Improving the Accuracy of System Performance Estimation by Using Shards

Nicola Ferro
ferro@dei.unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Mark Sanderson
mark.sanderson@rmit.edu.au
Computer Science, School of Science, RMIT University
Melbourne, Australia

ABSTRACT
We improve the measurement accuracy of retrieval system performance by better modeling the noise present in test collection scores. Our technique draws its inspiration from two approaches: one, which exploits the variable measurement accuracy of topics; the other, which randomly splits document collections into shards. We describe and theoretically analyze an ANalysis Of Variance (ANOVA) model able to capture the effects of topics, systems, and document shards as well as their interactions. Using multiple TREC collections, we empirically confirm theoretical results in terms of improved estimation accuracy and robustness of found significant differences. The improvements compared to widely used test collection measurement techniques are substantial. We speculate that our technique works because we do not assume that the topics of a test collection measure performance equally.

CCS CONCEPTS
• Information systems → Test collections; Retrieval effectiveness;

KEYWORDS
effectiveness model; ANOVA; multiple comparison

ACM Reference Format:

1 INTRODUCTION
Measuring the difference in performance between two Information Retrieval (IR) systems using offline test collection data has long been recognized as noisy. Attempts to improve the accuracy of such measurement are extensive and diverse. Techniques explored include multiple evaluation measures; different significance tests; alternate acquisitions of relevance judgments; and determining the ideal number of topics. Surveys [30] and descriptions of best practice [29] detail such attempts. There are, however, less explored approaches to improving performance measurement accuracy.

Robertson and Kanoulas [26] pointed out a common assumption in the use of test collections namely “all topics are considered equally valuable”. They examined this assumption by measuring (via bootstrapping) the confidence intervals of each topic score and of each system. The intervals were found to be variable across topics but largely independent of system. The researchers concluded that some topics measure performance more accurately than others.

Ferro and Sanderson [11] examined splitting the documents of a test collection into shards, measuring the performance of systems on each shard. They used an ANOVA model to understand if system performance changed across shards. The authors mentioned that significant differences between systems on shards were more common than on unsharded. However, the reasons for the result was not explored as the experiment was designed to address a different research question. Voorhees et al. [39] randomly split a collection in half. The authors stated that the two resulting shards allowed more accurate performance measurement. However, it was reported that splitting the collection further did not improve accuracy; reasons for no improvement were not examined in detail.

We describe research that takes the Robertson and Kanoulas view that topics have unequal value and combines it with the ANOVA approach of Ferro and Sanderson [11] and the sharding method of Voorhees et al. [39]. We ask: Can the unequal value of topics be exploited to improve measurement of system performance accuracy on a test collection? We make the following contributions:

- We validate an ANOVA model via a theoretical examination, showing why explicitly accounting for differences across topics yields accuracy improvements.
- We experimentally show that the model identifies notable numbers of significant differences between systems.
- We experimentally show that the differences are not due to measurement error of the significance formulas.

Next, related work is described followed by ANOVA models and their properties. The setup and report of experimental findings are described before conclusions and future work are detailed.

2 RELATED WORK
We review three research areas: topics with few relevance judgments, ANOVA modeling, and the sharding of collections.

2.1 Topics with few relevance judgments
There is an assumption, in test collection based evaluation, that all topics are valuable equally. Performance is measured by taking the arithmetic mean of topics scores. Swanson [33] described such a process in 1960. When the mean is taken, each topic score contributes equally regardless of the accuracy of that measure. The potential
for error was described by Voorhees [36]: "When [topics] have very few relevant documents (fewer than five or so), summary evaluation measures such as average precision are themselves unstable; tests that include many such queries are more variable". Soboroff [32] pointed out that rank cut off evaluation measures (e.g. precision at 10) will have an upper bound < 1 for topics with few relevant documents.

The notion that not all topics have equal value was implicitly exploited in work identifying a subset of test collection topics that rank systems similarly to a full topic set [13, 19]. To the best of our knowledge, however, Cormack and Lynam [9] were the first to incorporate an unequal view of topics into test collection measurement. They treated each topic as a "separate test", calculating topic confidence intervals using a bootstrap approach. Topics with ≤ 5 relevant documents were subject to a "Small-R Correction" to overcome measurement instability.

Robertson [25] considered the broader question of what is the "per-topic noise or error" present in the topics of a test collection. The paper considered if evaluation measures could be adapted to cope with an unequal view of topics. Later, Robertson and Kanoulas [26] measured the variance of topic scores by bootstrapping from the document collection. The researchers found that topics showed different levels of variance, but the variance was relatively consistent across systems. The researchers described a significance test that incorporated topic score variation. Comparisons between the new test and the commonly used t-test showed some differences in the conclusions one might draw when comparing systems.

More recently, Yang et al. [41] examined how much rankings of systems were affected by per-topic score variance and if there was any impact on significance tests. They found that the variance did not affect overall rankings notably, but that the number of significant differences observed between systems dropped.

Note, there is much research on subjects such as query difficulty prediction [42], topic score normalization [40], average average precision [20], GMAP [24], etc. Such work focuses on so-called difficult topics, we focus on topics for which measurement is variable.

### 2.2 ANOVA modeling

ANOVA can decompose the data of an IR experiment into a model of factors, into interactions between those factors, and into a level of unmodeled error. Tague-Sutcliffe and Blustein [34] described an example of this approach by comparing the variation in performance across two factors: topics and systems. The former was found to be larger than the latter. Measurement of interaction between topics and systems was not possible owing to a lack of replicates of topic*system measures. Banks et al. [2] approximated such an interaction, suggesting it would be strong and significant. Later, Bodoff and Li [3] used a test collection with multiple relevance assessments to obtain the required replicates. The authors reported that the magnitude of the topic*system interaction factor was less than the topic factor, but greater than the system factor.

Both Ferro and Sanderson [11] and Voorhees et al. [39] generated replicates by sharding a document collection. This enabled them to measure the topic*system effect. We describe that work next.

#### 2.3 Sharding

Voorhees et al. [39] used a bootstrap ANOVA approach that drew on a sample of the scores of topics measured across different systems and shards. The researchers tested on the TREC-3, TREC-8, and 2006 Terabyte track collections. Success of the approach was measured by counting the number of significant differences found between systems submitted to TREC tracks. The researchers found substantially more such differences were measured than with conventional approaches. Two shards were used. When three or five shards were tried, the researchers found the number of significant differences dropped, the reasons for which were not examined in detail. The relative impact of each component of the technique – bootstrap ANOVA, the approach to multiple comparisons, and sharding method – was not described.

