Formalizing Evaluation in Music Information Retrieval

A Look at the MIREX Automatic Mood Classification Task

Sturm, Bob L.

Published in:
Proceedings of the 10 th International Symposium on Computer Music Multidisciplinary Research

Publication date:
2013

Document Version
Accepted author manuscript, peer reviewed version

Formalizing Evaluation in Music Information Retrieval: A Look at the MIREX Automatic Mood Classification Task

Bob L. Sturm*

Audio Analysis Lab, AD:MT, Aalborg University Copenhagen
A.C. Meyers Vænge 15, DK-2450 Copenhagen SV, Denmark
bst@create.aau.dk

Abstract. We develop a formalism to disambiguate the evaluation of music information retrieval systems. We define a "system," what it means to "analyze" one, and make clear the aims, parts, design, execution, interpretation, and assumptions of its "evaluation." We apply this formalism to discuss the MIREX automatic mood classification task.

Keywords: Evaluation, systems analysis music information research

1 Introduction

While a considerable amount of work has contributed to standardizing the evaluation of solutions to problems in music information retrieval (MIR), e.g., [6–8,10,15,17,22], and some work contributes to the critical discussion of evaluation in MIR, e.g., [1,4,5,12,18,20–30], very little work contributes to formalizing evaluation in MIR, i.e., disambiguating evaluation to make clear its aims, parts, design, execution, interpretation, and assumptions. Much of the formalism underlying evaluation in MIR has been adopted, unknowingly or without much question, from that of machine learning and information retrieval. However, this formalism remains for the most part implicit in MIR evaluation. In this work, we aim to make it explicit.

To be sure, the massive and concerted efforts of MIREX (Music Information Retrieval Evaluation eXchange) [6,7,10] enables a systematic and rigorous evaluation of MIR systems proposed for such tasks as beat tracking, chord and onset detection, melody extraction, and genre and emotion recognition. Inspired by the Text Retrieval Conference (TREC) [8], MIREX aims to standardize MIR *benchmarks*: realistic information needs and problem formulations (e.g., appropriately

* This work is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd; and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. Part of this work was undertaken during a visit to the Centre for Digital Music at Queen Mary University of London, supported by EPSRC grant EP/G007144/1 (Plumbley). This publication only reflects the authors' views.

describing the mood of recorded music), test data collections (e.g., MIREX music mood dataset [15]), and performance measurement (e.g., accuracy of labels to "ground truth"). That MIREX has had an impact is indisputable: the trials, techniques, data, and/or software of MIREX has between 2005 – 2010 been cited in over 500 journal articles, conference papers, master's theses and PhD dissertations [6]. However, it has only been recently that the methodologies of some evaluation in MIREX have begun to be questioned.

Criticism of evaluation practices of MIR systems is not new, but it is certainly not as widespread as work proposing new MIR systems [30]. In 2003, Downie [9] argued for standardizing evaluation practices in MIR, and thus coordinated three workshops on the topic [8], and developed MIREX [17]. In 2005, Pampalk et al. [18] discovered a source of bias in the evaluation of music similarity algorithms: performance is significantly better when train and test datasets contain material from the same artist or album, than when such tracks are isolated to only one of the datasets. In 2006, Flexer [12] addressed the lack of but necessity for formal statistical testing in MIR evaluation. In the same year, McKay and Fujinaga [16] argued that the problem and evaluation of solutions for music genre recognition must be rethought, from building the datasets used for training and testing, to defining the overall goal in a more realistic way. In 2007, Craft et al. [4, 5] pinpoint problems with music datasets having naturally vague labels, and propose a slightly different approach to evaluation than what has been used.

More recently, four articles of eight in a forthcoming special issue entitled, "MIRrors: The past of Music Information Research reflects on its future",[1] are critical of current evaluation practices in MIR [1, 20, 26, 30]. Aucouturier and Bigand [1] point out that the extent to which research and evaluation in MIR is goal-oriented situate it outside the practice of science. Schedl et al. [20] highlight the near absence of the consideration of users in typical MIR system evaluations. Urbano et al. [30] observe that current MIR evaluation practices play little role in the development of MIR systems: problems and success criteria are often artificial with no real use contexts; evaluation procedures are often weak, nonuniform, irreproducible and incomparable; and little to no effort is made to learn why and/or why not a system worked, and how to then improve it. In our article [26], we show how the standard approach used most to evaluate systems for music genre recognition lacks the scientific validity for making any useful conclusion. This finding applies also to evaluation in music emotion recognition [24], and by extension, to autotagging since the majority of tags are genre and emotion [3].

