Int J Multimed Info Retr (2013) 2:59–70
DOI 10.1007/s13735-012-0030-4

REGULAR PAPER

Minimal test collections for low-cost evaluation of Audio Music Similarity and Retrieval systems

Julián Urbano · Markus Schedl

Received: 2 July 2012 / Revised: 14 September 2012 / Accepted: 11 December 2012 / Published online: 1 January 2013

Abstract Reliable evaluation of Information Retrieval systems requires large amounts of relevance judgments. Making these annotations is not only tedious but also complex for many Music Information Retrieval tasks. As a result, performing such evaluations usually requires too much effort. A low-cost alternative is the application of Minimal Test Collections algorithms, which offer very reliable results while significantly reducing the required annotation effort. The idea is to represent effectiveness scores as random variables that can be estimated, iteratively selecting which documents to judge so that we can compute accurate estimates with a certain degree of confidence and with the least effort. In this paper we show the application of Minimal Test Collections to the evaluation of the Audio Music Similarity and Retrieval task, run by the annual MIREX evaluation campaign. An analysis with the MIREX 2007, 2009, 2010 and 2011 data shows that with as little as 2 % of the total judgments we can obtain accurate estimates of the ranking of systems. We also present a method to rank systems without making any annotations, which can be successfully used when little or no resources are available.

Keywords Music information retrieval · Evaluation · Experimentation · Test collections · Relevance judgments

J. Urbano (✉)
University Carlos III of Madrid, Madrid, Spain
e-mail: jurbano@inf.uc3m.es; urbano.julian@gmail.com

M. Schedl
Johannes Kepler University, Linz, Austria
e-mail: markus.schedl@jku.at

1 Introduction

The evaluation of Information Retrieval (IR) systems requires a test collection, usually containing a set of documents, a set of task-specific queries, and a set of annotations that provide information as to what results a system should return for each query [10,22]. Depending on the task, the set of queries may comprise the collection of documents itself, and the type of annotations can differ widely. In the field of Music IR (MIR), building these collections is very problematic due to the very nature of the musical information, legal restrictions upon the documents, etc. [7]. In addition, annotating a test collection is a very time-consuming and expensive process for some MIR tasks. For instance, annotating a single clip for Audio Melody Extraction can take several hours. As a result, test collections for MIR tasks use to be very small, biased, and unlikely to change from year to year, posing serious problems for the proper evolution of the field [17].

The annual Music Information Retrieval Evaluation eXchange (MIREX) started in 2005 as an international forum to promote and perform evaluation of MIR systems for various tasks [8]. MIREX was developed following the principles and methodologies that have made the Text REtrieval Conference (TREC) [24] such a successful forum for evaluating Text IR systems [6,23]. However, since its inception in 2005, the MIREX campaigns have evolved in parallel to TREC, practically ignoring all recent developments in the evaluation of IR systems [10,17]. In fact, the last 5 years have witnessed several works on low-cost, yet reliable evaluation techniques, allowing the number of queries used to grow up to as many as 40,000 [5]. One of these works is the development of algorithms for evaluation with Minimal Test Collections (MTC) [1–3].

The idea behind MTC is that the results of an IR evaluation experiment may be estimated with high confidence even if

Table 1 Summary of MIREX AMS editions

Year	Teams	Systems	Queries	Results	Judgments	Overlap
2006	5	6	60	1,800	3×1,629	10%
2007	8	12	100	6,000	4,832	19%
2009	9	15	100	7,500	6,732	10%
2010	5	8	100	4,000	2,737	32%
2011	10	18	100	9,000	6,322	30%

In the 2006 edition three different assessors provided annotations for every query-document pair. The task did not run in 2008

the set of annotations is very incomplete. In a typical setting, it means that we do not need to judge all documents retrieved for a query, but only a small fraction of it, to estimate with high confidence which of two systems is better. In this paper we study the application of MTC to the evaluation of Audio Music Similarity and Retrieval (AMS) systems, as it is one of the tasks that most closely resembles the ad hoc Text IR scenario: for a given audio clip (the query), an AMS system returns a list of music pieces deemed to be similar to it. AMS is one of the most important tasks in MIR, and it has been run in MIREX in five of the seven editions so far (see Table 1).

Each edition of the AMS task requires the work of dozens of volunteers to perform similarity judgments, telling how similar two 30 s audio clips are. In the last edition, in 2011, 6,322 of these judgments were needed, meaning that at least 53 h of assessor time were needed to complete the judging task. In practice, though, collecting all these judgments takes several days, even weeks [11]. But along with the Symbolic Melodic Similarity (SMS) task, AMS is one of the couple of exceptions for which a new set of queries and relevance judgments are put together every year. Most of the MIR tasks just use the same collections over and over again because they are too expensive to build, especially in terms of judging or annotation effort. Therefore, the study of low-cost evaluation methodologies is imperative for the development of proper test collections to reliably evaluate MIR systems and properly advance the state of the art [17].

Developing low-cost evaluation methodologies is essential for private, in-house evaluations too. A researcher investigating several improvements of an existing MIR technique is not really interested in knowing how well they perform for the task (which is highly dependent on the test collection anyway), but in which one performs better. That is, she is interested in the *comparative* evaluation of systems. MTC is specifically designed for these cases: it minimizes the annotation effort needed to find a difference between systems, iteratively selecting for judging those documents that are more informative to figure out the difference between systems, and reusing previous judgments when available.

2 AMS evaluation

Audio Music Similarity and Retrieval systems are evaluated according to an effectiveness measure that assesses how well they would satisfy an arbitrary user for a given query [18]. In order to generalize the results of an evaluation experiment to an arbitrary query, the MIREX evaluations use a random sample $\mathcal{Q}$ of 100 queries. Each system is run for every query, returning a list of all documents in the collection $\mathcal{D}$, ranked by their similarity to the query. The effectiveness measure used in MIREX is Average Gain of the top k documents retrieved ($AG@k$), with $k = 5$ [8,19]. For an arbitrary system A, $AG@k$ is defined as:

$$AG@k = \frac{1}{k} \sum_{i \in \mathcal{D}} G_i \cdot I(\mathsf{A}_i \leq k)$$

where G_i is the gain of document i, A_i is the rank at which system A retrieved document i, and $I(x)$ is a boolean indicator function that evaluates to 1 if the expression x is true and to 0 otherwise. Therefore, the summation adds the gain of all documents in the collection that were ranked by A in the top k.

The gain of a document is a measure of how much information the user will gain from inspecting that result. In MIREX, there are two different scales [11,19]: the Broad scale is a 3-point graded scale where a document is considered either not similar to the query (gain 0), somewhat similar (gain 1) or very similar (gain 2); and the Fine scale, where the gain of a document ranges from 0 (not similar at all) to 100 (identical to the query)[1]. These gain scores are assigned by humans, who make similarity judgments between queries and documents. After all the judging is done, every system gets an $AG@k$ score for each query, and then they are ranked by their mean score across all queries.