As part of a study on the interaction between different types of shards and system scores, Ferro and Sanderson [11] described a series of ANOVA models tested on the TREC-7 and TREC-8 adhoc test collections. Like the previous research, the value of these models was quantified by the number of significant differences measured between systems. The researchers showed that a more sophisticated ANOVA model produced the highest number of significant differences measured between systems. However, the shards were very skewed in size.

The research described shows that the topics of test collections can produce scores of different variance, which can impact the measurement of significance between systems. There is, as yet, not an extensive body of research examining such topic variability. Most work has explored bootstrap approaches from document collections to assess the variance. The recent examination of sharding has not been explored in conjunction with the work on topic variability. We explore the connection between these two lines of inquiry examining the style of ANOVA modeling used by Ferro and Sanderson [11]. We also measure the accuracy of the model across a range of sharding configurations that have not been examined before.

### 3 METHODOLOGY

Suppose we have T topics, R systems, and S shards and thus N = T · R · S total samples. We can form the following six ANOVA models:

\[
y_{ijk} = \mu + \tau_i + \alpha_j + \beta_k + \epsilon_{ijk} \quad \text{(MD1)}
\]

**Main Effects**

\[
y_{ijk} = \mu + \tau_i + \alpha_j + \epsilon_{ijk} \quad \text{(MD2)}
\]

**Main Effects**

\[
y_{ijk} = \mu + \tau_i + \alpha_j + (\tau\alpha)_ij + \epsilon_{ijk} \quad \text{(MD3)}
\]

**Main Effects + Interaction Effects**

\[
y_{ijk} = \mu + \tau_i + \alpha_j + \beta_k + (\tau\beta)_ijk + \epsilon_{ijk} \quad \text{(MD4)}
\]

**Main Effects + Interaction Effects**

\[
y_{ijk} = \mu + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_ij + (\alpha\beta)_jk + \epsilon_{ijk} \quad \text{(MD5)}
\]

**Main Effects + Interaction Effects**

\[
y_{ijk} = \mu + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_ij + (\tau\beta)_ijk + (\alpha\beta)_jk + \epsilon_{ijk} \quad \text{(MD6)}
\]
Where:

- $y_{ijk}$ is the performance score of three factors, the $i$-th topic ($i = 1, \ldots, T$) retrieving on the $j$-th system ($j = 1, \ldots, R$) from the $k$-th shard ($k = 1, \ldots, S$);
- $\mu$ is the grand mean;
- $\tau_i = \mu - \mu$ is the effect of the $i$-th topic, where $\mu$ is the marginal mean of the topic;
- $\alpha_j = \mu_j - \mu$ is the effect of the $j$-th system, where $\mu$ is the marginal mean of the system;
- $\beta_k = \mu_k - \mu$ is the effect of the $k$-th shard, where $\mu$ is the marginal mean of the shard;
- $(\tau_\alpha)_{ij} = \mu_{ij} - \mu_i$ is the interaction between topics and systems, where $\mu_i$ is the marginal mean of the interaction between the $i$-th topic and $j$-th system;
- $(\tau_\beta)_{ik} = \mu_{ik} - \mu_k$ is the interaction between topics and shards, where $\mu_k$ is the marginal mean of the interaction between the $i$-th topic and $k$-th shard;
- $(\alpha_\beta)_{jk} = \mu_{jk} - \mu_j - \mu_k$ is the interaction between systems and shards, where $\mu_{jk}$ is the marginal mean of the interaction between the $j$-th system and $k$-th shard; and
- $\epsilon_{ijk}$ is the error of the model in predicting $y_{ijk}$.

Model (MD1) was used by Tague-Sutcliffe and Blustein [34] and Banks et al. [2]. It can be viewed as a classic approach to measuring significance on a test collection, as in this form, it is operationally similar to a t-test. The model components have two subscripts ($i$, $j$) because the collection does not have shards. Model (MD2) is model (MD1) but with shards. While model (MD1) has only one performance score for each (topic, system) pair, in model (MD2) the shards provide replicates scores for the pairs when estimating the model parameters.

The presence of replicates is exploited in model (MD3) by adding a topic*system interaction factor. Model (MD3) was used by Robertson and Kanoulas [26] and Voorhees et al. [39], though Voorhees et al. did not rely on classical ANOVA, instead adopting a bootstrap approach [10]. Model (MD4) explicitly accounts for a shard factor and model (MD5) adds the interaction between systems and shards. Both models are close to models proposed by Ferro and Sanderson [11], but they omitted the topic*system interaction in their models.

Model (MD6) adds a topic*shard interaction, by leveraging the presence of more replicates for each (topics, shard) pair - there are as many replicates as the number of used systems $R$. It is the focus on our work here.

3.1 Exploiting topic variability with the model

How does improved measurement accuracy arise from a more sophisticated ANOVA model applied over a test collection whose documents are randomly split into shards? Models add more factors with the goal of better fitting the data. Since the total Sum of Squares (SS) is the same for all models, each new factor should explain a further part of the total SS. As a consequence, there is a reduction of the error SS, i.e. the leftover unexplained by the model, and, broadly speaking, this leads to a more accurate estimate.

How does model (MD6) exploit the variable measurement of topics? With random even sized shards, the probability of having relevant documents in a shard is uniform across the shards. This probability is smaller for topics with fewer relevant documents and greater for topics with more relevant documents. Therefore, for each topic, the number of shards without any relevant documents is proportional to the number of relevant documents for that topic. Model (MD6) accounts for this by explicitly considering $(\tau_\beta)_{ik}$, i.e. the topic*shard interaction effect. When there are no relevant documents for a topic on a given shard, we set the score to undefined for all the systems with respect to that topic on that shard. The more shards without relevant documents for a topic, the more undefined values there are, which is reflected in the estimation of the $(\tau_\beta)_{ik}$ factor. Therefore, the estimation of the SS of the topic*shard interaction factor directly removes from the total SS the variability due to these intrinsic differences among topics, reducing the error SS and giving us the possibility of a more accurate estimation of the differences among systems. Instead of seeing shards as a mere “technical trick” to obtain replicates, we can look at them as a form of “diagnostic tool”, which allows us to systematically probe measurement differences across topics and to account for the differences in a model.

We next consider a series of questions about model (MD6):

- How do different models affect the significant differences among systems, accounting for multiple comparisons?
- How do we compute confidence intervals from the model?
- How do we estimate effect size?
- Is it legitimate to use undefined values?