Clearly, though the efforts in standardizing MIR evaluation have been very successful, the current practices of MIR evaluation sit upon a foundation that might be adequate for solving some problems, but not others. To answer which might be the case in MIR, there is thus a need to answer several critical questions: What does it mean to evaluate a system? What is the aim of an evaluation? What are its parts? What is its design? How is it executed? How can the results be interpreted and presented? What assumptions are made? Where can problems

[1] This will appear in the Journal of Intelligent Information Systems, with the special issue editors Perfecto Herrera-Boyer and Fabien Gouyon.

arise, and how can they be solved? Our work here attempts to contribute to this discussion through formalizing evaluation. We first review the concept of a system, and what it means to analyze and evaluate one. We then review the design and analysis of experiments, and discuss relevance and scientific validity. Finally, we use this formalism to look at and critique the evaluation performed in the MIREX automatic mood classification task (MIREX-AMC).

2 Formalization of System Evaluation

In this section, we define a system, and describe its analysis and evaluation. We then review formalized experimental design and analysis. We finally discuss relevance and scientific validity.

2.1 Systems and System Analysis

A *system* is a connected set of interacting and interdependent components that together address a goal [19]. Of the many systems in which MIR is interested, there are four essential kinds of components: *operator(s)* (agent(s) that employ the system); *instructions* (a specification for the operator(s), like an application programming interface); *algorithm(s)* (each a set of describable and ordered operations to transduce an input into an output); and *environment* (connections between components, external databases, the space within which the system operates, its boundaries). A system can fail (to meet a goal) from any combination of errors on the parts of its components.

An *analysis* of a system addresses questions and hypotheses related to its past, present and future. About its past, one can study the system history: why and how it came to be, the decisions made, past implementations and applications, its goals, its successes and failures, etc. About its present, one can study its current implementation and applications, evaluate its success with respect to a goal, compare it to alternative existing systems, etc. About its future, one can study ways to improve it with respect to a goal, adapt it for a goal and predict its success, perform a cost-benefit analysis, etc. All of these involve breaking the system into its components (which also include the decisions made within the components and their connection), breaking the components into their components (which also include the decisions made within the components and their connection), evaluating and tuning components, and so on.

An essential part of a system analysis is an *evaluation*: a "fact-finding campaign" intended to address a number of relevant questions and/or hypotheses related to the goal of a system. (The intention is to seek "truth," whether or not it exists.) Through evaluation, the analysis of a system seeks to improve, adapt, and advertise it. Evaluating a system or a component with respect to its goal is to scientifically address questions and hypotheses relevant to that goal by the design, implementation and analysis of relevant and valid experiments.

2.2 Experimental Design and Analysis

An *experiment* consists of assigning and applying treatments to units, and measuring responses to determine the real effects of those treatments.[2] A *treatment* is the thing being evaluated. *Units* are the materials of the experiment. An *experimental unit* is a group of materials to which a treatment is applied, and an *observational unit* (or *plot*) is a group of materials from which one measures a response. An *experimental design* specifies how treatments are assigned to plots. A *response* is the "real" effect of a treatment, and its determination is the goal of an experiment. A *measurement* is a quantitative description of that response, relevant to the question or hypothesis being tested. The *analysis* of measurements involves the application of statistics to facilitate valid conclusions, implemented to carefully control all sources of variation and bias, in view of the hypothesis. An experiment is *valid* for a question or hypothesis when it can logically answer that question or hypothesis (whether or not the result is really "true").

Formally, the set of all N plots in an experiment is notated Ω, and the set of all t treatments is notated $\mathcal{T}$. The experimental design, $T : \Omega \to \mathcal{T}$, is a function that maps one plot to one treatment. For plot $\omega \in \Omega$, a measurement made of its response to a treatment is y_ω. The response of the treatment applied to ω is $\tau_{T(\omega)}$. Thus, the measurement of plot ω thus produces y_ω, which is related in some way to the response $\tau_{T(\omega)}$. From the measurements then, one wishes to estimate the responses, and thereby quantify and compare the treatments.

