Limits on the Application of Statistical Correlations to Continuous Response Data

MARL
NYU Music and Audio Research Laboratory

Finn Upham

NYU Steinhardt
Steinhardt School of Culture, Education, and Human Development

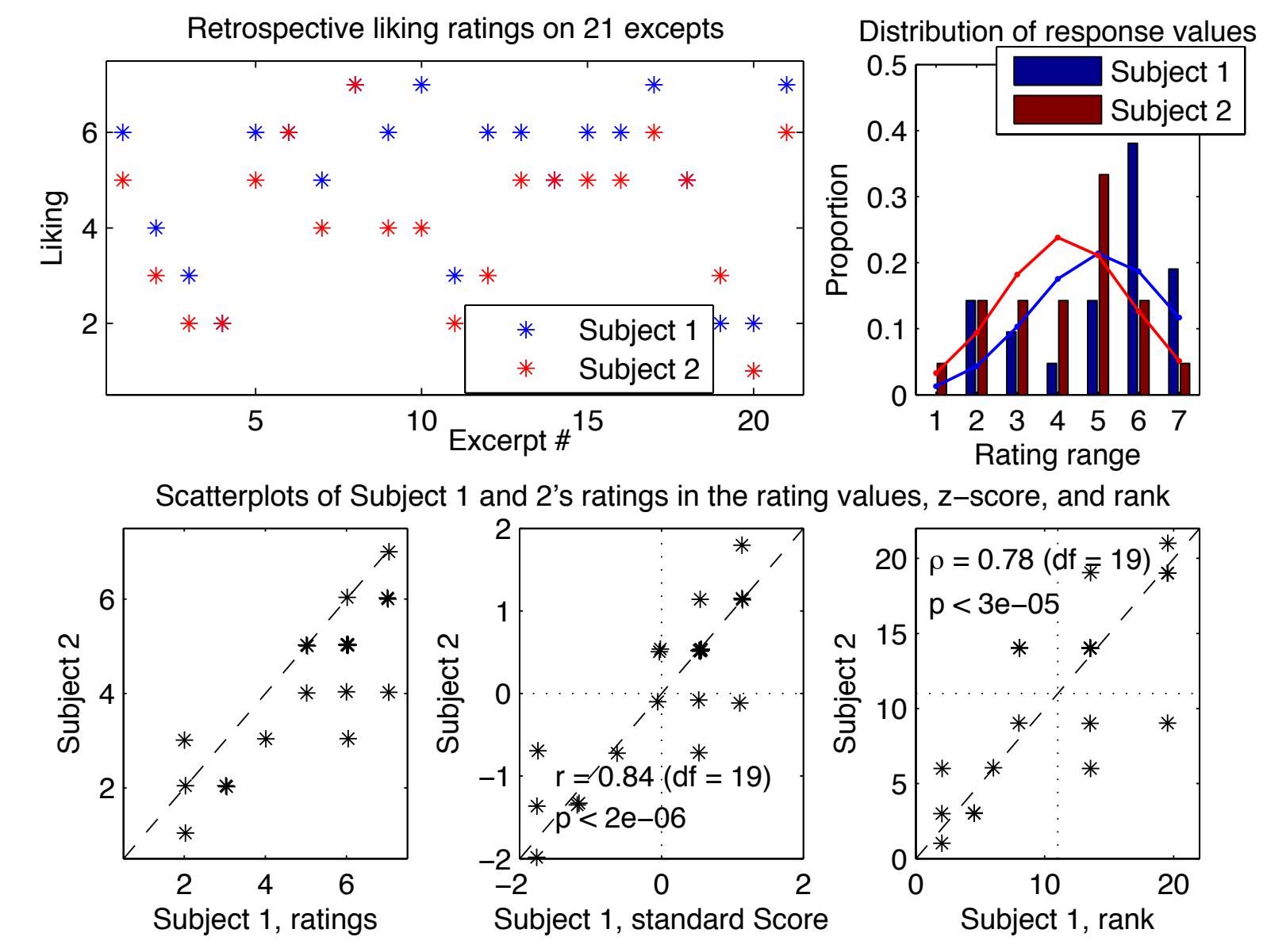

Figure 1: Correlation between two listeners' retrospective ratings of liking of 22 excerpts of music. In the top row, the rating of each except by the subjects beside the distribution of values reported by each subject and the normal distribution suggested by each distributions' mean and variance. Below are three scatter plots of the subjects' discrete ratings: first (from the left) in the original values, second in values normalized to have zero mean and unit standard diviation, and last in tied rank. These last two plots report the Pearson and Spearman correlation coefficients, respectively, and their significance according to the student T estimate.