The following sections will answer the questions by showing that model (MD6) provides benefits in all these areas and, most importantly, makes estimations concerning the system factor independent of undefined values due to the sharding process.

4 MULTIPLE COMPARISONS

If one simultaneously compares multiple system pairs, the probability of committing a Type I error increases and the Family-wise Error Rate (FWER) (the probability of committing at least one Type I error) is $\text{FWER} = 1 - (1 - \alpha)^c$, where $c$ is the total number of comparisons to be performed [15, pp. 7–8]. It is crucial to control Type I errors when performing multiple comparisons [4, 6, 29].

Tukey [35] proposed the Honestly Significant Difference (HSD) test, which creates confidence intervals for all pairwise differences between factor levels, while controlling the FWER. Two systems $u$ and $v$ are considered significantly different when:

$$|t_k| = \frac{|\hat{\mu}_u - \hat{\mu}_v|}{\sqrt{\frac{\text{MSError}}{T}}} > Q^\alpha_{R,d_{\text{error}}}$$

where: $\hat{\mu}_u$ and $\hat{\mu}_v$ are the marginal means of the systems $u$ and $v$ as estimated from the actual data; $d_{\text{error}}$ are the Degrees of Freedom (DF) of the error; $\text{MSError}$ is the Mean Squares (MS) of the error, i.e. an estimation of the variance left unexplained; and $Q^\alpha_{R,d_{\text{error}}}$ is the upper 100 × (1 − $\alpha$)-th percentile of the studentized range distribution [22]. Note, that in the case of the model (MD1) the denominator of eq. (1) becomes just $T$, since the whole corpus is constituted by a single shard and thus $S = 1$. 

We have also examined different sharding approaches and how they impact the effect size of ANOVA model factors [12]. That paper does not examine in detail the impact of the model on significance tests.
Figure 1: Studentized range distribution $Q_{R,d_{error}}$ for different numbers of systems to be compared and different degrees of freedom of the error. The lines in each plot correspond to different DF of the error $Q^a_{R,d_{error}}$ for $a = 0.05$; red lines are for $100$ DF, blue lines are for $500$ DF. Solid lines are for $R = 5$ systems; dashed lines are for $R = 25$ systems; and, dotted lines are for $R = 75$ systems.

Figure 1 shows the Cumulative Density Function (CDF) of the Studentized range distribution for different numbers of compared systems and different values of the DF of the error. The DF lines are almost superimposed on each other. The values of $Q^a_{R,d_{error}}$ are equal, apart from the lower values of DF where they are marginally different. The main difference across plots is that increasing the number of systems to be compared shifts the CDF to the right. In a typical IR setting where $R$ systems are compared, the factor $Q^a_{R,d_{error}}$ in eq. (1) is practically constant. As a consequence, even if models from (MD1) to (MD6) lead to different values $d_{error}$, the models “see” the same value of $Q^a_{R,d_{error}}$ and, therefore, the size of the interval needed to consider two systems as significantly different mostly depends on the factor $\sqrt{MS_{error}/S}$. In models (MD2) to (MD6), the marginal means $\hat{\mu}_u$ and $\hat{\mu}_v$ of the compared systems are the same as well as the $T \cdot S$ factor; therefore, differences in the sizes of the intervals are due only to the $\sqrt{MS_{error}/S}$ factor. Since the typical benefit of having richer models is to reduce the size of the error, we expect $MS_{error}$ to decrease and, consequently, the test statistic $|tk|$ increases, allowing us to detect more significant differences. The increasingly richer models lead to a more accurate estimate of the actual differences among systems. Moreover, the $MS_{error}$ is further divided by $T \cdot S$, which suggests that, for a given number of topics $T$, increasing the number of shards $S$ should provide further benefits.

The test statistic $|tk|$ allows us to compute the $p$-value

$$p = P\left(Q_{R,d_{error}} \geq |tk|\right)$$

of observing a more extreme value of the Studentized range distribution. We can then compare this $p$-value to the desired significance level $\alpha$ and, if it is $\leq \alpha$, the two systems $u$ and $v$ are significantly different, still controlling the FWER. Eqs. (1) and (2) are two equivalent ways to perform multiple comparisons controlling the FWER.

$^2$Strictly, the SS of the error decreases because the additional factors in a model explain more of the total SS, leaving less to the SS of the error. However, $MS_{error} \approx \frac{SS_{error}}{S}$, if a richer model causes a drop in $d_{error}$, this decreased denominator may lead to a greater $MS_{error}$, even if $SS_{error}$ is decreased. However, as a first approximation, it is enough to consider both quantities as decreasing as we add factors to a model.

5 CONFIDENCE INTERVALS

We consider three types of confidence interval.

5.1 Tukey

The Tukey HSD test of eq. (1) allows us to define exact confidence intervals for the system main effects, still controlling the FWER. Hochberg and Tamhane [15] suggest creating a half-width confidence interval around the marginal mean of a system $u$

\[
\hat{\mu}_u \pm \frac{1}{2}Q_{R,d_{error}}^a \sqrt{\frac{MS_{error}}{T \cdot S}}
\]

Systems $u$ and $v$ are significantly different, according to the Tukey HSD test of eq. (1), if and only if their confidence intervals of eq. (3) do not overlap [15, p. 116]. From model (MD2) to (MD6), we expect that confidence intervals will reduce as $MS_{error}$ decreases.

5.2 Standard Error of the Mean

The confidence interval of eq. (3) differs from the typical confidence interval based on the Standard Error of the Mean (SEM):

\[
\hat{\mu}_u \pm t^*_{R-1} \sqrt{\frac{\hat{\sigma}^2}{T \cdot S}}
\]

where $\hat{\sigma}^2_u = \frac{1}{T \cdot S - 1} \sum_{i=1}^{T} \sum_{j=1}^{S} (y_{ijk} - \hat{\mu}_u)^2$ is the sample variance of the $u$-th system and $t^*_{R-1}$ is the upper 100th $(1 - \alpha/2)$-th percentile of the Student’s $t$ distribution with $T \cdot S - 1$ degrees of freedom. Note, these are the confidence intervals used by Ferro and Sanderson [11] when showing the improved accuracy due to the use of shards.

Differently from the confidence interval of eq. (3), those of eq. (4) do not depend on any of the more accurate ANOVA models, they just depend on the underlying data. Moreover, they do not account for any multiple comparison adjustment since they consider each system in isolation. While the confidence intervals of eq. (3) have the same size for all systems as they need to control for FWER, the confidence intervals of eq. (4) change size from system to system as they depend on the sample variance of each system.