To estimate responses, one must *model* measurements. A typical model is linear, where the measurements are assumed to be realizations of random variables. For a plot ω, let its measurement be modeled by the random variable Y_ω arising from the non-random response of the treatment $\tau_{T(\omega)}$, and a random variable Z_ω encompassing measurement error, i.e., effects contributed by the plot independent of the treatment, and other factors. The measurement y_ω, a realization of Y_ω, thus includes things unrelated to the treatment, z_ω, a realization of Z_ω. With the linear model, one decomposes the measurement as $y_\omega = \tau_{T(\omega)} + z_\omega$, and models it with $Y_\omega = \tau_{T(\omega)} + Z_\omega$.

Given t treatments and N measurements, an experiment is modeled by

$$\mathbf{Y} = \boldsymbol{\tau} + \mathbf{Z} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_t]_{N \times t} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \end{bmatrix} + \mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z} \tag{1}$$

where $\mathbf{Y}$ is a vector of N measurements, $\mathbf{Z}$ is a length-N random vector, and $\boldsymbol{\tau}$ is a vector of the N responses to the t treatments $\mathcal{T}$. The matrix $\mathbf{X} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_t]$ is the *plan*, or experimental design: column i specifies which plots are treated with treatment i. Finally, the vector $\boldsymbol{\beta}$ contains the responses of the t treatments.

If the true means of $\mathbf{Y}$ and $\mathbf{Z}$ are known, the responses can be found exactly by solving $\mathbf{X}\boldsymbol{\beta} = E[\mathbf{Y}] - E[\mathbf{Z}]$. This knowledge is typically not possessed (an

[2] We use the terminology and notation of Bailey [2].

experiment would not be necessary then), and so one must build models of $\mathbf{Y}$ and $\mathbf{Z}$. From these, the responses can be estimated, notated $\hat{\boldsymbol{\beta}}$, *and* the relationship between them can be found, e.g., the bias and variance of the estimator, a bound on $|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|^2$, positive correlation, and so on. Only then can one test hypotheses and answer questions subject to the strict assumptions of the selected models.

The *simple textbook model* [2] makes the assumption $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N\sigma^2)$, i.e., that $\mathbf{Z}$ is multivariate Gaussian with mean $E[\mathbf{Z}] = \mathbf{0}_N$ (a length-N vector of zeros), and $Cov[\mathbf{Z}] = \mathbf{I}_N\sigma^2$ for $\sigma^2 > 0$, where $\mathbf{I}_N$ is an identity matrix of size N. The choice of this model makes the assumption that the measurements of the responses are affected only by independent and identically distributed zero-mean white noise of power σ^2. In this case, $\mathbf{X}\boldsymbol{\beta} = E[\mathbf{Y}]$, and one knows the measurements will vary due to the noise as $Cov[\mathbf{Y}] = \mathbf{I}_N\sigma^2$.

The simple textbook model may not fit the measurements when Ω is very heterogeneous, and/or when its units are selected randomly from a population In such cases, "group effects" can bias the measurements, which then make estimates of $\boldsymbol{\beta}$ using the simple textbook model consistently poor. The *fixed-effects model* [2] decomposes $\mathbf{Z} = \boldsymbol{\alpha} + \mathbf{W}$ where the vector $\boldsymbol{\alpha}$ describes the contribution of each plot to each measurement invariant of the treatment, and $\mathbf{W}$ is a random vector, where $E[\mathbf{W}] = \mathbf{0}_N$ and $Cov[\mathbf{W}] = \mathbf{I}_N\sigma^2$. In this case, $\mathbf{X}\boldsymbol{\beta} = E[\mathbf{Y}] + \boldsymbol{\alpha}$, and $Cov[\mathbf{Y}] = \mathbf{I}_N\sigma^2$. The *random-effects model* [2] decomposes $\mathbf{Z} = \mathbf{U} + \mathbf{W}$, where both $\mathbf{U}$ and $\mathbf{W}$ are random vectors. As in the fixed-effects model, $\mathbf{U}$ describes the contribution of each plot to each measurement invariant of the treatment, but it takes into account the uncertainty in the random sampling of the experimental units from the population. In this case, one assumes $E[\mathbf{W}] = E[\mathbf{U}] = \mathbf{0}_N$ and $Cov[\mathbf{W}] = \mathbf{I}_N\sigma^2$, and so $\mathbf{X}\boldsymbol{\beta} = E[\mathbf{Y}]$, but $Cov[\mathbf{Y}] = Cov[\mathbf{U}] + Cov[\mathbf{W}] = Cov[\mathbf{U}] + \mathbf{I}_N\sigma^2$. There are also *mixed-effects models*, incorporating both random- and fixed-effects [31]. Since there are two sources of variance in these cases, the error of the response estimates can vary to a high degree.

