

WHEN IS AN EVENT IN A TIME-SERIES SIGNIFICANT?

Emery Schubert

Music and Music Education
University of New South Wales, Sydney, Australia

ABSTRACT

Collecting continuous, self-report cognitive responses to music and other kinds of performance is growing in popularity. However, some methods of time-series analysis are complex and difficult to implement. For example, the probabilistic elements in the kinds of time-series that occur in response to such cognitive tasks make it difficult to know which regions of a series are reliable or 'significant' as representations of underlying regularities, and which regions are not. This paper describes a method which I call 'the second order standard deviation threshold', which is relatively simple to use when responses from several participants, responding to the same temporally unfolding stimulus, are available. The method can be implemented on standard spreadsheet packages and offers researchers dealing with these kinds of time-series data a form of visual display from which meaningful conclusions can be reported. Advantages and disadvantages of this method are discussed.

1. INTRODUCTION

Suppose one has collected continuous emotional responses from several participants, responding to the same piece of music. To identify the underlying nature of the emotional response, one might take the average value across all the participants at each sampled moment of time, and so create a new 'summary' time-series. There may be several questions about these data one wishes to address, but here we will examine one question: How do we know which regions, if any, of the averaged time-series are important or reliable? In other words, how can we distinguish those regions of the series that reflect some 'true' underlying emotional response ('signal') from those which are purely random fluctuations ('noise')? This paper provides a conceptual description of one possible solution to this problem.

2. THE PROBLEM

Researchers with knowledge of inferential statistical methods may resolve the above question by using such approaches as ANOVA, F-tests or t-tests. Indeed, such methods have been used [1, 2]. Using repeated measures ANOVA one can compare two time points in a time-series and examine the difference between them for statistical significance, in the expectation that this will also reflect a difference of practical significance. This approach is reasonable, but has two drawbacks. One is that it provides information only about a significant difference between two (arbitrarily selected) points. It does not provide an 'absolute' significance level of any point in the series with respect to some general criterion. So while determining that one point may be significantly different from another point, we cannot be sure that either one of those points is therefore signal and not noise. Second, correct use of standard ANOVA rests on assumptions

about the distribution of the data points (e.g. that they are normally and independently distributed). This is unfortunate, because time-series data very commonly violate these assumptions [3]. More seriously, the actual distributional properties of time series data are, in general, unknowable *a priori*.

So, our problem reduces to the following: How can we determine statistically which regions of a time-series, made up of the average of responses from multiple experimental participants, are reliable or significant, without relying on standard parametric approaches?

3. ASSUMPTIONS

In order to proceed, we need to make *some* assumptions about the nature of the time-series we are dealing with. In the present case we are dealing with continuous responses, sampled once in equally spaced time intervals (e.g. once per second) with each sample consisting of a dependent variable rating scale (e.g. a rating made on a bipolar scale, coded from -100 [very sleepy] to +100 [very aroused]), as assessed by the participant (Fig. 1 shows a plot of the averaged-across-participant, sample by sample values obtained from these kinds of ratings). We also assume that there will be at least three participants (in the Fig. 1 example we see the averaged results of 67 participants), allowing us to make some kind of assessment of how reliable the underlying central tendency of responses are—that is, to make an estimate of the 'true', underlying mean of the scale (in the example, what the 'true' emotional arousal rating is as the piece of music unfolds in time).

On top of this, some more complex and empirical assumptions need to be made. The data from multiple participants responding to the same stimulus will allow calculation of central tendency estimates which are sometimes good representations of a 'true' mean response, and sometimes not—signal and noise respectively. By having a large sample size from a carefully selected subpopulation (e.g. skilled musicians), we can start approaching a better estimate of the true mean. This is a way of managing 'vertical noise' – the error at any given point in time.

