Stochastic Simulation of Test Collections: Evaluation Scores

Julián Urbano
Delft University of Technology
The Netherlands
urbano.julian@gmail.com

Thomas Nagler
Technical University of Munich
Germany
thomas.nagler@tum.de

ABSTRACT

Part of Information Retrieval evaluation research is limited by the fact that we do not know the distributions of system effectiveness over the populations of topics and, by extension, their true mean scores. The workaround usually consists in resampling topics from an existing collection and approximating the statistics of interest with the observations made between random subsamples, as if one represented the population and the other a random sample. However, this methodology is clearly limited by the availability of data, the impossibility to control the properties of these data, and the fact that we do not really measure what we intend to. To overcome these limitations, we propose a method based on vine copulas for stochastic simulation of evaluation results where the true system distributions are known upfront. In the basic use case, it takes the scores from an existing collection to build a semi-parametric model representing the set of systems and the population of topics, which can then be used to make realistic simulations of the scores by the same systems but on random new topics. Our ability to simulate this kind of data not only eliminates the current limitations, but also offers new opportunities for research. As an example, we show the benefits of this approach in two sample applications replicating typical experiments found in the literature. We provide a full R package to simulate new data following the proposed method, which can also be used to fully reproduce the results in this paper.

KEYWORDS

Evaluation, Test Collection, Simulation, Distribution, Copula

ACM Reference Format:
Julián Urbano and Thomas Nagler. 2018. Stochastic Simulation of Test Collections: Evaluation Scores. In *Proceedings of The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210043

1 INTRODUCTION

Much research on Information Retrieval (IR) investigates alternative methods to better evaluate systems. Some of these seek a higher correlation between offline system measures and online user measures, more power to discriminative between systems, or reduction of judgment pool incompleteness. In this type of research we typically ask "what if" questions, like what if we use non-expert relevance assessors Bailey et al. [5], what if we use a different parameterization of *DCG* [20], or what if we evaluate on a different document collection [24]?

One class of such problems is concerned with the reliability of our evaluation experiments and its trade-off with efficiency [29], that is, the extent to which evaluation results can be replicated, maybe under budget restrictions. Typical questions in this class are what if we change the topic set [37], what if we use *P*@10 instead of *AP* [8], what if we use the Wilcoxon test instead of the *t*-test [32], or how many topics will we need to achieve a certain level of confidence [12]? Unfortunately, research questions of this nature pose fundamental problems that make it impossible to find a direct answer or, at the very least, limit our ability to do so:

- Finite data. Evaluation research is often of empirical nature and uses existing data from archives like TREC. However, these data are limited to dozens of systems and topics for a given task, so the precision and generalizability of our results are severely constrained. We usually overcome this limitation by resampling the existing data, as a way to simulate new topic sets.
- Unknown truth. In many cases the researcher needs to know some underlying property of the systems, such as their true mean score and variance over the possibly infinite population of topics, which are of course impossible to know. The workaround usually consists in making random splits of the topic set, considering one as representing the population of topics and the other one as a random sample from it. However, and because of the limited amount of data, these splits are not really independent samples.
- Lack of control. Very often we want to study systems of predefined characteristics, such as systems with the same mean or variance, or with a certain degree of dependence. But we can not control these properties: the systems and topics in the existing data are what they are; we can not change them. Sometimes we work around this limitation with artificial modifications of the effectiveness scores, but they result in unrealistic data (eg. shifting scores by some quantity).

Consider for instance the problem of choosing an appropriate number of topics for a new test collection [30]. The IR literature contains data-based approaches that repeatedly split the topic sets to calculate some statistic like Kendall τ [8, 25, 40]. Extrapolating to larger topic sets, we find empirical results on the reliability of various topic set sizes. However, these studies are clearly limited by the small data sources, they calculate statistics between two samples of topics as opposed to between a sample and a population, and in the end we do not have full knowledge of the true system distributions anyway, so we can not really assess whether the extrapolation is accurate or not. There are also statistical approaches that make various assumptions like normal distributions or homoskedasticity, which clearly do not hold in IR evaluation data [12, 23, 31]. Still, the extent to which these assumptions pose

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210043

a practical problem, remains largely unknown because we can not control these properties of the data.

These limitations, however unfortunate, are present in our everyday research. As a consequence, one may wonder the extent to which past and current results really hold in practice. It is our firm believe that the IR community should seek experimentation methods that help us remove these barriers and allow us to study this kind of questions directly and under desired conditions. For this, we propose the use of stochastic simulation for the generation of evaluation data to serve as input in evaluation research. The idea of using simulation in IR research is of course not new. In the early days simulation was attempted to build collections, and in particular simulate judgments [14, 28]. More recent work has been devoted for instance to simulating queries [3], document scores [22] and various aspects of user interaction [4].

For problems pertaining to reliability we find few cases, such as [9, 10] and [30], which do simulate effectiveness scores. However, simulations therein are rather limited. In the former the scores are drawn from Beta and Uniform distributions without adjusting parameters and correlations based on existing data. The latter work does this to some degree, but the resulting simulations are still unrealistic in terms of support (eg. they are only continuous) and still do not allow us to have full control over system properties.

Building upon these works, in this paper we propose a method for the stochastic simulation of effectiveness scores that effectively avoids the three limitations discussed above. The general idea is to build a model for the joint distribution of system scores given by some effectiveness measure, such that we can simulate endlessly from it. It is important to note that we do not aim at creating a model of *the* systems that generated the given data, but rather a realistic model of *a* set of systems similar to those. For this model to be realistic, we implement several alternatives to model the marginal system distributions in a way that the underlying properties of the measure are preserved (eg. the support), as well as several alternatives to model the full dependence structure among systems. By fully specifying the effectiveness distributions, the model is also useful because we have complete knowledge of properties of the data such as the expected mean and variance. Furthermore, by separating the modeling of margins from the modeling of dependencies, we have a high level of customization that allows us to control aspects such as the levels of homoskedasticity and correlation. A full implementation of the proposed method is open-sourced as an R package, available at `https://github.com/julian-urbano/simIReff`. The results of the paper can be fully reproduced with data and code available at `https://github.com/julian-urbano/sigir2018-simulation`.

Sections 2 and 3 discuss how to model system dependencies and score distributions. Section 4 evaluates the proposed method, and Section 5 presents two application uses cases replicating past experiments. Section 6 presents the conclusions and future work.

2 MODELING SYSTEM DEPENDENCIES

In order to make realistic simulations, we need to build a joint stochastic model for the effectiveness of a set of systems. An appropriate model should reflect the behavior of the individual system

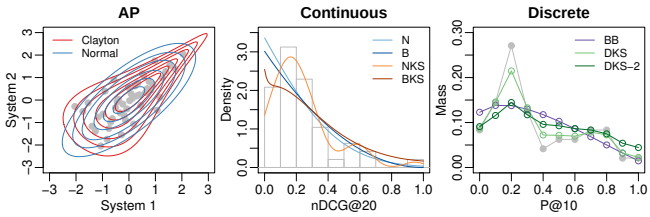

Figure 1: Sample bivariate copulas fitted to the *AP* scores of two TREC 2010 Web Ad hoc systems, and sample distributions of *nDCG*@20 and *P*@10 scores. Original data in grey.

scores as well as the dependence among them. But classical multivariate models like the multivariate Gaussian distribution are not flexible enough to describe the kind of data we have in IR evaluation (see Figure 1 for examples). A solution to this problem are copula models [19], which allow us to separate the modeling of marginal distributions (i.e. the individual distribution of each system, regardless of the others), and the dependence among systems.