With a model, and estimates of the responses, hypothesis testing becomes possible [2]. This involves specifying a null hypothesis and comparing its probability, to a pre-specified limit of statistical significance α, given all assumptions of the model. For example, a null hypothesis can be that the responses of all treatments are equivalent, i.e., that $\tau_1 = \tau_2 = \ldots = \tau_t$. One then computes the probability p of observing the estimates of the responses given the null hypothesis and model assumptions are true. If $p < \alpha$, then the null hypothesis is rejected, i.e., it is unlikely with respect to α given the model and observations that there is no differences between the responses of the t treatments.

2.3 Relevance and Scientific Validity

Even when estimates of responses have small error in some model, and null hypotheses are rejected, that does not mean an experiment has answered *the right question.* When the aims of the researcher are incompatible with the design, implementation and analysis of an experiment, an "error of the third kind" has occurred [14]: "giving the right answer to the wrong question." Furthermore, there is no assurance that, though something may be accurately measurable, it

is thus relevant to the question of interest.[3] Experiments must thus be designed after the hypothesis is formulated, and the scientific and statistical questions are "deconstructed" [14].

The relationships between an experiment to what is intended center upon the concept of validity. One kind of validity is that a meaningful or scientific conclusion can come from the experiment. If the variance in the estimates of responses are too large for any meaningful conclusion about treatments, then the experiment has no *conclusion validity*. Another kind of validity is that the estimated responses come entirely from the treatments. When they come in part from unrelated but confounded factors not taken into account in the measurement model, then a causal inference cannot logically be made between the treatments and the responses in an experiment, and it has no *internal validity*. Such problems arise from, e.g., a biased selection of units, the experimental design (mapping of treatments to units), the choice of measurements, who makes the measurements and how, and so on. A third kind of validity is the logical generalization of experimental results to the population, of which the experimental units are a subset. If the units are not a random sample of the population, or if an experiment has no conclusion and internal validity, then the experiment has no *external validity*.

3 The MIREX Automatic Mood Classification Task

We now apply the formalism above to the MIREX automatic mood classification task (MIREX-AMC), which has been run the same way since 2007 [15, 17]. A submitted algorithm labels a music recording with one of five mood labels. Performance is assessed by the number of "ground truth" labels an algorithm reproduces in the private MIREX-AMC dataset. This dataset consists of 600 30-s excerpts of music, with a "ground truth" generated with human assessment.

3.1 Systems and System Analysis

In MIREX-AMC, a participant submits a *machine learning algorithm*, which is composed of a feature extraction algorithm, and a training and classification algorithm. The inputs and outputs of these algorithms are specified by the instructions of MIREX-AMC.[4] A MIREX-AMC organizer — i.e., the operator — then integrates the machine learning algorithm with the environment: a computer, the private MIREX-AMC dataset, etc. This produces one system. Since MIREX-AMC uses 3-fold stratified cross-validation, each submitted machine learning algorithm produces three systems. Each system is built of the same machine learning algorithm, but trained with different data.