5.3 ANOVA

We can define the following confidence interval [29, p. 57], which falls between those of eq. (3) and those of eq. (4)

\[
\hat{\mu}_u \pm t^*_{d_{error}} \sqrt{\frac{MS_{error}}{T \cdot S}}
\]

As with eq. (3), the interval depends on the ANOVA model and its ability to explain the data. As with eq. (4), the interval does not adjust for multiple comparisons. Different from eq. (4) but similar to eq. (3), the interval has the same size for all systems. As above, the term $t^*_{d_{error}}$ is practically constant, following the discussion about eq. (3), we expect the confidence interval of eq. (5) to reduce either as the ANOVA models become richer or if we use more shards.

The difference between eq. (3) and eq. (5) is the replacement of $t^*_{R-1}$ with $t^*_{d_{error}}$. The former is typically 2-3 times bigger than the latter. The bigger the difference, the bigger the number of systems $R$ to be compared. This lets us understand the magnitude of adjustment needed to keep the FWER controlled. Consequently, the confidence intervals of eq. (3) are bigger than those of eq. (5).
Shard F
i
undefined
ements are missing, a whole "row" is filled in with
topic
topic. Similarly, topic
and, therefore, all the systems have the
and
ferences, the calculation of confidence intervals, and the effect size
both approaches fail as the number of shards increase.

In Figure 2, we have \( X_1 = \{1, 4\} \) and \( X_2 = \{3, 4\} \). Note that, for
any shard \( k \), there are \(|X_k| \cdot R \) undefined values and, in total, there are
\( R \sum_{k=1}^{\infty} |X_k| \) undefined values.

Proposition 7.2. Given models from (MD2) to (MD6) and a sys-
tem \( j \in [1, R] \), its estimated marginal mean is given by:

\[
\hat{\mu}_{j} = \frac{1}{T \cdot S} \sum_{k=1}^{S} \sum_{i \in X_k} y_{ijk} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k| \tag{8}
\]

Therefore, for any pair of systems \( u \in [1, R] \) and \( v \in [1, R], u \neq v \), the difference of their estimated marginal means \( \hat{\mu}_u - \hat{\mu}_v \) is independent of the undefined values.

Note, that the first element \( \hat{\mu}_j \) of eq. (8) is the estimated marginal mean of the system factor ignoring the undefined values. This is not the estimated marginal mean removing undefined values, since the denominator \( T \cdot S \) still accounts for all the values, both defined and undefined. The second element is the contribution to estimated marginal mean due only to the undefined values. It is constant and equal for all the systems. Therefore, the regularity in the pattern of undefined values allows us to separate the contributions due to the systems from those due to undefined values, which are the same for all the systems. Proposition 7.2 has three consequences:

1. The numerator of eq. (1), i.e. the multiple comparison among systems, is not affected by the undefined values.
2. Eq. (8) shows that the shift due to undefined values is the same for all the systems and, therefore, does not affect the Rankings of Systems (RoS), i.e. the ordering of the system by their estimated marginal mean. If a Kendall’s \( \tau \) correlation [17] was measured between the RoS on the whole corpus and the RoS when using shards, \( \tau \) is not affected by the undefined values.
3. For each shard we could have at worst \(|X_k| = T\), i.e. a shard for which no topic has relevant documents. However, test collections generally have at least one relevant document for each topic and, since shards are a partition of the whole corpus, it follows that \(|X_k| < T\). Therefore \( \frac{1}{T} \sum_{k=1}^{S} |X_k| \) is always strictly < 1. The effect of the undefined value is to shift the estimated marginal mean of the system factor by a fraction of that undefined value. From this perspective, setting \( x = 0 \), our choice in the experimentation, is not lowering the mean system performance but just leaving them at their level.

Proposition 7.3. Given models from (MD2) to (MD6), the SS of the system factor and, as a consequence, the MS of the system factor are independent of the undefined values.

Proposition 7.4. Given model (MD6), the residuals \( e_{ijk} \) are independent from the undefined values. Therefore, the SS of the error and, as a consequence, the MS of the error are independent of the undefined values.

Note that Proposition 7.4 holds only in the case of model (MD6) and only thanks to the topic × shard interaction \((\tau \beta)_{ik}\) factor.
Indeed, as shown in the appendix, all the estimated marginal means have a form similar to eq. (8), i.e. a mean contribution due to defined values plus a mean contribution due to undefined values. However, only the topic*shard interaction \((\tau \beta)_{ik}\) has the form

\[
\hat{\mu}_{ik} = \begin{cases} 
\hat{\mu}^*_{ik} & \text{if } i \notin X_k \\
\hat{\mu}^*_{ik} x_i & \text{if } i \in X_k 
\end{cases}
\]

which cancels out the undefined values when \(y_{ijk} = x\) and makes the residuals \(e_{ijk}\) independent from them. In this sense, in Section 3.1, we said that the topic*shard interaction \((\tau \beta)_{ik}\) is the factor dealing with the intrinsic differences among topics, since it is able to separate defined from undefined values. As we discussed, the number of undefined values is proportional to the number of relevant documents for a topic and, therefore, the topic*shard interaction \((\tau \beta)_{ik}\) factor accounts for the unequal value of topics.

Therefore, model (MD6) is not only a more precise model because, thanks to the more factors it considers, it is able to explain more variance than all the other models, leading to more accurate cause, thanks to the more factors it considers, it is able to explain more variance than all the other models.

\[\text{ Indeed, as shown in the appendix, all the estimated marginal means have a form similar to eq. (8), i.e. a mean contribution due to defined values plus a mean contribution due to undefined values. However, only the topic*shard interaction } \( (\tau \beta)_{ik} \) \text{ has the form } \]

\[
\hat{\mu}_{ik} = \begin{cases} 
\hat{\mu}^*_{ik} & \text{if } i \notin X_k \\
\hat{\mu}^*_{ik} x_i & \text{if } i \in X_k 
\end{cases}
\]

which cancels out the undefined values when \(y_{ijk} = x\) and makes the residuals \(e_{ijk}\) independent from them. In this sense, in Section 3.1, we said that the topic*shard interaction \((\tau \beta)_{ik}\) is the factor dealing with the intrinsic differences among topics, since it is able to separate defined from undefined values. As we discussed, the number of undefined values is proportional to the number of relevant documents for a topic and, therefore, the topic*shard interaction \((\tau \beta)_{ik}\) factor accounts for the unequal value of topics.

