On comparative statistics for labelling tasks: What can we learn from MIREX ACE 2013?

Burgoyne, J.A.; de Haas, W.B.; Pauwels, J.

Publication date
2014

Document Version
Final published version

Published in
Proceedings of the 15th International Society for Music Information Retrieval Conference

License
CC BY

Citation for published version (APA):
ON COMPARATIVE STATISTICS FOR LABELLING TASKS: WHAT CAN WE LEARN FROM MIREX ACE 2013?

John Ashley Burgoyne  
Universiteit van Amsterdam  
j.a.burgoyne@uva.nl

W. Bas de Haas  
Universiteit Utrecht  
w.b.dehaas@uu.nl

Johan Pauwels  
STMS IRCAM–CNRS–UPMC  
johan.pauwels@gmail.com

ABSTRACT

For MIREX 2013, the evaluation of audio chord estimation (ACE) followed a new scheme. Using chord vocabularies of differing complexity as well as segmenting measures, the new scheme provides more information than the ACE evaluations from previous years. With this new information, however, comes new interpretive challenges. What are the correlations among different songs and, more importantly, different submissions across the new measures? Performance falls off for all submissions as the vocabularies increase in complexity, but does it do so directly in proportion to the number of more complex chords, or are certain algorithms indeed more robust? What are the outliers, song-algorithm pairs where the performance was substantially higher or lower than would be predicted, and how can they be explained? Answering these questions requires moving beyond the Friedman tests that have most often been used to compare algorithms to a richer underlying model.

We propose a logistic-regression approach for generating comparative statistics for MIREX ACE, supported with generalised estimating equations (GEEs) to correct for repeated measures. We use the MIREX 2013 ACE results as a case study to illustrate our proposed method, including some of interesting aspects of the evaluation that might not apparent from the headline results alone.

1. INTRODUCTION

Automatic chord estimation (ACE) has a long tradition within the music information retrieval (MIR) community, and chord transcriptions are generally recognised as a useful mid-level representation in academia as well as in industry. For instance, in an academic context it has been shown that chords are interesting for addressing musicological hypotheses [3,13], and that they can be used as a mid-level feature to aid in retrieval tasks like cover-song detection [7,10]. In an industrial setting, music start-ups like Riffstation 1 and Chordify 2 use ACE in their music teaching tools, and at the time of writing, Chordify attracts more than two million unique visitors every month [6].

In order to compare different algorithmic approaches in an impartial setting, the Music Information Retrieval Evaluation eXchange (MIREX) introduced an annual ACE task in 2008. Since then, between 11 and 18 algorithms have been submitted each year by between 6 and 13 teams. Despite the fact that ACE algorithms are used outside of academic environments, and even though the number of MIREX participants has decreased slightly over the last three years, the problem of automatic chord estimation is nowhere near solved. Automatically extracted chord sequences have classically been evaluated by calculating the chord symbol recall (CSR), which reflects the proportion of correctly labelled chords in a single song, and a weighted chord symbol recall (WCSR), which weights the average CSR of a set of songs by their length. On fresh validation data, the best-performing algorithms in 2013 achieved WCSR of only 75 percent, and that only when the range of possible chords was restricted exclusively to the 25 major, minor and “no-chord” labels; the figure drops to 60 percent when the evaluation is extended to include seventh chords (see Table1).

MIREX is a terrific platform for evaluating the performance of ACE algorithms, but by 2010 it was already being recognised that the metrics could be improved. At that time, they included only CSR and WCSR using a vocabulary of 12 major chords, 12 minor chords and a “no-chord” label. At ISMIR 2010, a group of ten researchers met to discuss their dissatisfaction. In the resulting ‘Utrecht Agreement’, it was proposed that future evaluations should include more diverse chord vocabularies, such as seventh chords and inversions, as the 25 -chord vocabulary was considered a rather coarse representation of tonal harmony. Furthermore, the group agreed that it was important to include a measure of segmentation quality in addition to CSR and WCSR.

