary objective is to measure how well can we capture the marginal distribution of scores, of a given system. To this end, for each effectiveness measure, we select a random system from a random collection and perform the aforementioned split-half approach using the system's scores on all (50) topics. In total, $250{,}000$ splits were performed; $50{,}000$ for each effectiveness measure. This amount of splits seems to be sufficient, judging from the narrow confidence intervals we obtain for most of our results. For each split, we calculate a corresponding Δ_{exp} (Equation 3.5). Then, using the data in the first half of the split, we fit all possible models according to Table 2.1 and compute their corresponding Δ_{obs} (Equation 3.4).

Figure 3.5 shows the Δ_{obs} (left) and Δ_{exp} (right) for two example random splits. The approach used for computing these values is slightly different depending on whether the data are continuous (AP, nDCG@20, ERR@20) or discrete (P@10, RR). For the case of continuous measures, we use an estimation approach by averaging the absolute differences between the two curves across 1000 equally spaced x-values in the $[0, 1]$ range. For the case of discrete measures, we average the absolute differences between the two curves across all possible x-values. P@10 has 11 possible values: $\{0, 0.1, 0.2, \ldots, 1\}$, whereas RR has 1001. In the first example (Figure 3.5a), the effectiveness measure was AP (continuous metric) and the Truncated Normal Kernel Smoothing distribution was selected. The model performed worse than the expectation (0.064 > 0.048). In the second example (Figure 3.5b), the effectiveness measure was P@10 (discrete metric) and the Beta-Binomial distribution was selected. The model performed better than our expectation (0.04 < 0.047).

In Figure 3.6 we report the results across all $250{,}000$ random splits, separately for each effectiveness measure and family of distribution. All model selections were made according to AIC, as done in [31]. The left plot shows the mean values of Δ_{obs} and Δ_{exp}. In general, models that fit the data well should give us Δ_{obs} values about the same as their corresponding expected value Δ_{exp}. In order to make it easier to visually interpret the results, we have defined a goodness-of-fit metric, denoted as *GoF*, that



(a) Example random split using the AP scores of system 91 on the 50 topics of the Ad-hoc 2008 test collection.



(b) Example random split using the P@10 scores of system 107 on the 50 topics of the Ad-hoc 2008 test collection.

Figure 3.5: Visualization of Δ_{obs} (left) and Δ_{exp} (right) for: (a) the continuous case using AP scores, and (b) the discrete case using P@10 scores.

combines Δ_{obs} and Δ_{exp} in a single formula, as the (negative[1] of the) percentage deviation of Δ_{obs} from the corresponding expected value Δ_{exp}:

$$\text{GoF} = -\frac{\Delta_{obs} - \Delta_{exp}}{\Delta_{exp}} \tag{3.6}$$

The interpretation of GoF is quite straightforward. For example, when GoF is -0.333 (Figure 3.5a), this means that Δ_{obs} is 33.3% higher than the expectation. Similarly, when GoF is 0.149 (Figure 3.5b), this means that Δ_{obs} is 14.9% lower than the expectation.

Overall, our results show that the average GoF is slightly less than zero across the board, with the exception of the Beta Kernel Smoothing distribution family. Ideally, we want to observe values close to zero. These results suggest that the models provide a moderately good fit, however there is certainly room for improvement. Moreover, we notice that the models fit discrete metrics (P@10 and RR) noticeably better than continuous (AP, nDCG@20 and ERR@20).

Figure 3.7 shows the frequency with which each candidate family is selected. One side-effect of employing a split-half approach is that the data are being halved, which means that our marginal models are actually fitted on only 25 topic-scores, instead of the entire set of 50 topic-scores. However, we are truly interested in measuring the goodness-of-fit of models that are fitted on 50 scores, as opposed to 25. For this reason, our plot shows the frequency with which each candidate family is selected, when the models are fitted on *i)* 25, as well as *ii)* 50 scores. It appears that the models are selected very similarly in both cases, except for two Discrete Kernel Smoothing variants in the case of P@10. This increases

[1]This is because the goodness-of-fit of a model and its deviation from the ground truth are inversely related.

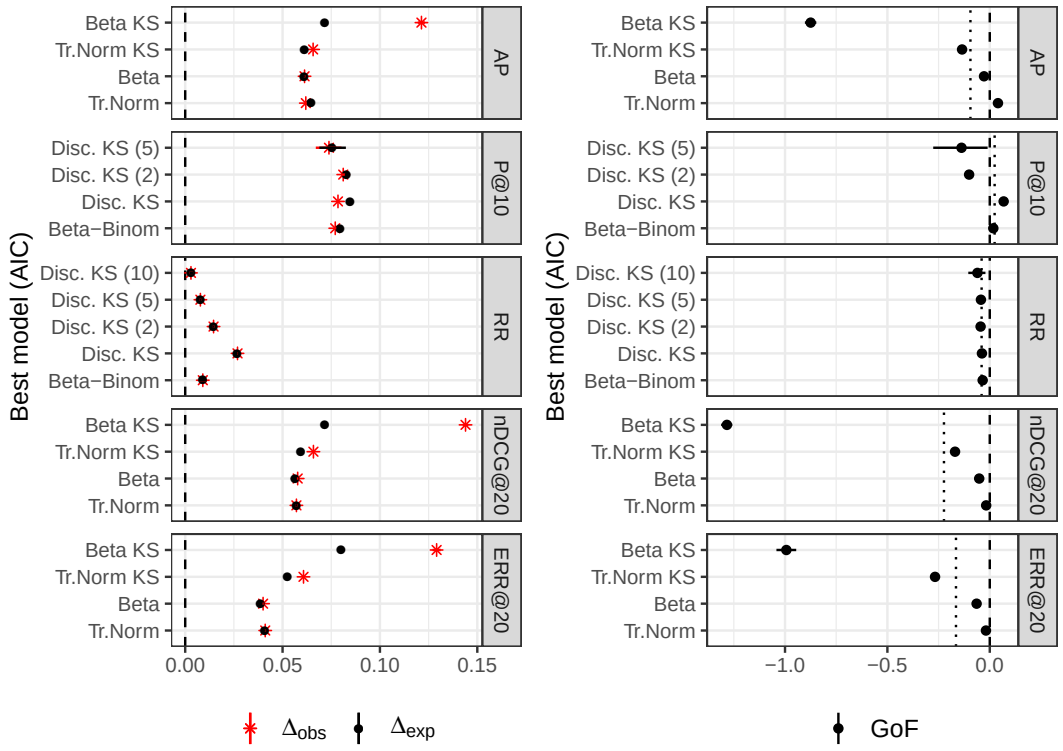


Figure 3.6: How well does each family of marginal distributions perform, when it is selected by AIC? Mean values are reported, along with 95% bootstrap confidence intervals. The dotted vertical lines (on the right plot) indicate the overall means across each effectiveness measure.

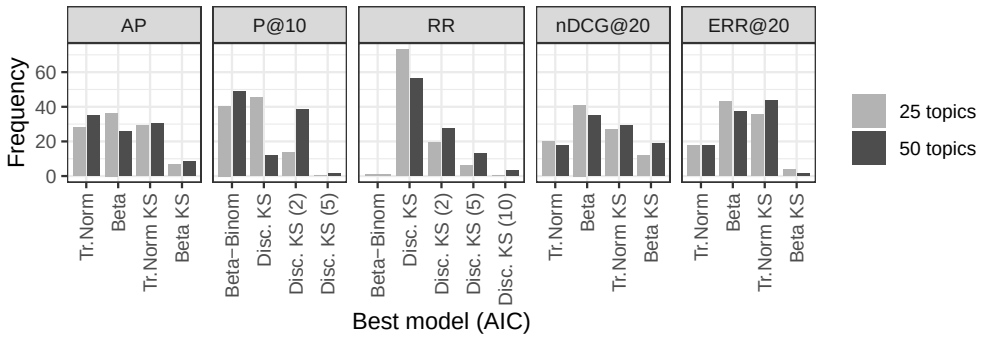


Figure 3.7: Frequency with which each candidate family of marginal distributions is selected by AIC. We consider the case where models are fitted on 25 and 50 topics respectively.

our level of confidence regarding the accuracy of our estimates of goodness-of-fit. Furthermore, since all candidate families get selected, and no particular family gets chosen with a significantly higher frequency than the rest; this reaffirms the idea that IR data are quite complex, as it implies that a variety of marginal models is required, to describe IR data.

The Beta KS distribution family is an obvious outlier with an average GoF of about -1; which means that the average Δ_{obs} is twice what we expected. This would not be a problem, if that average Δ_{obs} was a very low value, but as we can see in the left plot this is not the case. It is actually the highest across all (continuous) effectiveness measures. There are two possible explanations for this. One possible explanation is that our model selection criterion (AIC) made a poor choice by choosing this particular distribution family, which means that some other candidate family would have performed better if it had been properly selected. Another possible explanation is that this particular set of random splits is a corner case where none of the candidate families that are incorporated in the simulation would have performed well, and Beta KS just happened to be the lesser bad choice. As we can see in Table

	25 topics	*50 topics*
AP	6.43%	8.58%
nDCG@20	11.9%	18.6%
ERR@20	3.96%	1.56%
Overall	7.44%	9.59%

Table 3.2: How often is Beta KS selected by AIC? We consider the case where models are fitted on 25 and 50 topics respectively.

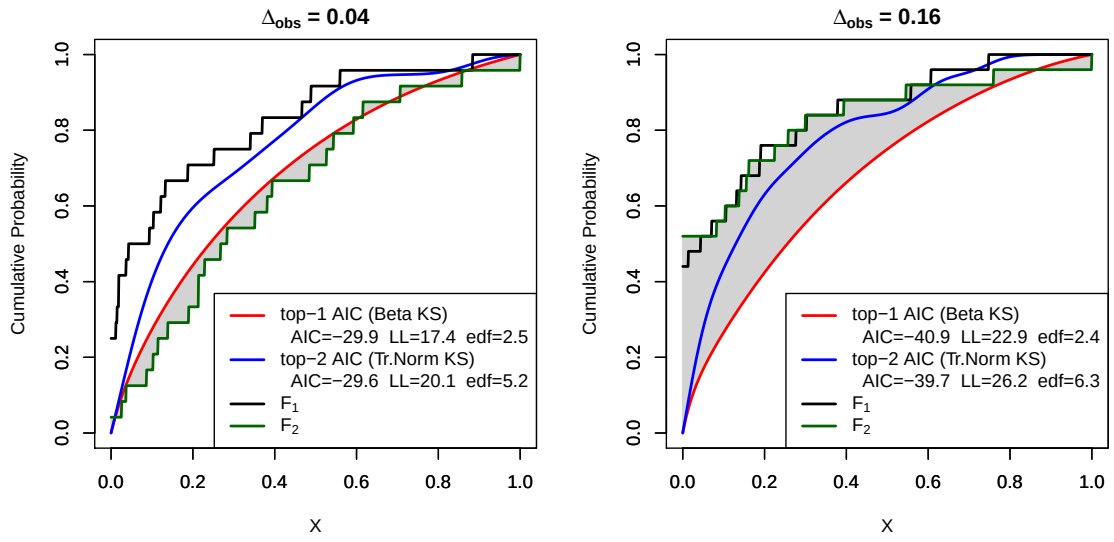


Figure 3.8: Visualization of two example Beta KS fits: a good fit (left), and a bad fit (right).

