Are survival analyses suitable to examine latency data?

Pablo Gomez (1), Javier Breithaupt (2), Manuel Perea(2 3), Jeff Rouder (4)

(1) DePaul University, Chicago, USA

(2) Universitat de Valencia, Valencia, Spain

(3) BCBL, Basque Center on Cognition, Brain, and Language, San Sebastian, Spain

(4) University of Missouri

Abstract

The uncovering of the time course of the influence of different factors in human performance is one of the principal topics of research in cognitive psychology/neuroscience. Over the past decades, researchers have proposed several methods to tackle this question using latency data. We examined a recently proposed procedure that employs survival analyses on latency data to provide "precise estimates" of the timing of the first discernible influence of a given factor on performance (e.g., word frequency on lexical access). Because the method is promising, an exploration of its strengths and its potential weaknesses is in order. Here we report the results of systematic simulations directed to parameter recovery that revealed that this method tends to over-estimate the divergence point with a realistic number of observations per condition. We conclude that divergence points can be informative under a very limited set of circumstances.

Perhaps the most common cognitive psychology experiment is one in which participants are presented with a stimulus that has been subject to an experimental manipulation. The stimulus elicits a response, and researchers often measure latencies to make inferences about mental processes depending on the pattern of differences found across conditions. This form of mental chronometry is widely used across very different tasks. Choice response times, naming times, and eye-fixation times are examples of latency-based measurements.

Often, the obtained effects are examined through null-hypothesis testing focusing on the mean latencies. Although this approach can be appropriate when the goal is to obtain evidence for the existence of a phenomenon per se, it also has an important disadvantage: the distributional information is lost, and hence, researchers may miss valuable information regarding what component of processing is affected by the experimental manipulation.

The shortcomings of focusing only on the mean latencies are well known (Balota & Yap, 2011; Heathcote, Popiel, & Mewhort, 1991; Ratcliff, 1979), and there are indeed

different methods to explore distributional properties of latency measurements. Each of these methods has a level of theoretical commitment regarding the processes generating the empirical latencies. One approach is to perform functional form analyses, which assumes that latencies can be described by distributions with known functional forms (e.g., ex-Gaussian, Weibull, among others). Notably, this idea is often not explicitly stated, for example, if one is to fit an ex-Gaussian distribution to data, the underlying assumption is that the latencies were generated by a convolution of a normally distributed process and an exponentially distributed process. Statistical inferences can be carried out on the parameters of the functional form (i.e., μ, τ, and σ in the ex-Gaussian distribution, see Heathcote, Popiel, & Mewhort, 1991).

A second approach that analyzes distributional information is based on process modeling. Here, the researcher assumes that the task is well accounted for by a mathematical or computational model, like the diffusion model for choice response times (Ratcliff, 1978) or the EZ-reader model for eye fixation durations during reading (Reichle, Pollatsek, Fisher, & Rayner, 1998). The best fitting parameters of the model being used turn into the dependent variables for statistical inference purposes (see Gomez, Perea, & Ratcliff, 2013, for a recent example).

Is there a procedure that goes beyond mean latency analyses without making strong theoretical commitments? In this article, we explore a method that might fulfill that void in the literature, namely, the divergence point analysis procedure. Its goal is to provide an estimate of the onset of the influence of a given variable (the divergence point) on the basis of latency data (e.g., response times, eye fixation times). There are three potential advantages to this method: i) it goes beyond the mean latency by exploring the full distributions of latencies; ii) it does not make assumptions about functional form and is not subject to parameter misspecification; and iii) it does not make a strong commitment to a process model.

In the context of latency measurements, a survival function on time (t) is defined as S(t)=P(T¿t); where the survival at time t is equal to the proportion of latencies that have not occurred by time t; hence at $t = 0$, $S(0) = 1$, while at the longest latency $S(max[t]) = 0$. There have been attempts to use survival and hazard functions in latency analyses. Notably, Van Zandt (2002) examined several of these procedures and concluded: "Serious hazard function analysis would use samples of at least a few hundred observations" (p. 482). Recently, Sheridan, Reingold, and colleagues (Sheridan, 2013; see also Ando, Matsuki, Sheridan, & Jared, 2015; Reingold & Sheridan, 2014; Reingold, Reichle, Glaholt, & Sheridan, 2012; Sheridan & Reingold, 2013; Sheridan, Rayner, Reingold, 2013) proposed a method that promises to overcome the limitations of traditional hazard function analyses by using a computationally intensive bootstrapping procedure. This procedure compares the latency distributions of two conditions by estimating the divergence point. The divergence point corresponds to the shortest latency value at which a manipulation has a significant impact. Thus, the divergence point offers an estimate of the timing of the first discernible influence of a given variable (e.g., word frequency in a word recognition task). Furthermore, this method can also inform us whether the effect of a given factor has an earlier onset (and for how long) than the effect of another factor.