At approximately the same time, Christopher Harte proposed a formalisation of measures that implemented the aspirations indicated in the Utrecht agreement [8]. Recently, Pauwels and Peeters reformulated and extended Harte’s work with the precise aim of handling differences in chord vocabulary between annotated ground truth and algorithmic
Table 1. Number of supported chord types, inversion support, training support, and an excellent ‘one size fits all’ approach that can be applied with minimal regard to the underlying distribution of the data, but these benefits come with costs. Like any non-parametric test, Friedman’s ANOVA can be less powerful than parametric alternatives where the distribution is known, and the rank transformation can obscure information inherent to the underlying measurement, magnifying trivial differences and neutralising significant inter-correlations.

But there is no need to pay the costs of Friedman’s ANOVA for evaluating chord estimation. Fundamentally, WCSR is a proportion, specifically the expected proportion of audio frames that an estimation algorithm will label correctly, and as such, it is naturally into logistic regression (i.e., a logit model). Likewise, HHD is constrained to fall between 0 and 100 percent, and thus it is also suitable for the same type of analysis. The remainder of this paper describes how logistic regression can be used to compare chord estimation algorithms, using MIREX results from 2013 to illustrate four key benefits: easier interpretation, greater statistical power, built-in correlation estimates for identifying relationships among algorithms, and better detection of outliers.

## 2. LOGISTIC REGRESSION WITH GEES

Proportions cannot be distributed normally because they are supported exclusively on $[0, 1]$, and thus they present challenges for traditional techniques of statistical analysis. Logit models are designed to handle these challenges without sacrificing the simplicity of the usual linear function relating parameters and covariates [1, ch.4]:

$$\pi(x; \beta) = \frac{e^{\mathbf{x}^\top \mathbf{\beta}}}{1 + e^{\mathbf{x}^\top \mathbf{\beta}}},$$

or equivalently

$$\log \frac{\pi(x; \beta)}{1 - \pi(x; \beta)} = \mathbf{x}^\top \mathbf{\beta},$$

where $\pi$ represents the relative frequency of ‘success’ given the values of covariates in $x$ and parameters $\beta$. In the case of a basic model for MIREX ACE, $\mathbf{x}$ would identify the algorithm and $\pi$ would be the relative frequency of correct chord labels for that algorithm (i.e., WCSR). In the case of data like ACE results, where there are proportions $p_i$ of correct labels over $n_i$ analysis frames rather than binary successes or failures, $i$ indexing all combinations of individual songs and algorithms, logistic regression assumes that each $p_i$ represents the observed proportion of successes among $n_i$ conditionally-independent binary observations, or more formally, that the $p_i$ are distributed binomially:

$$f_{P|N,X}(p | n, x, \beta) = \binom{n}{p} \pi^p (1 - \pi)^{(1 - p)n}.$$  \hspace{1cm} (3)

The expected value for each $p_i$ is naturally $\pi_i = \pi(x_i; \beta)$, the overall relative frequency of success given $x_i$:

$$E[P | N, X] = \pi(x; \beta).$$  \hspace{1cm} (4)

Logistic regression models are most often fit by the maximum-likelihood technique, i.e., one is seeking a vector $\hat{\beta}$ to maximise the log-likelihood given the data:

$$\ell_{P|N,X}(\beta; p, n, X) = \sum_i \left[ \log \left( \frac{n_i}{p_i n_i} \right) + p_i n_i \log \pi_i + (1 - p_i) n_i \log (1 - \pi_i) \right].$$  \hspace{1cm} (5)
One thus solves the system of likelihood equations for \( \beta \), whereby the gradient of Equation 5 is set to zero:

\[
\nabla_{\beta} \log f_{P|N,X}(\beta; p, n, X) = \sum_i (p_i - \pi_i)n_i x_i = 0
\]

and so

\[
\sum_i p_i n_i x_i = \sum_i \pi_i n_i x_i .
\]

In the case of MIREX ACE evaluation, each \( x_i \) is simply an indicator vector to partition the data by algorithm, and thus \( \hat{\beta} \) is the parameter vector for which \( \pi_i \) equals the song-length-weighted mean over all \( p_i \) for that algorithm.