To minimize random effects due to the particular sample of queries chosen, the Friedman test is run with the Average Gain scores of every system to look for significant differences, and the Tukey's HSD test is then used to correct the experiment-wide Type I error rate [19]. The grand results of the evaluation are therefore scale-dependent pairwise comparisons between systems, telling which one is better for the current set of queries $\mathcal{Q}$, and whether the observed difference was found to be statistically significant.

[1] In early editions of MIREX it was defined from 0 to 10, with one decimal digit. Both definitions are equivalent.

3 Evaluation with incomplete judgments

The evaluation methodology used in MIREX is expensive in the sense that a complete set of similarity judgments is needed: the top k documents retrieved by every system have to be judged for every query. However, we may investigate how to compare systems so that we do not need to judge all documents and still be confident about the result of an evaluation experiment.

The idea is to use random variables to represent gain scores. The upside is that their value can be estimated fairly well for most documents; the downside is that these estimates will have some degree of uncertainty. The goal of MTC is to select for judging those documents that allow us to compute good estimates of the difference between systems with very few judgments.

3.1 $AG@k$ as a random variable

Let G_i be a random variable representing the gain of document i. The distribution of G_i is multinomial and depends on the similarity scale used: for the Broad scale G_i can take one of 3 values, and for the Fine scale it can take one of 101 values. The expectation and variance of G_i are as follows:

$$\begin{aligned} E[G_i] &= \sum_{l \in \mathcal{L}} P(G_i = l) \cdot l \\ \mathrm{Var}[G_i] &= \sum_{l \in \mathcal{L}} P(G_i = l) \cdot l^2 - E[G_i]^2 \end{aligned} \tag{1}$$

where $\mathcal{L}$ is the set of possible relevance levels:

$$\begin{aligned} \mathcal{L}_{\text{Broad}} &= \{0, 1, 2\} \\ \mathcal{L}_{\text{Fine}} &= \{0, 1, \ldots, 100\} \end{aligned}$$

Whenever document i is judged and assigned a gain l, its expectation and variance are fixed to $E[G_i] = l$ and $\mathrm{Var}[G_i] = 0$; that is, no uncertainty about G_i. Given this definition of the gain of an arbitrary document, we can now define the $AG@k$ of an arbitrary system as a random variable too.

Under the assumption that the gain of one document is independent of the others, the expectation and variance of $AG@k$ are defined as:

$$\begin{aligned} E[AG@k] &= \frac{1}{k} \sum_{i \in \mathcal{D}} E[G_i] \cdot I(\mathsf{A}_i \leq k) \\ \mathrm{Var}[AG@k] &= \frac{1}{k^2} \sum_{i \in \mathcal{D}} \mathrm{Var}[G_i] \cdot I(\mathsf{A}_i \leq k) \end{aligned} \tag{2}$$

Having $AG@k$ defined this way allows us to estimate its value from an incomplete set of judgments. With no judgments at all, the variance of the estimator would be maximum, but as judgments are made the variance decreases. With all k documents judged, the variance is zero and the estimate equals the true $AG@k$ score.

3.2 Difference in $AG@k$

Using Eq. (2) we can estimate the $AG@k$ score of a system. But we are really interested in knowing which of two systems performs better, that is, the sign of their difference in $AG@k$. For arbitrary systems A and B:

$$\begin{aligned} \Delta AG@k &= \frac{1}{k} \sum_{i \in \mathcal{D}} G_i \cdot I(\mathsf{A}_i \leq k) - \frac{1}{k} \sum_{i \in \mathcal{D}} G_i \cdot I(\mathsf{B}_i \leq k) \\ &= \frac{1}{k} \sum_{i \in \mathcal{D}} G_i \cdot (I(\mathsf{A}_i \leq k) - I(\mathsf{B}_i \leq k)) \end{aligned} \tag{3}$$

If $\Delta AG@k$ is positive, we can conclude system A performed better than system B (worse if negative) for the query. We can see that only documents retrieved by one system and not by the other will contribute to $\Delta AG@k$: documents retrieved by both systems will contribute $G_i - G_i = 0$. Therefore, judging these documents will not tell us anything about the difference. Thus, the larger the overlap between the systems' outputs, the fewer the judgments necessary to figure out which one is better. Because the two systems are independent of each other, the expectation and variance are[2]:

$$\begin{aligned} E[\Delta AG@k] &= \frac{1}{k} \sum_{i \in \mathcal{D}} E[G_i] \cdot (I(\mathsf{A}_i \leq k) - I(\mathsf{B}_i \leq k)) \\ \mathrm{Var}[\Delta AG@k] &= \frac{1}{k^2} \sum_{i \in \mathcal{D}} \mathrm{Var}[G_i] \cdot (I(\mathsf{A}_i \leq k) - I(\mathsf{B}_i \leq k))^2 \end{aligned} \tag{4}$$

Now that we can compute an estimate of the difference for one query, let us generalize to a set of queries $\mathcal{Q}$, computing the mean of the $\Delta AG@k$ scores for all of them. As they are sampled randomly[3] [8,19], queries are independent of each other, so the expectation and variance are:

$$\begin{aligned} E\left[\overline{\Delta AG@k}\right] &= \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} E[\Delta AG@k_q] \\ \mathrm{Var}\left[\overline{\Delta AG@k}\right] &= \frac{1}{|\mathcal{Q}|^2} \sum_{q \in \mathcal{Q}} \mathrm{Var}[\Delta AG@k_q] \end{aligned} \tag{5}$$

With these estimates we can rank all systems by their difference in $AG@k$. In addition, for a given set of judgments, we can compute $P\left(\overline{\Delta AG@k} \leq 0\right)$, that is, the probability of system A performing worse than system B. If $P\left(\overline{\Delta AG@k} \leq 0\right) \leq \alpha$ then we can conclude that system A performs worse than B with α confidence ($1 - \alpha$ confidence of B being worse than A). If, while judging documents, we

[2] The indicator functions are squared in the variance so all documents have a positive contribution to the total variance.

[3] Note that this is rarely true in Text Information Retrieval.

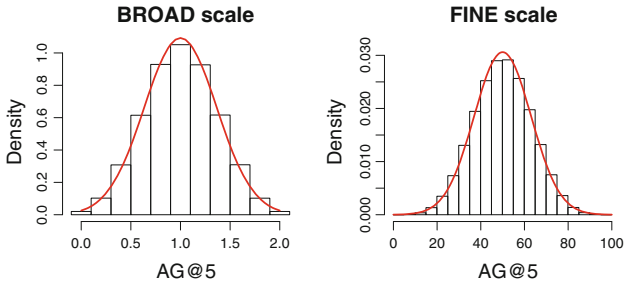

Fig. 1 Distribution of $AG@5$ assuming a uniform distribution of gains for the Broad (*left*) and Fine (*right*) scales. The *red lines* are normal distributions with means $E[AG@5]$ and variances $\mathrm{Var}[AG@5]$.

reach a certain confidence in the sign, say 95 %, we can stop judging.