In the bootstrapping method under consideration, in each iteration the latencies for each participant and condition are randomly re-sampled with replacement. These sampled

(bootstrapped) latencies are used to compute each participant?s survival curves, which in turn are averaged across subjects *a la* Vincentile. Next, for each time bin t the difference between conditions: $\Delta_{t,i}$ are computed (i^{th} iteration from 1 to 10,000), and then sorted. The range between the 5^{th} and the $9,995^{th}$ value becomes the confidence interval $CI(\Delta t)$ and the divergence point is defined as the shortest t at which the $CI(\Delta t)$ does not include 0. Aiming for a high temporal resolution, Sheridan and Reingold use $1 - ms$ bins (see Figure 1 for an example). We should note that in a recent paper, Reingold and Sheridan (2014) proposed some changes on the original version of the procedure by using confidence intervals rather than point estimates, but the both the conceptual and the technical limitations outlined in the present note apply equally to the original and the modified versions of the procedure.

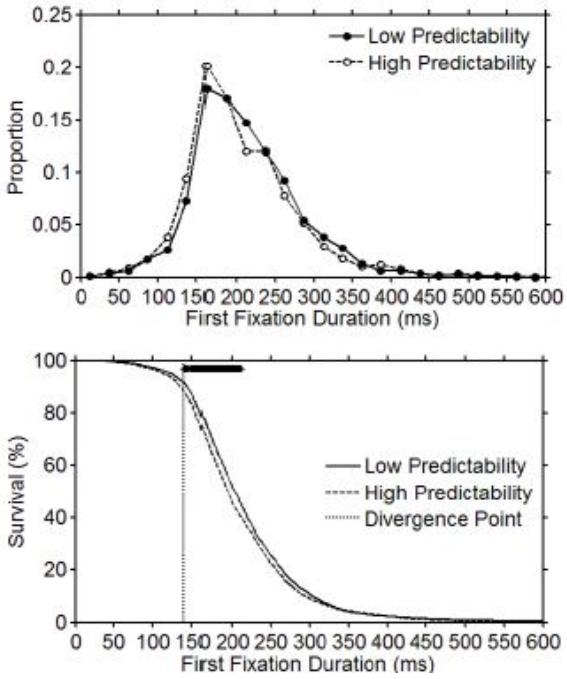

Figure 1. The figure (modified from Sheridan, 2013, page 27) shows the distributions of first-fixation duration on target words in the low and high predictability conditions in the top panel, and the survival curves in the bottom panel. The row of points at the top of the survival curves indicates the time bins with a significant difference between the low and high predictability curves using the method being examined in the present note

Although this method seems promising and useful for researchers interested in exploring the time course of empirical effects, there are some important technical and conceptual issues that severely limit its applicability.

Limitations of divergence point estimations in survival analyses

The divergence point method is a way to estimate the point at which distributions corresponding to two different experimental conditions separate. The implication is, therefore, that before that time, the two distributions of latencies under comparison have equal

densities. Figures 2 and 3 display the cumulative density functions generated with an ex-Gaussian distribution in which there are effects on μ (Figure 2) and on τ (Figure 3). The interpretation of both figures is quite straightforward: the distributions are always separated. The divergence point is at the starting point of the latency distribution any time there is stochastic dominance – note that this is the case for the overwhelming majority of experimental manipulations in psychology. Stochastic dominance refers to the probability of observations smaller than x being greater for one variable than the other for all values of x (see Heathcote, Brown, Wagenmakers, & Eidels, 2010). The true divergence point is equal to the shortest latency, which has a very important practical implication: with enough observations, the divergence point will approach the shortest latency (as we will demonstrate below).

Only in very specific situations do divergence points occur at times other than the shortest latency. Figure 4 shows one such scenario. In this case, up to a specific point in time (the divergence point), the latencies are generated by one mechanism in the two conditions. After this divergence point the latencies are generated by separate and distinct processes in each condition. The divergence point is, hence, the time at which the second component of processing kicks in. Although one could think of (somewhat contrived) situations in which this might occur, we do not feel that those situations are plausible. For example, one might assume a non-cascaded process model in which latencies are generated only after a common initial stage shared by the two conditions has been terminated, and in which the duration of a later stage is affected by the experimental manipulation. However, cognitive neuroscience has taught us that interactive cascaded models are more biologically plausible than serial, staged models (see Carreiras, Armstrong, Perea, & Frost, 2014, for a recent review).

