# 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 segmentation 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 2million unique visitors every month [6].

In order to compare different algorithmic approaches in an impartial setting, the Music Information Retrieval Evaluation eXchange (MIREX) introducted 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; thefi gure 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 of12 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', [3] it was proposed that future evaluations should include more diverse chord vocabularies, such as seventh chords and inversions, as the25 -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

Johan Pauwels is no longer affiliated with STMS. Data and source code to reproduce this paper, including all statistics andfi gures, are available from http://bitbucket.org/jaburgoyne/ismir-2014.



[1] http://www.riffstation.com/
[2] http://chordify.net
[3] http://www.music-ir.org/mirex/wiki/The_Utrecht_Agreement_on_Chord_Evaluation

| Algorithm | # Types | Inversions? | Training? | I | II | III | IV | V | VI | VII | VIII |
|---|---|---|---|---|---|---|---|---|---|---|---|
| KO2 | 7 | | • | 76 | 74 | 72 | 60 | 58 | 84 | 79 | 89 |
| NMSD2 | 10 | | | 75 | 71 | 69 | 59 | 57 | 82 | 79 | 86 |
| CB4 | 13 | | • | 76 | 72 | 70 | 59 | 57 | 85 | 80 | 90 |
| NMSD1 | 10 | | | 74 | 71 | 69 | 58 | 56 | 83 | 79 | 86 |
| CB3 | 13 | | | 76 | 72 | 70 | 58 | 56 | 85 | 81 | 89 |
| KO1 | 7 | | | 75 | 71 | 69 | 54 | 52 | 83 | 80 | 88 |
| PP4 | 5 | | | 69 | 66 | 64 | 51 | 49 | 83 | 78 | 87 |
| PP3 | 2 | | | 70 | 68 | 65 | 50 | 48 | 83 | 82 | 84 |
| CF2 | 10 | • | | 71 | 67 | 65 | 49 | 47 | 83 | 83 | 83 |
| NG1 | 2 | | | 71 | 67 | 65 | 49 | 46 | 82 | 79 | 86 |
| NG2 | 5 | | | 67 | 63 | 61 | 44 | 43 | 82 | 81 | 83 |
| SB8 | 2 | | | 9 | 7 | 6 | 5 | 5 | 51 | 92 | 35 |

**Table1** . Number of supported chord types, inversion support, training support, and MIREX results on the *Billboard* 2013test set for all2013 ACE submissions. I: root only; II: major-minor vocabulary; III: major-minor vocabulary with inversions; IV: major-minor vocabulary with sevenths; V: major-minor vocabulary with sevenths and inversions; VI: mean segmentation score; VII: under-segmentation; VIII: over-segmentation. Adapted from the MIREX Wiki.

output on one hand, and among the output of different algorithms on the other hand [15]. They also performed a rigorous re-evaluation of all MIREX ACE submissions from 2010 to2012 . As of MIREX 2013, these revised evaluation procedures, including the chord-sequence segmentation evaluation suggested by Harte [8] and Mauch [12], have been adopted in the context of the MIREX ACE task.

MIREX ACE evaluation has also typically included comparative statistics to help determine whether the differences in performance between pairs of algorithms are statistically significant. Traditionally, Friedman's ANOVA has been used for this purpose, accompanied by Tukey's Honest Significant Difference tests for each pair of algorithms. Friedman's ANOVA is equivalent to a standard two-way ANOVA with the actual measurements (in our case WCSR or directional Hamming distance [DHD], the new segmentation measure) replaced by the rank of each treatment (in our case, each algorithm) on that measure within each block (in our case, for each song) [11]. The rank transformation makes Friedman's ANOVA an excellent 'one sizefi ts 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, itfi ts naturally into *logistic regression* (i.e., a *logit model*). Likewise, DHD 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]:

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

or equivalently

$$\log \frac{\pi(\mathbf{x};\boldsymbol{\beta})}{1-\pi(\mathbf{x};\boldsymbol{\beta})} = \mathbf{x}'\boldsymbol{\beta} , \tag{2}$$

where $\pi$ represents the relative frequency of 'success' given the values of covariates in $\mathbf{x}$ and parameters $\boldsymbol{\beta}$. 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:

$$f_{P|N,\mathbf{X}}(p \mid n,\mathbf{x};\boldsymbol{\beta}) = \binom{n}{pn} \pi^{pn}(1-\pi)^{(1-p)n} . \tag{3}$$

The expected value for each $p_i$ is naturally $\pi_i = \pi(\mathbf{x}_i;\boldsymbol{\beta})$, the overall relative frequency of success given $\mathbf{x}_i$:

$$\mathbf{E}[P \mid N,\mathbf{X}] = \pi(\mathbf{x};\boldsymbol{\beta}) . \tag{4}$$