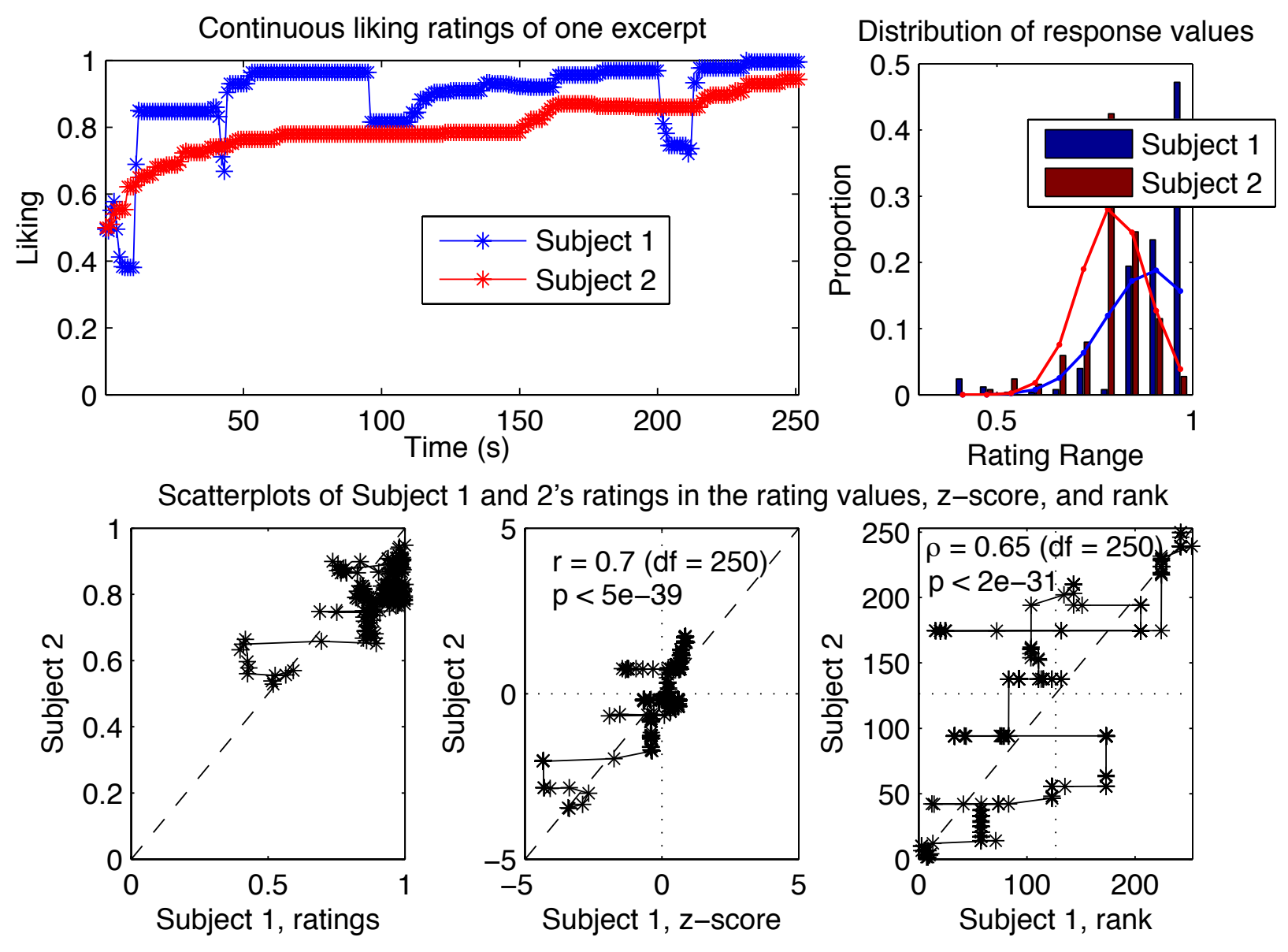

Figure 2: Correlation between continuous ratings of liking from two subjects for the same musical excerpt, presented in the same form as Figure 1. Notice how the values bunch and spread in the scatterplots. The significance estimates are for demonstration, and not valid.