Denote the effectiveness of system s on some topic by X_s, the marginal distribution of system s by F_s, and the joint distribution of all m system scores as

$$F(x_1, \ldots, x_m) = P(X_1 \leq x_1, \ldots, X_m \leq x_m). \quad (1)$$

By Sklar's theorem [27], we can decompose this distribution as

$$F(x_1, \ldots, x_m) = C\big(F_1(x_1), \ldots, F_m(x_m)\big). \quad (2)$$

The function C is called the copula of F, and it captures the dependence between systems. It is a distribution function of the random vector $(U_1, \ldots, U_m)$, where $U_s = F_s(X_s) \rightarrow U_s \sim \mathit{Uniform}$. Taking the derivative on both sides of (2) gives us a decomposition of the corresponding joint density:

$$f(x_1, \ldots, x_m) = c\big(F_1(x_1), \ldots, F_m(x_m)\big) \cdot f_1(x_1) \cdots f_m(x_m), \quad (3)$$

where $f_1, \ldots, f_m$ are the marginal densities and c is the density corresponding to the copula C.

A convenient property of copulas is that they can be fitted separately from the margins. The most common procedure is:

(1) Fit the marginal distributions F_s of each system.
(2) Use the fitted margins to transform the observed scores X_s to *pseudo-observations* of the copula: $U_s = F_s(X_s)$.
(3) Fit the copula model to the pseudo-observations.

The procedure to simulate the scores on a new random topic is:

(1) Generate a pseudo-observation $(R_1, \ldots, R_m)$ from the copula.
(2) Compute $Y_s = F_s^{-1}(R_s)$.

By construction then, we have $Y_s \sim F_s$. In reality, a copula models the dependence between quantiles of the margins (the pseudo-observations), not the actual raw observations. This way, the same copula can be used to describe the dependence among systems, but we retain full control over individual system distributions just by plugging different margins into the copula.

2.1 Gaussian Copulas

There is a variety of parametric models for the copula C, the most popular of which is the Gaussian copula [19]. It is derived from the multivariate Gaussian distribution by inversion of (2):

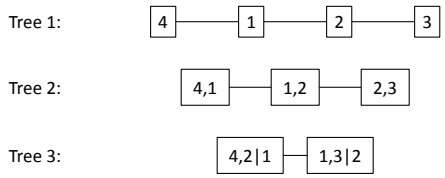

Figure 2: An R-vine tree sequence for four systems.

$$C_{gaussian}(u_1, \ldots, u_m) = \Phi_\Sigma\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m)\big), \quad (4)$$

where Φ is the univariate standard Gaussian cumulative distribution function, and Φ_Σ is the distribution function of a multivariate Gaussian with mean 0 and correlation matrix Σ. Since the Gaussian copula is parameterized by Σ, it allows to control the strength of dependence between each pair of variables. This is a major advantage over other parametric families such as Archimedean copulas [19], which only have 1 or 2 parameters in total.

2.2 Vine Copulas

Because the Gaussian copula is derived from the multivariate Gaussian distribution, the dependence between each pair of variables is constrained to be highly symmetric. In reality though, dependence is often asymmetric (see Figure 1-left for an example). The dependence may be stronger for small values of the system scores and weaker for large values, or the other way around. An even more flexible class of copula models are *vine copulas* [1]. A vine copula model is graphically represented as a set of linked trees where the edges of tree i are the nodes of tree $i+1$. The full dependence structure is built with individual two-dimensional copulas attached to each of the edges.

A special kind of vine are *regular vines* or *R-vines*, where two edges in tree i are joined by a node in tree $i+1$ only if they share a common node. An example on four systems is shown in Figure 2. The vertices in the second tree are the edges in the first tree. For instance, the edge $\{4,1\}$ encodes the dependence between X_4 and X_1. The edge $\{\{4,1\},\{1,2\}\}$ encodes the dependence between X_4 and X_2 conditional on X_1.[1] This edge is then represented in tree 3 with the vertex $\{4,2|1\}$. The combination of all conditional pair-copulas in the vine fully determines the dependence structure of the vector $(X_1, \ldots, X_4)$.

The concept of vine copulas naturally extends to higher dimensions. One merely has to attach an edge with a new vertex to each tree in the R-vine. The trees are not restricted to be paths as in Figure 2 (e.g. vertex $\{3\}$ could be connected to vertex $\{1\}$ instead of $\{2\}$), but the edge sets have to fulfill certain conditions that ensure that the vine copula results in a valid and fully specified dependence structure [6]. An algorithm for selecting an appropriate structure in a data-driven manner was introduced by Dissmann et al. [16].

The main advantage of vine copulas is that one can select from a large variety of parametric copula families for each pair-copula, symmetric and asymmetric ones. This makes the models very flexible, but also very complex. A vine copula on m systems consists of $m(m-1)/2$ pair-copulas, and each of them requires us to select a family and estimate its parameters. Fortunately, it is possible to do this sequentially, one pair-copula at a time (for details, see Aas et al. [1]). In higher dimensions ($m \geq 10$), it is common to *truncate* the model [7]: only the pair-copulas in the first few trees are specified, and all subsequent pairs are considered independent. This allows us to fit and select a vine copula model in a matter of seconds, even in high dimensions, using mainstream computing resources.

[1] The conditioning variable is found by intersecting node indexes: $\{4,1\} \cap \{1,2\} = \{1\}$.

3 MODELING SYSTEM EFFECTIVENESS DISTRIBUTIONS

As discussed in the previous section, we can model the marginal system effectiveness distributions separately from their dependence. The obvious choice would be to use the empirical distribution of the given data, but then a value that has not occurred in these data would never come up in a simulation. Instead, in this section we describe various parametric and non-parametric alternatives to model the margins. The first distinction we make is between continuous and discrete distributions. In principle, all measures are discrete because they calculate a score based on a finite document list, a finite set of judgments and a discrete relevance scale, so the possible set of outcomes is also finite. This is evident in measures like *P*@10, where only 11 different values are possible. However, for more fine-grained measures like *AP* or *nDCG*, the set of possible outcomes is fairly large and we can comfortably assume they follow a continuous distribution.

In the following, let us change notation and refer to the set of scores for a system over n topics as $\{X_1, \ldots, X_n\}$.

3.1 Continuous Distributions

When assuming a continuous distribution of effectiveness, we have simple and well-known options such as the Normal or Beta distributions. However, they are fairly restricted as to the shape that their density function can take, so we also consider kernel smoothing. Next, we present four alternatives to model a continuous distribution of effectiveness, based on the Normal and Beta.