3.2, Beta KS is not selected very frequently. Overall, it is selected 9.59% of the times where it is eligible for selection. Even though this not a particularly high percentage, dealing with this outlier could considerably improve the quality of the simulation. It is therefore meaningful to explore this further.

In an effort to explain why the Beta KS distribution performs poorly on average, even though AIC ranks it 1st, we experimented by comparing a few cases where it performed well, with cases where it performed poorly, at the extremes. Figure 3.8 illustrates two of those example cases, where Beta KS provided a good fit (left) and a bad fit (right), respectively. These examples were hand-selected to demonstrate a noticeable pattern that we observed through our exploratory experimentation. In both examples, we see that Beta KS has trouble fitting the data, when the number of zero values (in this case, nDCG@20 scores) is too large. This is because the fitted models tend to be too simple, as evident by the low *effective degrees of freedom* (edf) of 2.5 and 2.4 respectively. In other words, Beta KS is not a complex enough model to capture the high appearance of zeros in the data. Despite this, in both examples, AIC selected these simple yet seriously underfitted Beta KS models. Looking at the right-hand side plot, we see that Beta KS does not describe the ground truth data well, measuring $\Delta_{\text{obs}} = 0.16$. In contrast, Log-Likelihood would have made a better choice, namely Truncated Normal KS, that would have measured $\Delta_{\text{obs}} = 0.07$. However, since none of the model selection criteria are perfect, this is not entirely surprising. Looking at the left-hand side plot, we see that the reason why Beta KS sometimes performs well, in this case measuring $\Delta_{\text{obs}} = 0.04$, is simply due to chance. More specifically, due to the randomness of splitting the data in half, sometimes one of the data halves contains many more zero values than the other half. If the zeros are present mostly in the first half, the underfitted Beta KS model describes the ground truth data well, due to chance. In summary, we found that in certain corner cases (i.e., when the appearance of zeros in the data is high), AIC tends to select Beta KS due to its low complexity, despite the model being underfitted, which often results in high Δ_{obs} measurements.

Based on our exploratory experimentation, we speculated that the large number of zero scores in the data causes outliers in our results. To further verify this, in Figure 3.9 we compare those specific

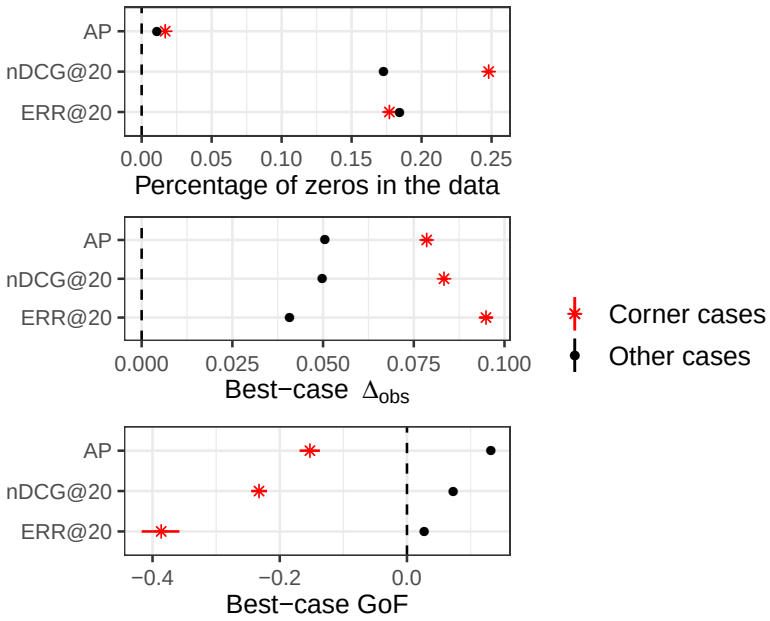


Figure 3.9: Comparing the cases where AIC ranked Beta KS 1st, with all other cases.

random splits where AIC selected Beta KS (we refer to those cases as 'corner cases'), with all other cases. In the top plot, we see that for the case of nDCG@20 (which is actually the most common case, as per Table 3.2), the number of zeros in the data is significantly larger in these corner cases, compared to all other cases. This is in line with our previous findings. However, this is not true for AP and ERR@20, which means that there are other particularities about these corner cases, beyond the high appearance of zeros, which we did not manage to identify through our exploratory experimentation. In the middle and bottom plots, we show the theoretical minimum Δ_{obs} and maximum GoF, that would have been achieved if the candidate models had been selected in an optimal manner. This was done by simply selecting the model with the lowest Δ_{obs} in each random split. It appears that even the best possible candidate would not have performed well in the corner cases; in both absolute terms (Δ_{obs}), and terms relative to the expectation (GoF). This suggests that additional candidate models would have been required, to achieve a good fit. One possible addition of such candidate could be a mixture model, that models the zero scores separately from non-zeros. However, we leave this for future work.

In Figure 3.10, we see a clear correlation between the appearance of zeros in the data, and goodness-of-fit; both in absolute terms (Δ_{obs}), and terms relative to the expectation (GoF). This further verifies our hypothesis that data containing a high number of zeros are not modeled properly.

In an attempt to correct the outliers in our results, we continue to focus on these corner cases where AIC determined that Beta KS was the best candidate model. One possible starting point (that does not involve expanding the list of candidate models) is to investigate if the removal of the Beta KS distribution from the list of candidates, would have resulted in an overall improvement, in terms of overall mean Δ_{obs}. This is the equivalent of selecting the 2nd best model according to AIC, since we are now only focusing on the specific splits where Beta KS was ranked 1st. We also include the 3rd best model in the comparison, to verify if it is indeed worse than the 2nd best. Figure 3.11a verifies that the 2nd best model is indeed better than the 3rd best, by a significant margin. However, Beta KS does not provide a good fit on average. In fact, surprisingly, both the 2nd best and 3rd best models provide a better fit, with only one exception in the case of ERR@20. Moreover, we see that the 2nd best model measures a Δ_{obs} that is about as good as it could have been ('best-case Δ_{obs}'), without the addition of more models to the current list of candidates. In Figure 3.11b we break down the results based on the alternative model that would have been selected. It appears that, on average, Beta KS is the worst candidate model, except for only Beta models that are ranked 3rd. In summary, these results suggest that the removal of the Beta KS distribution from the list of candidates would have resulted in an overall improvement, in terms of overall mean Δ_{obs}. In Figure 3.12, we visualize this improvement. We see that it is mostly noticeable for the case of nDCG@20 (which is the most frequent case, as per Table 3.2).

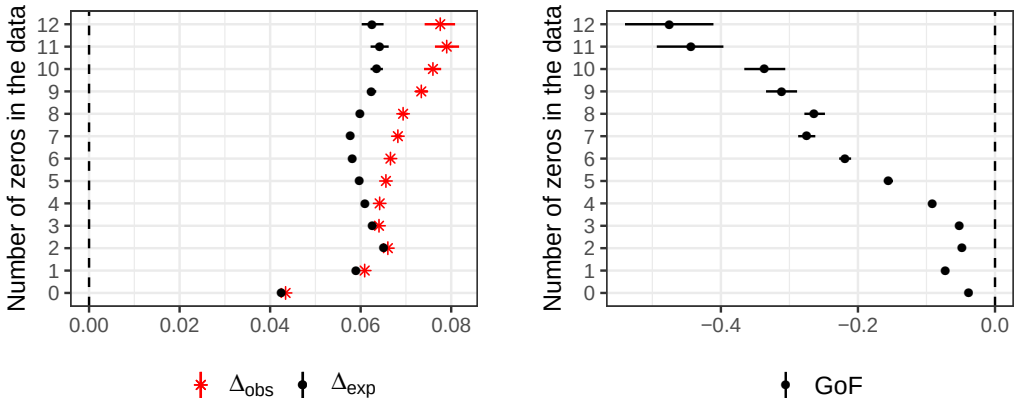
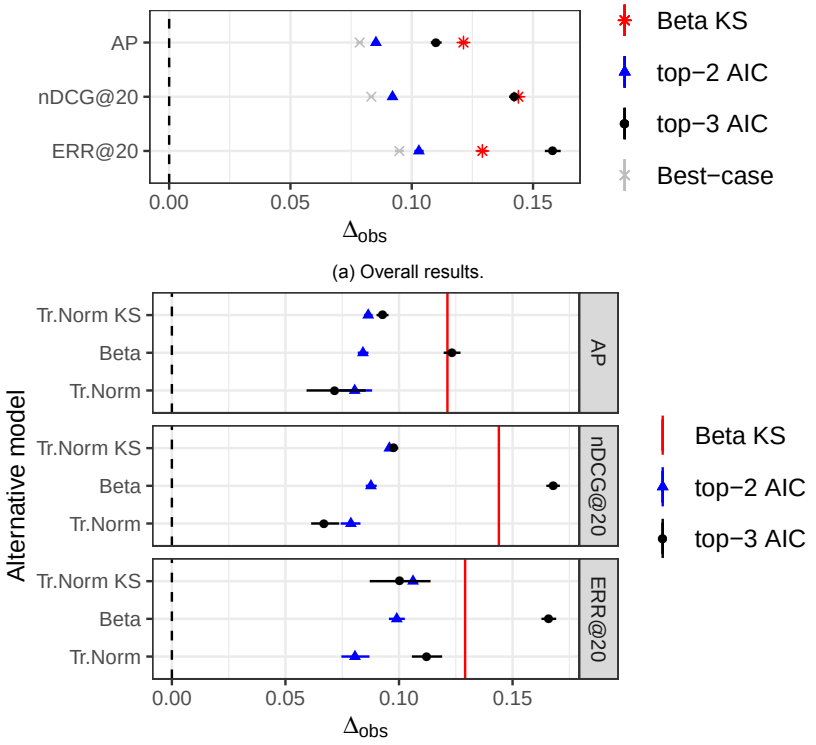


Figure 3.10: Correlation between the appearance of zeros in the data, and goodness-of-fit. Best model was selected according to AIC.



(a) Overall results.

(b) Breakdown of the results based on the alternative candidate model that would have been selected.

Figure 3.11: For those cases where AIC ranked Beta KS 1st, how would the 2nd or 3rd best models perform, if they had been selected?