Therefore, model (MD6) is not only a more precise model because, thanks to the more factors it considers, it is able to explain more variance than all the other models, leading to more accurate estimates of the differences among systems. But, especially, it is also the model with the most desirable properties, thanks to the presence of the topic*shard interaction \((\tau \beta)_{ik}\) factor. Indeed, Proposition 7.4 has two consequences:

1. The denominator of eq. (1) is independent of the undefined values. This, jointly with Proposition 7.2, means that undefined values do not affect the identification of significantly different systems. Consequently, the confidence intervals of eq. (3) are independent from the undefined values. The same holds for the confidence intervals of eq. (5).
2. Recall that the F-statistic of the system factor is given by

\[ F_{\text{system}} = \frac{MS_{\text{system}}}{MS_{\text{error}}} \]

where \(MS_{\text{system}} = SS_{\text{system}} / df_{\text{system}}\) and \(MS_{\text{error}} = SS_{\text{error}} / df_{\text{error}}\). Since both \(SS_{\text{system}}\) (Proposition 7.3) and \(SS_{\text{error}}\) (Proposition 7.4) are independent from the undefined values, it follows that the F-statistic of the system factor is also independent from undefined values. They do not affect the significance of this factor. Moreover, it follows that the effect size of the system factor \(\hat{\omega}^2_{\text{system}}\) of eq. (6) is independent of the undefined values.

8 EXPERIMENTAL SETUP

To empirically test the analyses above, we experimented on the collections, topics, and system runs of the following datasets:

- Adhoc track T08 [38]: 528,155 documents of the TIPSTER disks 4-5 corpus minus congressional record (TIP); 50 topics, each with binary relevance judgments drawn from a pool depth of 100; 129 system runs retrieving 1,000 documents for each topic.
- Web track T09 [14]: 1,692,096 documents of the WT10g Web corpus; 50 topics, each with multi-graded relevance judgments and a pool depth of 100; 104 system runs retrieving 1,000 documents for each topic.
- Common Core track T27 [1]: 595,037 documents of the Washington Post corpus (WAPO); 50 topics, each with multi-graded relevance judgments; relevance judgments were obtained mixing depth-10 pools with multi-armed bandit [18, 37], stratified sampling [7] and move-to-front [8] approaches; 72 system runs retrieving 10,000 documents for each topic.

We any mapped multi-graded relevance judgments to binary by treating everything above not relevant as relevant.

For each corpus, we created \(S\) randomly formed even sized shards, where \(S \in \{2, 3, 4, 5, 10, 25, 50\}\). We label the shards of a corpus as \(<\text{corpus}_n>_{\text{RNDE}S}\); e.g., the WAPO corpus split into 5 shards is labeled \(\text{WAPO_RNDE}_5\). For each shard size, we re-sampled 10 times; i.e., in the case of \(\text{WAPO_RNDE}_5\) we have 10 independent sets of 5 random even size shards on the WAPO corpus. For space reasons, we report only some combinations of measures and tracks but the observed trends hold also for the other results.

For each corpus split into shards, system runs retrieving from the corpus were also sharded. A run was split into the same number of shards as the corresponding corpus. The random document split used to shard a corpus was the same split used to shard a run. Such splitting is a simulation of how a system would retrieve documents on each shard. Past empirical work showed the simulation to work well Sanderson et al. [31].

We consider the following evaluation measures: Average Precision (AP) [5], Precision at ten retrieved documents (P@10), Rank-Biased Precision (RBP) [21], and Normalized Discounted Cumulated Gain (nDCG) [16]. We calculated RBP by setting \(p = 0.8\) as persistence parameter while we use a log2 discounting function in nDCG, to consider not too impatient users. We considered \(\alpha = 0.05\) to determine if a factor is statistically significant. Our experimental source code is at: https://bitbucket.org/ffrncl/sigir2019-fs-code/.

9 EXPERIMENTS

We conduct three experiments.

9.1 Confidence Intervals

We study the three types of confidence intervals under different ANOVA models. Figure 3 compares the intervals on the whole corpus using model (MD1) and on three shards using model (MD6).

In the case of the whole corpus and model (MD1) in Figure 3a, we see that, as expected, the Tukey confidence intervals (eq. (3)) are larger than the ANOVA ones (eq. (5)) since the latter do not account for multiple comparisons. We also see that the Tukey intervals of eq. (3) are similar to the SEM intervals (eq. (4)), which are independent from any model of the data and just consider each system in isolation. The fact that model-dependent confidence intervals
(Tukey ones) look close to model-independent ones (SEM ones) suggests that the topic and system factors of model (MD1) are not enough to accurately explain the data.

When using the shards (Figure 3b), we note that both the Tukey and ANOVA confidence intervals are smaller than SEM, suggesting that model (MD6) better explains the underlying data thanks to the additional factors it considers.

Note, that this difference between model (MD1) and (MD6) is not due to the increased number of samples passing from the whole corpus to shards but to the better ability of model (MD6) to explain the data. Indeed, the additional beneficial effect of increasing the number of samples is apparent in Figure 3b from the fact that all the confidence intervals get smaller when using shards, but this would happen for whatever model.

Figure 4 shows how the Tukey confidence intervals change across different models. The black dotted line is the system performance (marginal mean of the system $a_j$ factor) on the whole corpus, i.e. the same line shown in Figure 3a in the case of AP. The continuous line is the system performance (marginal mean of the system $a_j$ factor) on shards, i.e. the same line shown in Figure 3b in the case of AP; note that the green line for model (MD2), the orange one for model (MD3), and the red one for model (MD6) are superimposed since the marginal mean of the system $a_j$ factor is the same in all these models. The shaded areas in the color of the line of each model represent the Tukey confidence interval for the corresponding model; for example, gray shaded area is for model (MD1) while the red shaded area is for model (MD6).

For all measures, the confidence interval using model (MD1) on the whole corpus is bigger than the confidence interval when using the other models. In particular, comparing the confidence intervals of models (MD1) and (MD2), which are computed without and with shards respectively. Comparing models (MD2), (MD3), and (MD6), we see the increasingly complex models improve the accuracy by shrinking the confidence interval. Moreover, comparing model (MD3) to model (MD6) we see that adding shard*system and topic*shard factors substantially reduce the intervals.