3.1.1 (Truncated) Normal Distribution (N). One of the simplest choices to model continuous data is the Gaussian or Normal distribution, parameterized by the mean μ and variance σ^2. However, a Normal distribution is supported on $(-\infty, +\infty)$, while effectiveness scores are typically supported on $[0,1]$. To solve this issue we can truncate the distribution. Let f, F and F^{-1} be the density, distribution, and quantile functions of some distribution (Gaussian in our case). The corresponding distribution truncated between a and b is

$$f_{trunc}(x) = \frac{f(x)}{F(b) - F(a)}, \quad (5)$$

$$F_{trunc}(q) = \frac{F(q) - F(a)}{F(b) - F(a)}, \quad (6)$$

$$F^{-1}_{trunc}(p) = F^{-1}\big(p\,(F(b) - F(a)) + F(a)\big), \quad (7)$$

where $a = 0$ and $b = 1$ for our purposes. Note that no assumptions are made about the original distribution, other than it being continuous. Therefore, we can follow the same idea to truncate other distributions and maintain the support of interest. For the Truncated Normal, μ and σ^2 are typically estimated numerically by maximizing the log-likelihood of the given data. Finally, the

expected value and variance are [17]:

$$\mathrm{E}[X] = \mu + \frac{\phi(a') - \phi(b')}{\Phi(b') - \Phi(a')}\sigma \tag{8}$$

$$\mathrm{Var}[X] = \sigma^2 \left[1 + \frac{a'\phi(a') - b'\phi(b')}{\Phi(b') - \Phi(a')}\right] - (\mathrm{E}[X] - \mu)^2 \tag{9}$$

$$a' = (a - \mu)/\sigma,\ b' = (b - \mu)/\sigma.$$

Figure 1 shows a truncated Normal (N) fitted to real TREC data.

3.1.2 Beta Distribution (B). A natural alternative for bounded data is the Beta distribution, which is already supported on the unit interval. In addition to a bell shape, its density function can also have J and U shapes, with custom degrees of asymmetry. It is typically described with two shape parameters $\alpha, \beta > 0$, and its expectation and variance are

$$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta}, \tag{10}$$

$$\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{11}$$

There is no close form solution for maximum likelihood estimation of the parameters, so they are typically estimated via numerical optimization. Figure 1 shows a Beta (B) fitted to real TREC data.

3.1.3 (Truncated) Normal Kernel Smoothing (NKS). Even though Normal and Beta distributions are simple and well-known, they are restricted to certain shapes. A more flexible alternative is Kernel Smoothing. In its general form, a kernel-smoothed density estimator is parameterized by a bandwidth $b > 0$ that controls the degree of smoothing, and a kernel k, which is a non-negative function that integrates to one. The kernel-smoothed distribution is defined as

$$f(x) = \frac{1}{nb}\sum_i k\left(\frac{x - X_i}{b}\right), \tag{12}$$

$$F(x) = \frac{1}{n}\sum_i K\left(\frac{x - X_i}{b}\right). \tag{13}$$

In our case, the kernel is $k = \phi$ and $K = \Phi$, that is, the standard Normal. In general, there is no close form expression for the quantile function F^{-1}. However, because F is continuous and monotonically increasing, a quantile p can be computed numerically by finding the solution to $F(x) = p$.

In order to select the bandwidth parameter b, we use the automatic procedure by Wand and Jones [34]. As with the Normal, we truncate the distribution between 0 and 1 following (6). The expected value and variance are calculated via numerical integration of the quantile function:

$$\mathrm{E}[X] = \int x f(x)\,dx = \int_0^1 F^{-1}(x)\,dx, \tag{14}$$

$$\mathrm{Var}[X] = \int (x - \mathrm{E}[X])^2 f(x)\,dx = \int_0^1 F^{-1}(x)^2\,dx - \mathrm{E}[X]^2. \tag{15}$$

The first expressions are the text-book formulas for a continuous random variable, but because we will ultimately simulate new data through the quantile functions, we use the second expressions for higher numerical precision. Figure 1 shows a truncated Normal Kernel-Smoothed (NKS) fitted to real TREC data.

3.1.4 Beta Kernel Smoothing (BKS). The kernel function can also be based on the Beta distribution to naturally bound the support. Here we use the kernel proposed by Chen [13], which yields the density function

$$f(x) = \frac{1}{n}\sum_i f_{Beta}\left(X_i\ ;\ \frac{x}{b+1}, \frac{1-x}{b+1}\right). \tag{16}$$

Note that the above expression is not a valid density because it is not guaranteed to integrate to one, but this is easily solved by normalization over the full support. The distribution and quantile functions are once again calculated numerically, and the expectation and variance are computed following (14) and (15). The bandwidth parameter is set to $b = n^{-2/5}$ by default [13]. Figure 1 shows a Beta Kernel-Smoothed (BKS) fitted to real TREC data.

3.2 Discrete Distributions

The typical method to fit a discrete distribution is to assign an integer rank to each possible value of the support, fit a well known distribution such as the Binomial or Poisson, and convert back to the original support when simulating new data. Let f be a probability mass function with support $\{s_1, \ldots, s_z\}$. The corresponding cumulative distribution and quantile functions, as well as the expectation and variance, are easily calculated from their very definitions for a discrete distribution:

$$F(q) = \sum\nolimits_{s \le q} f(s), \tag{17}$$

$$F^{-1}(p) = \inf\{s : F(s) \ge p\}, \tag{18}$$

$$\mathrm{E}[X] = \sum s \cdot f(s), \tag{19}$$

$$\mathrm{Var}[X] = \sum f(s) \cdot (s - \mathrm{E}[X])^2. \tag{20}$$

Below we propose a parametric and a non-parametric alternative.

3.2.1 Beta-Binomial (BB). The Beta-Binomial distribution is the Binomial distribution in which the probability of success in each trial is not fixed, but distributed according to a Beta. Similarly, its mass function can have several shapes, making it a suitable candidate for discrete data. It is parameterized by the $\alpha, \beta > 0$ parameters of the underlying Beta, and the number of trials $m > 0$ of the underlying Binomial. The support is $\{0, 1, \ldots, m\}$, so the original effectiveness scores need to be transformed into integers. For instance, a $P@k$ score X would be transformed to $k \cdot X$. In general,the i-th support value is transformed into $i - 1$.

Parameters are estimated by numerical optimization of the log-likelihood. However, note that m can be fixed to the size of the original support set minus one. In the case of $P@k$, there are only $k+1$ possible outcomes which correspond to the number of successes in $m = k$ trials. Therefore, only α and β need to be estimated. Figure 1 shows a Beta-Binomial (BB) fit to real TREC data.

3.2.2 Discrete Kernel Smoothing (DKS). The kernel-smoothed distributions in (12) and (16) can be adapted to discrete variables by using a kernel function with a discrete support [35]:

$$f(x) = \frac{1}{n}\sum_i k\left(x, X_i, b\right), \tag{21}$$

$$k(x, X_i, b) = \begin{cases} (1-b) & X_i = x \\ \frac{1}{2}(1-b)\cdot b^{|X_i - x|} & \text{otherwise.} \end{cases}$$

The kernel function k is designed for the case where X_i is integer valued. The discrete kernel smoother is not a proper probability mass function because it usually does not sum up to one, but this can easily be corrected by normalization. The bandwidth parameter b can be selected automatically by least-squares cross-validation [35].