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

$$\ell_{P|N,\mathbf{X}}(\boldsymbol{\beta};\mathbf{p},\mathbf{n},\mathbf{X}) = \sum_i \left[ \log \binom{n_i}{p_i n_i} + p_i n_i \log \pi_i + (1-p_i) n_i \log(1-\pi_i) \right] . \tag{5}$$

One thus solves the system of likelihood equations for $\boldsymbol{\beta}$, whereby the gradient of Equation 5 is set to zero:

$$\nabla_{\boldsymbol{\beta}} \ell_{P|N,\mathbf{X}}(\boldsymbol{\beta}; \mathbf{p}, \mathbf{n}, \mathbf{X}) = \sum_i (p_i - \pi_i) n_i \mathbf{x}_i = \mathbf{0} \quad (6)$$

and so

$$\sum_i p_i n_i \mathbf{x}_i = \sum_i \pi_i n_i \mathbf{x}_i \ . \quad (7)$$

In the case of MIREX ACE evaluation, each $\mathbf{x}_i$ is simply an indicator vector to partition the data by algorithm, and thus $\hat{\boldsymbol{\beta}}$ 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 $n_i$:

$$\text{var}\left[P \mid N, \mathbf{X}\right] = \left(\frac{1}{n}\right) \pi(1 - \pi) \ . \quad (8)$$

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' [14, p.125]. One 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

$$\text{var}\left[P \mid N, \mathbf{X}\right] = \left(\frac{\phi}{n}\right) \pi(1 - \pi) \ . \quad (9)$$

These models are known as quasi-likelihood models because one loses a closed-form solution for the actual probability distribution $f_{P|N,\mathbf{X}}$; one knows only that the $p_i$ behave something like binomially-distributed variables, with identical means but proportionally more variance. The parameter estimates $\hat{\boldsymbol{\beta}}$ and predictions $\pi(\cdot; \hat{\boldsymbol{\beta}})$ 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 $\hat{\phi}$ (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 Equation9 , 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 $\mathbf{p}_i$ 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 $\mathbf{p}_i$, 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,

$$\mathbf{E}\left[P_j \mid N, \mathbf{X}_j\right] = \pi(\mathbf{x}_j; \boldsymbol{\beta}) \ . \quad (10)$$

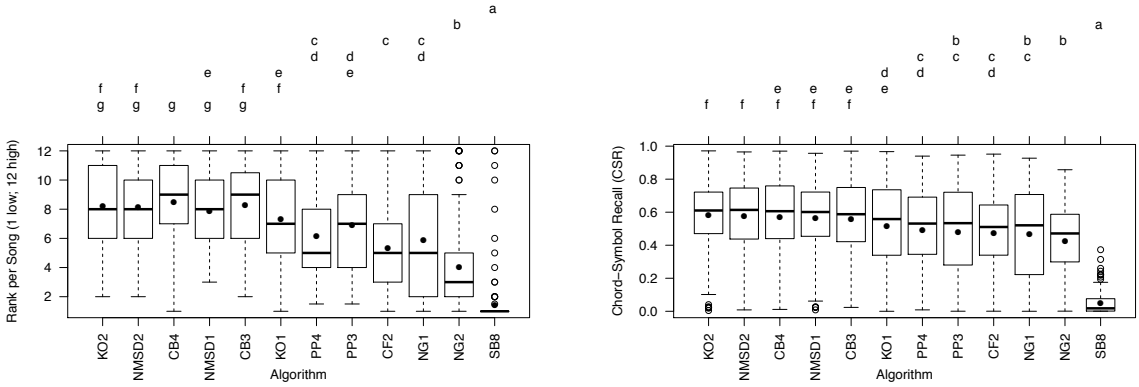
Likewise, analogous to the quasi-binomial variance,

$$\text{var}\left[P_j \mid N, \mathbf{X}_j\right] = \left(\frac{\phi}{n}\right) \pi_j(1 - \pi_j) \ . \quad (11)$$

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 $\boldsymbol{\beta}$ and $\phi$, the approach requires a further vector of parameters $\boldsymbol{\alpha}$ and an *a priori* assumption on the correlation structure of the $P_j$ in the form of a function $R(\boldsymbol{\alpha})$ 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 $B_{jj} = \text{var}\left[P_j \mid N, \mathbf{X}_j\right]$,

$$\text{cov}\left[\mathbf{P} \mid N, \mathbf{X}\right] = B^{1/2} R(\boldsymbol{\alpha}) B^{1/2} \ . \quad (12)$$

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. Thefi nal 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].