[3] R. Hamming, "You get what you measure", lecture at Naval Post- graduate School, June 1995. `http://www.youtube.com/watch?v=LNhcaVi3zPA`

[4] `http://www.music-ir.org/mirex/wiki/2013:Audio_Classification_(Train/Test)_Tasks`

MIREX-AMC analyzes a system by comparing its output labels in a dataset to the "ground truth" labels. This produces several figures of merit: mean accuracies (per class, per fold, and overall), and confusions. Such a test of a system addresses the question: How many "true" labels of the test dataset does this system produce? Combining the tests of three systems from a machine learning algorithm addresses the question: How many "true" labels of the test dataset does this machine learning algorithm produce? Finally, statistically comparing the test results of several machine learning algorithms addresses the question: is there a significant difference in the numbers of "true" labels of this dataset produced by these machine learning algorithms? Since MIREX-AMC only tests and compares systems in their production of "true" labels of a test dataset, no matter *how* that production happens, and does not address the motivations of the systems it analyzes — to automatically recognize emotion in recorded music — its system analysis is shallow. The analysis is not concerned with, e.g., how a system makes its decisions, whether it is doing so in an "acceptable manner" (with respect to some use case), whether any system is better than another (with respect to some use case), how a system can be improved (with respect to some use case), and so on.

3.2 Experimental Design and Analysis

MIREX-AMC evaluates a machine learning algorithm by an operator applying the three resulting systems (treatments) to three folds (experimental units) of the test dataset. The experimental design of MIREX-AMC maps one system to only one fold. If one system was mapped to another fold as well, then it would be tested using some of the data with which it was trained. The operator measures for each of the five labels in a fold the proportion of matching labels produced by a system. Denote the five class labels by a, b, c, d, e. Hence, an observational unit is the set of excerpts in a fold from the same class, i.e., all excerpts with label a. Notate the measurements of system i applied to fold i, $\widehat{\mathbf{Y}}_i = [\hat{y}_{i,a}, \hat{y}_{i,b}, \ldots, \hat{y}_{i,e}]^T$, where $\hat{y}_{i,a}$ is the number of excerpts in fold i that system i labels a correctly, divided by the number of excerpts in that fold with a "true" label a.

MIREX-AMC reports several figures of merit for a machine learning algorithm: the *ordered set of fold-specific mean classification accuracies* ($\mathbf{1}_n$ is a length-n vector of ones divided by n)

$$\widehat{\mathcal{Y}} := \{\mathbf{1}_5^T\widehat{\mathbf{Y}}_1, \mathbf{1}_5^T\widehat{\mathbf{Y}}_2, \mathbf{1}_5^T\widehat{\mathbf{Y}}_3\} \tag{2}$$

the *class-specific mean classification accuracies*

$$\widehat{\mathbf{S}} := (\widehat{\mathbf{Y}}_1 + \widehat{\mathbf{Y}}_2 + \widehat{\mathbf{Y}}_3)/3 \tag{3}$$

and the *mean classification accuracy*

$$\hat{y} := \mathbf{1}_5^T\widehat{\mathbf{S}}. \tag{4}$$

MIREX-AMC 2007 reports only mean classification accuracy and confusion table. These results are further analyzed in [15]. In all years since, MIREX-AMC

reports fold-specific mean classification accuracies, class-specific mean classification accuracies, mean classification accuracy, and confusion table. Statistical tests are also run in these years to determine if there are significant differences between fold-specific and class-specific accuracies for all systems.

Since 2008, MIREX-AMC tests two null hypotheses. First, for any of the three folds, the classification accuracies are the same for all L machine learning algorithms, i.e.,

$$\mathcal{H}_0^{(i)} : \widehat{\mathcal{Y}}_1(i) = \widehat{\mathcal{Y}}_2(i) = \ldots = \widehat{\mathcal{Y}}_L(i) \tag{5}$$

where $\widehat{\mathcal{Y}}_l(i)$ is the ith element of $\widehat{\mathcal{Y}}_l$, the ordered set of fold-specific mean classification accuracies of machine learning algorithm l. Second, for any of the five classes, k, the classification accuracies are the same for all L machine learning algorithms, i.e.,

$$\mathcal{H}_0^{(k)} : \mathbf{e}_k^T \widehat{\mathbf{S}}_1 = \mathbf{e}_k^T \widehat{\mathbf{S}}_2 = \ldots = \mathbf{e}_k^T \widehat{\mathbf{S}}_L \tag{6}$$