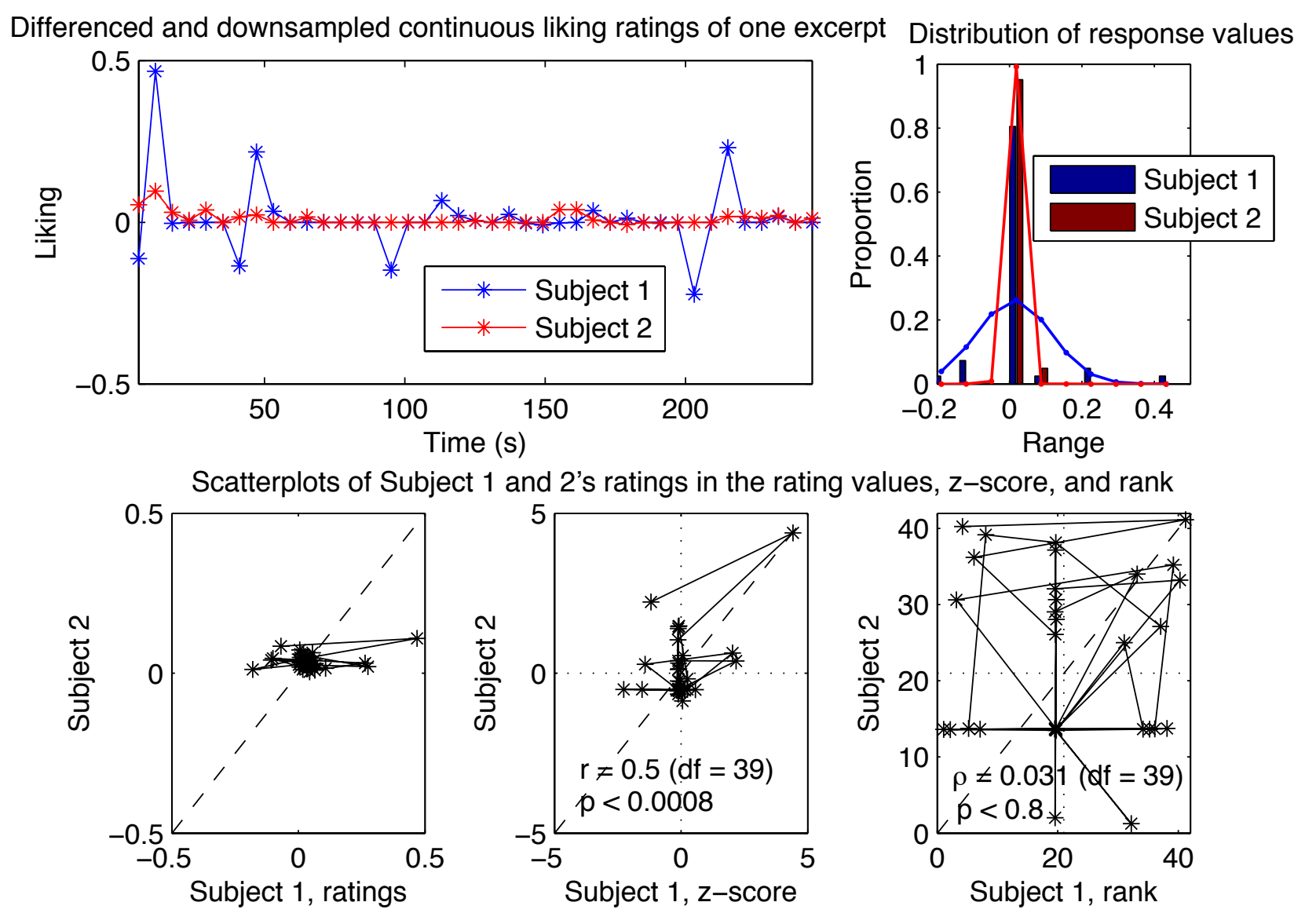

Figure 3: The same responses in Figure 2, processed to reduce autocorrelation, according to the aggregate results of many collections: downsampled to 0.167 Hz and differenced. The scatterplots show the contribution of these fewer samples to the overall correlation in Pearson and Spearman.

The data collections

This project includes analysis of many collections of continuous responses to music. The exprimental collections are thirty two sets of emotion ratings on a single dimension to a single musical stimulus. Though they come from several experiments, each set is sampled at 1 Hz, on a scale ranging from 0 to 1, and average 30 responses per collection. Twelve collections are of emotion perceived in the music, half valence, half arousal (collected concurrently), twelve are of emotion felt by the participants, half valence, half arousal, and eight contain ratings of emotional intensity. The stimuli are all concert music pieces, mostly of the classical and nationalistic eras.

For comparison, 32 random response collections were constructed from these experimental data sets by sampling randomly across collections. The resulting collections are composed of responses in different measures and of different stimuli, matching each experimental collection in number of responses. Note these collections are truncated to the shortest response included. These collections give a clue as to how unrelated response collection would perform under the same treatment, a check in inferring too much from the experimental data collections.