The conceptual limitation just discussed is highly related to a technical limitation: in most experimental manipulations there is stochastic dominance, and hence the true divergence point approximates the shortest latencies. Indeed, when generating latency distributions for two different conditions in evidence accumulation models (e.g., Ratcliff's diffusion model), there are differences between conditions from the start, regardless of whether the two conditions vary in "response caution" (distance to the response boundaries, a parameter), "quality of information" (drift rate parameter), or "non-decisional time" (time of encoding and response, Ter parameter) (i.e., the true divergence point would be shortest latency). Even when the experimental manipulation could potentially produce a divergence point, this tends not to occur; Ratcliff & Rouder (2000) showed that in masking manipulations, there is stochastic dominance and a stationary drift rate in the evidence accumulation process.

What are then the consequences of applying this method to data with stochastic dominance? The answer to this question reveals a technical limitation of the method, as will be explained below. We carried out a series of Monte Carlo simulations to explore the issue. In particular, we generated data from an ex-Gaussian distribution assuming that the experimental effect was either in the μ or τ parameters. We manipulated the number of hypothetical trials per condition and then applied the bootstrapping method to estimate the divergence point.

Simulation studies

For the first simulation, we explored if the method yields false positives when samples generated from an ex-Gaussian with parameters $\mu = 541$, $\sigma = 68$, and $\tau = 115$ (i.e., a null effect in which identical parameters generate the simulated data for the two conditions, these parameters were the average parameters in Heathcote et al, 1991). The results from this simulation were satisfactory: The method generates false divergence points less than .01% regardless of the number of items in the simulations.

In the second simulation, we generated data in which the difference between the two conditions was an effect on μ. The data for the baseline condition was generated from an ex-Gaussian distribution with $\mu = 541$, $\sigma = 68$, and $\tau = 115$. We generated data for three simulated experimental conditions by changing μ to 561, 581 and 621 ($\Delta\mu = 20, 40, 80$) There is, therefore, stochastic dominance of the baseline condition relatively to all of these other conditions, and the true divergence point is at the starting point of the distributions. The results from this simulation are not encouraging for the method, as the estimation of the divergence point is highly biased by the number of trials per condition ($n = 20, 30, 50, 100, 250, 500, 1000$). Figure 5 shows the average divergence point for each of the parameter combinations (μ and n) across 1000 simulations. As a consequence of increased statistical power due to larger sample size at the trial level, the larger the number of trials per condition, the shorter the estimated divergence point (i.e., there is a statistical bias dependent on sample size). For example, with $\Delta\mu = 80$, and an n of 50, the divergence point is about 100ms higher than for $n = 1000$. In fact, when the number of trials is below 100, the different conditions are undistinguishable from each other. While the use of confidence intervals proposed by Reingold and Sheridan (2014) might partially alleviate this problem, the existence of a bias dependent on sample size is inherent to the procedure.

In the third simulation, we generated latency data in which the difference between the two conditions was an effect on ?. The data for the baseline condition was the same as in the previous simulations: it was generated from an ex-Gaussian distribution with $\mu = 541$, $\sigma = 68$, and $\tau = 115$. We generated data for three simulated experimental conditions by changing τ to 135, 155 and 195 ($\Delta\tau = 20, 40, 80$). As shown in Figure 2, changes in τ also produce stochastic dominance of the baseline condition and the true divergence point is at the starting point of the distributions. The results from this simulation are very similar to those from Simulation 2: The estimation of the divergence point is severely biased by the number of trials per condition (n = 20, 30, 50, 100, 250 or 500). Figure 6 shows the average divergence point for each of the parameter combinations (τ and n) across 1000 simulations. In sum, the method of estimating divergence points from two latency distributions does not generate false positives. Indeed, as it stands, the method might be too conservative. More importantly, when two latency distributions truly differ, the method severely overestimates the point in time of the divergence if a number of trials per condition within the norms of the field is being used. As mentioned above, it becomes an issue of statistical power: when $n \rightarrow \inf$, the divergence point corresponds to the smallest latency.

Discussion

We have examined a method to estimate the onset of the temporal differences between two latency distributions. The method provides with a single data point: the time in

which the survival functions have diverged. The authors have recently made some extensions to change the mechanics of the estimation that reports a confidence interval for the divergence point instead of just one point estimate (Reingold & Sheridan, 2014); however, the shortcomings of the approach seem more conceptual in nature as we have argued in this note.