However, the data will also contain 'horizontal noise' which is indicative of one participant making a response to the stimulus at a given time slightly out of synchronization with another participant (for example, due to difference in reaction time). This causes a complication because a simple averaging of ratings made at a particular sample point will not be an averaging or response to the same thing: Some participants will be responding to something that happened one second ago, others to what happened a few seconds ago and so forth. That is, we need to recognise and to take as an assumption that the responses of a particular participant, to a stimulus being rated continuously, are neither instantaneous (there will be a reaction time delay) nor

will they be in synchrony with those of other participants (inter and intra individual response latency patterns).

One way of quantifying this assumption analytically is to take the average across a small window of time for each response. A simple method is the moving average (MA), which replaces the assessed values at a time point with the average of the time points in the proximity of the current point ('horizontally', i.e. over time). The question is how wide should this window be? A window too narrow (e.g. a single point, which would merely reproduce the original series – as in Fig 1a) will lead to a lot of noise produced because the non-synchronous responses were not captured by that window (notice the red mean line in Fig. 1a is more jagged than that in 1b or 1c). However a MA window will be too wide if changes occurring in the underlying means are accumulated and therefore obliterated (this is referred to as the 'regression to the mean' problem [4]). It should be possible to set an optimal window size in between these two extremes. For the present investigation, this will be set to three seconds. This value is based on the experimental finding that emotional arousal response/assessment latency to a given point in a piece of music typically varies from about 1 to 3 seconds [5]. Accepting the longer end of this range means greater likelihood of capturing spurious departures from synchronous response for up to 3 seconds, without losing too much 'true mean' response at that point in time. This is a rough guide, and the optimal window size under different circumstances remains to be calculated.

So, at this stage we generate 'smoothed' versions of the time-series, by replacing each sample's data point with a horizontal average along three consecutive seconds (Fig. 1b and c). We do this for the points which follow the data point to be replaced ('forward moving average') because that way we assume that responses made at that point are the fastest, with points following being due to slower response times. That is, we assume that participants will not be anticipating their response.

4. IDENTIFYING SIGNIFICANT REGIONS IN THE SERIES

We now turn to the problem at hand – of identifying which parts of a time-series are 'signal' (in the example, reliable values of emotional assessment of the music as the music progresses), and which are noise (random fluctuations, rather than reflection of the 'true' underlying mean). As discussed above, some simple common techniques can be used to investigate this. The time-series can be broken into regions, and each region could be compared to the other using multiple repeated measures analysis of variance (an extension of the technique used by [2]). However, with repeated testing (in the order of the factorial of the number of samples) the type 1 error will grow exponentially. Another approach might be to use a one sample t-test with a set test value, comparing each value in the time-series against the test value. This method has the problems associated with violation of assumptions, as already mentioned, and the difficulty in knowing

what test value to set. Further, t-tests would only identify when the series was significantly different from the test value. It would ignore, for example, values which happened to be in the vicinity of the test value, but were actually signal (if you like, 'significant' or reliable) values. In other words, such a t-test will identify extreme absolute values — those outside the confidence interval boundary, and not necessarily values which arise with good agreement.

The method described here does not solve all of the above problems, but aims to address some issues and to provide a relatively simple solution, without requiring a sophisticated statistics package. The method described can be implemented on a basic spreadsheet application, and analysed by visual inspection.

First, the (vertical) sample-by-sample central tendencies are each plotted. We will use means for convenience (however the median is also a very useful measure of central tendency in these circumstances). These are indicated as dark red lines in Fig. 1. Then an indicator of the (vertical) spread across the participants at each sample is calculated and (if possible) plotted on the same chart. In the present case we will use standard deviation as the indicator of spread (light red line in Fig. 1). By inspection of the plot of these two series (Mean and SD) we may already be able to see how certain regions of time can be assumed to be signal and how others might be unreliable noise. The points at which SD drops or maintains a low level we will assume to be 'significant' mean values (that is, reliable estimates of the mean). After applying a forward MA transformation of the mean and recalculating the SD for this smoothed signal (see two such examples in Fig. 1b and c), the jitter of the mean series should be attenuated, and it should be easier to identify significant 'true' mean regions.