The discrete kernel-smoothed distribution is very appealing for measures with non-standard support, such as Reciprocal Rank. In the typical case with an evaluation cutoff $k = 1,000$, the support of an *RR* score is $\{0, 1/1000, 1/999, \ldots, 1/1\}$. Most values are thus concentrated near 0, so a parametric model like the Beta-Binomial will not provide a good fit. The kernel-smoothed distribution in (21) is flexible enough to adapt, but it can easily overfit with rare supports like this. To alleviate this issue we introduce a bandwidth multiplier $h > 1$ such that the actual bandwidth is $bh \leq 1$. Figure 1 shows the Discrete Kernel Smoothing (DKS) fit with the initially selected bandwidth (i.e. $h = 1$), and the variant with $h = 2$ (DKS-2).

3.3 Model Selection

Each of the above distribution models can be fitted for a given set of effectiveness scores, so in practice we will need to choose the best distribution from a set of candidates. An obvious criterion for selection is the log-likelihood (LL)

$$\text{LL} = \sum\nolimits_i \log f(X_i), \tag{22}$$

which is maximized by the best model. However, the log-likelihood can be made arbitrarily large by making the model more complex, so it favors distributions that overfit the data. The Akaike Information Criterion [2] and the Bayesian Information Criterion [26] are two classical criteria that correct for this by penalizing the number of parameters θ of a model. They are defined as

$$\text{AIC} = -2\text{LL} + 2\theta, \tag{23}$$

$$\text{BIC} = -2\text{LL} + \theta\log(n), \tag{24}$$

where n is the sample size (the number of topics in our case). The best fitting model minimizes the AIC or BIC. If $n > 8$, the BIC puts a stronger penalty on the number of parameters, so it favors more parsimonious models.

For the parametric distributions, the number of parameters is straightforward: the Normal, Beta and Beta-Binomial distributions have two parameters each (recall that m is fixed upfront). For nonparametric distributions the number of parameters is not defined, but there is an equivalent concept called *effective degrees of freedom* that can be used in the formulas of AIC and BIC. The most common definition arises from the concept of linear estimators [18, 21]: for any density estimator that can be written in the form $f(x) = \sum_i g(x, X_i)$, the effective degrees of freedom are

$$edf = \sum_i \frac{g(X_i, X_i)}{\sum_j g(X_i, X_j)}. \tag{25}$$

This is possible to calculate for all three kernel-smoothed distributions presented in the previous sections.

3.4 Ensuring Expectations

As mentioned above, there are many situations in which researchers want to experiment with systems whose expected values differ by some predefined amount, or systems that have the same expected value but a test collection says otherwise. In some other cases we may want a fitted distribution to have the same expected value as the observed mean score in the know data. Unfortunately, these restrictions are now guaranteed by construction in the above models.

Let F be a cumulative distribution function with mean μ. A natural way to ensure a expected value μ^* is to shift all observations by subtracting $\mu - \mu^*$, but this has two problems: the resulting distribution would not be bounded by $[0, 1]$, and if it is discrete, the shifted observations are not guaranteed to have a correct support anymore. For instance, a $P@10$ score of 0.2 could easily become something like 0.17, which is clearly an invalid value for that measure. We propose an alternative way to modify F that ensures that the new mean is μ^* and the support remains unchanged.

For any increasing function $T\colon [0,1] \to [0,1]$, the function $\tilde{F}(x) = T(F(x))$ is again a distribution function and has the same support as F. The goal is to find a transformation T such that the resulting mean of $\tilde{F}$ is $\tilde{\mu} = \mu^*$. To this end, we first restrict the transformation to the family of Beta distribution functions $\mathcal{B} = \{F_B(\cdot\,;\,\tilde{\alpha}, \tilde{\beta})\} : \tilde{\alpha} > 0, \tilde{\beta} > 0\}$. Then we can solve numerically for a combination of $\tilde{\alpha}$ and $\tilde{\beta}$, such that $\tilde{\mu} = \mu^*$ according to (14). This gives us a new distribution $\tilde{F}$ that has the desired mean μ^*, and whose distribution functions are:

$$\tilde{f}(x) = f_B\big(F(x)\,;\,\tilde{\alpha}, \tilde{\beta}\big)\cdot f(x), \tag{26}$$

$$\tilde{F}(q) = F_B\big(F(q)\,;\,\tilde{\alpha}, \tilde{\beta}\big), \tag{27}$$

$$\tilde{F}^{-1}(p) = F^{-1}\big(F_B^{-1}\big(p\,;\,\tilde{\alpha}, \tilde{\beta}\big)\big). \tag{28}$$

Essentially the same procedure may be used to ensure a predefined variance instead of mean, thus allowing us to control the homoskedasticity of the model. If on the other hand we want to enforce both mean and variance, the parametric models could be instantiated to already meet the constraints because they have precisely two parameters. In the general case, bivariate optimization can be attempted, but this goes beyond the scope of this paper.

4 EVALUATION

In this section we evaluate the proposed simulation method from the perspectives of the effectiveness distributions, the copulas, and the simulated data. In particular, we will build and evaluate distribution and copula models using data from the Ad hoc submissions to the TREC Web track between 2010 and 2014. Each of these collections contains about 50 topics and between 30 and 88 systems, for a total of 12,924 system-topic pairs. In terms of effectiveness measures, we use Average Precision, *nDCG*@20 and *ERR*@20 as exemplars of measures with continuous support, and Reciprocal Rank, *P*@10 and *P*@20 as exemplars of measures with discrete support.

4.1 System Effectiveness Distributions

The first step towards a successful simulation is to fit a model to each of the marginal system effectiveness distributions, following the principles in Section 3. Therefore, our first point of interest is which of the various alternatives provide the best fit to real

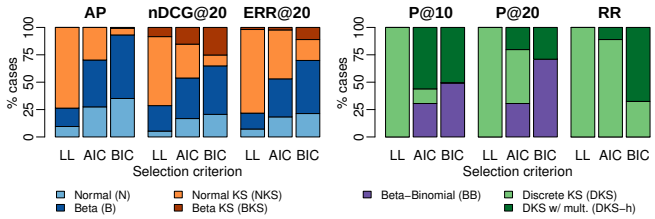

Figure 3: Distributions of models selected for the marginal system distributions, *before* μ transformation.

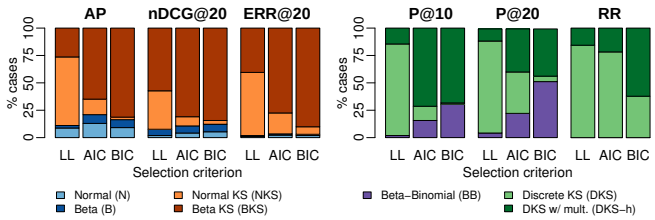

Figure 4: Distributions of models selected for the marginal system distributions, *after* μ transformation.

evaluation data. For each collection and measure, we start from the $n \times m$ matrix with the scores by all m systems on all n topics. For measures with continuous support we fit 4 different models: truncated Normal (N), Beta (B) truncated Normal Kernel Smoothing (NKS), and Bounded Kernel Smoothing (BKS). For measures with discrete support we fit up to 5 different models: Beta-Binomial (BB), Discrete Kernel Smoothing with multiplier $h = 1$ (DKS), and DKS with multiplier $h \in \{2, 5, 10\}$ (DKS-h)[2].