### 2.1 Quasi-Binomial Models

Under a strict logit model, the variance of each \( p_i \) is inversely proportional to \( n_i \):

\[
\text{var} \left[ P \mid N, X \right] = \left( \frac{1}{n} \right) \pi (1 - \pi).
\]

Equation 8 only holds, however, if the estimates of chord labels for each audio frame are independent. For ACE, this is unrealistic: only the most naïve algorithms treat every frame independently. Some kind of time-dependence structure is standard, most frequently a hidden Markov model or some close derivative thereof. Hence one would expect that the variance of WCSR estimates should be rather larger than the basic logit model would suggest.

This type of problem is extremely common across disciplines, so much so that is has been given a name, over-dispersion, and some authors go so far as to state that ‘unless there are good external reasons for relying on the binomial assumption [of independence], it seems wise to be cautious and to assume that over-dispersion is present to some extent unless and until it is shown to be absent’. With a standard approach to handling over-dispersion is to use a so-called quasi-likelihood [1, §4.7]. In case of logistic regression, this typically entails a modification to the assumption on the distribution of the \( p_i \) that includes an additional dispersion parameter \( \phi \). The expected values are the same as a standard binomial model, but

\[
\text{var} \left[ P \mid N, X \right] = \left( \frac{\phi}{n} \right) \pi (1 - \pi).
\]

These models are known as quasi-likelihood models because one loses a closed-form solution for the actual probability distribution \( f_{P|N,X} \); one knows only that the \( p_i \) behave something like binomially-distributed variables, with identical means but proportionally more variance. The parameter estimates \( \hat{\beta} \) and predictions \( \hat{\pi}(\cdot; \hat{\beta}) \) for a quasi-binomial model are the same as ordinary logistic regression, but the estimated variance-covariance matrices are scaled by the estimated dispersion parameter \( \phi \) (and likewise the standard errors are scaled by its square root). The dispersion parameter is estimated so that the theoretical variance matches the empirical variance in the data, and because of the form of Equation 9, it renders any scaling considerations for the \( n_i \) moot.

Other approaches to handling over-dispersion include beta-binomial models [1, §13.3] and beta regression [5], but we prefer the simplicity of the quasi-likelihood model.

### 2.2 Generalised Estimating Equations (GEEs)

The quasi-binomial model achieves most of what one would be looking for when evaluating ACE for MIREX: it handles proportions naturally, is consistent with the weighted averaging used to compute WCSR, and adjusts for over-dispersion in a way that also eliminates any worries about scaling. Nonetheless, it is slightly over-conservative for evaluating ACE. As discussed earlier, quasi-binomial models are necessary to account for over-dispersion, and one important source of over-dispersion in these data is the lack of independence of chord estimates from most algorithms within the same song. MIREX exhibits another important violation of the independence assumption, however: all algorithms are tested on the same sets of songs, and some songs are clearly more difficult than others. Put differently, one does not expect the algorithms to perform completely independently of one another on the same song but rather expects a certain correlation in performance across the set of songs. By taking that correlation into account, one can improve the precision of estimates, particularly the precision of pair-wise comparisons [1, §10.1].

A relatively straightforward variant of quasi-likelihood known as generalised estimating equations (GEEs) incorporates this type of correlation [1, ch.11]. With the GEE approach, rather than predicting each \( p_i \) individually, one predicts complete vectors of proportions \( p \), for each relevant group, much as Friedman’s test seeks to estimate ranks within each group. For ACE, the groups are songs, and thus one considers the observations to be vectors \( p_i \), one for each song, where \( p_{ij} \) represents the CSR or segmentation score for algorithm \( j \) on song \( i \). Analogous to the case of ordinary quasi-binomial or logistic regression,

\[
E \left[ P_j \mid N, X_j \right] = \pi(x_j; \beta).
\]

Likewise, analogous to the quasi-binomial variance,

\[
\text{var} \left[ P_j \mid N, X_j \right] = \left( \frac{\phi}{n} \right) \pi_j (1 - \pi_j).
\]