3.3 Distribution of $\Delta AG@k$

To compute the confidence in the sign, we need to know the distribution of $\overline{\Delta AG@k}$. For a relevance scale with only two levels (similar and not similar), $AG@k$ is basically the same as $P@k$ (precision at k), which can be approximated by a normal distribution under a binomial or uniform prior distribution of G_i [2]. In our case, the Broad scale has 3 possible levels, and the Fine scale has 101 levels.

Let us define Γ^k as the set of all $|\mathcal{L}|^k$ possible assignments that can be made for k documents. The probability of $AG@k$ being equal to a value z is:

$$P(AG@k = z) := \sum_{\gamma^k \in \Gamma^k} P\left(AG@k = z|\gamma^k\right) \cdot P\left(\gamma^k\right)$$

that is, if we can compute the probability of making each γ^k assignment, we can just sum the probabilities of those that lead to $AG@k = z$. In our case, there are $3^5 = 243$ possible assignments of relevance with the Broad scale and $101^5 \approx 10.5$ billion assignments with the Fine scale. However, we still need information about the distribution of each G_i in order to compute $P\left(\gamma^k\right)$.

But $AG@k$ turns out to be a special case. Let G be a random variable representing the gain of the top k documents retrieved by a system for all possible queries, and let the set $\{AG@k_1, \ldots, AG@k_{|\mathcal{Q}|}\}$ be a random sample of size $|\mathcal{Q}|$ where each $AG@k_q$ is the average gain of k documents sampled from G. By the Central Limit Theorem, as $|\mathcal{Q}| \to \infty$ the distribution of the sample mean $\overline{AG@k} = \sum AG@k_q/|\mathcal{Q}|$ approximates a normal distribution, regardless of the underlying distribution of G. Therefore, with a large number of queries $\overline{\Delta AG@k}$ can be approximated by a normal distribution, because it is the sum of two variables approximately normal themselves.

The left plot in Fig. 1 shows the histogram of possible $AG@5$ scores with the Broad scale assuming a uniform distribution of assignments; and the right plot shows the scores observed in a random sample of 1 million assignments with the Fine scale. The red lines are normal distributions with means $E[AG@k]$ and variances $\mathrm{Var}[AG@k]$. We can see that the normal distributions do indeed approximate very well.

Therefore, we can use the normal cumulative density function Φ to approximate the probability of A being worse than B as:

$$P\left(\overline{\Delta AG@k} \leq 0\right) = \Phi\left(\frac{E\left[\overline{\Delta AG@k}\right]}{\sqrt{\mathrm{Var}\left[\overline{\Delta AG@k}\right]}}\right) \tag{6}$$

which measures the area under the curve that is to the left of zero. From here we can define the confidence C_{AB} in the sign of $\overline{\Delta AG@k}$ as the maximum between the probability of it being positive and it being negative:

$$C_{\mathsf{AB}} = \max\left(P\left(\overline{\Delta AG@k} \leq 0\right), 1 - P\left(\overline{\Delta AG@k} \leq 0\right)\right) \tag{7}$$

Whenever we pass a threshold on confidence, say $C_{\mathsf{AB}} \geq 95\,\%$, we can stop judging and conclude which system is better based on the sign of $E\left[\overline{\Delta AG@k}\right]$.

3.4 Document selection

Equations (4) and (5) can be used to estimate the difference between two systems with an incomplete set of judgments, but the problem is: which documents should we judge? Ideally, we want to judge only those that are most informative to know the sign of the difference in $AG@k$. For just two systems it is obvious from Eq. (3) only documents retrieved by one system but not by the other one are informative. For an arbitrary number of queries, we can just refer to a query-document pair as a single document (i.e. the gain of a document *for* a particular query).

However, with an arbitrary number of systems a particular document could be informative for more than just one of the pairwise comparisons. We can assign a weight w_i to every query-document i, equal to the number of pairwise system comparisons for which judging query-document i would affect the estimate of $\Delta AG@k$. Being $\mathcal{S}$ the set of all system pairs, the weight of an arbitrary document i is defined as:

$$w_i = \sum_{(\mathsf{A},\mathsf{B}) \in \mathcal{S}} \left(I\left(\mathsf{A}_i \leq k\right) - I\left(\mathsf{B}_i \leq k\right)\right)^2 \tag{8}$$

At all times, we will want to judge those documents with the largest weight because they will have the largest effect on the ranking. Algorithm 1 lists MTC to rank a set of systems $\mathcal{S}$ with $1 - \alpha$ confidence.

For the stopping condition we compute the mean confidence across all system pairs: if it is sufficiently large, we stop judging altogether. We call this the *confidence in the ranking*. We note though that MTC can be used with a different stopping condition. For instance, we may require *at least* 95 %

Algorithm 1 MTC for $\Delta AG@k$

while $\frac{1}{|\mathcal{S}|}\sum_{(\mathsf{A},\mathsf{B})\in\mathcal{S}} C_{\mathsf{AB}} \leq 1-\alpha$ **do**
  $i^* \leftarrow argmax_i w_i$ for all unjudged query-document pairs
  judge query-document i^* (obtain true $gain_{i^*}$)
  $E[G_{i^*}] \leftarrow gain_{i^*}$
  $Var[G_{i^*}] \leftarrow 0$
end while

confidence in *all* comparisons, as opposed to an *average* of 95 % as we do here. In such cases, the definition of w_i could differ from that in Eq. (8). For instance, we could consider just the system pairs for which $C_{\mathsf{AB}} < 1-\alpha$, and make their contribution to w_i proportional to C_{AB}. We could further modify the algorithm by considering the magnitude of the difference between systems instead of just its sign [18]. This would allow us to estimate system differences from the perspective of expected user satisfaction, for instance by computing $P\left(\overline{\Delta AG@k} \leq -0.3\right)$ instead of $P\left(\overline{\Delta AG@k} \leq 0\right)$.

4 Estimation of gain scores

Equations (6) and (7) allow us to compute the confidence in the sign of the difference between two systems. But tracking back to Eq. (1), we still need to know what the distribution of G_i is; that is, what $P(G_i = l)$ is for each of the labels in the similarity scale used. There are two immediate choices: a fixed distribution for each document i, maybe estimated from judgments in previous MIREX editions; or a distribution for each document as returned by a model fitted with various features.

4.1 Distribution of gain scores

A simple choice is to assume that every similarity assignment is equally likely [3,20]. For the Broad scale, all three assignments would have probability 1/3, while for the Fine scale each assignment would have probability 1/101. According to Eq. (1), an arbitrary unjudged document would have expectation 1 and variance 2/3 in the Broad scale, and in the Fine scale it would have expectation 50 and variance 850.