We report the Kendall’s $\tau$ correlation between the RoS on the whole corpus and on shards in the title of the plots in Figure 4. We can see that in three of the four plots, $\tau > 0.9$, the empirical threshold used to consider to ranking equivalence [36]. This suggests that we are not only improving accuracy but also maintaining coherence with what happens in traditional analyses.

Figure 5 compares the Tukey confidence intervals of eq. (3) for different shard numbers using model (MD6) in the case of AP on T08. As expected, the confidence intervals tend to reduce as the number of shards increases, due to the increased number of measurements on the shards. Kendall’s $\tau$ remains $> 0.9$, suggesting that the increased number of shards does not substantially deteriorate the agreement of the RoS on the whole corpus.

### 9.2 Multiple Comparisons

Table 1 reports summary statistics for multiple comparison analyses on T08 using different splits for AP. We observe a large system effect size ($\hat{\omega}^2_{(sys)}$). We also can see a drop in $\hat{\omega}^2_{(sys)}$ passing from model (MD1), i.e. the whole corpus, to model (MD2), i.e. the same model but using shards. The shards appear to introduce a new factor, which interacts with the other factors and thus the size of $\hat{\omega}^2_{(sys)}$ reduces. However, as the models account for more factors ((MD2)-(MD6)), $\hat{\omega}^2_{(sys)}$ increases, suggesting that the more a model explains the data, the more prominent $\hat{\omega}^2_{(sys)}$ becomes. In the case of model (MD6) and for fewer shards, $\hat{\omega}^2_{(sys)}$ can be notably bigger than on the whole corpus.

Considering the number of significantly different pairs (columns Sig and NotSig), we see how moving from (MD1) – a classic significance testing approach – to any shard-based model always increases the number of pairs. More shards also means more significantly different pairs. However, there is a limited gain in using more shards: in the case of model (MD2) passing from two to five shards gives a 13.28% increase in the number of pairs but passing from five to ten produces only a 0.15% gain. More complex models are less sensitive to the increase in the number of shards, since they detect almost all the significantly different pairs already at a low number of shards. For example, in the case of model (MD6) passing from two to five shards gives just a 0.78% increase in the number of significantly different pairs while passing from five to ten produces a 0.20% increase.

The more sophisticated a model, the more significant differences are detected. However, not all models are equally impactful. From model (MD2) to (MD3), i.e. adding the topic*system interaction, produces notable increases while passing from model (MD3) to (MD4) and (MD5), do not provide substantial benefits. However, model (MD6), i.e. adding the topic*shard interaction, makes another substantial increase in the number of significant differences, confirming the importance of this factor.

If we consider the group of the systems insignificantly different from the top performing system (column TopG), we can appreciate another benefit of using shards. The number of systems in the top group drops from 7 when using the whole corpus to 1 when using shards and the more descriptive models, suggesting that the increased accuracy in estimating differences among systems allows us to detect that the top system is actually different from others.

### 9.3 Robustness to Shard Sampling

Table 2 show the summary of the analyses for AP across different shard sizes when using ten samples for each shard size. The Kendall’s $\tau$ column reports the average value of $\tau$ over the samples and its 95% confidence interval. For all the tracks, the $\tau$ values are quite high with small confidence intervals. This suggest that the RoS is quite stable and does not depend much on the specific random shards. Similar considerations hold also in the case of the Tukey confidence interval, which gets smaller as the shard size increases and whose values are similar across shard samples. This suggests that the detection of significantly different systems is not affected much by the specific random shards at hand.

The total number of significantly different pairs support this hypothesis since we can see how the confidence interval around this value is small, indicating that their number does not change much when the shard sample changes. The final column reports the fraction of significant pairs found in common across all 10 samples. Here, there is a notable level of consistency across the samples.
10 CONCLUSIONS AND FUTURE WORK

At the start of the paper, we asked: can an unequal value of topics be exploited to improve measurement of system performance accuracy on a test collection?

We described and validated, theoretically and empirically, an ANOVA model combined with a random sharding technique. We showed that the model (MD6) measures substantially more significant differences between IR systems than conventional approaches, as represented by model (MD1). While it is true that a more sophisticated ANOVA model is expected to reduce measurement error, the scale of improvement seen with (MD6) is perhaps less expected. We showed that model (MD6) agrees well with the RoS taken from conventional test collection measurement and that the increased significance is not due to measurement error.

Past work has examined the question of whether the variability of topic measurement can be exploited to improve the accuracy of IR system measurement, we contend that our research shows that...
### Table 1: Comparing models for three shard sizes across 8256 system pairs, AP, track T08.

| Model | vs Model | $\omega^2_{y|x}$ | $\omega^2_{y|x}$ | Sig | Not Sig | Top G | Sig | Not Sig | Top G |
|-------|----------|------------------|------------------|-----|---------|--------|-----|---------|--------|
| TIP_RNDE_05 | TIP_RNDE_02 | 0.05991 | 3425 | 4833 | 0.05991 | 3425 | 4833 | 0.05991 | 3425 | 4833 |
| MD1 | MD2 | -12.29% +18.81% -15.33% -42.86% -35.95% -34.59% -24.50% -71.43% -60.02% -33.74% -23.64% -71.43% |
| MD3 | MD4 | 0.5657 | 5175 | 3081 | 0.3495 | 5135 | 3123 | 0.1840 | 4831 | 3425 |
| MD5 | MD6 | 0.5657 | 5175 | 3081 | 0.3495 | 5135 | 3123 | 0.1840 | 4831 | 3425 |