Because the GEE approach is concerned with vector-valued estimates rather than point estimates, it also involves estimating a full variance-covariance matrix. In addition to \( \beta \) and \( \phi \), the approach requires a further vector of parameters \( \alpha \) and an a priori assumption on the correlation structure of the \( P_j \) in the form of a function \( R(\alpha) \) that yields a correlation matrix. (One might, for example, assume that that the \( P_j \) are exchangeable, i.e., that every pair shares a common correlation coefficient.) Then if \( B \) is a diagonal matrix such that \( B_{ij} = \text{var} \left[ P_j \mid N, X_j \right] \),

\[
\text{cov} \left[ P \mid N, X \right] = B^{1/2} R(\alpha) B^{1/2}.
\]

If all of the \( P_j \) are uncorrelated with each other, then this formula reduces to the basic quasi-binomial model, which assumes a diagonal covariance matrix. The GEE step of GEE estimation adjusts Equation 12 according to the actual correlations observed in the data, and as such, GEEs are quite robust in practice even when the a priori assumptions about the correlation structure are incorrect [1, §11.4.2].
Vocabulary V (seventh chords and inversions), weighted by song length. Bold lines represent medians and dashed lines represent quartile observations), but in general, observations across all algorithms are insufficient to distinguish among algorithms sharing a letter, correcting to hold the FDR at \( \alpha = .005 \). Although Friedman’s ANOVA detects 2 more significant pairwise differences than logistic regression (45 vs. 43), it operates on a different scale than CSR and misorders algorithms relative to WCSR.

3. ILLUSTRATIVE RESULTS

MIREX ACE 2013 evaluated 12 algorithms according to a battery of eight rubrics (WCSR on Uve harmonic vocabularies and three segmentation measures) on each of three different data sets (the Isophonics set, including music from the Beatles, Queen, and Zweeck [12] and two versions of the McGill Billboard set, including music from the American pop charts [4]). There is insufficient space to present the results of logistic regression on all combinations, and so we will focus on a single one of the data sets, the Billboard 2013 test set. In some cases, logistic regression allows us to speak to all measures (11,592 observations), but in general, we will also restrict ourselves to discussing the newest and most challenging of the harmonic vocabularies for WCSR: Vocabulary V (1932 observations), which includes major chords, minor chords, major sevenths, minor sevenths, dominant sevenths, and the complete set of inversions of all of the above. We are interested in four key questions.

1. How do pairwise comparisons under logistic regression compare to pairwise comparisons with Friedman’s ANOVA? Is logistic regression more powerful?
2. Are there differences among algorithms as the harmonic vocabularies get more difficult, or is the drop in performance uniform? In other words, is there a benefit to continuing with so many vocabularies?
3. Are all ACE algorithms making similar mistakes, or do they vary in their strengths and weaknesses?
4. Which algorithm-song pairs exhibited unexpectedly good or bad performance, and is there anything to be learned from these observations?

3.1 Pairwise Comparisons

The boxplots in Figure 1 give a more detailed view of the performance of each algorithm than Table 1. The figure is restricted to Vocabulary V, with the algorithms in descending order by WCSR. Figure 1a comes from Friedman’s ANOVA weighted by song length, and thus its y-axis reflects not CSR directly but the per-song ranks with respect to CSR. Figure 1b comes from quasi-binomial regression estimated with GEEs, as described in Section 2. Its y-axis does reflect per-song CSR. Above the boxplots, all significant pairwise differences recorded as a compact letter display. In the interest of reproducible research, we used a stricter \( \alpha = .005 \) threshold for reporting pairwise comparisons with the more contemporary false-discovery-rate (FDR) approach of Benjamini and Hochberg, as opposed to more traditional Tukey tests at \( \alpha = .05 \). Within either of the subfigures, the difference in performance between two algorithms that share any letter in the compact letter display is not statistically significant. Overall, Friedman’s ANOVA found 2 more significant pairwise differences than logistic regression.