A better alternative is to estimate the gain score of each document individually [1,2,4]. The problem reduces then to fitting a model that, given certain features about a query-document, allows us to estimate its gain score. We may consider two frameworks for creating such a model: classification and regression. The classification approach is not appropriate because it ignores the order of the labels. In the Broad scale, for instance, it means that if the true gain of a document were 0, an estimation of 1 would be as good as an estimation of 2, while the latter is clearly worse. Linear regression is not appropriate either, because the predicted gains could be well outside the limits [0–2] and [0–100]. This could be solved with truncated regression [13], but we would still need to make assumptions about its underlying distribution. Multinomial regression has the same problem as classification, namely that it ignores the order of the levels in the outcome.

Ordinal logistic regression is the most appropriate framework [4,12]. The dependent variable is modeled as an ordinal variable and, as opposed to classification and multinomial regression, the order of the levels is therefore taken into account. For an arbitrary similarity scale $\mathcal{L} = \{l_1, \ldots, l_{|\mathcal{L}|}\}$, the model for our ordinal variable is:

$$\log \frac{P(G_i \geq l_j | f_i)}{P(G_i < l_j | f_i)} = \alpha_j + \sum_{k=1}^{|f_i|} \beta_k \cdot f_{ik} \tag{9}$$

where β_k are the parameters to fit, α_j is the fitted intercept for the particular level l_j, and f_i is the feature vector for document i. Once the model is fitted, we can use the inverse logit function to compute $P(G_i \geq l_j | f_i)$. Then, the probability of G_i being equal to some similarity level l_j is computed as[4]:

$$P(G_i = l_j | f_i) = P(G_i \geq l_j | f_i) - P(G_i \geq l_{j+1} | f_i) \tag{10}$$

This proportional odds model is generalized by the Vector Generalized Additive Model (VGAM) [26], which is implemented in standard statistical packages such as R [25] and facilitate the above calculations.

Therefore, the ordinal logistic framework allows us to estimate the distribution $P(G_i = l)$ in Eq. (1), which in turn enables the computation of expectation and variance as usual. As opposed to using the uniform distribution, this model is expected to produce estimates closer to the true score and with reduced variance. As a result, the confidence calculations as per Eq. (7) are expected to be more reliable and require fewer judgments to pass a threshold like 95 %.

4.2 Features used and fitted models

We consider two types of features to use in the above model in order to estimate gain scores: output-based features and judgment-based features.

4.2.1 Output-based features

This set of features represent different aspects of the system outputs, so they can still be used when there are no judgments at all. For an arbitrary document d and query q:

- *pSYS*: percentage of systems that retrieved d for q. Intuitively, the more systems retrieve d, the more likely for it to be similar to q.

[4] Note that $P(G_i \geq l_1 | f_i)$ is always 1.

- *pTEAM*: percentage of research teams participating in MIREX that retrieved d for q. Systems by the same team are likely to return similar documents, so the effect of *pSYS* could be biased if teams participate with a large number of systems. *pTEAM* can be used to reduce this bias.
- *OV*: degree of overlap between systems, to calibrate inherent similarities among systems when using the *pSYS* and *pTEAM* features.
- *aRANK*: average rank at which systems retrieved d for q. Documents retrieved closer to the top of the results lists are expected to be more similar to q.
- *sGEN*: whether the musical genre of d is the same as q's (either 1 or 0), as documents of the same genre are usually considered similar to each other [14].
- *pGEN*: percentage of all documents retrieved for q that belong to the same musical genre as d does.
- *pART*: percentage of all documents retrieved for q that belong to the same artist as d does. Note that a feature like *sGEN* for artists does not make sense because all retrieved documents by q's artist are filtered out [8,9].

4.2.2 Judgment-based features

This set of features takes advantage of known judgments to produce better predictions:

- *aSYS*: average gain score obtained by the systems that retrieved d for q. Intuitively, a document retrieved by good systems is likely to be a good result.
- *aDOC*: average gain score of all the other documents retrieved for q. Likewise, this feature models query difficulty: if documents retrieved for q are not similar, d is not likely to be similar either.
- *aGEN*: average gain score of the documents retrieved for q that belong to the same genre as d does.
- *aART*: average gain score of the documents retrieved for q and by the same artist as d's.

4.2.3 Fitted models

We used data from the MIREX 2007, 2009, 2010 and 2011 editions of the Audio Music Similarity and Retrieval task to fit the models following the regression framework described in Sect. 4.1. Starting with a saturated model, we simplified to a model, called L_{judge}, using the features *pTEAM*, *OV*, *aSYS* and *aART*. All these features showed a very significant effect on the response ($p < 0.0001$). While other features did improve the model, they did so very marginally, so we decided to keep it as simple as possible. The coefficient of determination R^2 can be used to assess the goodness of fit, measuring the proportion of variability in the outcome that is accounted for by the model. The predictions of L_{judge} are particularly good, with an adjusted R^2 score of approximately 0.9 (the value $R^2 = 1$ means that the model offers a perfect fit of the data).

Even though L_{judge} produces very good results, we can only use it to estimate the G_i scores of documents for which we can compute both *aSYS* and *aART*. However, because our goal is to reduce the amount of judging as much as possible, we will not be able to estimate the gain scores for most of the documents until we have made a fair amount of judgments. Therefore, we decided to fit another model, called L_{output}, that only uses output-based features. With this model, we can always estimate G_i scores, even when there are no judgments available at all.

Proceeding as before, we simplified to a model using the features *pTEAM*, *OV*, *pART*, *sGEN*, *pGEN* and the *sGEN:pGEN* interaction. Despite all features showed again a significant effect ($p < 0.0001$), the predictions were significantly worse than with L_{judge}, resulting in an adjusted R^2 score of approximately 0.35.

When fitting the models for the Fine scale, we further simplified by breaking the scale down to 10 levels rather than the original 101. Therefore, we actually use the scale $\{0, 11, 22, \ldots, 99\}$. In order to avoid overfitting, when estimating the gain scores for one MIREX edition we excluded all data from that edition when fitting the model. Therefore, we actually fitted L_{judge} and L_{output} for each scale and each edition. See the appendix for more details regarding the models.

4.3 Estimation errors in practice

To check the accuracy of the G_i estimates we again used the similarity judgments collected in MIREX 2007, 2009, 2010 and 2011 (see Table 2). First, we computed the Root Mean Square Error (RMSE) between every document's true gain score and its estimation. The errors with the uniform prior distribution are ≈ 0.8 with the Broad scale and ≈ 30 with the Fine scale. Both regression models consistently produce less error, with the L_{judge} model having an error of ≈ 0.27 with the Broad scale and ≈ 8.9 with the Fine scale; that is, the error is reduced to about one third.