Introduction

Correlations are a common and powerful statistic, designed to capture shared variance in two measures of the same samples without constraints of common units. In Figure 1, a prosaic application of Pearson and Spearman correlations on a two subjects discrete ratings of liking for a set of music excerpt, the scatterplots show the information used for each statistic. Figure 2 shows the same isolation of information of these statistics when applied to continous ratings of liking by the same two subjects to a single work. The interpretation of correlations of continous rating data are complicated by three issues:

1. Auto Correlation
2. Non-Parametric distributions of values with repeats.
3. Arbitrary sample number and estimating significance

Despite these complications, researchers have been trying to employ correlations as a measure of cohesiveness and to validate the average time series as representative of a collection of responses. One favoured approach to make standard interpretation of correlations possible is doing away with auto correlation. The following evaluates the effectiveness of two methods, down-sampling [Chapin et al., 2010] and differencing [Schubert, 2002], on the average inter-response correlation and proposes alternative methods of assessing coherence in continuous response collections.

Mission: Reduce Autocorrelation

Autocorrelation correlation can be assessed by correlating a series with itself delayed by one sample. To test these proposed solutions, Figure 4 shows results from 32 collections of rating responses: the distributions of the average autocorrelations of these collections are compared for sample rates going from 1 Hz to 0.1 Hz (once a second to once every ten seconds).

Undifferenced data (top of figure 4) does decrease in autocorrelation with downsampling, but is far from zero even at 0.1 Hz. Differencing the data does dramatically decrease autocorrelation, however these collections do not distribute evenly around zero average autocorrelation without also downsampling to 0.167 Hz, or once every 6 seconds.

If our primary concern is to eliminate serial correlation from these analyses (rather than compensate using more complicated autoregression models), **these data suggest that correlations should be assessed on first-order differenced series which are sampled no faster than once every five seconds**.