(a) Friedman's ANOVA

(b) Logistic Regression

**Figure 1**. Boxplots and compact letter displays for the MIREX ACE 2013 results on the *Billboard* 2013 test set with vocabulary V (seventh chords and inversions), weighted by song length. Bold lines represent medians and filled dots means. $N = 161$ songs per algorithm. Given the respective models, there are insufficient data 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 onfi ve harmonic vocabularies and three segmentation measures) on each of three different data sets (the Isophonics set, including music from the Beatles, Queen, and Zweieck [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 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 Table1 . Thefigure is restricted to Vocabulary V, with the algorithms in descending order by WCSR. Figure1 a 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. Figure1 b comes from quasi-binomial regression estimated with GEEs, as described in Section2 . Its $y$-axis does reflect per-song CSR. Above the boxplots, all significant pairwise differences are 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 $\alpha = .05$[2 ,9 ]. 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 2more significant pairwise differences than logistic regression.

### 3.2 Effect of Vocabulary

To test the utility of the new evaluation vocabularies, we ran both Friedman ANOVAs (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 × Complexity interaction, $F(22,9440) = 3.21$, $p < .001$. The logistic regression model identified a significant three-way Algorithm × Complexity × 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-

| Algorithm | KO2 | NMSD2 | CB4 | NMSD1 | CB3 | KO1 | PP4 | PP3 | CF2 | NG1 | NG2 | SB8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KO2 | – | .07 | .11 | −.05 | .10 | .03 | −.41* | −.44* | −.03 | −.35* | .05 | −.01 |
| NMSD2 | .25* | – | −.01 | .49* | −.25* | −.20 | −.19 | −.36* | .00 | −.33* | .02 | −.06 |
| CB4 | .41* | .39* | – | .12 | .47* | −.46* | −.30* | −.48* | .09 | −.38* | .08 | −.09 |
| NMSD1 | .30* | .60* | .53* | – | −.17 | −.45* | −.08 | −.45* | .27* | −.44* | .17 | −.10 |
| CB3 | .34* | .10 | .76* | .42* | – | −.19 | −.26* | −.14 | −.08 | −.17 | −.16 | −.08 |
| KO1 | −.04 | −.42* | −.51* | −.51* | −.29* | – | −.10 | .42* | −.41* | .50* | −.52* | .05 |
| PP4 | −.22 | .08 | −.16 | .06 | −.07 | −.05 | – | .37* | −.03 | .00 | .05 | −.03 |
| PP3 | −.49* | −.46* | −.61* | −.53* | −.37* | .68* | .22 | – | −.48* | .66* | −.48* | .04 |
| CF2 | .09 | .19 | .24* | .42* | .17 | −.49* | .06 | −.51* | – | −.48* | .48* | −.14 |
| NG1 | −.54* | −.42* | −.60* | −.56* | −.41* | .68* | .04 | .85* | −.47* | – | −.40* | −.10 |
| NG2 | .09 | .17 | .17 | .16 | −.03 | −.50* | −.09 | −.54* | .50* | −.40* | – | −.11 |
| SB8 | −.32* | −.44* | −.44* | −.52* | −.46* | .00 | −.32* | .08 | −.33* | .08 | −.16 | – |

**Table2** . 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 $\alpha = .005$, 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.

logous to Friedman's ANOVA) in the upper triangle, and in the lower triangle, as estimated from logistic regression with GEEs. Significant correlations are marked, again controlling the FDR at $\alpha = .005$. Positive correlations do not necessarily imply that the algorithms perform similarly; rather it implies that theyfi nd the same songs relatively easy or difficult. Negative correlations imply that songs that one algorithm finds difficult are relatively easy for the other algorithm.

### 3.4 Outliers

To identify outliers, we considered all evaluations on the *Billboard* 2013 test set and examined the distribution of residuals. Chauvenet's criterion for outliers in a sample of this size is to lie more than4. 09 standard deviations from the mean [16, §6.2 ]. Under Friedman's ANOVA, Chauvenet's criterion identified 7 extreme data points. These are all for algorithm SB8, a submission with a programming bug that erroneously returned alternating C- and B-major chords regardless of the song, on songs that were so difficult for most other algorithms that the essentially random approach of the bug did better. Under the logistic regression model, the criterion identified 26 extreme points. Here, the unexpected behaviour was primarily for songs that are tuned a quartertone off from standard tuning (A4 = 440 Hz). The ground truth necessarily is 'rounded off' to standard tuning in one direction or the other, but in cases where an otherwise high-performing algorithm happened to round off in the opposite direction, the performance is markedly low.