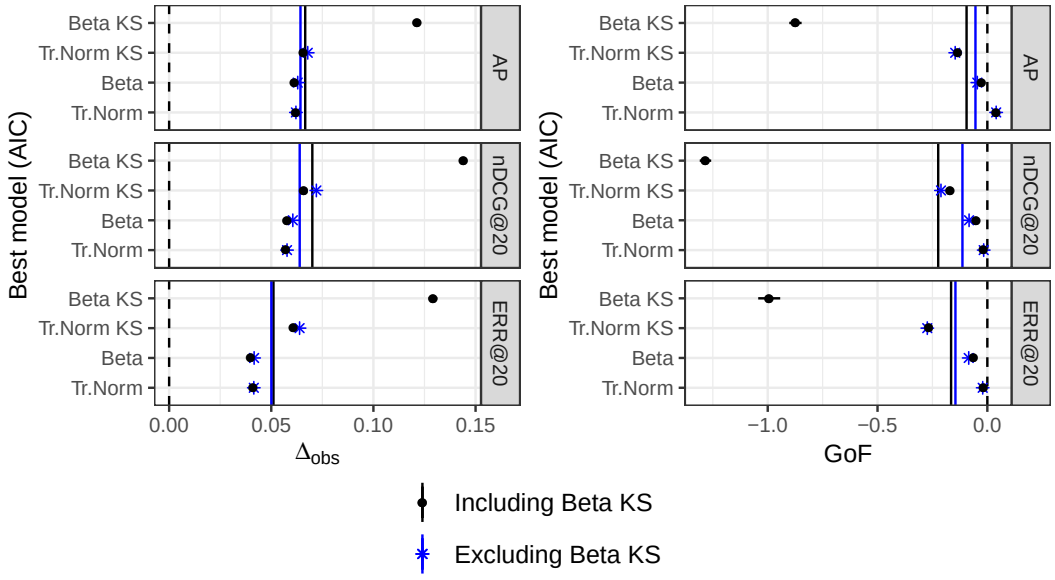


Figure 3.12: How would the goodness-of-fit of the marginal models be affected, if we had removed the Beta KS distribution from the list of candidates? The vertical lines indicate overall means across measures.

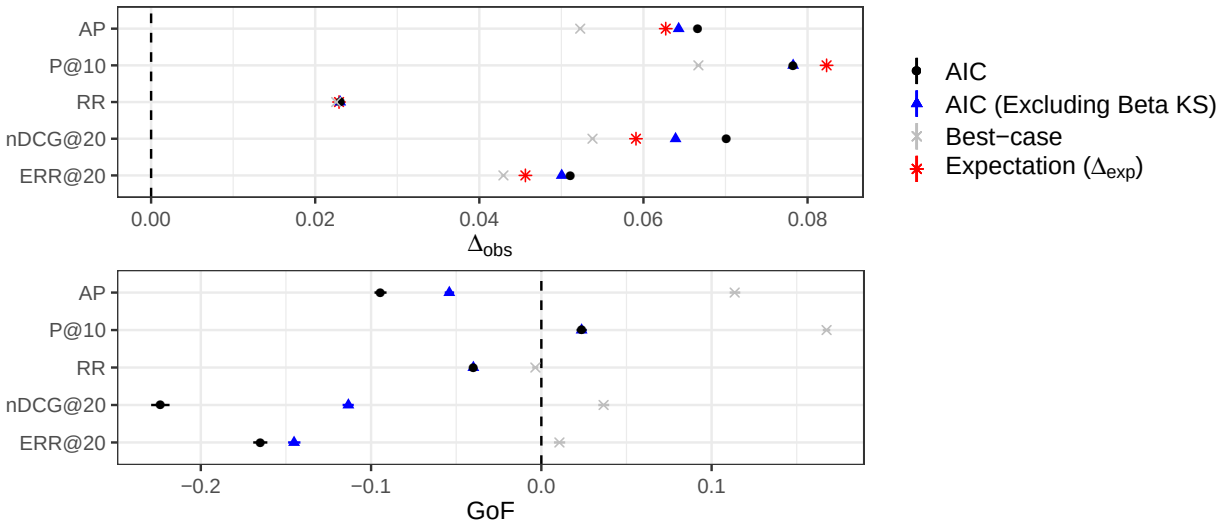


Figure 3.13: The overall mean Δ_{obs} and GoF that is measured, when the models are selected by *i)* AIC, *ii)* AIC with the removal of Beta KS from the list of candidates, and *iii)* optimally, by choosing the model with the lowest Δ_{obs} in each random split.

3.2.1. Comparing Model Selection Criteria

So far we have solely focused on AIC as a model selection criterion, as done in [31]. This criterion provides a balance between simple and underfitted models, and complex and overfitted models. However, we found that this selection is not always satisfactory, as we identified a specific set of corner cases where AIC tends to make poor choices. Moreover, our findings regarding the quality of the margins do suggest that there could be room for improvement. In the top plot of Figure 3.13, we see that our overall Δ_{obs} measurements, even with the improvement that we achieved by removing Beta KS from the list of candidates, is higher than the expectation across all continuous metrics. More specifically, looking at the bottom plot, we see that the Δ_{obs} is about 10% higher than the expectation, for the case of continuous metrics. Ideally we would have hoped for values slightly closer to zero, which means that the quality can be improved. Moreover, we demonstrate that *if* the models had been selected optimally (shown as 'best-case'), the GoF would have been much better. Of course, selecting models in an optimal manner is virtually impossible, however, it does show some potential for improving the quality of the margins, without adding new distributions families to the current list of candidate models. It is therefore meaningful to explore alternative ways of selecting the models that we already have.

In this context, we developed and experimented with a new model selection criterion, which we propose in this thesis. Our proposed criterion, is inspired by the split-half approach, and we denote it as SHC (Split-Half Criterion). In Figure 3.14, we illustrate how this criterion works, through an example. Starting with a set of scores (in this case 25 AP scores), we repeatedly split the data in half, n times. If the original data contain 25 scores, then the first half will contain 12 scores, and the second half will contain the remaining 13. For each split, we fit all possible models on the first half of the data, and then, using the empirical distribution of the second half of the data as ground-truth, we compute a Δ_{obs} for each model. For example, in the first trial split, we observed a delta of 0.21 for Beta KS, and 0.18 for the Truncated Normal distribution respectively. In total, we repeat this process n times. In the end, to compute the SHC of a model that is fitted on the entire set of 25 scores, say Beta KS, we simply average the Δ_{obs} values that we measured for Beta KS across all n splits. For example, for the case of Beta KS, we compute its value like so: $SHC_{\text{Beta KS}} = (0.21 + ... + 0.05)/n$. Essentially, the purpose of each split, is to give us an estimate of what the performance might look like for a given marginal distribution family, on the *entire* set of 25 scores. The more splits we perform, the better our estimate should be. In the end, the best model according to SHC, is the one with the lowest SHC value.

In Figure 3.15, we compare the performance of well established model selection criteria, namely LL, AIC and BIC (Equations 2.4-2.6), in terms of the overall mean GoF that we measured with them. On top of these criteria, we also include our newly proposed criterion, SHC, and set its parameter for the number of splits at $n = 10$. This comparison was performed on the same $250{,}000$ random splits that we previously created. The results for Δ_{obs}, are completely analogous, however we plot GoF values to

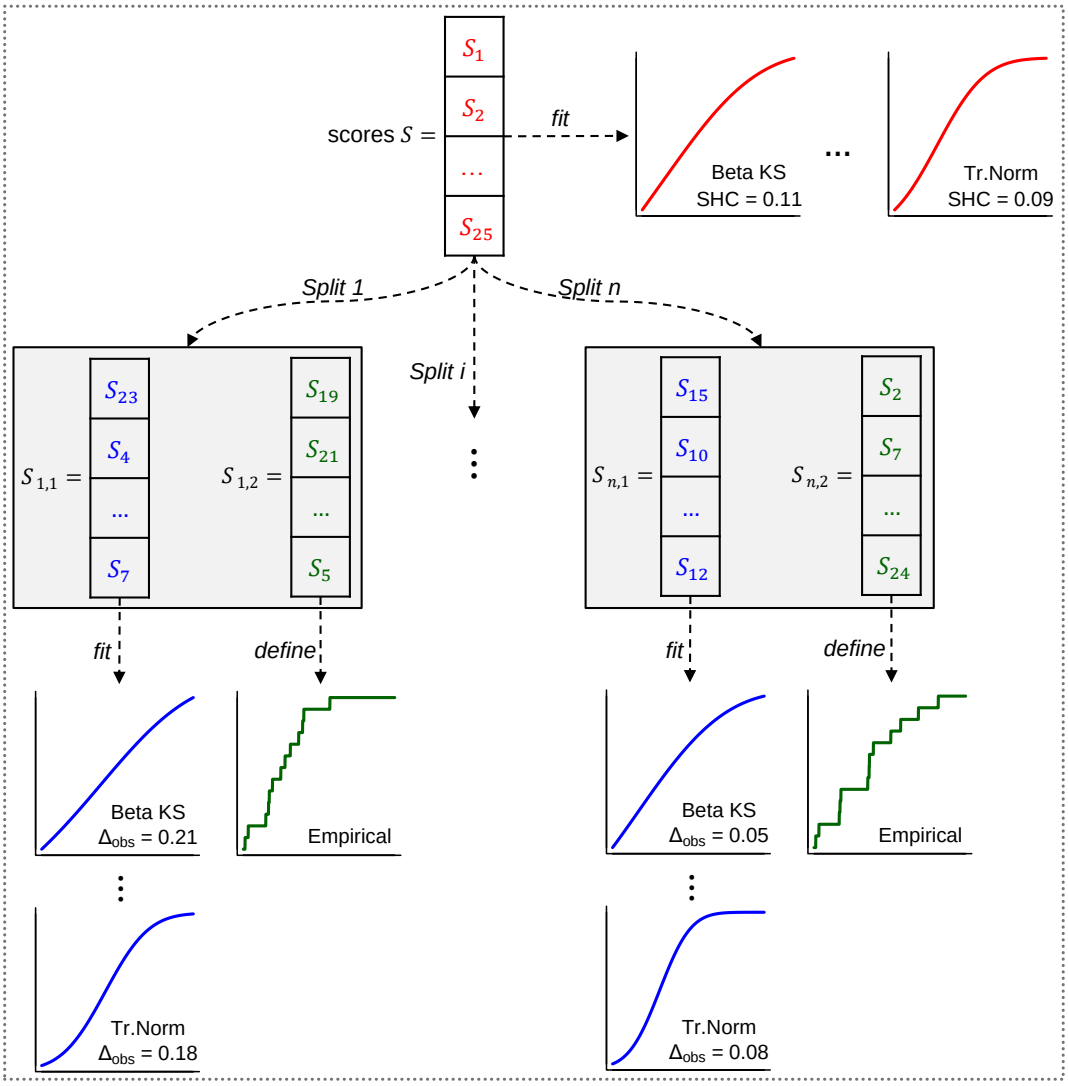
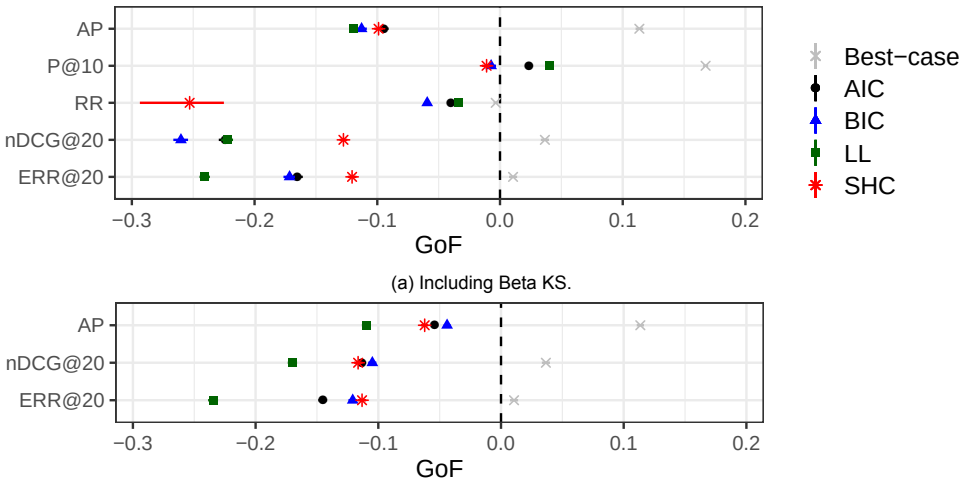