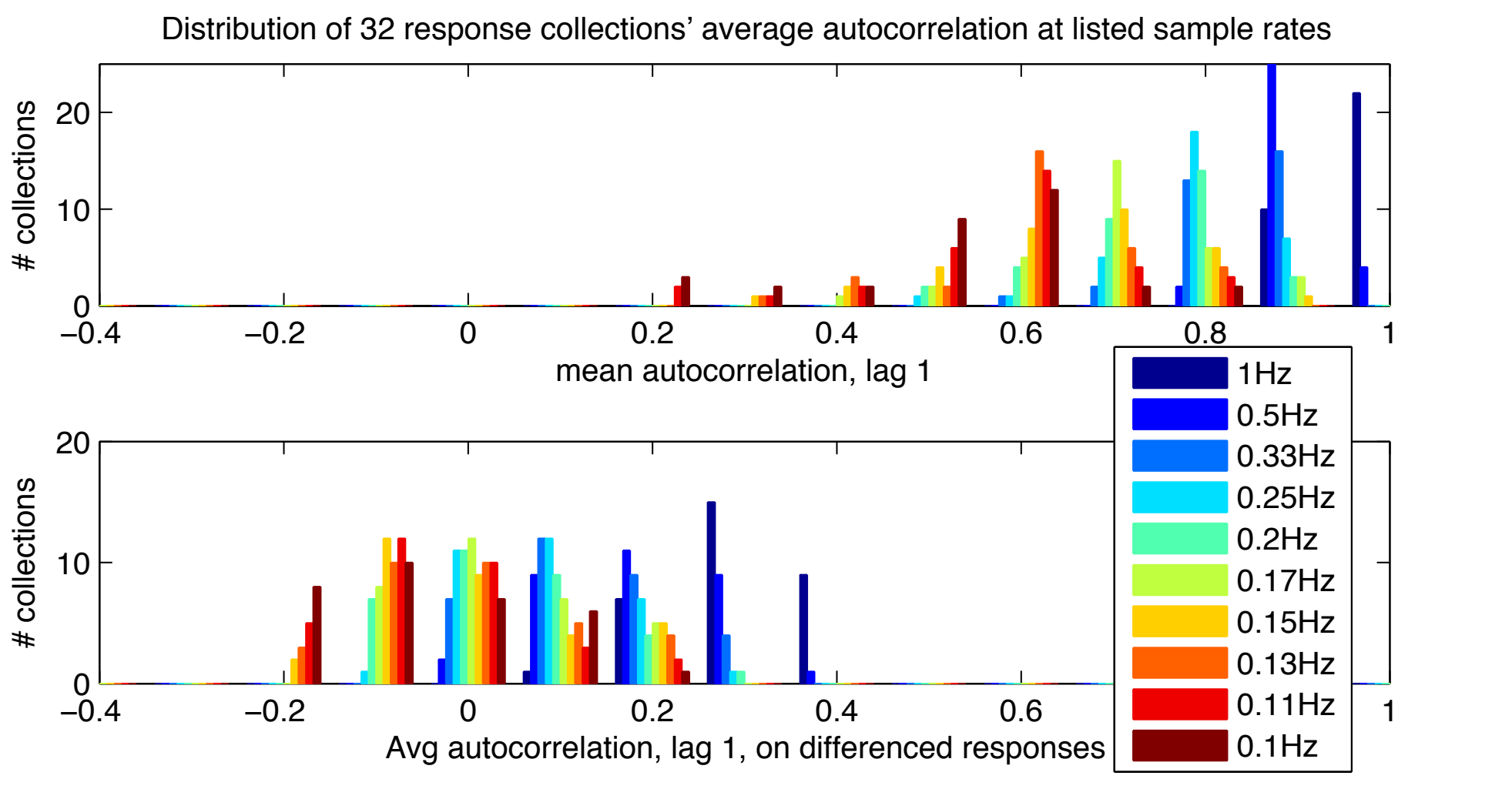

Figure 4. The distributions of average auto-correlation for 32 collections of continuous ratings of emotion to music on the original ratings and their first order difference series at different sample rates.

To give a sense of what these two transformations do to the average inter-response correlation, figure 5 shows the distribution of this popular cohesion measure on these collection.

Downsampling barely affects this statistic on the original rating data; this underlines the concern that the number of samples in these time series far exceeds the quantity of information they contain.

The bottom graph of figure 5 shows the average inter-response correlations of the differenced responses increasing as the sample rate goes down; this happens because each sample is representing a larger time window over which some change of rating may take place. High sample rates also have a larger proportion of zero values on these rating data, which can cause further problems for statistical interpretation.