3.2 Effect of Vocabulary

To test the utility of the new evaluation vocabularies, we ran both Friedman ANOVA (ranked separately for each vocabulary) and logistic regressions and looked for significant interactions among the algorithm, inversions (present or absent from the vocabulary) and the complexity of the vocabulary (root only, major-minor, or major-minor with 7ths). Under Friedman’s ANOVA, there was a significant Algorithm \( \times \) Complexity interaction, \( F(22, 9440) = 3.21, p < .001 \). The logistic regression model identified a significant three-way Algorithm \( \times \) Complexity \( \times \) Inversions interaction, \( \chi^2(12) = 37.35, p < .001 \), but the additional interaction with inversions should be interpreted with care: only one algorithm (CF2) attempts to recognise inversions.

3.3 Correlation Matrices

Table 2 presents the inter-correlations of WCSR between algorithms, rank-transformed (Spearman’s correlations, ana-
Table 2. Pearson’s correlations on the coefficients from logistic regression (WCSR) for the Billboard 2013 test set with vocabulary V (lower triangle); Spearman’s correlations for the same data (upper triangle). N = 161 songs per cell. Starred correlations are significant at α = 0.05, controlling for the FDR. A set of algorithms (viz., KO1, PP3, NG1, and SB8) stands out for negative correlations with the top performers; in general, these algorithms did not attempt to recognise seventh chords.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>KO2</th>
<th>NMSD2</th>
<th>CB4</th>
<th>NMSD1</th>
<th>CB3</th>
<th>KO1</th>
<th>PP4</th>
<th>PP3</th>
<th>CF2</th>
<th>NG1</th>
<th>NG2</th>
<th>SB8</th>
</tr>
</thead>
<tbody>
<tr>
<td>KO2</td>
<td>–</td>
<td>0.07</td>
<td>0.11</td>
<td>–0.05</td>
<td>0.10</td>
<td>0.03</td>
<td>–0.41</td>
<td>–0.44</td>
<td>–0.03</td>
<td>–0.35</td>
<td>0.05</td>
<td>–0.01</td>
</tr>
<tr>
<td>NMSD2</td>
<td>0.25</td>
<td>–</td>
<td>–0.01</td>
<td>0.49</td>
<td>–0.25</td>
<td>–0.20</td>
<td>–0.19</td>
<td>–0.36</td>
<td>0.00</td>
<td>–0.33</td>
<td>0.02</td>
<td>–0.06</td>
</tr>
<tr>
<td>CB4</td>
<td>0.41</td>
<td>0.39</td>
<td>–</td>
<td>0.12</td>
<td>0.47</td>
<td>–0.45</td>
<td>–0.39</td>
<td>–0.48</td>
<td>0.09</td>
<td>–0.38</td>
<td>0.08</td>
<td>–0.09</td>
</tr>
<tr>
<td>NMSD1</td>
<td>0.30</td>
<td>0.60</td>
<td>0.53</td>
<td>–0.17</td>
<td>–0.45</td>
<td>–0.08</td>
<td>–0.45</td>
<td>–0.27</td>
<td>–0.44</td>
<td>0.17</td>
<td>–0.08</td>
<td></td>
</tr>
<tr>
<td>CB3</td>
<td>0.34</td>
<td>0.10</td>
<td>0.76</td>
<td>–0.32</td>
<td>–0.42</td>
<td>–0.19</td>
<td>–0.26</td>
<td>–0.14</td>
<td>–0.08</td>
<td>–0.17</td>
<td>–0.16</td>
<td>–0.08</td>
</tr>
<tr>
<td>KO1</td>
<td>–0.04</td>
<td>–0.42</td>
<td>–0.51</td>
<td>–0.51</td>
<td>–0.29</td>
<td>–0.26</td>
<td>–0.14</td>
<td>–0.08</td>
<td>0.17</td>
<td>–0.16</td>
<td>–0.08</td>
<td></td>
</tr>
<tr>
<td>PP4</td>
<td>–0.22</td>
<td>0.08</td>
<td>–0.16</td>
<td>0.06</td>
<td>–0.07</td>
<td>–0.05</td>
<td>–0.37</td>
<td>–0.03</td>
<td>0.00</td>
<td>–0.05</td>
<td>–0.03</td>
<td></td>
</tr>
<tr>
<td>PP3</td>
<td>–0.49</td>
<td>–0.46</td>
<td>–0.61</td>
<td>–0.53</td>
<td>–0.37</td>
<td>–0.68</td>
<td>–0.22</td>
<td>–0.48</td>
<td>0.66</td>
<td>–0.48</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>CF2</td>
<td>0.09</td>
<td>0.19</td>
<td>0.24</td>
<td>0.42</td>
<td>0.17</td>
<td>–0.49</td>
<td>–0.06</td>
<td>–0.51</td>
<td>–0.48</td>
<td>–0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NG1</td>
<td>–0.54</td>
<td>–0.42</td>
<td>–0.60</td>
<td>–0.56</td>
<td>–0.41</td>
<td>–0.68</td>
<td>–0.04</td>
<td>0.85</td>
<td>–0.47</td>
<td>–0.40</td>
<td>–0.10</td>
<td></td>
</tr>
<tr>
<td>NG2</td>
<td>0.09</td>
<td>0.17</td>
<td>0.17</td>
<td>0.16</td>
<td>–0.03</td>
<td>–0.50</td>
<td>–0.09</td>
<td>–0.54</td>
<td>0.50</td>
<td>–0.40</td>
<td>–0.11</td>
<td></td>
</tr>
<tr>
<td>SB8</td>
<td>–0.32</td>
<td>–0.44</td>
<td>–0.44</td>
<td>–0.52</td>
<td>–0.46</td>
<td>0.00</td>
<td>–0.32</td>
<td>0.08</td>
<td>–0.33</td>
<td>0.08</td>
<td>–0.16</td>
<td></td>
</tr>
</tbody>
</table>