Figure 3.14: Diagrammatic representation of how the Split-Half Criterion (SHC) is computed.



(a) Including Beta KS.

(b) Excluding Beta KS.

Figure 3.15: Comparison of the model selection criteria, in terms of overall mean GoF that is measured.

TREC Track	Years	#Systems	#Topics	Effectiveness Measures
Terabyte	2006	61	149	{AP, RR, P@10}

Table 3.3: Some descriptive statistics about the Terabyte collection.

make the visual interpretation of the results easier, due to the fact that changes in absolute Δ_{obs} can be difficult to distinguish. In the top plot, Beta KS was included in the list of candidates, whereas in the bottom plot it was excluded; meaning, that in any event where some selection criteria ranked Beta KS 1$^{\text{st}}$, we simply selected the 2$^{\text{nd}}$ best model. The exclusion of Beta KS improves the overall mean GoF, regardless of the criterion that was used. In the top plot, we see that for the case of discrete metrics (P@10 and RR) all criteria perform well, with only one exception for SHC in the case of RR. For the case of continuous metrics (AP, nDCG@20 and ERR@20), SHC performs the best overall. In the bottom plot, we see that when Beta KS is excluded (which is something that affects all continuous metrics), all criteria perform very similarly, with the exception of LL. In summary, these results suggests that, the overall goodness-of-fit is maximized by excluding Beta KS from the list of candidates, in combination with using AIC or BIC. Furthermore, we can conclude that, since we did not observe a significant improvement with any of these criteria, to further improve the goodness-of-fit of the margins, beyond the exclusion of Beta KS, would likely require to enrich to the list of candidates with additional models.

3.2.2. Extrapolating Results to Larger Topic Set Sizes

Even though the split-half approach has the advantage that it remains true to the data, it suffers from the fact that the data are being halved. For example, in our case, the main goal is to measure the goodness-of-fit of marginal models that are fitted on the *entire* set of (50) topic-scores, however, in a split-half approach, half of the data are held-out to provide an estimate of the ground truth. This means that we actually measure the goodness-of-fit of models that are fitted on only *half* (25) of those topic-scores. For this reason, we want to know if we underestimate or overestimate the goodness-of-fit, and by how much. In other words, we want to extrapolate our findings to a larger topic set size. To accomplish this, we require data collections that contain more topics. One of those collections is the one from the Terabyte TREC Track of 2006 (Table 3.3), which contains the scores of various systems on 149 topics. The large number of topic-scores that is present in this collection, allows us to repeatedly split the data in three sets of scores, as shown in Figure 3.16. The first two sets are obtained by randomly splitting the entire set of scores in two, so that one split contains 50 scores (S_{50}), and the other split contains the remaining 99 scores (S_{99}). The third set, S_{25}, is obtained by randomly sampling 25 scores from S_{50}. By doing so, we can fit one model (F^*_{25}) on S_{25} and another model (F^*_{50}) on S_{50}. Then, using the empirical distribution of S_{99} as an estimate of the ground truth, we can compute Δ^{25}_{obs} and Δ^{50}_{obs}, respectively, for the two models.

In Figure 3.17 we report the Δ^{25}_{obs} and Δ^{50}_{obs} values that we measured, in $150{,}000$ trials. All models were selected according to AIC. Overall, our results show that when models are fitted on 50 topics, as opposed to 25, they measure a Δ_{obs} that is slightly lower. This suggests that our split-half approach actually tends to underestimate the goodness-of-fit of the marginal models, by a small amount. This occurrence seems to be consistent across all distribution families, however, we notice an outlier, that once again involves the Beta KS distribution. In those particular cases where this distribution is chosen (in the case of 25 topics), the performance is underestimated by a large amount. Interestingly, in this particular dataset, Beta KS is actually never selected in the case of 50 topics; although, this is not the case with other datasets, as we saw earlier in Figure 3.7. Table 3.4 shows that, on average, for the case of AP and P@10, Δ^{50}_{obs} is 7.6% and 2.8% smaller than Δ^{25}_{obs}, respectively. For the case of RR, it is 63.8% larger; however, in absolute terms, this difference is quite negligible, since both Δ^{50}_{obs} and Δ^{25}_{obs} tend to be very low. In summary, these results show that our split-half approach tends to slightly underestimate the goodness-of-fit of the marginal models.

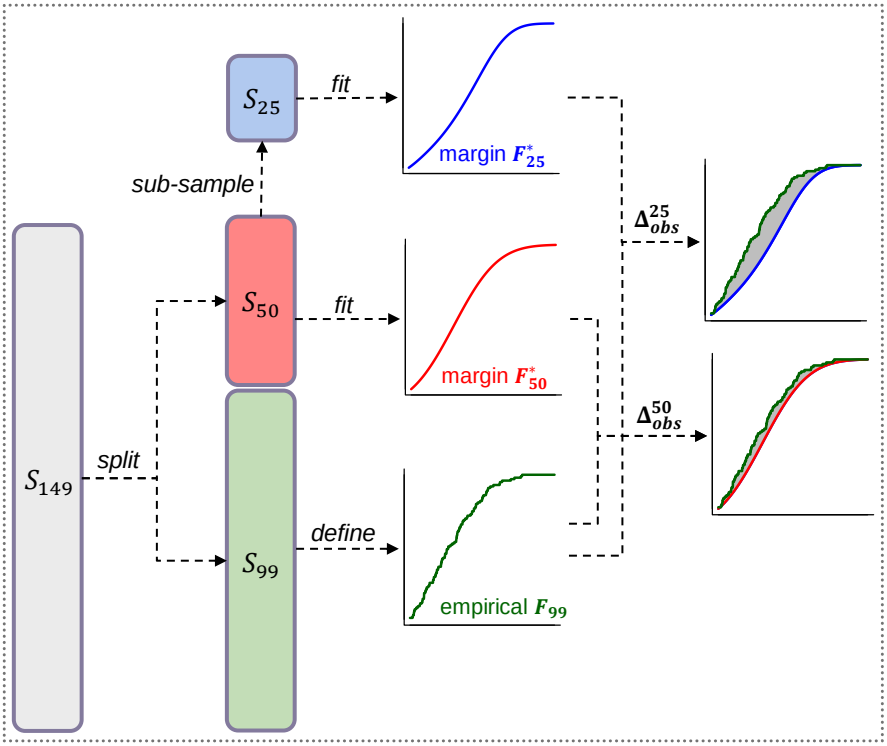


Figure 3.16: Diagrammatic representation of the approach used for computing Δ^{25}_{obs} and Δ^{50}_{obs}.

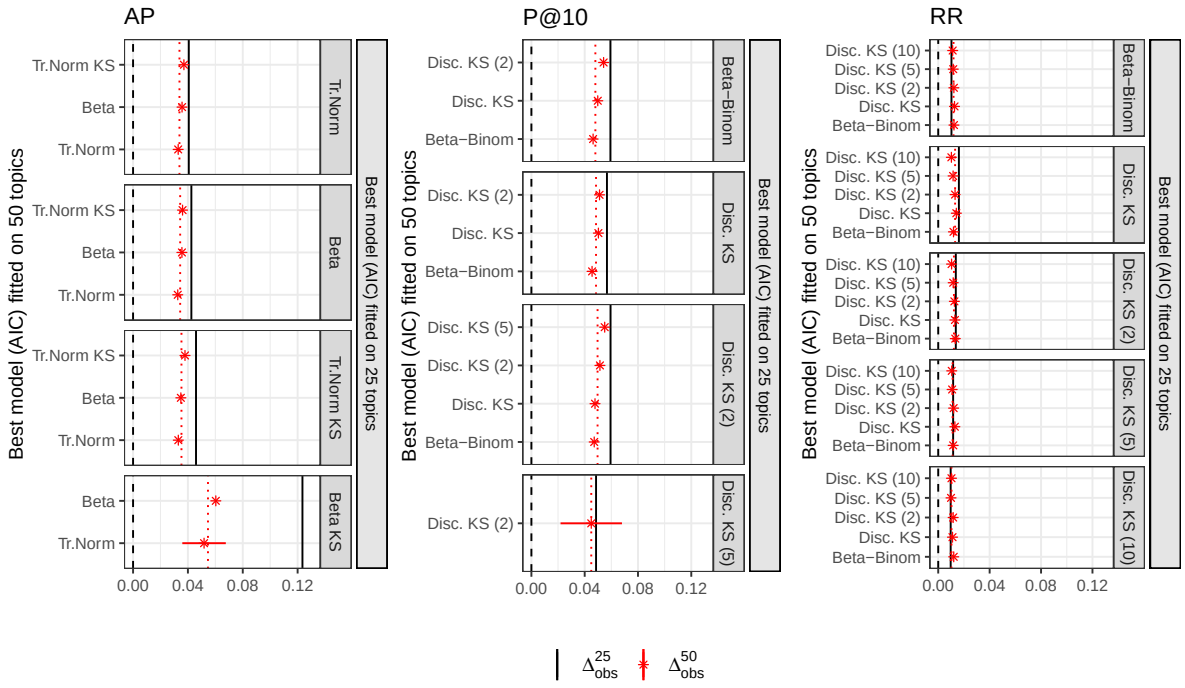


Figure 3.17: How different would our Δ_{obs} measurements be, if the marginal models had been fitted on 50 topics, as opposed to 25?