We must now establish a criterion that will determine that the mean displayed at any time point is significant (or reliable, or 'signal'). Rather than assuming knowledge of the (population) distribution of the time-series *a priori*, we shall work with empirically-generated sample deviations. We can estimate these by taking the (horizontal) SD of the (already calculated, vertical) SD scores. Let us call this the 'second order standard deviation'. The calculation also requires a window of time, which we will for now assume to be the entire length of the piece (230 samples in Fig. 1). This value provides us with an indication of the spread of SD scores across the time-series. We also calculate the central tendency of this (vertical) SD time-series (again, we will use the mean) to provide a reference frame about which the (horizontal) SDs fluctuate. If the (vertical) SD falls below a threshold level (regions shown with yellow shading in Fig. 1), we conclude that the mean response is significant (or reliable or signal). The threshold is some level below the mean SD, which might be set to a whole second-order (horizontal) SD below the mean SD (the yellow shaded regions in Fig. 1a and b), or perhaps one half of a second-order SD below (yellow shaded region in Fig. 1c).



Figure 1. Three second-order SD plots of Continuous Arousal Response to Slavonic Dance Op. 46 No. 1 by Antonin Dvorak. Yellow segments indicate regions identified as being statistically significant. Dark red line indicates averaged response of 67 participants from moment to moment. Light red line is the standard deviation at each sample (across participants or ‘verticle’ SD). X-axis is time, with precisely repeated portion of piece indicated by the two regions surrounded by the 5 second increment time interval labels within the square brackets under Fig. 1c (one commencing at t=8s, and the other at t=134s, duration 46s). ‘m’ refers to the measure or bar number at the corresponding time.

5. ASSESSMENT OF THE METHOD

To assess rigorously the efficacy of the foregoing procedure would require having available a series with known distinctions between signal and noise. In investigating human aesthetic and emotional response to temporally dependent art forms such as music and dance, finding such examples produces some challenges. One approach, used in the example in this paper, is to examine a piece of music that has an identically repeated section. We would hypothesise that the mean values, and the points identified as being significant, should be identical in both sections. See now Fig. 1, the arousal response to Dvorak’s Slavonic Dance Op. 46 No. 1 (for more information about the stimulus, see [6]), A section that is repeated identically is indicated by a highlighted block of tags on the x-axis (surrounded by square brackets under Figure 1c). We can make some preliminary assessments about the second order SD threshold method of identifying significant points in a time-series that are repeated. A comparison of the repeated sections in each of the three plots demonstrates that there are generally more significant (yellow shaded) regions in the repeated section. This is primarily because measurements made near the beginning of the piece are susceptible to greater fluctuations due to task learning, adjustment and orientation. This degradation effect seems to settle after 30 to 60 seconds. The second order SD could be weighted with a function such as a

negative exponential to reduce what may be an inflated SD at the start of the piece. However, it may also be that responses at the early stage of data collection will be less reliable, regardless of whether the section is repeated. Note, too, that the mean responses at the two parts of the series are not identical, but that the similarity improves near the end of the bracketed sections – this applies to the shape of mean time series and to the shape of the SD time series. So the method of comparing responses to identical sections must be treated with caution. The exposition of the section to be repeated needs to start later to reduce the larger SD occurring in responses at the beginning of the piece.

To summarise the second-order SD method of identifying significant moments in a time series has the following disadvantages and advantages.

Disadvantages:

- Because it rests on a post-hoc criterion, the method will treat a certain percentage of responses as significant, and a certain percentage as not, regardless of whether the samples were highly reliable, or highly unreliable. At worst it is a relative measure of significance—the method can only identify those regions which are more likely to be significant than others within the series (other methods described earlier face the same problem). For example, if the data points happened to be normally distributed, then a

1 SD threshold below the mean SD will identify 16% of the points in the time-series as being significant, regardless of the series under investigation. From this point of view, the method may be better thought of as a *ranking of significance*. With 1 SD threshold, the method identifies the (very crudely) top 16 percent of samples that are most likely to be significant responses. By contrast, a 0.5 SD threshold identifies 35% of samples in this way. These statements are, of course, only very crude indications, since the normality assumption is usually violated in time-series of the kind being discussed.