In MIREX 2006 three different assessors provided judgments for each query-document pair [8,11]. If we consider one assessor's judgments as the truth, and the other's as mere estimates, we find that the average RMSE among assessors was 0.795 with the Broad scale and 31.2 with the Fine scale. We note that these errors are extremely similar to the errors of the L_{output} model (see Table 2), and quite larger than the errors of the L_{judge} model. Therefore, we argue that the errors we make when using MTC or ranking without judgments are comparable to the differences we should expect just by having a different human assessor in the first place [11,21].

Table 2 Average error and variance of the G_i estimates computed with the uniform distribution and regression models

Year	Broad scale						Fine scale					
	Uniform		L_{output}		L_{judge}		Uniform		L_{output}		L_{judge}	
	RMSE	Var	RMSE	Var	RMSE	Var	RMSE	Var	RMSE	Var	RMSE	Var
2007	0.813	0.667	0.639	0.436	0.260	0.067	31.9	850	24.3	601	8.83	70
2009	0.812	0.667	0.632	0.454	0.254	0.069	31.1	850	23.4	626	8.76	73
2010	0.794	0.667	0.706	0.394	0.283	0.07	30.2	850	26.1	549	8.94	73
2011	0.789	0.667	0.690	0.390	0.304	0.078	29.6	850	25.2	561	9.36	72

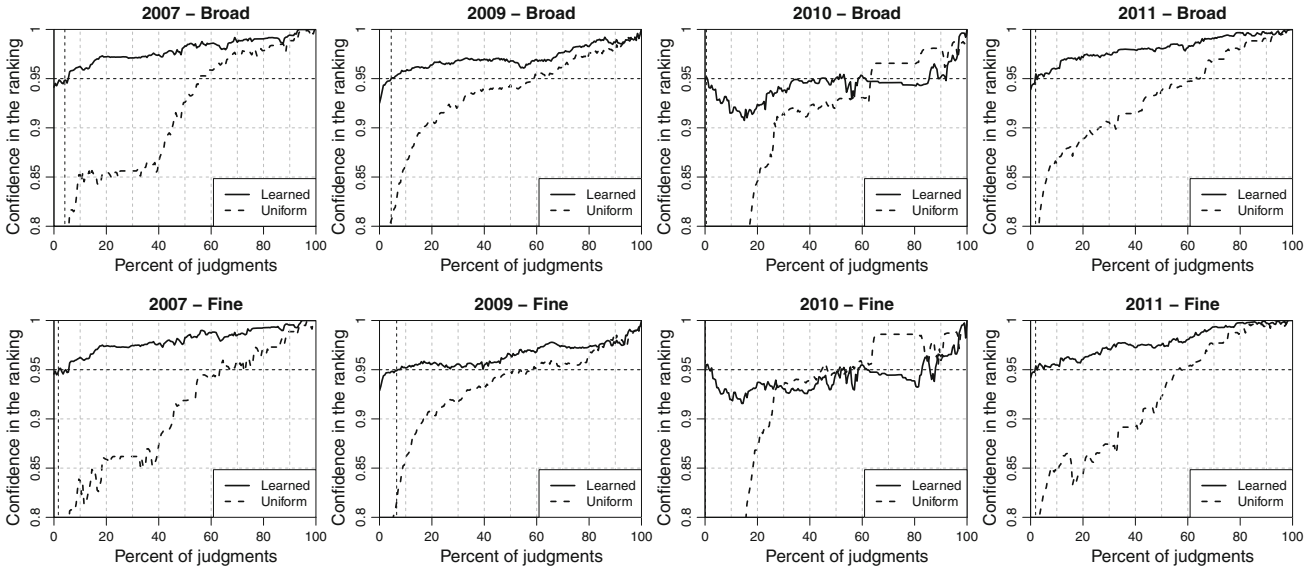

Fig. 2 Confidence in the ranking of systems as the number of judgments increases. The *dashed lines* mark the point at which 95 % confidence is reached for the first time

The MIREX evaluations assume arbitrary final users, so these errors can be ignored for all practical purposes. If no arbitrary users were assumed, but specific users were considered for instance in personalization [18], then our estimates would be erroneous to the degree reported here.

We also compared the average variance of the estimates. In Sect. 4.1 we saw that the variance in the uniform estimates is 2/3 with the Broad scale and 850 with the Fine scale. As Table 2 shows, the regression models improve the estimates also in terms of variance. The L_{judge} model reduces variance by one order of magnitude: $\approx$0.07 with Broad judgments and $\approx$72 with Fine judgments. Thus, the regression models provide better estimates and reduce variance to achieve high confidence in the sign differences earlier in the process.

5 Results

We simulated the use of MTC to evaluate all systems from the MIREX 2007, 2009, 2010 and 2011 Audio Music Similarity and Retrieval task (see Table 1). The number of pairwise system comparisons are 66, 105, 28 and 153, respectively. Recall that the L_{output} and L_{judge} models for one edition are fitted ignoring all information from that same edition, thus avoiding overfitting. When using MTC with the regression models, all G_i scores are estimated at the beginning with L_{output}, and updated every 20 judgments, when possible, with L_{judge}.

Figure 2 shows how the confidence in the ranking of systems increases as more judgments are made. This confidence in the ranking can be interpreted as the expected confidence in the sign of $\overline{\Delta AG@k}$ of any two systems picked at random. MTC with the estimates based on the uniform distribution need about 60 % of the judgments to reach 95 % confidence in the ranking. However, it is clearly outperformed by MTC with the learned distribution. As Table 3 shows, the judging effort is dramatically reduced: the median percentage of judgments needed with the Broad scale is 3 %, and as little as 1.8 % with the Fine scale. Considering that a single MIREX

Table 3 Judgments needed by MTC to reach 95 % confidence in the ranking of systems and accuracy of the sign estimates at that point

All G_i scores are estimated with L_{output} and L_{judge}

Year	Total judgments	Broad scale			Fine scale		
		Judgments	Accuracy	τ	Judgments	Accuracy	τ
2007	4,832	200 (4.1 %)	0.955	0.909	80 (1.7 %)	0.955	0.909
2009	6,732	300 (4.5 %)	0.971	0.943	440 (6.5 %)	0.952	0.905
2010	2,737	13 (0.5 %)	0.893	0.786	2 (0.1 %)	0.857	0.714
2011	6,322	120 (1.9 %)	0.941	0.882	120 (1.9 %)	0.941	0.882