4. DISCUSSION

We were surprised to find that in terms of distinguishing between algorithms, Friedman’s ANOVA was in fact more powerful than logistic regression, detecting a few extra significant pairs. Nonetheless, the two approaches yield substantially equivalent broad conclusions: that a group of top performers – CB3, CB4, KO2, NMSD1, and NMSD2 – are statistically indistinguishable from each other, with KO1 also indistinguishable from the lower end of this group. Moreover, having now benefited from years of study, WCSR is a reasonably intuitive and well-motivated measure of ACE performance, and it is awkward to have to work on the Friedman’s rank scale instead, especially since it ultimately ranks the algorithms’ overall performance in a slightly different order than the headline WCSR-based results.

Friedman’s ANOVA did exhibit less power for our question about interactions between algorithms and differing chord vocabularies. Again, WCSR as a unit and as a concept is highly meaningful for chord estimation, and there is a conceptual loss from rank transformation. Given the rank transformation, Friedman’s ANOVA can only be sensitive to reconstructions of relative performance as the vocabularies become more difficult; logistic regression can also be sensitive to different effect sizes across algorithms even when their relative ordering remains the same.

It was encouraging to see that under either statistical model, there was a benefit to evaluating with multiple vocabularies. That encouraged us to examine the inter-correlations for the performance of the algorithms. Figure 2 summarises the original correlation matrix in Table 2 more visually by using the correlations from logistic regression as the basis of a hierarchical clustering. Two clear groups emerge, both from the clustering and from minding negative correlations in the original matrix: one relatively low-performing group including KO1, PP3, NG1, and SB8, and one relatively high-performing group including all others but for perhaps PP4, which does not seem to correlate strongly with any other algorithm. The shape of the equivalent tree based on Spearman’s correlations is similar but for joining PP4 with SB8 instead of the high-performing group. Table 1 uncovers the secret behind the low performers: KO1 excepted, none of the low-performing algorithms attempt to recognise seventh chords, which comprise 29 percent of all chords under Vocabulary V. Furthermore, we performed an additional evaluation of seventh chords only, in the style of [15] and using their software available online. From the resulting low score of KO1, we can deduce that this algorithm is able to recognise seventh chords in theory, but that it was most likely trained on the relatively seventh-poor Isophon.

https://github.com/jpauwels/MusOOEvaluator