- Setting the threshold of the second order SD is arbitrary. In recent experiments, values of 0.5 SD and 1 SD below the mean of the (vertical) SD series have been used.

Advantages

- Few assumptions need to be made about the distribution of the data, making the method suitable for investigating continuous emotional and other responses to music, dance and other artistic, temporally dependent performances.
- Simple, commonly packaged statistical calculations are employed (Mean and SD only).
- The method is graphically oriented, allowing greater ease of analysis.

6. CONCLUSIONS

Time-series data are associated with complexities in nature and analytic techniques. With computer technology now enabling fairly easy collection of elaborate time-series data, there is a need for researchers to have access to simple methods of analyzing the substantial data sets, without making naïve mistakes sometimes attributed to ‘interocular’ testing [7] and other oversimplifying approaches. The method of second order standard deviation threshold described in this paper attempts to find a balance between simplicity on the one hand and validity on the other. The method, and similar approaches have recently been applied to continuous audience response to dance performances [8-12]. By generating a MA mean response time-series, its corresponding standard deviation times series, then using the mean and SD of that time-series, I have described a method that will allow identification of relatively significant, or reliable, responses. Further work will investigate what additional levels of sophistication may provide significantly greater analytic efficiencies. In the mean time, inquiry into the merits of various second order SD thresholds needs to be undertaken, so as to improve upon the rule-of-thumb element of the approach described.

7. ACKNOWLEDGEMENT

I am most grateful to Eric Sovey for his generous comments and suggestions about the method described in this paper. The research was supported by an Australian Research Council Grant DP0452290.

8. REFERENCES

1. McAdams, S., et al., Influences of Large-Scale Form on Continuous Ratings in Response to a Contemporary Piece in a Live Concert Setting. *Music Perception*, 2004. 22(2): p. 297-350.
2. Sloboda, J.A. and A.C. Lehmann, Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception*, 2001. 19(1): p. 87-120.
3. Ostrom, C.W., *Time series analysis regression techniques*. . 1990, Newbury Park, CA: Sage.
4. Shaughnessy, J.J. and E.B. Zechmeister, *Research Methods in Psychology*. 1990, New York: McGraw-Hill.
5. Schubert, E., Continuous measurement of self-report emotional response to music, in *Music and emotion: Theory and research*, P.N. Juslin and J.A. Sloboda, Editors. 2001. p. 393-414.
6. Schubert, E., Modeling perceived emotion with continuous musical features. *Music Perception*, 2004. 21: p. 561-585.
7. Gottman, J.M., *Time-series analysis: A comprehensive introduction for social scientists*. 1981, Cambridge: Cambridge University Press.
8. Vincs, K., Schubert, E., Stevens, C., Engagement and the ‘gem’ moment: How do dance students view and respond to dance in real time? *Proceedings of the 17th Annual Meeting of the International Association for Dance Medicine and Science*. Canberra, Australia, October 25-28, 2007 (in press).
9. Schubert, E., Stevens, C., Healey, S., & Haszard Morris, R.. Perceptual structure in Fine Line Terrain: Mapping continuous perceived emotional responses onto choreographic plan. In M. Atherton (Ed.) *Proceedings of CAESS Conference: Scholarship & Community*. Sydney: University of Western Sydney, 2005.
10. Stevens, K., with Landy, L. Measuring audience reactions to contemporary dance. Invited talk and workshop/demonstration, *Electroacoustic Music Studies Network (EMS07)*, De Montfort University, Leicester, UK, June 12, 2007.
11. Stevens, K. (2007). Audience development: techniques for measuring cognitive and emotional response to live performance. Invited paper for Workshop on Music and the Brain, Warsaw, June 22, 2007.
12. Schubert, E., Stevens, C., Gordon, M., Chen, J.. Continuous response to contemporary dance: An analysis of audience reactions to Albert David's 'Silent Heartbeat'. Manuscript in preparation.