A grand total of 5,425 models were fitted for all 1,572 combinations of measure and system. For each combination, the best model was selected according to LL, AIC and BIC (see Section 3.3). Figure 3 shows the distributions of best models for each effectiveness measure and *before* μ-transformation. Across measures we can see that log-likelihood favors the more complex non-parametric models NKS and DKS. With continuous measures, the parametric models N and B are chosen only about 25% of the times, and the discrete BB is in fact never selected by the log-likelihood criterion. As expected though, AIC and BIC penalize this complexity and tend to select the simpler models. The exception is Reciprocal Rank, where the BB model is still never selected because it is not able to adapt to the non-standard support as well as the non-parametric DKS. What we do appreciate in this case is the selection of DKS-h by AIC and BIC, because of the lower effective degrees of freedom.

As discussed in Section 3.4, we often need to transform system distributions to make sure their expectation equals some predefined value. In principle, the best model before transformation is not necessarily the best one after transformation, so we attempted to transform all 5,425 models such that their expected values equal the mean score observed in the given data (within a 10^{-5} threshold), and then performed model selection again. The transformation was

[2]Not all DKS-h models can be fitted in all cases. Whether a particular value of h produces a valid fit depends on the pre-selected bandwidth b.

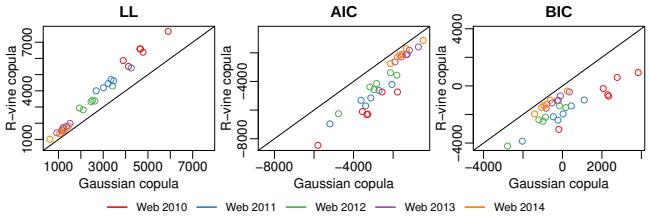

Figure 5: LL (higher is better), AIC and BIC (lower is better) of the fitted Gaussian vs. R-vine copulas.

successful in 5,003 cases (92%), and Figure 4 shows the distributions of selected models. Compared to the untransformed case, we can see that non-parametric models are almost always selected because of their flexibility. In fact, the most successful model in the continuous case is BKS, which was seldom selected before.

Although these results do not tell us in and on themselves about the quality of the fitted models, they suggest that the best options are generally those that are neither too complex like kernel smoothing, nor too simple like basic Normal or Beta distributions. Furthermore, they suggest that there is no single best model for all cases, not even within measures, and that if transformations are required, model selection should be performed afterwards.

4.2 Copulas

The second step towards a successful simulation is to fit a copula model to the pseudo-observations, following the principles in Section 2. Therefore, our second point of interest is which alternative provides the best dependence model to real evaluation data.

In order to study whether Gaussian copulas are appropriate to model the dependence among systems, we first fit bivariate copulas between every pair of systems in the same collection, selecting the best candidate according to log-likelihood. From the total of 39,627 system pairs, Gaussian copulas are selected only in 2.7% of the cases, thus evidencing that simple correlation is not enough to model the dependence found among real systems. The most common copula is the BB8 copula, selected 30% of the times, followed by the Tawn 1 (16%), Tawn 2 (16%), BB7 (12%), t (9%), BB1 (4.8%), Clayton (4.2%), Frank (4%) and others. Figure 1-left shows an example.

We then fit full Gaussian and R-vine copulas to all systems and topics in an effectiveness matrix, performing selection based on LL, AIC and BIC. Suggested by the previous results on pair-copulas, we first check whether the extra complexity of vine copulas actually helps us capture the full dependence structure in a better way. Figure 5 shows the goodness of fit of all 30 Gaussian copulas compared to their R-vine counterparts. According to all selection criteria, R-vine copulas do indeed provide a significantly better fit.

Another aspect to analyze is the possibility to truncate the vines in order to speed up the fitting process at the cost of reducing the goodness of fit. One indicator to decide whether to continue to the next vine level or not, is the Kendall τ correlation observed in the pair-copulas of the current level: if the correlation is low, we may decide to stop and truncate, and if the correlation is high we may decide to continue because there appears to be some degree of dependence to further incorporate into the vine. Figure 6

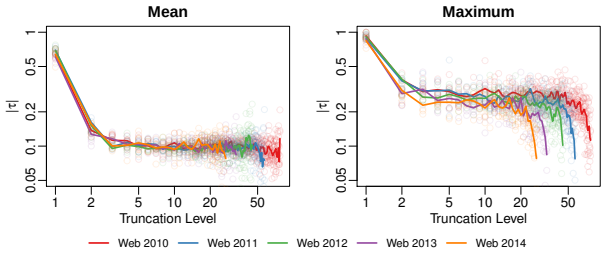

Figure 6: Mean and maximum absolute value of the Kendall τ (log-scaled) at each truncation level (log-scaled) of the R-vine copulas. Lines represent within-collection means.

shows the mean and maximum absolute value of the τ correlation scores at each level of the R-vine copulas. In the first level we can see very high correlations, but already in the second level they drop significantly. The average correlation remains around 0.1 throughout truncation levels, but the maximum decreases very slightly between 0.3 and 0.2 until nearly the end, indicating that some specific pair-copulas are still capturing a relevant amount of dependence. Overall, these results suggest that there is no obvious point to truncate the vines, and that in any case the truncation level of course depends on the number of systems. Even though the number of pair-copulas fitted in a full vine grows quadratically with the number of systems, a mainstream computer fitted the full models in under one minute each, so our suggestion is not to truncate.

4.3 Simulated Scores

The third and final step towards a successful simulation is to generate random pseudo-observations from the copula, and from there the final effectiveness scores via the marginal distributions. The mean and variance of the simulated scores should meet the values predefined in these distributions. This is the ultimate goal of using the proposed method: simulating new data from a model for which we know the full characteristics in advance.

We start with the marginal distributions and copulas fitted in the previous section, and proceed as follows. For each copula we simulate 1,000 random observations, and record the observed mean and variance for each of the m systems. This is repeated 1,000 times, yielding $1,000 \cdot m$ sample means and sample variances. For each of these we compute the deviation from the expected values, that is, $\mu - \overline{X}$ and $\sigma^2 - s^2$. Over repetitions, we expect these deviations to be centered at zero, meaning that the simulated scores are unbiased. Figure 7 shows at the top the distributions of deviations. We can first appreciate that all distributions are indeed centered at zero, indicating that the simulated scores are not biased and therefore the μ and σ^2 scores predefined by the margins can be trusted. We can also observe higher variability in the discrete measures, because their support is much less fine-grained than the continuous measures, leading to less stability [8]. In terms of magnitude of the deviations, we can see that in most cases they are within 0.01 of the mean and 0.002 of the variance.