	$\overline{\left(\frac{\Delta^{50}_{\text{obs}}-\Delta^{25}_{\text{obs}}}{\Delta^{25}_{\text{obs}}}\right)}$
AP	-0.076
P@10	-0.028
RR	0.638

Table 3.4: How different is Δ^{50}_{obs} compared to Δ^{25}_{obs}?

3.2.3. Summary of Results

Summarizing our results, we found that the marginal models fit the data moderately well, when they are selected by AIC, with the exception of the Beta Kernel Smoothing distribution. The goodness-of-fit of the models that are fitted on discrete metrics (P@10 and RR) is noticeably better than those fitted on continuous metrics (AP, nDCG@20 and ERR@20). Also, the models are selected similarly in 25 topics, compared to 50 topics, which gives us some confidence regarding the accuracy of our goodness-of-fit estimates. Moreover, due to the fact that all candidate families get selected, and no particular family gets chosen with a significantly higher frequency than the rest; this implies that IR data are indeed complex.

The Beta Kernel Smoothing distribution is an obvious outlier, that on average, measures a Δ_{obs} twice that of our expectation. We explored this further and found that part of the explanation behind this is the high appearance of zero scores in the data. Although, there are other particularities about those corner case, which we did not manage to detect. It appears that Beta KS models tend to be too simple and underfitted to capture these types of data, yet they are still selected by AIC.

In an attempt to correct this outlier, we found that the exclusion of Beta KS from the list of candidates, notably improves the overall GoF, from -0.22 to -0.11, for the case of nDCG@20, which is the case where Beta KS is selected the most. Interestingly, we discovered that even if the models were selected optimally, none of the candidate families would have performed up to standard, in those particular cases where AIC selected Beta KS. This implies that in order to further improve the quality of the margins in those specific cases, beyond the exclusion of Beta KS, more candidate models would have to be considered. One of such models, could be a mixture model that models the zero scores separately from non-zeros. We leave this for future work.

In an attempt to refine the quality of the margins, instead of focusing on expanding the list of candidate models with additional ones, we experimented with alternative ways of selecting them, including AIC, BIC and LL. We also proposed a new selection criterion that is inspired by the split-half approach, which we denote as SHC (Split-Half Criterion). We found that for the case of continuous metrics (especially nDCG@20), SHC selects models considerable better than the other criteria, although, for the case of discrete metrics (especially RR), it performs considerably worse. However, we found that the best approach for maximizing the overall goodness-of-fit of the margins, is to simply exclude Beta KS from the list of candidates, and select models based on either AIC or BIC. This approach works more consistently across the different effectiveness metrics.

In a separate, smaller scale experiment, we determined that our estimates regarding Δ_{obs} are underestimated by 7.6%, 2.8% for AP and P@10 respectively. For the case of RR we overestimate by 63.8%, however, in absolute terms, it is a negligible value.

4

Copulas

In this chapter, we explore how well the dependence among systems is modeled by the copulas, regardless of their marginal distributions. The methodologies used in this chapter are analogous to Chapter 3. In this thesis, we focus specifically on the case of only two systems, as done in [31]. This means that we deal with the specific case of *bi-variate* copulas.

4.1. Experiments

In order to measure the goodness-of-fit of the copulas, we once again utilize the split-half approach, as shown in Figure 4.1. The approach is slightly different in the case of copulas, compared to the case of the margins, for two reasons. Firstly, the distributions are joint instead of univariate. Joint distributions are visually represented as surfaces instead of curves. For this reason, Δ_{obs} and Δ_{exp} are computed by measuring the volume between surfaces, as opposed to the area between curves. Secondly, the scores need to be converted to pseudoscores, as we previously discussed (see Figure 2.1).

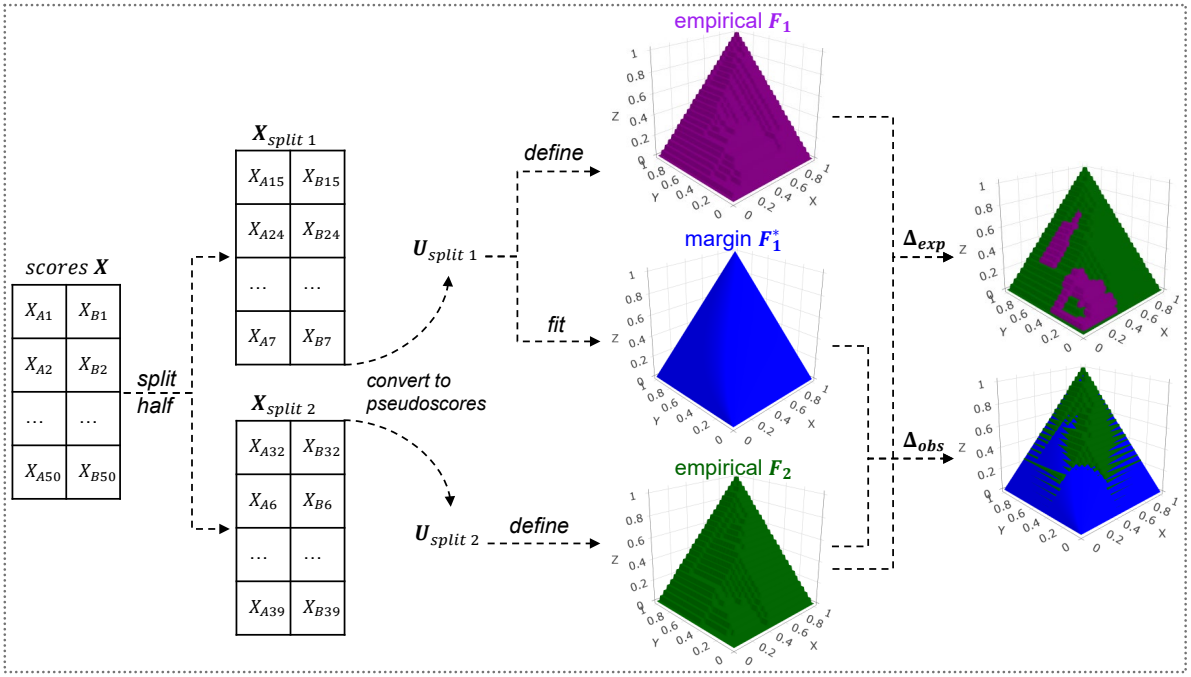


Figure 4.1: Diagrammatic representation of the split-half approach for the case of bi-variate copula models. The Δ_{obs} and Δ_{exp} are computed by measuring the volume between the two surfaces.

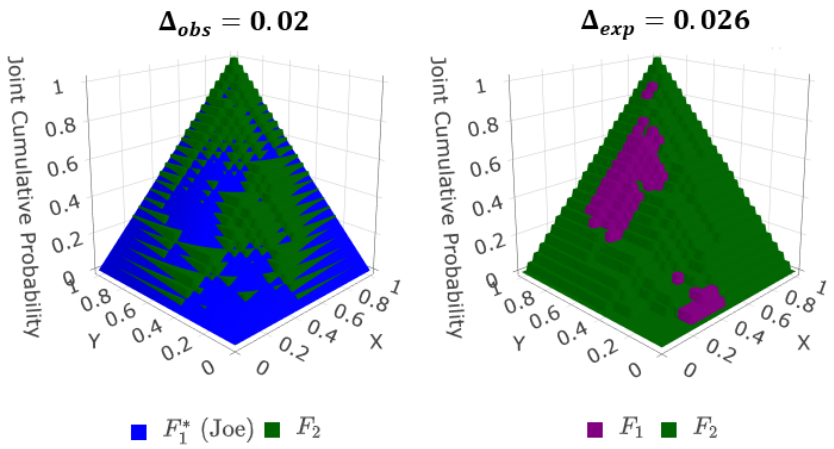


Figure 4.2: Visualization of Δ_{obs} (left) and Δ_{exp} (right) for an example random split that was performed on the AP scores of systems 16 and 43, on the 50 topics of the Ad-hoc 2007 test collection. Each Δ is the volume between two surfaces.

Our primary objective is to measure how well can copulas capture the dependence between two given systems. To this end, for each effectiveness measure, we select two random systems from a random collection and perform the aforementioned split-half approach using the scores of those systems on the entire set of 50 topics. Once again, a total of $250{,}000$ splits were performed; $50{,}000$ for each effectiveness measure. This amount of splits seems to be sufficient, judging from the narrow confidence intervals we obtain for most of our results. The data collections we used are the same as those in the case of the margins (Table 3.1). For each split, we calculate a corresponding Δ_{exp}. Then, using the data in the first half of the split, we fit all copulas (according to Table 2.1), and compute their corresponding Δ_{obs}.

Figure 4.2 shows the Δ_{obs} (left) and Δ_{exp} (right), in one example random split. The approach we use for computing these Δ_{obs} and Δ_{exp} values, is to average the absolute differences between the two surfaces (in the Z-axis), across $100x100$ equally spaced points. In this example, the best copula model according to AIC, was a Joe copula, that measured $\Delta_{obs} = 0.02$. The expected Δ was measured at 0.026.

In Figure 4.3, we report the results across all $250{,}000$ random splits, separately for each effectiveness measure and copula family. All model selections were made according to AIC. In the right plot, we have removed the copulas BB1 and BB6, to make it easier to view the results, due to the fact that those copulas are outliers in terms of their GoF. This occurrence is simply due to chance. In those particular splits, where BB1 or BB6 are selected, the Δ_{exp} values are much lower than usual. This implies that in those particular random splits, the two data halves are much more similar than usual. In those cases, the GoF value is low, due to the division with Δ_{exp}. This phenomenon is mostly inconsequential, because, as we can see in Table 4.1, the copulas BB1 and BB6 are almost never selected. Overall, the copula families measure a Δ_{obs} that is strictly larger than Δ_{exp}, on average. To make it easier to interpret the results, looking at the right plot, we see that GoF is mostly negative, at around -0.23, which implies that, on average, we measure Δ_{obs} values that are about 23% higher than the expectation. This is somewhat higher than we would ideally wish for, although it is still within reason. Moreover, in contrast to the case of the margins, we do not detect any obvious outliers, as the performance appears to be spread somewhat evenly across the different copula models.

Figure 4.4, shows the frequency with which each candidate copula family is selected, when the models are fitted on *i)* 25, as well as *ii)* 50 topics. It appears that the copulas are selected somewhat similarly between these two cases. This adds a degree of confidence regarding the accuracy of our estimates of goodness-of-fit. Moreover, it reaffirms the idea that IR data are quite complex, since all candidate families get selected, and no particular family gets chosen with a disproportionately higher frequency than the rest. In other words, this implies that a variety of copula models is required, to describe IR data.