The issues explored here, have practical implications that might potentially mislead researchers into making inaccurate inferences. This method (if it worked) could extend other forms of analyses such as ERPs – note that ERP experiments are designed to track the time course of an effect. Could the divergence point method offer information of the time course of an effect that is, to some degree, complementary to that offered by ERP waves? In a recent masked priming lexical decision experiment that examined phonological priming and word frequency effects, Ando et al. (2014) collected ERPs and also conducted estimations of divergence points in their latency data. The ERP data revealed a phonological priming effect starting in the 200 - 250 ms time window and a word-frequency effect starting in the 250 - 300 ms time window (i.e., an earlier effect of phonology than of word frequency). The estimations of the divergence points were substantially longer and in the opposite direction: 552 ms and 496 ms for the phonological priming effect and the word-frequency effect, respectively. Given that a skilled adult reader can identify a written word in 150-250 ms (refs), it seems unclear what are the cognitive processes reflected by divergence points above 450 or 500 ms. To explain the discrepancy between the ERP data and the latency data, Ando et al. (2014) indicated that lexical decision times "reflect relatively later processing" and argued that this was consistent with the fact that when examining a late 400600 ms time window, the ERP data reflected a word-frequency effect in the 400500 ms time window, whereas the masked phonological priming effect occurred in the 500 - 600 ms time window. However, Ando et al. (2014) did not report any statistical analyses to support this conclusion (e.g., an analysis of the relationship between the ERP effects with the divergence points). Furthermore, there is empirical and modeling evidence that shows that lexical decision times may tap into early encoding processes during word identification (see Gomez et al., 2013).

In short, survival function analyses might have a place in the researcher toolbox, however, just like any other statistical method, a new procedure needs to be vetted and examined before is being used in situations that do not fulfill the assumptions of the method. For the method in question here (the divergence point analysis procedure), our take home message is that its usefulness is limited to situations in which there is no stochastic dominance. Even in scenarios in which there is indeed a divergence point, as in Figure 4, the number of trials needed is out of reach in most experimental situations, as Van Zandt (2002) stated.

References

Ando, E., Matsuki, K., Sheridan, H., & Jared, D. (2015). The locus of KatakanaEnglish masked phonological priming effects. Bilingualism: Language and Cognition, 18, 101-117.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. Current Directions in Psychological Science, 20, 160-166.

Carreiras, M., Armstrong, B.C., Perea, M. & Frost, R. , (2014) The What, When, Where, and How of Visual Word Recognition. Trends in Cognitive Sciences (TICS), 18, 90-98.

Gomez, P., Perea, M., & Ratcliff, R. (2013). A diffusion model account of masked vs. unmasked priming: Are they qualitatively different? Journal of Experimental Psychology: Human Perception and Performance, 39, 1731-1740. DOI: 10.1037/a0032333

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. Psychological Bulletin, 109, 340.

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. Journal of Mathematical Psychology, 54, 454-463.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85, 59108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. Psychological bulletin, 86(3), 446-461.

Ratcliff, R., & Rouder, J.N. (2000). A diffusion model account of masking in two-choice letter identification. Journal of Experimental Psychology: Human Perception and Performance, 26, 127-140.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological Review, 105, 125 -157.

Reingold, E. M., & Sheridan, H. (2014). Estimating the divergence point: a novel distributional analysis procedure for determining the onset of the influence of experimental variables. Frontiers in Psychology, 5.

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. Cognitive psychology, 65, 177-206.

Sheridan, H. (2013). The Time-course of Lexical Influences on Fixation Durations during Reading: Evidence from Distributional Analyses. Doctoral Dissertation, University of Toronto. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/35996/6/Sheridan_Heathe_201306_PhD_thesis.pdf

Sheridan, H., & Reingold, E. M. (2013). Distinct stages of word identification during reading: Evidence from eye movements. Journal of Vision, 13, 511-511.

Sheridan, H., Rayner, K., & Reingold, E. M. (2013). Unsegmented text delays word identification: Evidence from a survival analysis of fixation durations. Visual Cognition, 21, 38-60.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Ed.), Stevens' Handbook of Experimental Psychology (3rd ed., pp. 461-516). San Diego, CA: Academic Press.