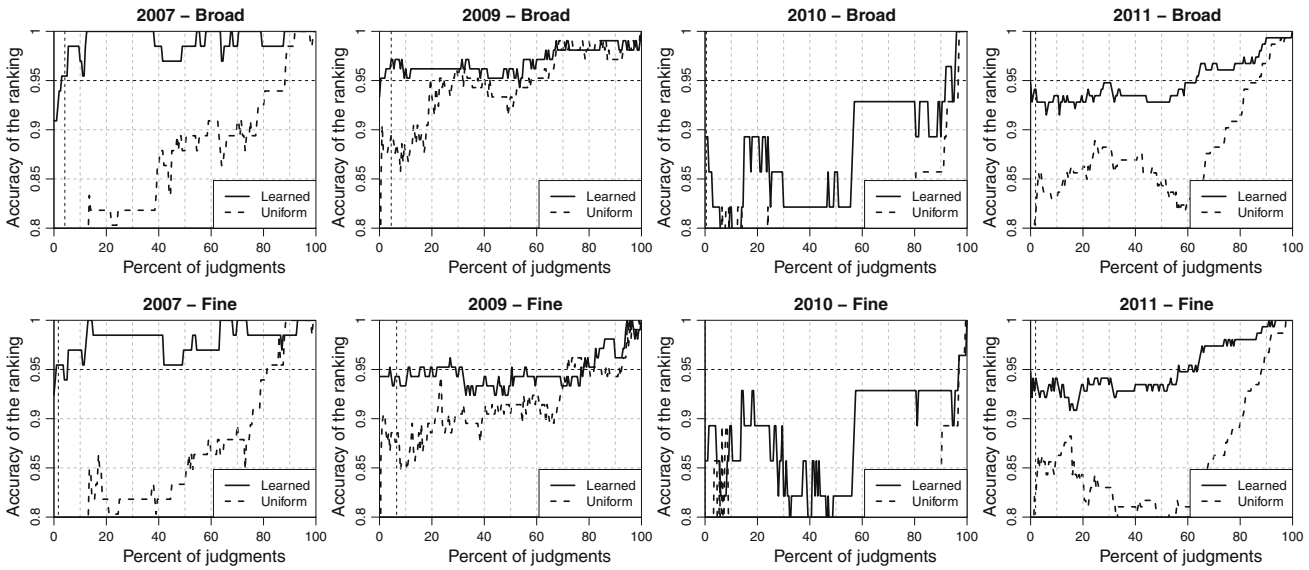

Fig. 3 Accuracy of the ranking of systems as the number of judgments increases. The *dashed lines* mark the point at which 95 % confidence is reached for the first time

assessor makes about 220 judgments per edition [8,11], the use of MTC would significantly reduce the required manpower to just 1 or 2 assessors.

We can see that very high confidence levels can be achieved with considerably fewer judgments, but how good are the estimates of the sign of $\overline{\Delta AG@k}$? Figure 3 shows how the accuracy of the estimated ranking tends to increase as more judgments are made, where accuracy is defined as the proportion of sign estimates that are correct across all systems pairs:

$$\text{Accuracy} = \frac{\text{correct}}{\text{total}}$$

In particular, Table 3 reports the performance of MTC when judging until the average confidence achieved is 95 %. The accuracy is above 0.95 for the 2007 and 2009 collections, and as high as 0.941 for 2011. However, for 2010 it drops below 0.9 for 2010. Nonetheless, in no case is an estimate wrong between two systems for which the true $\overline{\Delta AG@k}$ is statistically significant.

Another traditional way of comparing the estimated ranking and the true ranking is to compute Kendall's τ correlation coefficient between the two, defined as:

$$\tau = \frac{\text{correct} - \text{incorrect}}{\text{total}}$$

Kendall's τ ranges between 1 (exact same rankings) and -1 (opposite rankings), with 0 meaning that half of the pairs are swapped. Rankings with correlations above 0.9 are usually considered equivalent if we account for the effect of having one or another assessor make the judgments [11,21]. Formally, 0.9 Kendall correlation is achieved with 5 % of incorrect estimates, which corresponds to 0.95 accuracy. As Table 3 shows, correlations are above 0.9 in the 2007 and 2009 collections, but a little below in 2011 and, especially, in 2010. However, we note that with only 28 system pairs in 2010, just a single incorrect estimate would drop τ to $26/28 = 0.929$; so low correlations are expected with this collection. This dramatic effect of one single erroneous estimate can be easily seen in Fig. 3. Nonetheless, the median

Table 4 Accuracy versus confidence in the sign estimates when running MTC to 95 % confidence in the ranking

Conf.	Broad scale		Fine scale	
	In bin	Acc.	In bin	Acc.
[0.50, 0.60)	7 (2.0 %)	0.714	13 (3.7 %)	0.615
[0.60, 0.70)	15 (4.3 %)	0.733	13 (3.7 %)	0.846
[0.70, 0.80)	11 (3.1 %)	0.818	7 (2.0 %)	0.714
[0.80, 0.90)	24 (6.8 %)	0.833	24 (6.8 %)	0.833
[0.90, 0.95)	15 (4.3 %)	0.733	15 (4.3 %)	0.667
[0.95, 0.99)	31 (8.8 %)	1.000	22 (6.2 %)	0.909
[0.99, 1)	249 (70.7 %)	0.992	258 (73.3 %)	0.996

Table 5 Confidence and accuracy of the estimated ranking when no judgments are made

Year	Broad scale			Fine scale		
	Conf.	Acc.	τ	Conf.	Acc.	τ
2007	0.941	0.909	0.818	0.946	0.924	0.848
2009	0.925	0.933	0.867	0.929	0.943	0.886
2010	0.947	0.893	0.786	0.949	0.857	0.714
2011	0.939	0.948	0.895	0.942	0.948	0.895

correlation across collections is as high as 0.896 with the Broad scale and 0.894 with the Fine scale. We note again that all mistakes are produced between systems that are not significantly different anyway.

5.1 Accuracy of the individual estimates

Despite the average confidence in the ranking generally corresponds to the average accuracy of the sign estimates, there can be the case where the average confidence is biased by a few comparisons for which we are extremely confident. The question now is: how trustworthy are each of the individual estimates? We ran MTC with all four collections and the two similarity scales, and stopped judging when the average confidence was at least 95 %. The 352 system pairs from all four collections were divided by confidence in the sign of the individual $E\left[\overline{\Delta AG@k}\right]$.

Ideally, we would want accuracy to correspond to confidence (e.g. 0.80 accuracy in all pairs with 0.80 confidence), and Table 4 shows that this is generally the case. However, confidence seems slightly overestimated in the range [0.90–0.99], though we note again that there are just too few occurrences in that range to compute a reliable accuracy score. Nonetheless, over 70 % of the times confidence is larger than 0.99, where almost all estimates are indeed correct. On the other hand, having such a high proportion of very confident estimates seemingly tends to overestimate the average confidence in the ranking, which is here used as stopping condition in Algorithm 1.