## 4. DISCUSSION

We were surprised tofi nd 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 reconfigurations 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 2summarises 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. [4] 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-

[4] https://github.com/jpauwels/MusOOEvaluator

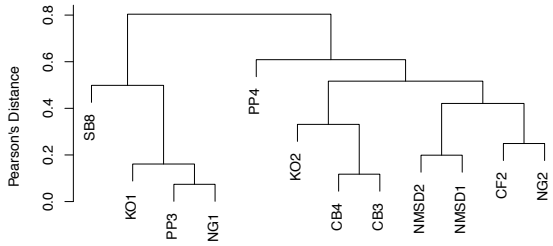



**Figure2** . Hierarchical clustering of algorithms based on WCSR for for the *Billboard* 2013 test set with vocabulary V, Pearson's distance as derived from the estimated correlation matrix under logistic regression, and complete linkage. The group of algorithms that is negatively correlated with the top performers appears at the left. PP4 stands out as the most idiosyncratic performer.

ics corpus (only 15 percent of all chords). KO2 is the same algorithm trained directly on the MIREX *Billboard* training corpus, and with that training, it becomes a top performer.

Our analysis of outliers again showed Friedman's ANOVA to be less powerful than logistic regression, as one would expect given the range restrictions on rank transformation. But here also the more important advantage of logistic regression is the ability to work on the WCSR scale. Outliers under the logistic regression model are also points that have an unusually strong effect on the reported results. In our analysis, they highlight the practical consequences of the well-known problem of atypically-tuned commercial recordings. Although we would not propose deleting outliers, it is sobering to know that tuning problems may be having an outsized effect on our headline evaluationfi gures. It might be worth considering allowing algorithms their best score in keys up to a semitone above or below the ground truth.

Overall, we have shown that as ACE becomes more established and its evaluation more thorough, it is useful to use a subtler statistical model for comparative analysis. We recommend that future MIREX ACE evaluations use logistic regression in preference to Friedman's ANOVA. It preserves the natural units and scales of WCSR and segementation analysis, is more powerful for many (although not all) statistical tests, and when augmented with GEEs, it allows for a detailed correlational analysis of which algorithms tend to have problems with the same songs as others and which have perhaps genuinely broken innovative ground. This is by no means to suggest that Friedman's test is a bad test in general – its near-universal applicability makes it an excellent choice in many circumstances, including many other MIREX evaluations – but for ACE, we believe that the extra understanding logistic regression can offer may help researchers predict which techniques are most promising for breaking the current performance plateau.

## 5. REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley, New York, 2nd edition,2007.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*,1(57):289–300, 1995.

[3] J. A. Burgoyne. *Stochastic Processes and Database-Driven Musicology*. PhD thesis, McGill U., Montréal, QC,2012.

[4] J. A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground-truth set for audio chord recognition and music analysis. In *Proc. Int. Soc. Music Inf. Retr.*, pages633–38, Miami, FL,2011.

[5] S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *J. Appl. Stat.*,31(7):799–815, 2004.

[6] W. B. de Haas, J. P. Magalhães, D. ten Heggeler, G. Bekenkamp, and T. Ruizendaal. Chordify: Chord transcription for the masses. Demo at the Int. Soc. Music Inf. Retr. Conf., Curitiba, Brazil,2012.

[7] W. B. de Haas, J. P. Magalhães, R. C. Veltkamp, and F. Wiering. Harmtrace: Improving harmonic similarity estimation using functional harmony analysis. In *Proc. Int. Soc. Music Inf. Retr.*, pages67–72 , Miami, FL,2011.

[8] C. Harte. *Towards Automatic Extraction of Harmony Information from Music Signals*. PhD thesis, Queen Mary, U. London,2010.

[9] V. E. Johnson. Revised standards for statistical evidence. *P. Nat'l Acad. Sci. USA*,110(48):19313–17 ,2013.

[10] M. Khadkevich and M. Omologo. Large-scale cover song identification using chord profiles. In *Proc. Int. Soc. Music Inf. Retr. Conf.*, pages233–38 , Curitiba, Brazil, 2013.

[11] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, Boston, MA,5 th edition,2005.

[12] M. Mauch. *Automatic Chord Transcription from Audio Using Computational Models of Musical Context*. PhD thesis, Queen Mary, U. London,2010.

[13] M. Mauch, S. Dixon, C. Harte, M. Casey, and B. Fields. Discovering chord idioms through Beatles and Real Book songs. In *Proc. Int. Soc. Music Inf. Retr. Conf.*, pages255–58 , Vienna, Austria,2007.

[14] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, FL,2 nd edition, 1989.

[15] J. Pauwels and G. Peeters. Evaluating automatically estimated chord sequences. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages749–53 , Vancouver, British Columbia,2013.

[16] J. R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, CA,2 nd edition,1997.