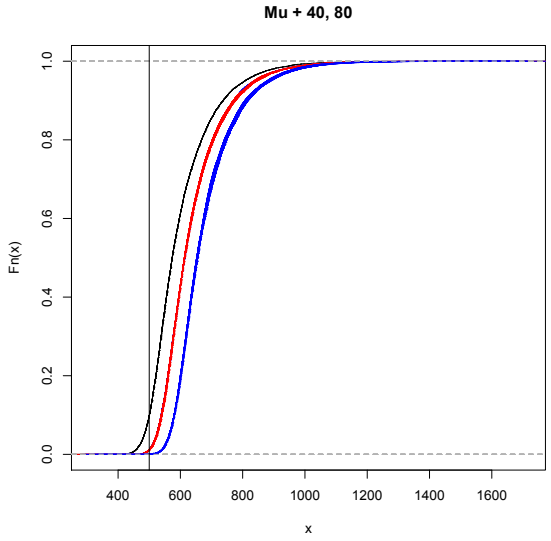

Figure 2. The figure shows the cumulative density functions generated with an ex-Gaussian distribution in which there are effects on μ

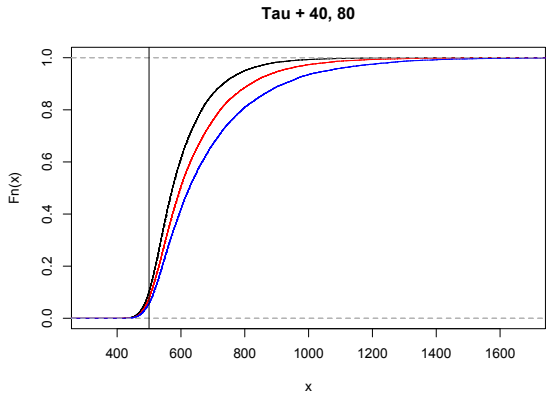

Figure 3. The figure shows the cumulative density functions generated with an ex-Gaussian distribution in which there are effects on τ

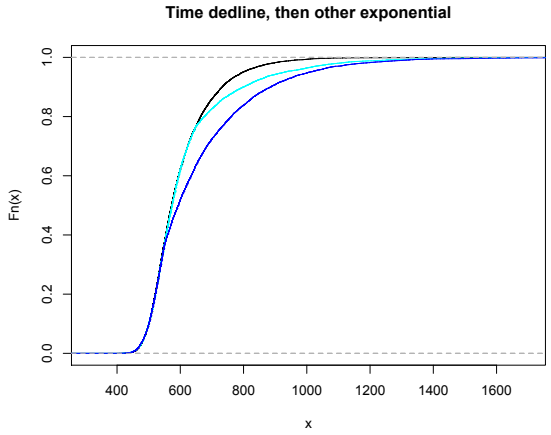

Figure 4. The figure shows the cumulative density functions generated with two sequential non-cascade processes

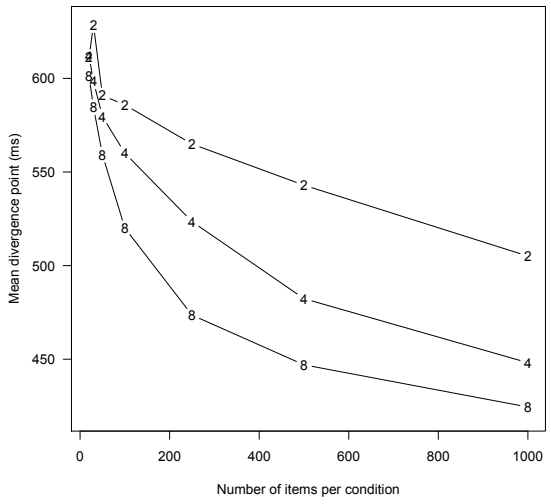

Figure 5. The figure shows average diverge point for simulation 2.

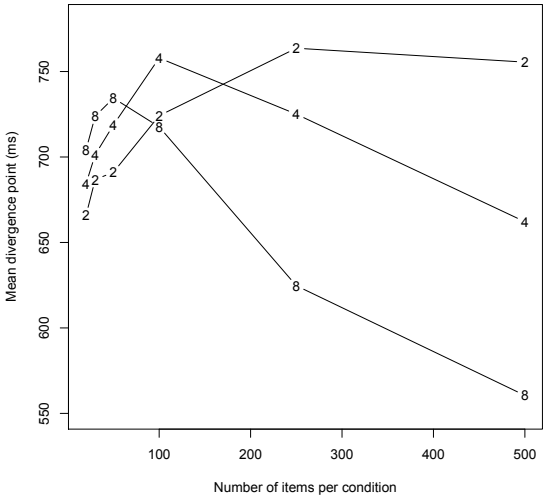

Figure 6. The figure shows average diverge point for simulation 2.