where $\mathbf{e}_k$ is the kth standard vector. To test these hypotheses, MIREX-AMC applies the *Method of Ranks* [13]. This approach builds a two-way table with the L treatments as columns and three (fold) or five (class) observational units as rows. Each measurement in a row is assigned a rank, with the largest value assigned L, the next largest $L-1$, and the smallest assigned 1.[5] If the classification accuracies are the same for all treatments, then the distribution of the ranks in the two-way table will be random. Assuming the measurements are mutually independent, the chi-squared test can then be used to test the null hypotheses. If either null hypothesis is rejected, then the fold-specific or class-specific mean classification accuracies show a dependence on the machine learning algorithm. MIREX-AMC also makes pairwise comparisons to test for significant differences between the measured accuracies of machine learning algorithms (classes) and systems (folds).

Through the figures of merit it reports, and the statistical tests it performs, MIREX-AMC implicitly assumes the simple textbook model. Furthermore, it assumes the same model applies to all machine learning algorithms it tests. A linear model explaining the measurements of MIREX-AMC is given by

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_1 \\ \boldsymbol{\tau}_2 \\ \boldsymbol{\tau}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \mathbf{Z}_3 \end{bmatrix} = \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \mathbf{Z}_3 \end{bmatrix} \tag{7}$$

where $\boldsymbol{\tau}_i = [\tau_{i,a}, \tau_{i,b}, \ldots, \tau_{i,e}]^T$ are the determinate responses of system i, and $\mathbf{Z}_i$ are contributions to the measurements that are not due to system i. The ordered set of fold-specific mean classification accuracies is thus modeled

$$\mathcal{Y} := \{\mathbf{1}_5^T(\boldsymbol{\tau}_1 + \mathbf{Z}_1), \mathbf{1}_5^T(\boldsymbol{\tau}_2 + \mathbf{Z}_2), \mathbf{1}_5^T(\boldsymbol{\tau}_3 + \mathbf{Z}_3)\}. \tag{8}$$

The expectation and variance of the ith member of this set are given by

$$E[\mathcal{Y}(i)] = \mathbf{1}_5^T(\boldsymbol{\tau}_i + E[\mathbf{Z}_i]) \tag{9}$$

$$Var[\mathcal{Y}(i)] = Var\left[\mathbf{1}_5^T E[\mathbf{Z}_i]\right]. \tag{10}$$

[5] From the MATLAB implementation `friedman`.

The class-specific mean classification accuracies are thus modeled

$$\mathbf{S} = \frac{1}{3}\sum_{i=1}^{3}\boldsymbol{\tau}_i + \frac{1}{3}\sum_{i=1}^{3}\mathbf{Z}_i. \tag{11}$$

and the expectation and covariance of this are

$$E[\mathbf{S}] = \frac{1}{3}\sum_{i=1}^{3}\boldsymbol{\tau}_i + \frac{1}{3}\sum_{i=1}^{3}E[\mathbf{Z}_i] \tag{12}$$

$$Cov[\mathbf{S}] = \frac{1}{9}Cov\left[\sum_{i=1}^{3}\mathbf{Z}_i\right]. \tag{13}$$

Finally, the mean classification accuracy is thus modeled

$$y = \frac{1}{3}\sum_{i=1}^{3}\mathbf{1}_5^T\boldsymbol{\tau}_i + \frac{1}{3}\sum_{i=1}^{3}\mathbf{1}_5^T\mathbf{Z}_i. \tag{14}$$

and its expectation and variance are

$$E[y] = \frac{1}{3}\sum_{i=1}^{3}\mathbf{1}_5^T\boldsymbol{\tau}_i + \frac{1}{3}\sum_{i=1}^{3}E[\mathbf{1}_5^T\mathbf{Z}_i] \tag{15}$$

$$Var[y] = \frac{1}{9}Var\left[\sum_{i=1}^{3}\mathbf{1}_5^T\mathbf{Z}_i\right]. \tag{16}$$

The responses $\boldsymbol{\beta}$ can be estimated from $\widehat{\mathbf{S}}$ if $E[\mathbf{Z}_i]$ is known for all i. Typically, these are not known and must be modeled. Finally, the error of the estimate of $\boldsymbol{\beta}$ depends upon $Cov[\mathbf{S}]$, which itself depends on the covariance of each $E[\mathbf{Z}_i]$.