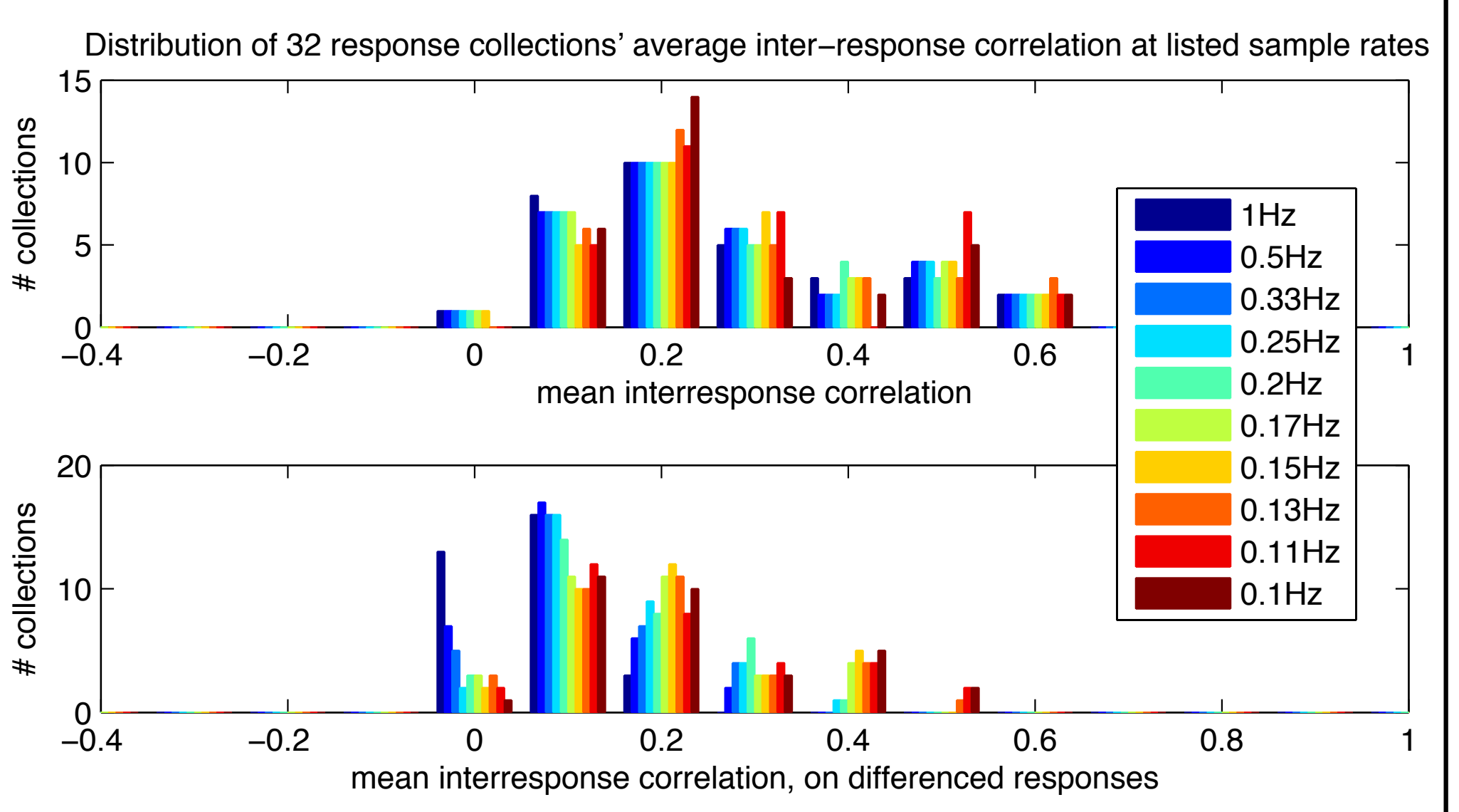

Figure 5. The distributions of average inter-response correlations for 32 collections of continuous ratings of emotion to music on the original ratings and their first order difference series at different sample rates.

Figure 3 shows the results of applying these reductions to the two responses discussed earlier. The top-left graph shows how the two series are flattened to zero, with variation when the original ratings changed values. Even when downsampled to 0.167 Hz, the distributions are strongly dominated by zero valued data points. The scatterplots show a much less convincing story of the relationship between the two series: the relatively large r value is strongly influenced by one data point, and this advantage is lost in the rank representation of these rating change series.

Acknowledgements

My thanks to Stephen McAdams, Mark Korhonen, Michael Grubb, and Mary Farbood for sharing their continuous rating data; to NSERC, SSHERC and other funding bodies which made the collection of these data possible, and the many people involved in the BSO and CARS experiments; to Stephen McAdams again for supervising the beginning of this research as part of my master's thesis, and to my current advisor, Mary Farbood, and the Steinhardt School at NYU for supporting its continuation.