We note that these are deviations for samples of size 1,000 topics; larger samples have of course smaller deviations. For comparison, the bottom plots in Figure 7 show the distributions of deviations for

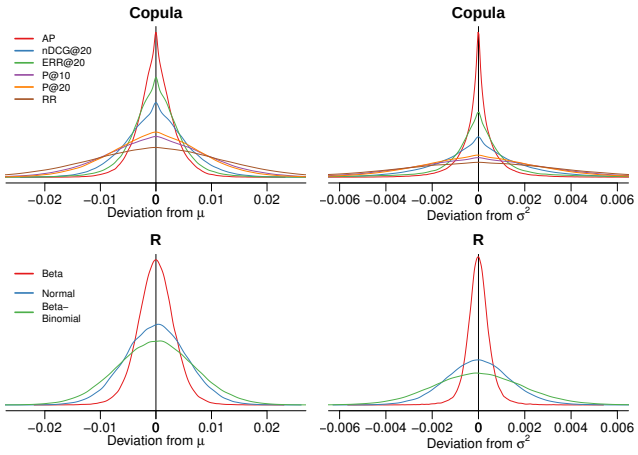

Figure 7: Deviations from the mean (left) and variance (right), in samples of size 1,000 simulated from the R-vines. The bottom plots show similar statistics for simulations of standard distributions as implemented in R.

various random number generators in the statistical software R. In particular, we simulated 100,000 samples of 1,000 observations each, and similarly recorded the deviations from the theoretical mean and variance. The data are simulated from three distributions with a random selection of parameters such that the simulated scores have dispersion similar to the original TREC data: Gaussian with $\sigma \in [0.15, 0.2]$, Beta with $\alpha, \beta \in [1, 20]$, and Beta-Binomial with $m = 10$ and $\alpha, \beta \in [1, 6]$ (normalized to the unit interval)[3]. As the plots show, the distributions of deviations in base R are indeed similar to the deviations in the simulated evaluation data.

Finally, Figure 8 (left) shows as an example the Spearman correlation matrix among the Web 2010 systems, as per *nDCG*@20. The second matrix presents the observed correlations in 500 random topics simulated with the Gaussian copula. As expected, they are very similar because the Gaussian copula models linear correlations. The third matrix presents the correlation in 500 topics simulated from the R-vine copula without truncation, which is much more faithful to the original data at least with respect to correlation (recall that vine copulas model dependence structures more complex than simple correlation). The last matrix shows similar results from an R-vine truncated at level 2, showing that even if we truncate this early the bulk of the dependence structure is accounted for.

Overall, the results suggest that the simulated data behaves as expected within reasonable precision bounds, and that it is indeed capable of capturing the dependence underlying the existing data. These models effectively provide an endless supply of realistic evaluation data with known and predefined characteristics, which can prove to be a very valuable resource for IR evaluation research.

5 SAMPLE APPLICATIONS

In this section we describe two sample applications of the proposed simulation method, showing how it can help us overcome the discussed limitations in IR evaluation research. In particular,

[3]These simulations used the standard functions `stats::rnorm`, `stats::rbeta` and `extraDistr::rbbinom`.

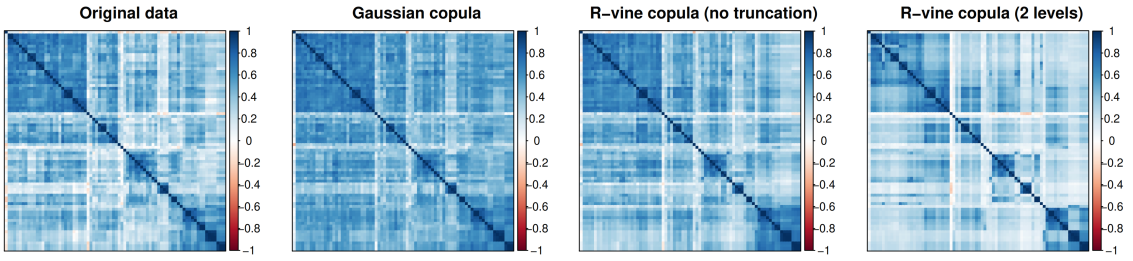

Figure 8: Spearman correlation matrices of the original Web 2010 *nDCG@20* scores and 500 observations simulated with the Gaussian copula, the R-vine copula without truncation, and truncated at 2 levels.

we repeat an experiment by Webber et al. [38] on the estimation of the variability of between-system score deltas for the purposes of power analysis, and an experiment by Voorhees [33] on hypothesis testing. The selection of these two works is by no means intended as criticism. On the contrary, they are selected as clear examples of how researchers have to make do with the available data and work around these limitations.

5.1 Statistical Power and Topic Set Size

A typical problem in IR evaluation is deciding an appropriate number of topics to compare two systems with a some degree of confidence. In Section 5.3 of [38] the authors describe an experiment using the TREC 2004 Robust data, emulating a researcher who builds a new test collection by iteratively adding new topics until a certain level of statistical power is achieved. In particular, they use only the 150 topics from TREC 6–8 and the 78 not description-only systems. Next we describe their original design and two alternatives using our simulation method (Figure 9 presents an outline).

Original Design (O). Repeat the following experiment 100 times:

(1) Randomly select a system from the second quartile of runs and another system from the top three quartiles.
(2) Compute the original per-topic system score deltas $D_1 \dots D_{150}$.
(3) Compute σ_{150} as the standard deviation observed in $D_1 \dots D_{150}$. Assume σ_{150} is the true population standard deviation.
(4) Compute the target δ equal to the minimum detectable difference by a t-test, with power 80% and significance level 5%, if using 100 topics.
(5) Repeat the following iterative process for $i = 1 \dots B$:
   (a) Start with an empty topic set.
   (b) Sample a new topic with replacement from the 150 available, compute the standard deviation $\sigma^{(i)}$ with the current topic set, and the detectable difference d assuming $\sigma^{(i)}$.
   (c) If $d > \delta$, go back to step 5.b. If not, stop and record the current number of topics $n^{(i)}$ and $\sigma^{(i)}$.
(6) Record the mean number of topics and standard deviation over the $B = 1,000$ trials: $n_B = \frac{1}{B}\sum n^{(i)}$ and $\sigma_B = \frac{1}{B}\sum \sigma^{(i)}$.

In this experiment, σ_B is intended to estimate the population σ, which is set to σ_{150}. Similarly, n_B is intended to estimate the required number of topics, set to 100. In their Figure 5 they plot each σ_B versus the corresponding σ_{150}, showing a clear bias. We repeated their experiment with 200 system pairs; plot O) in Figure 10 shows the same results. The standard deviations are always underestimated, by 5.3% on average, and the required number of

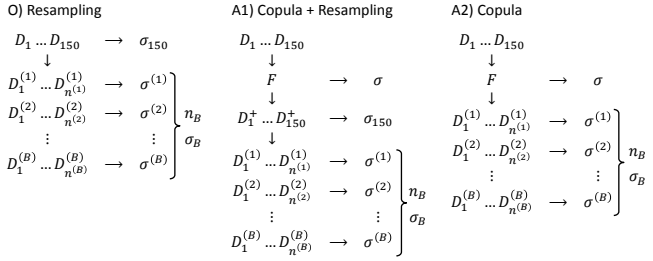

Figure 9: Outline of the experimental designs to show the effect of sequential testing: O) original design [38], A1) and A2) alternatives with the proposed simulation method.

topics is also underestimated to 93. Essentially, this experiment shows empirical evidence of the problem of sequential testing [39], further demonstrated in subsequent IR works [11, 36].