### Table 2: Summary of analyses for AP using 10 samples of each random split and model (MD6).

<table>
<thead>
<tr>
<th>Split</th>
<th>T08 – 8256 system pairs compared</th>
<th>CI Width</th>
<th>Sig. Pairs</th>
<th>Frac. Sig. Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIP_RNDE_02</td>
<td>0.9609</td>
<td>0.9504</td>
<td>3122.20</td>
<td>0.6225</td>
</tr>
<tr>
<td>TIP_RNDE_03</td>
<td>0.9745</td>
<td>0.9551</td>
<td>5083.90</td>
<td>0.6160</td>
</tr>
<tr>
<td>TIP_RNDE_04</td>
<td>0.9680</td>
<td>0.9546</td>
<td>5104.10</td>
<td>0.6182</td>
</tr>
<tr>
<td>TIP_RNDE_05</td>
<td>0.9689</td>
<td>0.9549</td>
<td>3051.20</td>
<td>0.6118</td>
</tr>
<tr>
<td>TIP_RNDE_10</td>
<td>0.9633</td>
<td>0.9520</td>
<td>3010.90</td>
<td>0.6060</td>
</tr>
<tr>
<td>TIP_RNDE_25</td>
<td>0.9418</td>
<td>0.9445</td>
<td>2522.80</td>
<td>0.6350</td>
</tr>
<tr>
<td>TIP_RNDE_50</td>
<td>0.9189</td>
<td>0.9351</td>
<td>5462.40</td>
<td>0.6616</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Split</th>
<th>T09 – 5356 system pairs compared</th>
<th>CI Width</th>
<th>Sig. Pairs</th>
<th>Frac. Sig. Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT06_RNDE_02</td>
<td>0.9069</td>
<td>0.9052</td>
<td>2288.30</td>
<td>0.5243</td>
</tr>
<tr>
<td>WT06_RNDE_03</td>
<td>0.9453</td>
<td>0.9717</td>
<td>2874.00</td>
<td>0.5366</td>
</tr>
<tr>
<td>WT06_RNDE_04</td>
<td>0.9300</td>
<td>0.9663</td>
<td>2947.70</td>
<td>0.5594</td>
</tr>
<tr>
<td>WT06_RNDE_05</td>
<td>0.9275</td>
<td>0.9657</td>
<td>3034.50</td>
<td>0.5666</td>
</tr>
<tr>
<td>WT06_RNDE_10</td>
<td>0.9037</td>
<td>0.9530</td>
<td>3426.80</td>
<td>0.6398</td>
</tr>
<tr>
<td>WT06_RNDE_25</td>
<td>0.8813</td>
<td>0.9389</td>
<td>3478.00</td>
<td>0.6998</td>
</tr>
<tr>
<td>WT06_RNDE_50</td>
<td>0.8675</td>
<td>0.9288</td>
<td>3891.60</td>
<td>0.7270</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Split</th>
<th>T27 – 25 system pairs compared</th>
<th>CI Width</th>
<th>Sig. Pairs</th>
<th>Frac. Sig. Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAPU_RNDE_02</td>
<td>0.9764</td>
<td>0.9646</td>
<td>1821.50</td>
<td>0.7126</td>
</tr>
<tr>
<td>WAPU_RNDE_03</td>
<td>0.9634</td>
<td>0.9495</td>
<td>1791.20</td>
<td>0.7008</td>
</tr>
<tr>
<td>WAPU_RNDE_04</td>
<td>0.9617</td>
<td>0.9485</td>
<td>1801.10</td>
<td>0.7043</td>
</tr>
<tr>
<td>WAPU_RNDE_05</td>
<td>0.9583</td>
<td>0.9480</td>
<td>1802.70</td>
<td>0.7053</td>
</tr>
<tr>
<td>WAPU_RNDE_10</td>
<td>0.9470</td>
<td>0.9460</td>
<td>1822.80</td>
<td>0.7131</td>
</tr>
<tr>
<td>WAPU_RNDE_25</td>
<td>0.9219</td>
<td>0.9410</td>
<td>1848.30</td>
<td>0.7231</td>
</tr>
<tr>
<td>WAPU_RNDE_50</td>
<td>0.8812</td>
<td>0.9337</td>
<td>1833.30</td>
<td>0.7251</td>
</tr>
</tbody>
</table>

### 11 ACKNOWLEDGMENTS

This research is supported in part by the Australian Research Council’s Discovery Projects Scheme (DP180102687). The work is also partially funded by the “Data Benchmark for Keyword-based Access and Retrieval” (DAKKAR) Starting Grants project sponsored by University of Padua and Fondazione Cassa di Risparmio di Padova e Rovigo.
REFERENCES


Improving the Accuracy of System Performance Estimation by Using Shards

Electronic Appendix

Nicola Ferro
ferro@dei.unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Mark Sanderson
mark.sanderson@rmit.edu.au
Computer Science, School of Science, RMIT University
Melbourne, Australia

CCS CONCEPTS
• Information systems → Test collections; Retrieval effectiveness;

KEYWORDS
effectiveness model; ANOVA; multiple comparison

A BACKGROUND ON ANOVA
Analysis Of Variance (ANOVA) [2, 3] attempts to explain data (the dependent variable scores) in terms of the experimental conditions (the model) and an error component. Typically, ANOVA is used to determine under which experimental condition do dependent variable score means differ and what proportion of variation in the dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variable(s).

A typical ANOVA model is \( y_{ij} = \mu + \alpha_i + \epsilon_{ij} \), where \( y_{ij} \) is the \( i \)-th subject’s dependent variable score in the \( j \)-th experimental condition, the parameter \( \mu \) is the grand mean of the experimental condition population means that underlies all subjects’ dependent variable scores, the parameter \( \alpha_i \) is the effect of the \( i \)-th experimental condition and the random variable \( \epsilon_{ij} \) is the error term, which reflects variation due to any uncontrolled source.

For a given model, the ANOVA table summarizes the outcomes of the ANOVA test indicating, for each factor, the Sum of Squares (SS), the Degrees of Freedom (DF), the Mean Squares (MS), the F statistics, and the p-value of that factor, which allows us to determine the significance of that factor.

When it comes to independent variables they can be either fixed effects – i.e., they have precisely defined levels, and inferences about its effect apply only to those levels – or random effects – i.e., they describe a randomly and independently drawn set of levels that represent variation in a clearly defined wider population. The latter case is a more sophisticated model which, in the estimation of the variance attributed to the different factors, also accounts for the additional randomness due to sampling of effect levels.

The experimental design determines how you compute the model and how you estimate its parameters. In particular, it is possible to have an independent measures design where different subjects participate to different experimental conditions (factors) or a repeated measures design, where each subject participates to all the experimental conditions (factors).

A final distinction is between crossed/factorial designs, where every level of one factor is measured in combination with every level of the other factors, and nested designs, where levels of a factor are grouped within each level of another nesting factor.

A.1 Estimating the Model
Figure 1 shows the experimental layout for the model (MD1). This is the typical Information Retrieval (IR) setting where you have a set of topics and a set of systems which are run against those topics; in ANOVA terms this is a crossed/factorial repeated measures design. Note that this is the same model used by Banks et al. [1] and Tague-Sutcliffe and Blustein [4] to analyse TREC data.

\[
y_{ij} = \mu + \tau_i + \alpha_j + \epsilon_{ij} \quad \text{(MD1)}
\]

where:
- \( \mu \) is the grand mean;
- \( \tau_i \) with \( i = 1, \ldots, T \) represents the effect of topics;
- \( \alpha_j \) with \( j = 1, \ldots, R \) represents the effect of systems;
- \( \epsilon_{ij} \) is the residual error.