It is clear that since MIREX-AMC tests (5) and (6), it implicitly assumes for all i, $E[\mathbf{Z}_i] = \mathbf{0}_5$ (where $\mathbf{0}_5$ is a length-5 vector of zeros), and $Cov[\mathbf{Z}_i] = \sigma^2\mathbf{I}_5$ for some σ^2 quite small. Otherwise, the measured fold-specific mean classification accuracies do not reflect the responses of the systems, and the measured class-specific mean classification accuracies do not reflect their "true" values for the machine learning algorithms. Furthermore, MIREX-AMC does not specify any bounds on the errors of the estimates, which come from the covariance of all $\mathbf{Z}_i$. Hence, only subject to the strict assumptions on this measurement model — which appears supported only by convenience and not evidence — can the experiments of MIREX-AMC provide any scientific knowledge [11].

3.3 Relevance and Scientific Validity

MIREX-AMC measures and compares how many "correct" labels the systems of machine learning algorithms produce for a private set of labeled data, regardless of how the systems select the labels. Whether or not it is acceptable to produce

"correct" labels by any means depends on a use case; and thus the relevance of the measurements and comparisons made in MIREX-AMC depend on a use case. If the goal is "classifying music by moods" [15], and *by* denotes using criteria relevant to mood, for example, music modality and tempo, and not possible confounds, like female voice, presence of piano, and the 2nd MFCC, then it is not relevant to measure the correct labels produced by the systems of a machine learning algorithm for a specific dataset, unless one assumes label selection can only be either random or by relevant criteria in that dataset, or the responses of the systems in the measurement model can be estimated with error bounds that are informative. In other words, classification accuracy is not enough for this goal [24,26]. With respect to the aim "classifying music by moods" then, MIREX-AMC has no conclusion validity (it does not unambiguously address whether a system is classifying music by mood), it has no internal validity (possible confounded factors are not controlled), and thus has no external validity.

4 Conclusion

We have attempted to contribute a formalism missing from the conversation of evaluation in MIR. The development of a solid formalism is especially important now considering the recent appearance of several works highly critical of current evaluation practices in MIR. Such a formalism can facilitate real and cooperative advancement of evaluation in MIR, by disambiguating its aims, parts, design, execution, interpretation, and assumptions. At a high level, a formalism is essential to scientifically address questions such as, "What accomplishment is being attempted?", "What must be done to achieve that?", "What was actually accomplished?", and "What does that accomplishment contribute?"

As a specific case, we discuss MIREX-AMC with this formalism. This shows how its system analysis is shallow: of interest is only the number of "true" labels produced by systems from a machine learning algorithm, and not how the systems work, whether they work by using criteria relevant to mood, where their weak points lie and how they can be improved, and so on. In terms of formalized experimental design and analysis, this formalism clarifies the evaluation applied in MIREX-AMC, and uncovers its implicit measurement model and its accompanying assumptions, as well as the critical lack of estimation error analysis. Finally, we see that the relevance of the measurements of MIREX-AMC depend on a use case. If the goal is classifying music by moods using criteria relevant to mood, then MIREX-AMC has no conclusion validity, no internal validity, and no external validity.

Acknowledgments Many thanks to Mathieu Barthet for inviting this paper, and to Nick Collins for the fun discussions.

References

1. Aucouturier, J.J., Bigand, E.: Seven problems that keep MIR from attracting the interest of cognition and neuroscience. J. Intell. Info. Systems (2013, in press)