However, this is a clear case in which researchers are restricted by the available data and they have to make certain assumptions to approximate the statistics of interest. Indeed, the ultimate goal of this experiment was to see how much bias the iterative sampling of topics introduces in the estimation of the true population σ. In step (3), the σ_{150} observed in the available data *is taken as the true population* σ, which is of course unknown in reality. The larger question though, is not how biased σ_B is with respect to σ_{150}, because it has some degree of error itself, but with respect to the true σ.

Alternative Design 1 (A1). The first alternative to overcome this problem using our simulation methodology, modifies step (2) to:

(2) Fit the margins and bivariate copula from $D_1 \dots D_{150}$, thus setting the true distribution F and σ. Simulate a new set $D_1^+ \dots D_{150}^+$ from the copula, which are used to compute σ_{150}.

The A1) plots in Figure 10 show the results. First, we see that σ_B is still always underestimating σ_{150} because of sequential testing, but when compared to the true and known σ, the underestimation is not as consistent. Still, on average σ is underestimated by 2.9%, and the required number of topics is 97.

Alternative Design 2 (A2). However, the best way to make full use of our proposed simulation methodology is to use the original 150 topics just to fit the copula, and always simulate a brand new topic from it in step (5.b). This allows us to compare σ_B with the true and known σ over truly independent topic sets, thus eliminating the limitations of the original design. In the A2) plot of Figure 10 we can observe that the bias due to sequential testing is still clearly present,

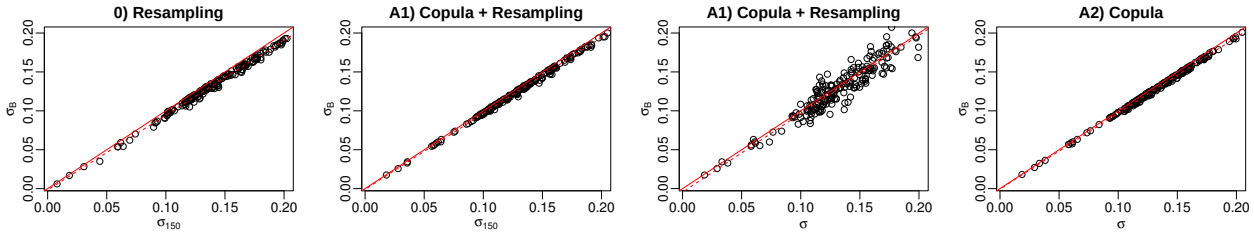

Figure 10: O) σ_B estimated via resampling vs. assumed σ_{150} from the original 150 topics (replication of Figure 5 in [38]). A1-left) same but from 150 new topics simulated from a copula. A1-right) σ_B estimated via resampling vs. the true σ prefixed by the copula. A2) σ_B estimated via true random sampling from the copula vs. true σ prefixed by the copula.

with an average underestimation of 2.9% on σ and an average of 97 topics required from the expected 100.

In summary, this first use case allows us to pinpoint two limitations in this kind of IR evaluation research, namely that we do not know the true characteristics of systems (population σ), and the limited availability of data (only 150 topics) that forces us to resample and make assumptions. The proposed methodology for stochastic simulation effectively eliminates these two limitations and allows us to study the research question directly and precisely. Indeed, our experiments reproduce the undesired effect of sequential testing.

5.2 Hypothesis Testing and Type I Errors

A recurring problem in IR evaluation concerns hypothesis testing and Type I errors. A sample work on this topic is [33], where the authors describe an experiment using the TREC 2004 Robust data. In particular, they employed the 100 topics from TREC 7–8, and only the top 83 systems according to mean AP score over the entire set. Next we describe their original design and three alternatives using our simulation method.

Original Design (O). The following was repeated 1,000 times:

(1) Randomly split the 100 topics in two halves T_1 and T_2.
(2) For each pair of systems A and B, run a t-test on the scores over T_1 and another test on T_2, both 2-tailed and at $\alpha = 0.05$.
(3) Record which of the two tests are significant, and mark the pair as *minor conflict* if the mean scores have different sign but only one test is significant, *major conflict* if the means have different signs and both tests are significant, or *no conflict* otherwise.

Similar experiments appear in the literature in reference to Type I error rates, sometimes incorrectly approximating it with the fraction of significant results that are in a conflict [15, 25, 32, 40]. In [33], authors reported that 2.8% of the significant results where part of a conflict when using *AP*, and 10.9% when using *P*@10. Other papers report similar findings for *AP*, somewhat suggesting that the t-test is too conservative with IR data and it makes about half as many Type I errors as it should, or twice as many in the case of *P*@10. Here we repeated the original experiment of [33] and found 2.91% of conflicts with *AP* and 12% with *P*@10, thus confirming their results to a large extent.

Alternative Design 1 (A1). This experiment is similarly constrained by the limited amount of topics, which are repeatedly split in two in order to simulate independent topic sets. This may introduce bias because the two sets are not really independent and always come from the same 100 topics. To assess the effect of this bias we carried out a similar experiment, but instead of the random split in step (1), we simulated two new random sets of 50 topics each from a bivariate copula. This means that every single comparison between a pair of systems is made with truly random and independent topic sets. The observed conflict rates are 2.67% in *AP* and 15% in *P*@10.

Another limitation of the original design is that the conflict rate is only a rough approximation to the Type I error rate, which is the statistic we are really interested in. In fact, there is no way of knowing if a test yielded a false positive or not because the true system mean scores are unknown, that is, whether the null hypothesis is true or not is unknown to begin with. Using our proposed methodology based on simulation, we have two options to avoid this limitation.

Alternative Design 2 (A2). The first alternative ensures that both systems have the same marginal distribution and hence the same expected values, making the null hypothesis true by definition. The Type I error rate can now be empirically estimated as the fraction of comparisons that yield a significant result. We proceed as follows. Given the two systems A and B, we fit the bivariate copula just like in the previous example, but use the same marginal distribution F_A for both systems. This way we simulate from systems with the same distribution but different dependence structure. For higher precision, the experiment was repeated 10,000 times for each pair of systems, and we found 4.88% of Type I errors with *AP* at $\alpha = 0.05$, and 0.9% at $\alpha = 0.01$. With *P*@10 we found 4.94% and 0.96% of errors respectively.

Alternative Design 3 (A3). The last alternative consists in using different marginal distributions, but transformed such that they have the same expected value (see Section 3.4). This presents a more realistic scenario than the previous alternative. We proceed as follows. Given a system A, we randomly select a system B from the 10 systems whose expected values are closest to A's, and transform F_B such that its expected value is μ_A. This way we simulate from systems with different distributions but same expected values, and different dependence structure. The experiment was again repeated 10,000 times for each pair of systems, and we found 4.88% and 0.9% of Type I errors with *AP* at $\alpha = 0.05$ and $\alpha = 0.01$. With *P*@10 we found 5% and 0.96% of errors respectively.