Model (MD1) has the following estimators:
We can compute the and F statistics as follows:

### A.2 Assessment of the Model

Therefore, the score predicted by the model is

\[ \hat{y}_{ij} = \hat{\mu}_i + \hat{\tau}_i + \hat{\alpha}_j + \hat{\beta}_{ij} + \hat{\epsilon}_{ijk} \]

and the estimated residuals are:

\[ \hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\hat{\mu}_i + \hat{\tau}_i + \hat{\alpha}_j + \hat{\beta}_{ij} + \hat{\epsilon}_{ijk}) \]

#### B EFFECT OF UNDEFINED VALUES ON ESTIMATION

Suppose we are dealing with \( T \) topics, \( R \) system (runs), and \( S \) shards and thus \( N = T \cdot R \cdot S \) total samples.

We consider the following ANOVA models:

\[ y_{ijk} = \mu + \tau_i + \alpha_j + \epsilon_{ijk} \]

(\textit{Main Effects})

\[ y_{ijk} = \mu + \tau_i + \alpha_j + \beta_{ij} + \epsilon_{ijk} \]

(\textit{Interaction Effects})

Note that:

\[ SS_{\text{total}} = SS_{\text{topic}} + SS_{\text{system}} + SS_{\text{error}} \]

We can then compute the critical value for the F statistics of a factor, i.e. \( F_{\text{crit}} = F(\overline{d}f_{\text{factors}}, \overline{d}f_{\text{error}}) \), and determine its significance if \( F_{\text{factor}} > F_{\text{crit}} \); this allows us also to obtain the \( p \)-value for that factor.
Therefore, for any pair of systems \( u \neq v \), the difference of their estimated marginal means \( \hat{\mu}_u - \hat{\mu}_v \) is independent of the undefined values.

Proof. The estimated grand mean is given by:

\[
\hat{\mu}_g = \frac{1}{T \cdot R \cdot S} \sum_{i=1}^{T} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} = \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \sum_{j=1}^{R} \sum_{i \in X_k}^{T} y_{ijk} = \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \sum_{i \in X_k}^{T} [X_k] \cdot R \cdot x
\]

Then, the difference of the estimated marginal means is given by:

\[
\hat{\mu}_u - \hat{\mu}_v = \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \sum_{i \in X_k}^{T} [X_k] \cdot R \cdot x
\]

Therefore, for any two systems \( u \) and \( v \), we have that:

\[
\hat{\mu}_u - \hat{\mu}_v = \left( \frac{1}{T \cdot R \cdot S} \sum_{i \in X_k}^{T} x \right) - \left( \frac{1}{T \cdot R \cdot S} \sum_{i \in X_k}^{T} x \right) = \hat{\mu}_u - \hat{\mu}_v
\]

Therefore, the difference of the estimated marginal means of the system factor is independent of the undefined values.
The SS of the system factor is given by:

$$SS_{\text{system}} = T \cdot S \sum_{j=1}^{R} \left( \bar{\mu}_{j} - \bar{\mu} \right)^2$$

$$= T \cdot S \sum_{j=1}^{R} \left( \bar{\mu}_{j} - \bar{\mu} \right)^2$$

$$= T \cdot S \sum_{j=1}^{R} \left( \bar{\mu}_{j} - \bar{\mu} \right)^2$$

Therefore, the SS of the system factor is independent of the undefined values, i.e. they do not change the amount of variance attributed to the system factor. Since \(MS_{\text{system}} = \frac{SS_{\text{system}}}{df_{\text{system}}} \), also the the MS of the system factor is independent of the undefined values.

**Proposition B.5.** Given model (MD6), the residuals \(\epsilon_{ijk} \) are independent of the undefined values. Therefore, the SS of the error and, as a consequence, the MS of the error are independent of the undefined values.

**Proof.** The estimated marginal mean of the topic factor is given by:

$$\hat{\mu}_{i..} = \frac{1}{R} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} = \frac{1}{R} \sum_{j=1}^{R} \left( \sum_{k=1}^{S} y_{ijk} + \sum_{k \in X_i} x \right)$$

$$= \frac{1}{R} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} + \frac{1}{R} \sum_{j=1}^{R} \left( \sum_{k \in X_i} x \right)$$

$$= \frac{1}{R} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} + \frac{1}{R} \sum_{j=1}^{R} \left( \sum_{k \in X_i} x \right)$$

The estimated marginal mean of the topic factor is given by:

$$\hat{\mu}_{i..} = \frac{1}{R} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} = \frac{1}{R} \sum_{j=1}^{R} \sum_{k \in X_i} x$$

The predicted score for (MD6) is given by:

$$\hat{y}_{ijk} = \hat{\mu}_{..} + (\hat{\mu}_{i..} - \hat{\mu}_{..}) + (\hat{\mu}_{j..} - \hat{\mu}_{..}) + (\hat{\mu}_{k..} - \hat{\mu}_{..})$$

$$= \hat{\mu}_{..} + \hat{\mu}_{i..} - \hat{\mu}_{..} + \hat{\mu}_{j..} - \hat{\mu}_{..} + \hat{\mu}_{k..} - \hat{\mu}_{..}$$

$$= \hat{\mu}_{..} + \hat{\mu}_{i..} - \hat{\mu}_{..} + \hat{\mu}_{j..} - \hat{\mu}_{..} + \hat{\mu}_{k..} - \hat{\mu}_{..}$$

$$= \hat{\mu}_{..} + \hat{\mu}_{i..} - \hat{\mu}_{..} + \hat{\mu}_{j..} - \hat{\mu}_{..} + \hat{\mu}_{k..} - \hat{\mu}_{..}$$
Thus, the estimated residuals for (MD6) are given by:

\[
\hat{\epsilon}_{ijk} = \hat{y}_{ijk} - \hat{\bar{y}}_{ijk} = \begin{cases} 
\hat{y}_{ijk} - \hat{\mu}_{i}^{'j} - \hat{\mu}_{i}^{'k} & \text{if } i \notin X_k \\
\hat{y}_{ijk} - \hat{\mu}_{i}^{'k} & \text{if } i \in X_k
\end{cases}
\]

Therefore, the estimated residuals \(\hat{\epsilon}_{ijk}\) are independent of the undefined values. As a consequence, the SS of the error and \(MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}}\) are independent of the undefined values as well.

\(\Box\)

REFERENCES