One possible explanation for these results could be that our list of candidate copulas contains an insufficient selection of asymmetric copula families. In Figure 4.4, we see that the Tawn copula is selected with the most frequency, almost consistently across all metrics. The Tawn copula is actually the

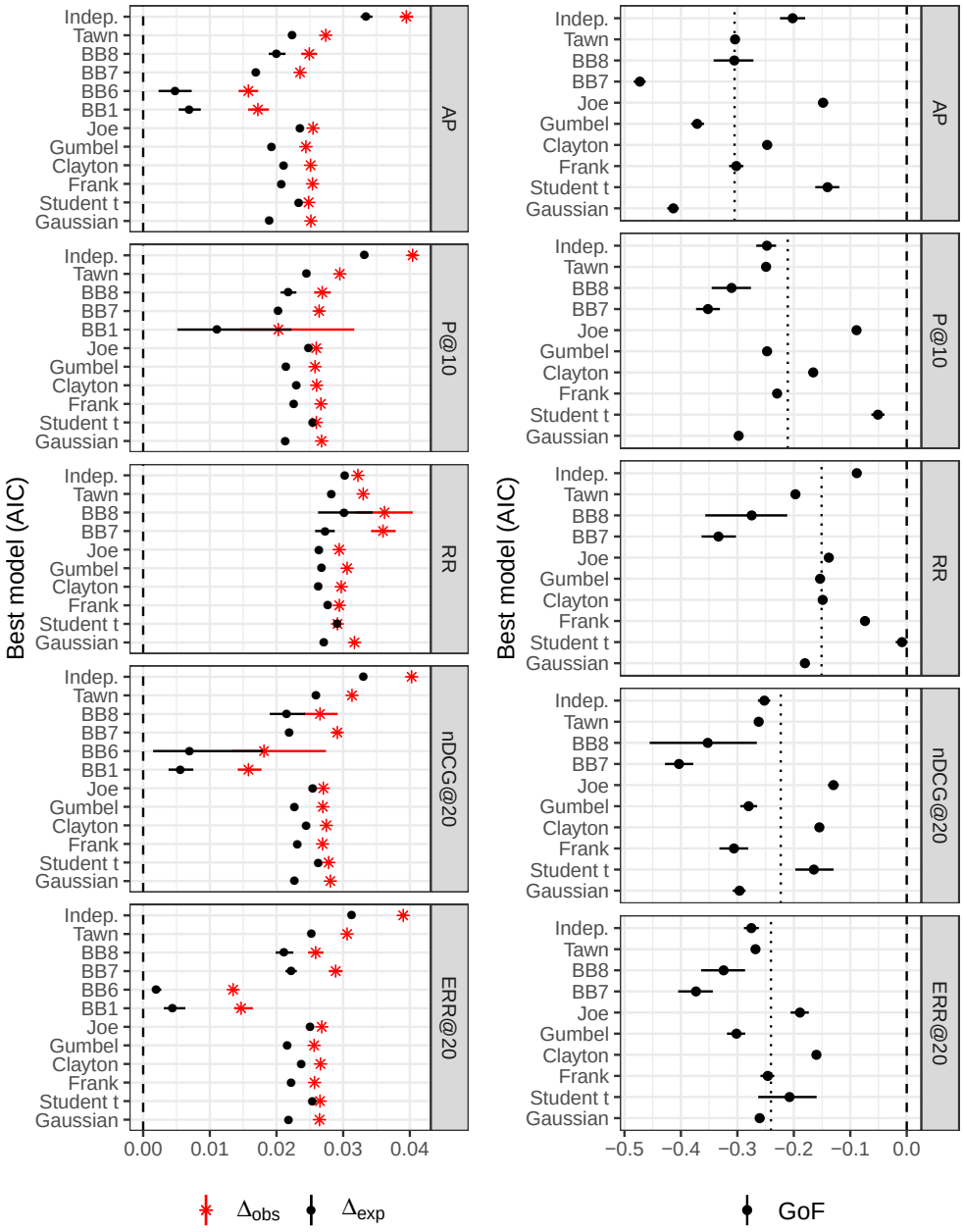


Figure 4.3: How well does each copula family perform, when it is selected by AIC? In the right plot, outliers BB1 and BB6, are excluded from the list of candidates.

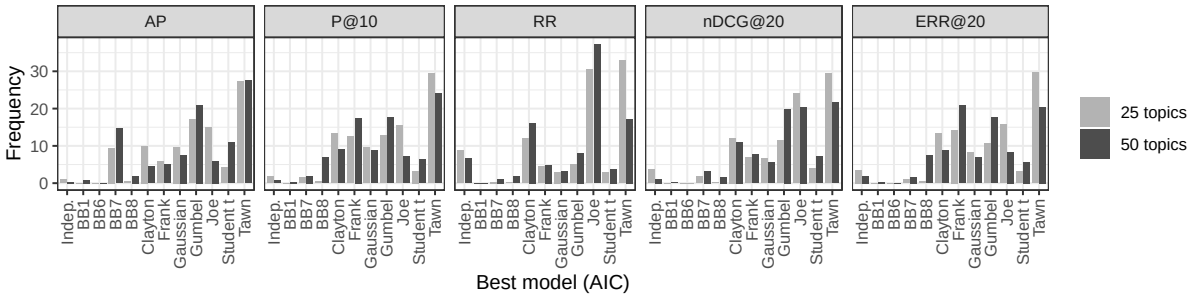


Figure 4.4: Frequency with which each candidate copula family is selected by AIC. We consider the case where models are fitted on 25 and 50 topics respectively.

BB1		BB6	
25 topics	*50 topics*	*25 topics*	*50 topics*
0.04%	0.36%	0.01%	0.02%

Table 4.1: How often is BB1 and BB6 selected by AIC? We consider the case where models are fitted on 25 and 50 topics respectively.

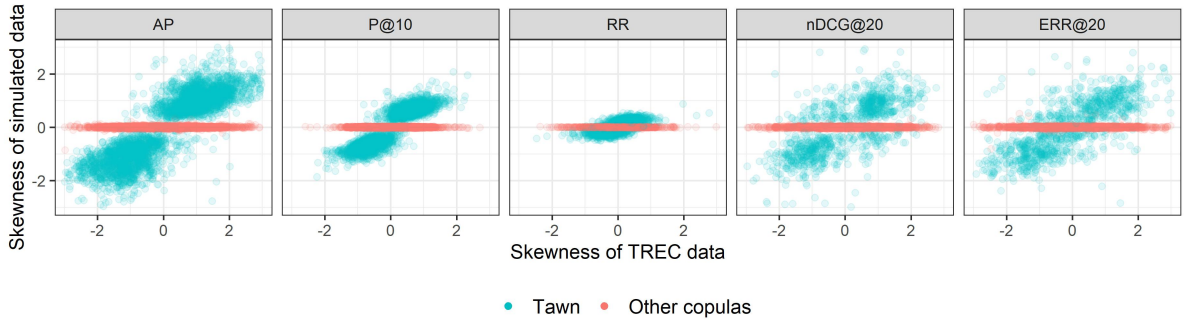


Figure 4.5: (Copy of Figure 5 from [30].) Skewness of the distribution of per-topic score differences, in the case of actual TREC data, compared to simulated data. Only the Tawn copula is able to produce skewed distributions.

only asymmetric copula that is included in our list of candidates. It could be the case that asymmetric copulas capture the dependence among systems well, but there is simply not a diverse enough selection of them in our list of candidates to cover all scenarios. Figure 4.5 illustrates that the distribution of per-topic score differences, in the case of actual TREC data, is often asymmetric. Furthermore, it shows that only the Tawn copula is able to produce skewed distributions. In many cases, the skewness that is present in actual TREC data is not maintained in the simulated data. This supports the idea that the simulation could potentially be improved by expanding the list of candidates with more asymmetric copulas.

4.1.1. Comparing Model Selection Criteria

In view of our findings regarding the lower than expected goodness-of-fit of the copula models, it is meaningful to experiment with alternative ways of selecting models, in an attempt to improve the overall performance. It is possible that some model selection criteria, other than AIC, might be able to select the candidate copulas more optimally.

In Figure 4.6, we compare the performance of LL, AIC, BIC, as well as our proposed criterion SHC, where we set its parameter for the number of splits at $n = 10$. This comparison was performed on the same $250{,}000$ random splits that we previously created. Interestingly, our criterion selects the models more optimally, consistently across all effectiveness measures; in both absolute terms (Δ_{obs}), and terms relative to the expectation (GoF). Although, all criteria perform quite similarly. AIC appears to be the second best criterion, followed by BIC, and finally by LL. Overall, selecting the copulas with SHC compared to AIC, improves the GoF from approximately -0.23 to -0.19. Because a variety of criteria was compared, this implies that in order to further improve the goodness-of-fit by a significant amount, beyond the use of SHC, would likely require to enrich the list of candidates with additional copula families. These results further support the hypothesis we previously proposed, regarding the lack of asymmetric copulas in the list of candidates.

In the left plot of Figure 4.7, we show how SHC selects models differently than AIC. In the right plot, we show how BIC selects models differently than AIC. We convert counts to percentages by averaging per row. We can see that AIC and BIC select models similarly, by looking at the diagonal values, where the percentage of agreement is consistently close to 1. In contrast SHC does appear to select models much differently. It appears to be favoring the Clayton, Joe, BB8 and Tawn copulas. This does show that our criterion is not redundant, and that it differs from other criteria.

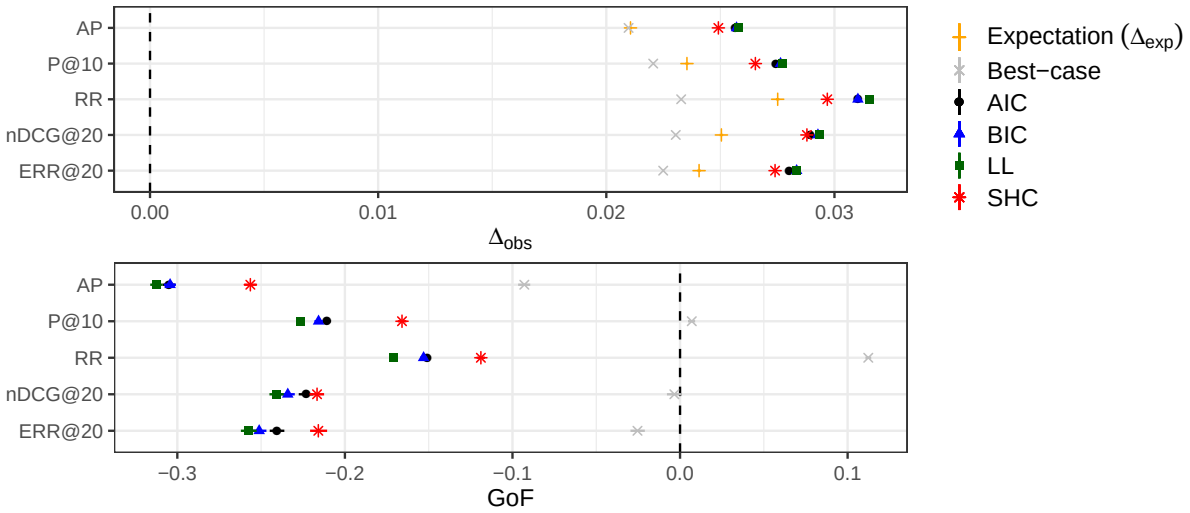


Figure 4.6: Comparison of the model selection criteria, in terms of overall mean Δ_{obs} and GoF that is measured.