In summary, this second use case serves us to pinpoint three main limitations in this kind of IR evaluation research, namely the limited availability of data (only 100 topics), the lack of control over these data (truth of the null hypothesis), and inability to measure the actual statistics of interest (Type I error rate). The proposed

methodology for stochastic simulation avoids these limitations again and allows us to study the research question directly. In this particular case, our experiments show that the empirical error rates are kept at the nominal α level, showing that the t-test is *not* a conservative test with IR data.

6 CONCLUSIONS

In this paper we make the case for stochastic simulation of evaluation data to support research on IR evaluation without the known limitations of current practice, namely the scarcity of real data for large experimentation, the lack of control and customization over these data, and the lack of full knowledge about certain properties such as the true distributions of system effectiveness over populations of topics. We propose a method based on R-vine copulas that rids of these limitations and allows us to simulate realistic data about new and random topics, with full knowledge and control over the underlying properties. As an example, we replicated two typical experiments of IR evaluation research to show the benefits of our proposal. In the first experiment we obtained empirical results that reproduce and confirm the sequential testing problem, and in the second experiment we carry out the first empirical and direct assessment of Type I error rates in IR experimentation.

However, the proposed method can be used only in a certain class of evaluation problems. For example, we fit and assume a model for the distribution of scores produced by each measure, so research questions involving the comparison of distributional properties of the measures can not be answered. Similarly. our method simulates effectiveness directly, so research questions regarding pooling methods can not be studied with our method either. In both these cases we would need to simulate, not final effectiveness scores, but rather system runs and judgments; note that the content of documents and topics is not needed for this. This is an exciting topic we plan to study as well.

We have several other plans for further research in this line. First, we will implement other alternatives to model the margins, with special care of evaluation cutoffs that produce censored data. Second, we will study optimal structures for the vines to gain new simulation capabilities such as simulating a new random system for a given set of topics. Third, we will study the inclusion of a third factor other than systems and topics. A fixed third factor such as effectiveness measure seems straightforward to model just by adding new dimensions to the model, but a random factor such as assessor is more challenging. Last but not least, we plan on replicating previous evaluation research to confirm results with experiments free of the limitations discussed in this paper.

ACKNOWLEDGMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. JU dedicates this work to the brave people of Namek.

REFERENCES

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken. 2009. Pair-copula Constructions of Multiple Dependence. *Insurance: Mathematics and Economics* 44, 2 (2009).
[2] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.
[3] L. Azzopardi, M. de Rijke, and K. Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. In *ACM SIGIR*.
[4] L. Azzopardi, K. Järvelin, J. Kamps, and M.D. Smucker. 2010. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum* 44, 2 (2010), 35–47.
[5] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does it Matter?. In *ACM SIGIR*. 667–674.
[6] T. Bedford and R.M. Cooke. 2002. Vines — a new graphical model for dependent random variables. *The Annals of Statistics* 30, 4 (2002), 1031–1068.
[7] E.C. Brechmann, C. Czado, and K. Aas. 2012. Truncated regular vines in high dimensions with application to financial data. *Canadian Journal of Statistics* 40, 1 (2012), 68–85.
[8] C. Buckley and E.M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *ACM SIGIR*. 33–34.
[9] B. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM TOIS* 30, 1 (2012).
[10] B. Carterette. 2015. Bayesian Inference for Information Retrieval Evaluation. In *ACM ICTIR*. 31–40.
[11] B. Carterette. 2015. The Best Published Result is Random: Sequential Testing and Its Effect on Reported Effectiveness. In *ACM SIGIR*. 747–750.
[12] B. Carterette, V. Pavlu, E. Kanoulas, J.A. Aslam, and J. Allan. 2009. If I Had a Million Queries. In *ECIR*. 288–300.
[13] S.X. Chen. 1999. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis* 31, 2 (1999), 131–145.
[14] M.D. Cooper. 1973. A simulation model of an information retrieval system. *Information Storage and Retrieval* 9, 1 (1973), 13–32.
[15] Gordon V. Cormack and Thomas R. Lynam. 2007. Validity and Power of t-test for Comparing MAP and GMAP. In *ACM SIGIR*. 753–754.
[16] J. Dissmann, E.C. Brechmann, C. Czado, and D. Kurowicka. 2013. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis* 59 (2013), 52–69.
[17] C. Forbes, M. Evans, N. Hastings, and B. Peacock. 2011. *Statistical Distributions*. Wiley.
[18] J. Friedman, T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*. Springer.
[19] H. Joe. 2014. *Dependence Modeling with Copulas*. Chapman & Hall/CRC.
[20] E. Kanoulas and J.A. Aslam. 2009. Empirical Justification of the Gain and Discount Function for nDCG. In *ACM CIKM*. 611–620.
[21] C. Loader. 2006. *Local regression and likelihood*. Springer.
[22] S. Robertson and E. Kanoulas. 2012. On Per-Topic Variance in IR Evaluation. In *ACM SIGIR*. 891–900.
[23] Tetsuya Sakai. 2015. Topic Set Size Design. *Information Retrieval Journal* 19, 3 (2015), 256–283.
[24] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. 2012. Differences in Effectiveness Across Sub-collections. In *ACM CIKM*. 1965–1969.
[25] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *ACM SIGIR*. 162–169.
[26] G. Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.
[27] A. Sklar. 1959. *Fonctions de Répartition à n Dimensions et Leurs Marges*.
[28] J. Tague, M. Nelson, and H. Wu. 1981. Problems in the Simulation of Bibliographic Retrieval Systems. In *ACM SIGIR*. 236–255.
[29] J. Tague-Sutcliffe. 1992. The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing and Management* 28, 4 (1992), 467–490.
[30] J. Urbano. 2016. Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation. *Information Retrieval Journal* 19, 3 (2016), 313–350.
[31] J. Urbano and M. Marrero. 2016. Toward estimating the rank correlation between the test collection results and the true system performance. In *ACM SIGIR*. 1033–1036.
[32] J. Urbano, M. Marrero, and D. Martín. 2013. A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation. In *ACM SIGIR*.
[33] E.M. Voorhees. 2009. Topic Set Size Redux. In *ACM SIGIR*. 806–807.
[34] M.P. Wand and M.C. Jones. 1994. Multivariate plug-in bandwidth selection. *Computational Statistics* 9, 2 (1994), 97–116.
[35] M.C. Wang and J.V. Ryzing. 1981. A Class of Smooth Estimators for Discrete Distributions. *Biometrika* 68, 1 (1981), 301–309.
[36] W. Webber, M. Bagdouri, D.D. Lewis, and D.W. Oard. 2013. Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness. In *ACM SIGIR*. 933–936.
[37] W. Webber, A. Moffat, and J. Zobel. 2008. Score Standardization for Inter-collection Comparison of Retrieval Systems. In *ACM SIGIR*. 51–58.
[38] W. Webber, A. Moffat, and J. Zobel. 2008. Statistical Power in Retrieval Experimentation. In *ACM CIKM*. 571–580.
[39] G.B. Wetherill and K.D. Glazebrook. 1986. *Sequential Methods in Statistics*. Chapman and Hill.
[40] J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments?. In *ACM SIGIR*. 307–314.