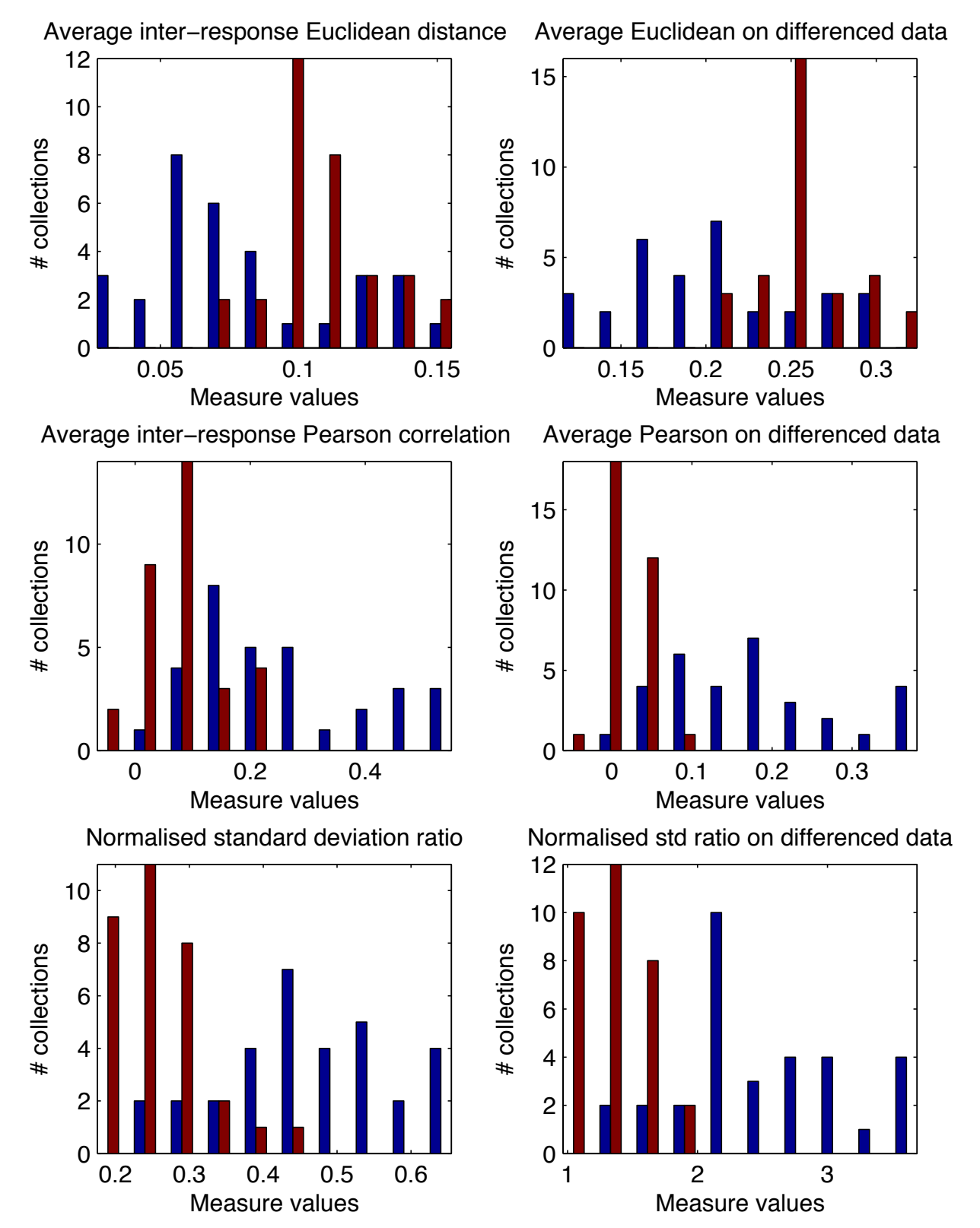

Figure 6. Distributions of coherence statistics on experimental collections of continuous responses and randomly assembled unrelated continuous response collections. Red bars show the distribution over the random response collections, while the blue show values from experimental collections. The left column shows these measures on rating data, the right column reports the measures as applied to differenced and downsampled responses.

Comparing Measures of Cohesion in Collections of Continuous Ratings of Emotion

Applying cohesion measures to both the experimental collections and the random collections demonstrates how effective they are at capturing what is shared between responses.

1. The average Euclidean inter-response distances: average rating range difference between pairs of responses per time point, averaged across all pairs of responses.

In figure 6, top row, the real experimental data sets and the random collections, to the left are the original rating series, to the right, the downsampled and differenced series. On 1Hz rating scale, more than half the experimental responses are more densely packed than the random collections, but there are many which spread out. The differenced ratings overlap less.

2. The average inter-response Pearson correlation, average correlation between all pairs of responses in the collection.

While some experimental collections far exceed the random collections in inter-response correlations, most fall in the same range of 0 to 0.2. The differenced data separate better, but the correlation values are very low.

3. The ratio of the standard deviation of the average response time series over the average standard deviation of the responses in the collection, adjusted for size of the collection.

Figure 6 shows the best separation yet on the undifferenced ratings, and the difference series pushed the random collections to the lower range, with every little overlap.

To demonstrate the effectiveness of this statistic, the normalized standard deviation ratio of mean- to- individual responses first-order difference series, Figure 7 shows three experimental collections which score high beside three which score within range of the random response collections, their averages plotted in black.