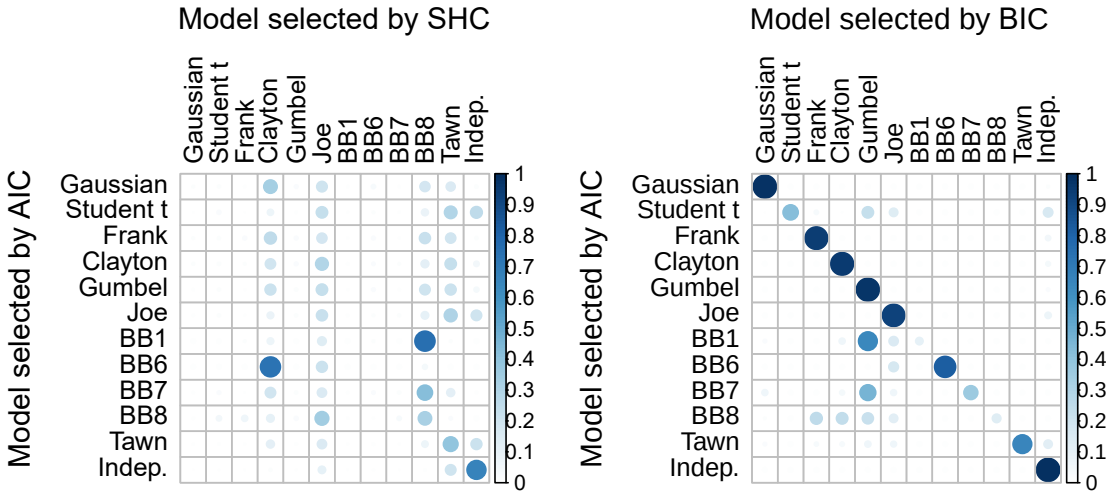


Figure 4.7: Left: How does SHC select models differently than AIC? Right: How does BIC select models differently than AIC? Percentages are computed per row.

4.1.2. Extrapolating Results to Larger Topic Set Sizes

Similar to the case of the margins, we want to know if we underestimate or overestimate the goodness-of-fit of the copulas, and by how much. This is because, due to the inherent limitations of a split-half approach, we fit copulas on only 25 topics, as opposed to 50, since half of the data are used to provide an estimate of the ground truth. We follow an approach that is analogous to the one shown in Figure 3.16 to repeatedly split the data in three sets of topics. A set of 25 topics, a set of 50, and a set of 99 topics. The first two sets are used to fit two copula models respectively: F^*_{25} and F^*_{50}. Using the final set of 99 topics as an estimate of the ground truth, we then compute a Δ^{25}_{obs} and Δ^{50}_{obs}, respectively, for the two copula models.

In Figure 4.8 we report the Δ^{25}_{obs} and Δ^{50}_{obs} values that we measured, in $150{,}000$ trials. All models were selected according to AIC. Overall, our results show that when models are fitted on 50 topics, as opposed to 25, they measure a Δ_{obs} that is slightly lower, which means that the goodness-of-fit of the copula models is underestimated. Moreover, we see that this occurrence is consistent across all copula families. These findings are consistent with the case of the margins.

Table 4.2 shows that, on average, Δ^{50}_{obs} is only 2.6% and 3.9% smaller than Δ^{25}_{obs}, for the case of AP and P@10 respectively. However, for the case of RR it is 10.3% smaller, which is notable.

In summary, these results show that our split-half approach tends to underestimate the goodness-of-fit of the copula models, only by a small amount.

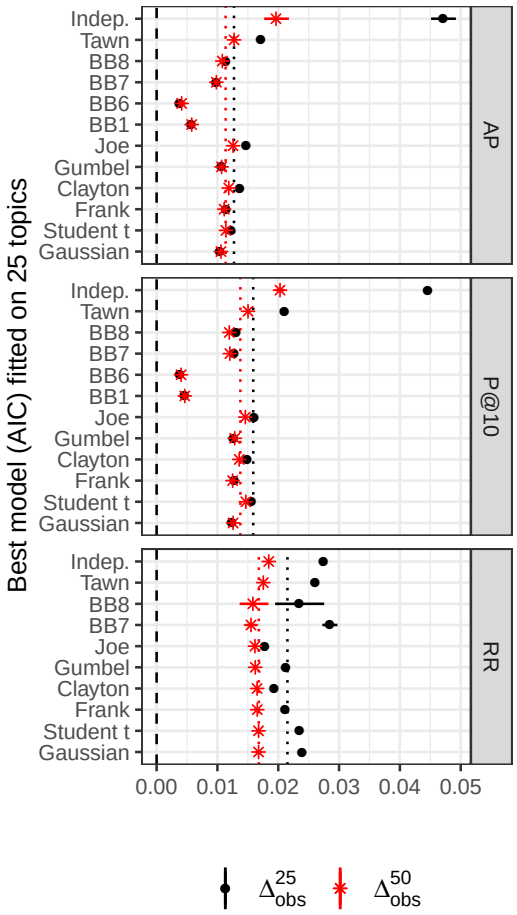


Figure 4.8: How different would our Δ_{obs} measurements be, if the copula models had been fitted on 50 topics, as opposed to 25?.

	$\overline{\left(\frac{\Delta^{50}_{obs}-\Delta^{25}_{obs}}{\Delta^{25}_{obs}}\right)}$
AP	-0.026
P@10	-0.039
RR	-0.103

Table 4.2: How different is Δ^{50}_{obs} compared to Δ^{25}_{obs}?

4.1.3. Summary of Results

Summarizing our results, we found that the copula models measure a Δ_{obs} that is about 23% higher than our expectation, when they are selected by AIC. These results are somewhat worse compared to the case of the margins, but still within reason. We did not detect any obvious outliers.

Looking at the frequency with which each copula family is selected, we see that they are selected quite similarly, between the case of 25 and 50 topics. The Tawn copula gets selected with the highest frequency. Due to the fact that this is the only asymmetric family that is incorporated in the simulation, coupled with the fact that the distribution of per-topic score differences tends to be skewed; we speculate that the goodness-of-fit of the copulas would likely improve, if additional asymmetric families were included in the list of candidates. However, we leave this for future work.

Looking for ways of refining the quality of the models, we experimented with alternative ways of selecting them, including AIC, BIC, LL as well as our criterion (SHC), and found that our criterion improves the overall mean GoF, consistently across all effectiveness measures. The overall change over AIC (the next best criterion) is from around -0.23 to -0.19, which is notable. Moreover, the selections that SHC makes are significantly different from those of AIC, BIC or LL. Since a large selection of

criteria was considered, it is likely that to further improve the goodness-of-fit, would require additional candidate copulas to be considered.

In a separate, smaller scale experiment, we determined that our estimates regarding Δ_{obs} are underestimated by 2.6%, 3.9% and 10.3% for AP, P@10 and RR respectively, which is quite low.

5

Conclusion

In this thesis, we made a first attempt at providing empirical evidence regarding the quality of the stochastic simulation used in [31]. This particular simulation approach, uses existing collections of system scores, to build a model for the joint distribution of system scores on topics, which can then be used to endlessly simulate scores by the same systems, but on random new topics. The simulation is based on two separate components: one *marginal* model for each system, which models the individual distribution of scores of the system, and a *copula* that models the dependence among systems. This allows us to study the marginal models and the copulas separately.

Measuring the quality of a stochastic simulation is not a trivial or one dimensional task. We focus on one, but highly important aspect of simulation, which is the goodness-of-fit of the models; meaning, how well do the models describe the data. Ideally, this would be measured by computing some similarity metric between a model F^* and the true distribution F. However, having knowledge of the true distribution of a system's scores is not feasible.

For this reason, we resort to the split-half approach, which works by repeatedly splitting the data in half, and treating the first half as *the sample*, and the second half as *the population*. Using the first half of the data to fit a model (F_1^*), and the empirical distribution of the second half (F_2) as an estimate of the ground truth, we can compute a measure of distance Δ between F_1^* and F_2. The goodness-of-fit can be defined as the negative of that Δ. In addition, we devised a method for calculating a reference value which we should approximately expect from a good fit, by computing the Δ between the two empirical distributions of the data. We performed a total of $250{,}000$ random splits for the margins and the copulas respectively, using a variety of IR data.

For the case of the margins, our results show that the models fit the data moderately well, when they are selected by AIC, with the exception of the Beta Kernel Smoothing distribution, which is an outlier. We explored this outlier further, and discovered that part of the explanation behind this is the high appearance of zero scores in the data. It appears that Beta KS models tend to be too simple and underfitted to capture those types of data, yet they are still selected by AIC. We found that the exclusion of Beta KS from the list of candidates, notably improves the overall goodness-of-fit, but not as much as we would have hoped. In fact, we discovered that none of the considered candidate families would have performed well enough in those particular corner cases, even if they were selected optimally. This implies that in order to further improve the quality of the margins in those specific cases, beyond the exclusion of Beta KS, more candidate models would need to be considered. One of such models, could be a mixture model that models the zero scores separately from non-zeros. We leave this for future work.

In view of the fact that there is room for improvement with regards to the quality of the margins, we shift our focus on refining it. In this work, instead of focusing on expanding the list of candidate models with additional ones, we chose to focus on how to select them in a more optimal manner. Our motivation behind this is the fact that we previously identified some cases where AIC tends to make poor choices, as well as the fact that the list of candidates is quite diverse as is; including a variety of both parametric and non-parametric distributions. We proposed a new selection criterion that is inspired by the split-half approach, which we denote as SHC (Split-Half Criterion), and also considered some other well established model selection criteria beyond AIC, such as BIC and LL. We found that for the

case of continuous metrics (especially nDCG@20), SHC selects models considerable better than the rest criteria; although, for the case of discrete metrics (especially RR), it performs considerably worse. However, we found that the best approach for maximizing the overall goodness-of-fit of the margins, is to simply exclude Beta KS from the list of candidates, and select models based on either AIC or BIC. This approach works more consistently across the different effectiveness metrics.

