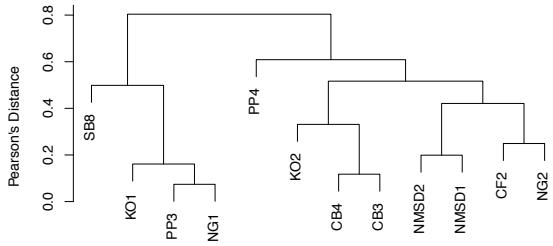

Figure 2 . Hierarchical clustering of algorithms based on WCSR for for the *Billboard* 2013 test set with vocabulary V, Pearson's distance as derived from the estimated correlation matrix under logistic regression, and complete linkage. The group of algorithms that is negatively correlated with the top 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.