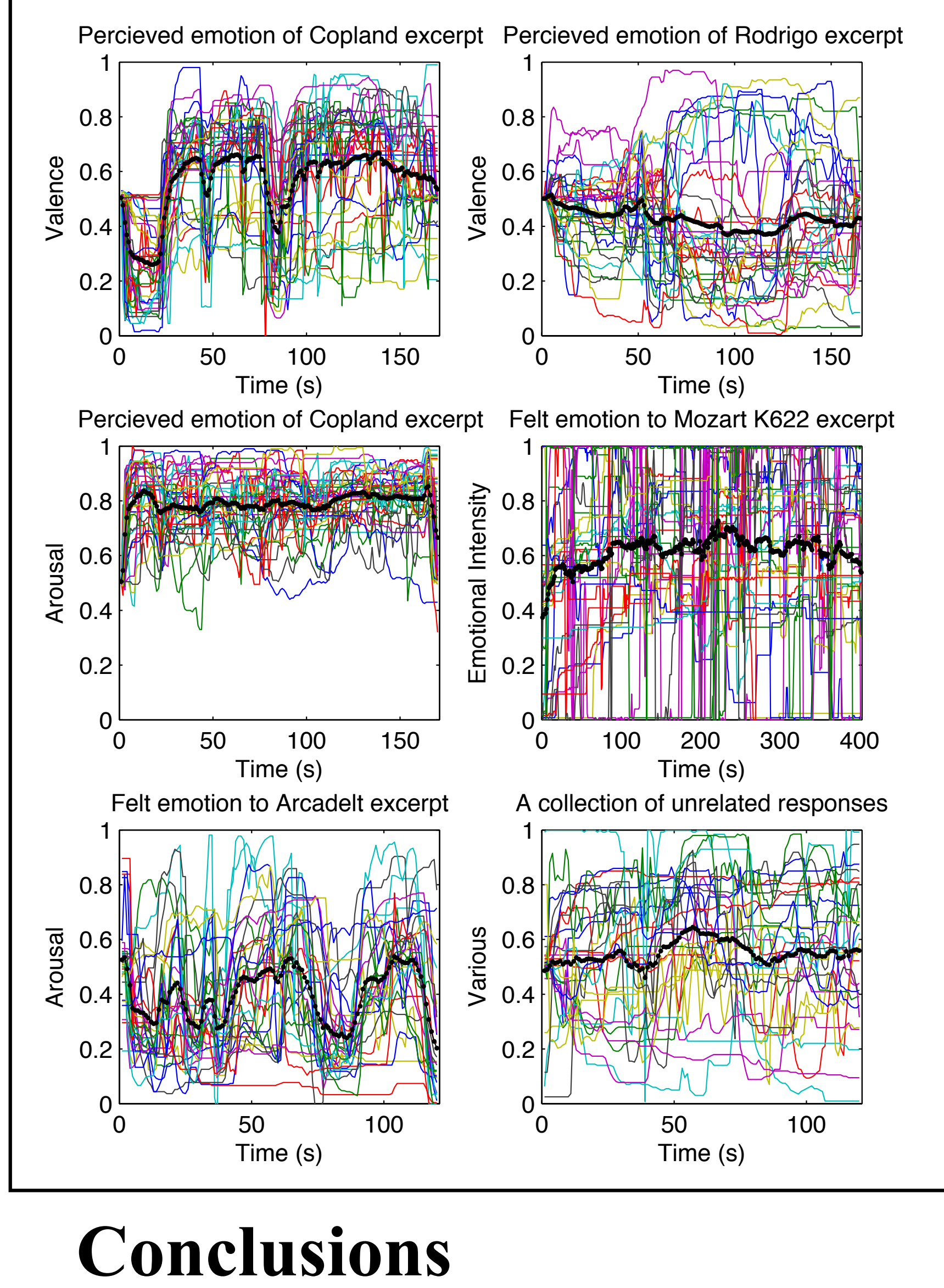

Figure 7. Collections of continuous response separated by the standard deviation ratio test on differenced data. To the left are collections with high cohesiveness, the right are collections which are no more cohesive than the unrelated response collections.

Conclusions

Continuous rating responses to music do not lend themselves to statistical significance tests of correlations, nor is this class of relatedness always the most relevant, particularly after compensating for serial correlation (see Figure 3). It is worth exploring measures of similarity or relatedness that are closer to our intuitions on these rating data.

Bibliography

Bartlett, M. (1935). Some aspects of the time-correlation problem in regard to tests of significance. *Journal of the Royal Statistical Society*, 98(3):536–543.

Chapin, H., Jantzen, K., Scott Kelso, J., Steinberg, F., Large, E., and Rodriguez-Fornells, A. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. PloS one, 5(12):169–200.

Dean, R. and Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. Empirical Musicology Review, 5(4):152–175.

Gregory, D. (1995). The continuous response digital interface: an analysis of reliability measures. Psychomusicology, 14:197–208.

Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. The Annals of Mathematical Statistics, 7(1):29–43.

Krumhansl, C. L. (1996). A perceptual analysis of mozart's piano sonata k. 282: Segmentation, tension, and musical ideas. Music Perception, 13(3):401–432.

Livingstone, S., Palmer, C., and Schubert, E. (2011). Emotional response to musical repetition. Emotion, Epub.

Lucas, B., Schubert, E., and Halpern, A. (2010). Perception of emotion in sounded and imagined music. Music Perception, 27(5):399–412.

Luck, G., Troiviainen, P., Erkkilä, J., Lartillot, O., Riikkilä, K., Mäkelä, A., Pyhäluoto, K., Raine, H., Varkila, L., and Värri, J. (2008). Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations. Psychology of Music, 36(1):25–45.

Rodgers, J. and Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59–66.

Schubert, E. (2002). Correlation analysis of continuous emotional response: Correcting for the effects of serial correlation. Musicae Scientiae, Special Issue 2001-2002:213–236.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. Music Perception, 21(4):561–585.

Vines, B., Krumhansl, C., Wanderley, M., and Levitin, D. (2006). Cross-modal interactions in the perception of musical performance. Cognition, 101(1):80–113.

Wöllner, C. and Auhagen, W. (2008). Perceiving conductors' expressive gestures from different visual perspectives. an exploratory continuous response study. Music Perception, 26(2):129–143.