For the case of the copulas, our results show that the models fit the data somewhat worse than the margins, but still within reason, when they are selected by AIC. We experimented with other criteria, such as BIC, LL as well as our proposed criterion. Interesting, we found that our criterion provides the best overall goodness-of-fit, consistently across all effectiveness measures. Although, the improvement over AIC is relatively small. To further improve the performance of the copulas, more candidate copula families would need to be considered. Due to the fact that the copula which is selected with the highest frequency (Tawn copula) happens to be the only asymmetric copula in the list of candidates, we speculate that the inclusion of more asymmetric copulas in the simulation, may improve the quality. We also found some evidence that support this hypothesis. This is something we do not experiment with however, leaving it for future work.

Due to the inherent limitation of the split-half approach, of halving the data; we can only measure the goodness-of-fit of models that are fitted on half the data. However, we are interested on models that are fitted on the entire set of data. For this reason, we also investigated if we underestimate or overestimate goodness-of-fit, and by how much. This was explored using a data collection that contains a larger set of topics, which allowed us to extrapolate our findings to larger topic set sizes. We found that we underestimate the performance of the margins as well as the copulas, but only slightly. Another consequence of halving the data, is that the candidate margin and copula distribution families, are not selected with the same frequency when the models are fitted on half the data, compared to the whole data. We explored this matter as well, and found that the differences are fairly minor. These findings suggest that our goodness-of-fit estimates should be quite accurate.

One highly important implication of our findings, is that due to the fact that both the marginal models, and the copula models (to a lesser extent), describe the data moderately well, this adds a high degree of reliability in the findings of Urbano et al. in [31]. More specifically, the conclusions reached, regarding the t-test and the permutation tests being the most optimal, and the sign, Wilcoxon and bootstrap-shift being the least optimal, for IR evaluation data.

At the same time, due to the fact that the concerns of Parapar et al. with regards to the quality of the simulation used by Urbano et al. have largely been addressed in this thesis, the question remains as to precisely why are their conclusions not in accordance. This requires further investigation. As suggested in [30], an important direction for future work is to compare the two simulation approaches, and their findings, in a controlled setting. This is important, not only for helping us determine which statistical significance test is optimal in IR and when, but also, for deepening our knowledge with regards to the properties of IR evaluation data.

Bibliography

[1] H. Akaike. “A New Look at the Statistical Model Identification”. In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. issn: 0018-9286. doi: 10.1109/TAC.1974.1100705.

[2] T. W. Anderson and D. A. Darling. “Asymptotic Theory of Certain ”Goodness of Fit” Criteria Based on Stochastic Processes”. In: *The Annals of Mathematical Statistics* 23.2 (1952), pp. 193–212. issn: 0003-4851. doi: 10.1214/aoms/1177729437.

[3] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. “Building Simulated Queries for Known-Item Topics”. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’07. New York, New York, USA: ACM Press, 2007, pp. 455–462. isbn: 9781595935977. doi: 10.1145/1277741.1277820.

[4] Leif Azzopardi et al. “Report on the SIGIR 2010 Workshop on the Simulation of Interaction”. In: *ACM SIGIR Forum* 44.2 (2011), pp. 35–47. issn: 0163-5840. doi: 10.1145/1924475.1924484.

[5] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. “Modeling behavioral factors ininteractive information retrieval”. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. CIKM ’13. New York, New York, USA: ACM Press, 2013, pp. 2297–2302. isbn: 9781450322638. doi: 10.1145/2505515.2505660.

[6] David Bodoff and Pu Li. “Test Theory for Assessing IR Test Collections”. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’07 July. New York, New York, USA: ACM Press, 2007, pp. 367–374. isbn: 9781595935977. doi: 10.1145/1277741.1277805.

[7] Ben Carterette. “Bayesian Inference for Information Retrieval Evaluation”. In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ICTIR ’15. New York, NY, USA: ACM, 2015, pp. 31–40. isbn: 9781450338332. doi: 10.1145/2808194.2809469.

[8] Ben Carterette. “But Is It Statistically Significant?: Statistical Significance in IR Research, 1995-2014”. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’17. New York, NY, USA: ACM Press, 2017, pp. 1125–1128. isbn: 9781450350228. doi: 10.1145/3077136.3080738.

[9] Ben Carterette et al. “If I Had a Million Queries”. In: *Proceedings of the 31st European Conference on IR Research*. ECIR 2009. 2009, pp. 288–300. doi: 10.1007/978-3-642-00958-7_27.

[10] Benjamin A. Carterette. “Multiple testing in statistical analysis of systems-based information retrieval experiments”. In: *ACM Transactions on Information Systems* 30.1 (2012), pp. 1–34. issn: 1046-8188. doi: 10.1145/2094072.2094076.

[11] Michael D Cooper. “A simulation model of an information retrieval system”. In: *Information Storage and Retrieval* 9.1 (1973), pp. 13–32. issn: 00200271. doi: 10.1016/0020-0271(73)90004-1.

[12] Gordon V. Cormack and Thomas R. Lynam. “Validity and Power of t-Test for Comparing MAP and GMAP”. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’07. New York, NY, USA: ACM Press, 2007, p. 753. isbn: 9781595935977. doi: 10.1145/1277741.1277892.

[13] David Hull. “Using Statistical Testing in the Evaluation of Retrieval Experiments”. In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’93. New York, NY, USA: ACM Press, 1993, pp. 329–338. isbn: 0897916050. doi: 10.1145/160688.160758.

[14] A N Kolmogorov. “Sulla determinazione empirica di una legge di distribuzione”. In: *Giorn. Inst. Italiano Attuari* 4 (1933), pp. 83–91.

[15] R. Manmatha, T. Rath, and F. Feng. “Modeling Score Distributions for Combining the Outputs of Search Engines”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR ’01. New York, New York, USA: ACM Press, 2001, pp. 267–275. isbn: 1581133316. doi: 10.1145/383952.384005.

[16] David Maxwell and Leif Azzopardi. “Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM ’16. New York, NY, USA: ACM, 2016, pp. 731–740. isbn: 9781450340731. doi: 10.1145/2983323.2983805.

[17] Richard von Mises. *Wahrscheinlichkeit Statistik und Wahrheit*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1928. isbn: 978-3-662-35402-5. doi: 10.1007/978-3-662-36230-3.

[18] Javier Parapar, David E. Losada, and Álvaro Barreiro. “Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation”. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC ’21. New York, NY, USA: ACM Press, 2021, pp. 655–664. isbn: 9781450381048. doi: 10.1145/3412841.3441945.

[19] Javier Parapar et al. “Using Score Distributions to Compare Statistical Significance Tests for Information Retrieval Evaluation”. In: *Journal of the Association for Information Science and Technology* 71.1 (2020), pp. 98–113. issn: 2330-1635. doi: 10.1002/asi.24203.

[20] C. J. van Rijsbergen. *Information Retrieval*. 2nd. Butterworth-Heinemann, 1979. isbn: 9781538633748. doi: 10.5555/539927.

[21] Stephen E. Robertson and Evangelos Kanoulas. “On per-topic variance in IR evaluation”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR ’12. New York, New York, USA: ACM Press, 2012, pp. 891–900. isbn: 9781450314725. doi: 10.1145/2348283.2348402.

[22] Tetsuya Sakai. “Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. SIGIR ’16. New York, NY, USA: ACM Press, 2016, pp. 5–14. isbn: 9781450340694. doi: 10.1145/2911451.2911492.

[23] Mark Sanderson and Justin Zobel. “Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability”. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’05 1979. New York, NY, USA: ACM Press, 2005, pp. 162–169. isbn: 1595930345. doi: 10.1145/1076034.1076064.

[24] Gideon Schwarz. “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2 (1978). issn: 0090-5364. doi: 10.1214/aos/1176344136.

[25] Abe Sklar. *Fonctions de Répartition à n Dimensions et Leurs Marges*. Vol. 8. Publications de l’Institut de statistique de l’Université de Paris, 1959, pp. 229–231.

[26] N. Smirnoff. “Sur les écarts de la courbe de distribution empirique”. In: *Matematicheskii Sbornik* 48.1 (1939), pp. 3–26.

[27] Mark D. Smucker, James Allan, and Ben Carterette. “A Comparison of Statistical Significance Tests for Information Retrieval Evaluation”. In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM ’07. New York, NY, USA: ACM Press, 2007, pp. 623–632. isbn: 9781595938039. doi: 10.1145/1321440.1321528.

[28] Jean Tague, Michael Nelson, and Harry Wu. “Problems in the Simulation of Bibliographic Retrieval Systems”. In: *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*. SIGIR ’80. Butterworth & Co., 1980, pp. 236–255. isbn: 0408107758. doi: 10.5555/636669.636684.

[29] Julián Urbano. “Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation”. In: *Information Retrieval Journal* 19.3 (2016), pp. 313–350. issn: 1573-7659. doi: 10.1007/s10791-015-9274-y.

[30] Julián Urbano, Matteo Corsi, and Alan Hanjalic. “How do Metric Score Distributions affect the Type I Error Rate of Statistical Significance Tests in Information Retrieval?” In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR ’21. New York, NY, USA: ACM Press, 2021, pp. 245–250. isbn: 9781450386111. doi: 10.1145/3471158.3472242.

[31] Julián Urbano, Harlley Lima, and Alan Hanjalic. “Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’19. New York, NY, USA: ACM Press, 2019, pp. 505–514. isbn: 0408709294. doi: 10.5555/539927. arXiv: 1905.11096.

[32] Julián Urbano, Mónica Marrero, and Diego Martín. “A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation”. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR ’13. New York, NY, USA: ACM Press, 2013, pp. 925–928. isbn: 9781450320344. doi: 10.1145/2484028.2484163.

[33] Julián Urbano and Thomas Nagler. “Stochastic Simulation of Test Collections: Evaluation Scores”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR ’18. New York, NY, USA, 2018, pp. 695–704. isbn: 9781450356572. doi: 10.1145/3209978.3210043.

[34] Ellen M. Voorhees and Chris Buckley. “The Effect of Topic Set Size on Retrieval Experiment Error”. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’02. New York, New York, USA: ACM Press, 2002, pp. 316–323. isbn: 1581135610. doi: 10.1145/564376.564432.

[35] Ryen W. White. “Using searcher simulations to redesign a polyrepresentative implicit feedback interface”. In: *Information Processing & Management* 42.5 (2006), pp. 1185–1202. issn: 03064573. doi: 10.1016/j.ipm.2006.02.005.

[36] Ryen W. White et al. “Evaluating implicit feedback models using searcher simulations”. In: *ACM Transactions on Information Systems* 23.3 (2005), pp. 325–361. issn: 1046-8188. doi: 10.1145/1080343.1080347.

[37] W. John Wilbur. “Non-parametric significance tests of retrieval performance comparisons”. In: *Journal of Information Science* 20.4 (1994), pp. 270–284. issn: 0165-5515. doi: 10.1177/016555159402000405.

[38] Justin Zobel. “How Reliable are the Results of Large-Scale Information Retrieval Experiments?” In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’98. New York, NY, USA: ACM Press, 1998, pp. 307–314. isbn: 1581130155. doi: 10.1145/290941.291014.