Table 6 Accuracy versus confidence in the sign estimates when ranking systems in all collections and with no judgments

Conf.	Broad scale		Fine scale	
	In bin	Acc.	In bin	Acc.
[0.50, 0.60)	16 (4.5 %)	0.500	16 (4.5 %)	0.625
[0.60, 0.70)	17 (4.8 %)	0.882	15 (4.3 %)	0.867
[0.70, 0.80)	15 (4.3 %)	0.800	15 (4.3 %)	0.733
[0.80, 0.90)	24 (6.8 %)	0.792	24 (6.8 %)	0.792
[0.90, 0.95)	16 (4.5 %)	0.875	13 (3.7 %)	0.846
[0.95, 0.99)	33 (9.4 %)	0.909	31 (8.8 %)	0.903
[0.99, 1)	231 (65.6 %)	0.996	238 (67.6 %)	0.996

5.2 Ranking systems without judgments

As discussed above, the confidence in the ranking is quite high with very few judgments, so next we ask the question: how well can we rank systems *with no judgments at all*? Soboroff et al. [16] first studied this problem with systems submitted to TREC, showing that randomly considering documents as relevant correlated positively with the true TREC rankings. Rather than using random judgments, we use the estimates provided by the L_{output} regression model. Note that the L_{judge} model cannot be used because it does require some known judgments.

Table 5 shows the confidence in the rankings when making no judgments at all. Confidence is very high across collections, with a median of 0.942. The accuracy of the rankings is again quite high: the medians are 0.921 with the Broad scale and 0.934 with the Fine scale, which correspond to median τ correlations of 0.843 and 0.867 respectively. The overall performance is worse than running MTC and making a few judgments, but it is still very good considering that no judgments are needed.

The next question is again: how trustworthy are each of the individual estimates? As in Tables 4, 6 bins all 352 individual system comparisons by confidence, showing the corresponding accuracy in each bin. Similarly, we see that confidence is slightly overestimated in the range [0.80–0.99] and that, in general, confidence tends to be lower than when running MTC. Nonetheless, about 66 % of the times confidence is again above 0.99, where virtually all estimates are correct. Therefore, estimating system differences with the gain scores predicted by L_{output} is a very reasonable method for developers to compare their systems when no judging resources are available. In particular, it can prove to be very useful at suggesting which systems perform very differently and which are very similar and thus require judging effort to gain more confidence.

6 Conclusions

We have shown how to adapt the Minimal Test Collections (MTC) family of algorithms for the evaluation of the MIREX Audio Music Similarity and Retrieval task. We showed that the distribution of $\overline{AG@k}$ scores is normally distributed, which allows us to look at it as a random variable whose expectation may be estimated with a certain level of confidence. This confidence is proportional to the number of similarity judgments available, and MTC ensures that the set of judgments we make to reach some confidence level is minimal.

Using data from the previous MIREX AMS evaluations, we fitted a model that allows us to predict gain scores when no judgments are available, and another model that considerably improves the predictions when judgments are available. Aided by these two models, MTC is shown to dramatically reduce the judging effort needed to rank systems with 95 % confidence. We simulated the MIREX AMS evaluations from 2007, 2009, 2010 and 2011, and showed that the average number of judgments needed is just 3 % with the Broad scale and 1.8 % with the Fine scale. The average accuracy of the estimated rankings is 0.948 with the Broad scale and 0.947 with the Fine scale, showing that MTC coupled with our models does not only require very little effort, but also produces accurate estimates. In fact, when systems show a statistically significant difference our estimates are always correct.

We further showed that these models can be used to rank systems without the need of making any judgments at all. Even though overall accuracy is slightly lower than when running MTC, we showed that the individual confidence scores can be trusted. Also, we showed that *the estimation errors are negligible in practice, because they compare to the disagreements produced by different human assessors.* This method can thus be employed to quickly check if there is a substantial difference between systems.

In general, the Fine scale seems to require fewer judgments than the Broad scale, while at the same time produces similarly accurate estimates. In previous work we also showed that the Fine scale is slightly more powerful and similarly stable as the Broad scale for a variety of measures [19], and that it is better correlated with final user satisfaction too [18]. Therefore, the evidence so far seems to indicate that the Fine scale works better than the Broad scale, suggesting its use alone in the MIREX AMS evaluations. Dropping the Broad scale would also lower the cost of the evaluations, at least in terms of judging time.

7 Future work

Two clear lines for future work can be identified. In this paper we used two sets of features to fit the regression models that allow us to predict gain scores: features based on the output of the systems and metadata, as well as features based on the known judgments. While these features work well in practice, a third set of features to consider could take advantage of the actual musical content used in the test collections, such as the similarity between the current document and those that have been judged as highly similar to the query. Unfortunately, the collection used in MIREX is not public, so we were not able to study these features here. Nonetheless, further research should definitely explore this line. Also, by no means are our models the only ones possible; other features or frameworks might prove better to predict gain scores. For instance, trying to predict gain scores on a per-system or per-query basis would probably improve the results.

The most important direction for further research is the study of low-cost evaluation methodologies for other MIR tasks. In accordance with previous work [19], we have shown here that the effort in evaluating a set of AMS systems can be greatly reduced, leaving open the possibility of building brand new test collections for other tasks for which making annotations is very expensive. For instance, the group of volunteers requested by MIREX for the annual evaluation of the AMS and SMS tasks could probably be better employed if some of them were instead dedicated to incrementally add new annotations for the other tasks in clear need of new collections [15].

Another clear setting for the application of low-cost methodologies is that of a researcher evaluating a set of systems with a private document collection, a scenario very common in MIR given the legal restrictions when sharing music corpora [7]. Those researchers, and in most cases public forums too, do not have the possibility of requesting large pools of external volunteers for annotating their collections. Thus, being able to evaluate systems with the minimal effort is paramount. To this end, low-cost evaluation methodologies must be investigated for the wealth of MIR tasks.

But in most of these tasks researchers rely on test collections annotated *a priori*, which can be very expensive and time consuming to build. However, we have seen that not all annotations are necessary to accurately rank systems. For instance, if two Audio Melody Extraction algorithms predict the same F0 (fundamental frequency) in a given audio frame, whether that F0 prediction is correct or not is not useful to know which of the two systems is better. The adoption of *a posteriori* evaluation methodologies such as MTC can take advantage of this idea to greatly reduce the annotation cost or allow the use of significantly larger collections. Getting to that point, though, requires a shift in the current evaluation practices. But given the benefits of doing so, both in terms of cost and reliability, we strongly encourage the MIR community to study these evaluation alternatives and progressively adopt them for a more rapid and stable development of the field.

Acknowledgments This research was supported by the Spanish Government (TSI-020110-2009-439, HAR2011-27540) as well as the Austrian Science Funds (FWF): P22856-N23.

Appendix

The models described in Sect. 4 to predict gain scores were fitted ignoring all data from the MIREX edition they were used for. For future editions though, we can use models fitted with all the available data from 2007, 2009, 2010 and 2011. Table 7 lists the fitted parameters, for both models and both similarity scales, for their use in future AMS evaluation experiments. Compared to the models fitted for each individual collection (see Table 2), these models produce similarly accurate estimates.