2. Bailey, R.A.: Design of comparative experiments. Cambridge University Press (2008)
3. T. Bertin-Mahieux, D. Eck, and M. Mandel, "Automatic tagging of audio: The state-of-the-art," in *Machine Audition: Principles, Algorithms and Systems* (W. Wang, ed.), IGI Publishing, 2010.
4. Craft, A.: The role of culture in the music genre classification task: human behaviour and its effect on methodology and evaluation. Tech. rep., Queen Mary University of London (Nov 2007)
5. Craft, A., Wiggins, G.A., Crawford, T.: How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In: Proc. Int. Soc. Music Info. Retrieval (2007)
6. Cunningham, S.J., Bainbridge, D., Downie, J.S.: The impact of MIREX on scholarly research. In: Proc. Int. Soc. Music Info. Retrieval (2012)
7. Downie, J., Ehmann, A., Bay, M., Jones, M.: The music information retrieval evaluation exchange: Some observations and insights. In: Advances in Music Information Retrieval, pp. 93–115. Springer Berlin / Heidelberg (2010)
8. Downie, J.S. (ed.): The MIR/MDL Evaluation Project White Paper Collection (2003), `http://www.music-ir.org/evaluation/wp.html`
9. Downie, J.S.: Toward the scientific evaluation of music information retrieval systems. In: Int. Soc. Music Info. Retrieval. Baltimore, USA (Oct 2003)
10. Downie, J.S.: The scientific evaluation of music information retrieval systems: Foundations and future. Computer Music Journal 28(2), 12–23 (2004)
11. E. R. Dougherty and L. A. Dalton, "Scientific knowledge is possible with small-sample classification," *EURASIP J. Bioinformatics and Systems Biology*, vol. 10, 2013.
12. Flexer, A.: Statistical evaluation of music information retrieval experiments. J. New Music Research 35(2), 113–120 (2006)
13. Friedman, M.: The use of ranks to avoid the assumption of normality in the analysis of variance. J. American Statistical Assoc. pp. 675–701 (1937)
14. Hand, D.J.: Deconstructing statistical questions. J. Royal Statist. Soc. A (Statistics in Society) 157(3), 317–356 (1994)
15. Hu, X., Downie, J.S., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX audio mood classification task: lessons learned. In: Proc. Int. Soc. Music Info. Retrieval (2008)
16. McKay, C., Fujinaga, I.: Music genre classification: Is it worth pursuing and how can it be improved? In: Proc. Int. Soc. Music Info. Retrieval. Victoria, Canada (Oct 2006)
17. MIREX: `http://www.music-ir.org/mirex` (2012)
18. Pampalk, E., Flexer, A., Widmer, G.: Improvements of audio-based music similarity and genre classification. In: Proc. Int. Soc. Music Info. Retrieval. pp. 628–233 (Sep 2005)
19. Rowe, W.: Why system science and cybernetics? IEEE Trans. Systems and Cybernetics 1, 2–3 (Nov 1965)
20. Schedl, M., Flexer, A., Urbano, J.: The neglected user in music information retrieval research. J. Intell. Info. Systems (2013, in press)
21. Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., Widmer, G.: Roadmap for Music Information ReSearch. Creative Commons (2013), `http://mtg.upf.edu/MIRES`
22. Sturm, B.L.: A survey of evaluation in music genre recognition. In: Proc. Adaptive Multimedia Retrieval. Copenhagen, Denmark (Oct 2012)

23. Sturm, B.L.: Two systems for automatic music genre recognition: What are they really recognizing? In: Proc. ACM MIRUM Workshop. Nara, Japan (Nov 2012)
24. Sturm, B.L.: Evaluating music emotion recognition: Lessons from music genre recognition? In: Proc. IEEE Int. Conf. Multimedia & Expo (2013)
25. Sturm, B.L.: The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv (2013), http://arxiv.org/abs/1306.1461
26. Sturm, B.L.: Classification accuracy is not enough: On the evaluation of music genre recognition systems. J. Intell. Info. Systems (2013, in press)
27. Urbano, J.: Information retrieval meta-evaluation: Challenges and opportunities in the music domain. In: Proc. Int. Soc. Music Info. Retrieval. pp. 609–614 (2011)
28. Urbano, J., McFee, B., Downie, J.S., Schedl, M.: How significant is statistically significant? The case of audio music similarity and retrieval. In: Proc. Int. Soc. Music Info. Retrieval (2012)
29. Urbano, J., Mónica, M., Morato, J.: Audio music similarity and retrieval: Evaluation power and stability. In: Proc. Int. Soc. Music Info. Retrieval. pp. 597–602 (2011)
30. Urbano, J., Schedl, M., Serra, X.: Evaluation in music information retrieval. J. Intell. Info. Systems (2013, in press)
31. Venables, W.N., Ripley, B.D.: Modern applied statistics with S. Statistics and Computing, Springer, 4 edn. (2002)