As an example, let us use L_{output} to estimate the Broad score of a document whose true score is 2 and has the following features: $pTEAM = 0.25$, $OV = 0.8053$, $pART = 0.0217$, $sGEN = 1$ and $pGEN = 0.8478$. Plugging these features and the parameters in Table 7 into Eq. (9):

Table 7 Parameters fitted for the regression models using all data from MIREX 2007, 2009, 2010 and 2011; and errors of the estimates (bottom)

Parameter	Broad scale		Fine scale	
	L_{output}	L_{judge}	L_{output}	L_{judge}
pTEAM	2.3677	2.0900	2.2223	1.4405
OV	1.9749	0.2420	2.0652	0.1139
pART	3.2041	–	2.9179	–
sGEN	1.9030	–	2.0174	–
pGEN	5.4144	–	5.4605	–
sGEN:pGEN	−2.9848	–	−3.4288	–
aSYS	–	1.1490	–	0.0115
aART	–	7.1853	–	0.2128
α_1	−3.2513	−5.5370	−1.7043	−2.1862
α_2	−5.3349	−12.2572	−2.6087	−4.6920
α_3	–	–	−3.2373	−6.9954
α_4	–	–	−3.7705	−9.2063
α_5	–	–	−4.2464	−11.2362
α_6	–	–	−4.8460	−13.5847
α_7	–	–	−5.5678	−15.8001
α_8	–	–	−6.6135	−18.2491
α_9	–	–	−8.4655	−21.2480
adjusted R^2	0.362	0.916	0.344	0.904
RMSE	0.651	0.275	24.1	8.97
Var	0.422	0.071	591	72

$$
\begin{aligned}
log\frac{P(G_i \geq 2)}{P(G_i < 2)} &= -5.3349 \\
&\quad + 2.3677 \cdot 0.25 + 1.9749 \cdot 0.8053 \\
&\quad + 3.2041 \cdot 0.0217 + 1.9030 \cdot 1 \\
&\quad + 5.4144 \cdot 0.8478 - 2.9848 \cdot 1 \cdot 0.8478 \\
&= 0.8798 \\
log\frac{P(G_i \geq 1)}{P(G_i < 1)} &= -3.2513 \\
&\quad + 2.3677 \cdot 0.25 + 1.9749 \cdot 0.8053 \\
&\quad + 3.2041 \cdot 0.0217 + 1.9030 \cdot 1 \\
&\quad + 5.4144 \cdot 0.8478 - 2.9848 \cdot 1 \cdot 0.8478 \\
&= 2.9634
\end{aligned}
$$

Next, we use the inverse logit function:

$$P(G_i \geq 2) = \frac{e^{0.8798}}{1 - e^{0.8798}} = 0.7068$$
$$P(G_i \geq 1) = \frac{e^{2.9634}}{1 - e^{2.9634}} = 0.9509$$

Plugging into Eq. (10):

$$P(G_i = 2) = 0.7068$$
$$P(G_i = 1) = 0.9509 - 0.7068 = 0.2441$$
$$P(G_i = 0) = 1 - 0.9509 = 0.0491$$

Finally, plugging into Eq. (1) we can compute the expectation and variance of G_i:

$$E[G_i] = 0.2441 + 0.7068 \cdot 2 = 1.6577$$
$$\text{Var}[G_i] = 0.2441 + 0.7068 \cdot 2^2 - 1.6577^2 = 0.3233$$

References

1. Carterette B (2007) Robust test collections for retrieval evaluation. In: International ACM SIGIR conference on research and development in information retrieval, pp 55–62
2. Carterette B (2008) Low-cost and robust evaluation of information retrieval systems. Ph.D. thesis, University of Massachusetts Amherst
3. Carterette B, Allan J, Sitaraman R (2006) Minimal test collections for retrieval evaluation. In: International ACM SIGIR conference on research and development in information retrieval, pp 268–275
4. Carterette B, Jones R (2007) Evaluating search engines by modeling the relationship between relevance and clicks. In: Annual conference on neural information processing systems
5. Carterette B, Pavlu V, Fang H, Kanoulas E (2009) Million query track 2009 overview. In: Text retrieval conference
6. Downie JS (2003) The MIR/MDL evaluation project white paper collection, 3rd edn. URL http://www.music-ir.org/evaluation/wp.html
7. Downie JS (2004) The scientific evaluation of music information retrieval systems: foundations and future. Comput Music J 28(2):12–23
8. Downie JS, Ehmann AF, Bay M, Jones MC (2010) The music information retrieval evaluation exchange: some observations and

insights. In: Zbigniew WR, Wieczorkowska AA (eds) Advances in music information retrieval. Springer, Berlin, pp 93–115
9. Flexer A, Schnitzer D (2010) Effects of album and artist filters in audio similarity computed for very large music databases. Comput Music J 34(3):20–28
10. Harman DK (2011) Information retrieval evaluation. Synth Lect Inf Concept Retr Serv 3(2):1–119
11. Jones MC, Downie JS, Ehmann AF (2007) Human similarity judgments: implications for the design of formal evaluations. In: International conference on music information retrieval, pp 539–542
12. Liu I, Agresti A (2005) The analysis of ordered categorical data: an overview and a survey of recent developments. Sociedad Estadística e Investigación Operativa Test 14(1):1–73
13. Long JS (1997) Regression models for categorical and limited dependent variables, 1st edn. Sage Publications, New York
14. Pohle T (2010) Automatic characterization of music for intuitive retrieval. Ph.D. thesis, Johannes Kepler University
15. Salamon J, Urbano J (2012) Current challenges in the evaluation of predominant melody extraction algorithms. In: International society for music information retrieval conference, pp 289–294
16. Soboroff I, Nicholas C, Cahan P (2001) Ranking retrieval systems without relevance judgments. In: International ACM SIGIR conference on research and development in information retrieval, pp 66–73
17. Urbano J (2011) Information retrieval meta-evaluation: challenges and opportunities in the music domain. In: International society for music information retrieval conference, pp 609–614
18. Urbano J, Downie JS, Mcfee B, Schedl M (2012) How significant is statistically significant? The case of audio music similarity and retrieval. In: International society for music information retrieval conference, pp 181–186
19. Urbano J, Martín D, Marrero M, Morato J (2011) Audio music similarity and retrieval: evaluation power and stability. In: International society for music information retrieval conference, pp 597–602
20. Urbano J, Schedl M (2012) Towards minimal test collections for evaluation of audio music similarity and retrieval. In: WWW international workshop on advances in music, information research, pp 917–923
21. Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. Inf Process Manag 36(5):697–716
22. Voorhees EM (2002) The philosophy of information retrieval evaluation. In: Workshop of the cross-language evaluation, forum, pp 355–370
23. Voorhees EM (2002) Whither music IR evaluation infrastructure: lessons to be learned from TREC. In: JCDL workshop on the creation of standardized test collections, tasks, and metrics for music information retrieval (MIR) and music digital library (MDL), evaluation, pp 7–13
24. Voorhees EM, Harman DK (2005) TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge
25. Yee T (2010) The VGAM package for categorical data analysis. J Stat Softw 32(10):1–34
26. Yee T, Wild C (1996) Vector generalized additive models. J R Stat Soc 58(3):481–493