

An Autocorrelated Mixture Model for Sequences of Response Time Data ^{1 2}

Peter F. Craigmile, *The Ohio State University*

Mario Peruggia, *The Ohio State University*

Trisha Van Zandt, *The Ohio State University*

Technical Report No. 778

April, 2006

**Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247**

¹ This material is based upon work supported by the National Science Foundation under Awards No. SES-0214574 and SES-0437251.

²The authors would like to thank and acknowledge assistance in the early stages of this project from Emily Johnson, Dartmouth College (REU student sponsored by the National Science Foundation under award No. DMS-9988006) and Maria Salotti, University of Wisconsin at Stevens Point, (REU student sponsored by the Department of Statistics and the College of Mathematical and Physical Sciences of The Ohio State University).

Abstract

Human response time (RT) data are widely used in experimental psychology to evaluate theories of mental processing. Typically, the data constitute the times taken by a subject to react to a succession of stimuli under varying experimental conditions. Because of the sequential nature of the experiments there are trends (due to learning, fatigue, fluctuations in attentional state, etc.) and serial dependencies in the data. The data also exhibit extreme observations that can be attributed to lapses, intrusions from outside the experiment, and errors occurring during the experiment. Any adequate analysis should account for these features and quantify them accurately, but current modeling practices (both frequentist and Bayesian) are lacking in this respect. We recognize that Bayesian hierarchical models are an excellent modeling tool, but note that most of the current literature is based on likelihood specifications that are mainly dictated by computational convenience. For this reason we focus on the elaboration of a realistic likelihood for the data and on a careful assessment of the quality of fit that it provides. We judge quality of fit in terms of how well the model captures the essential features of the data. Specifically, we validate our model by simulation, comparing the marginal and first order serial dependence properties of synthetic data generated from the posterior predictive distribution with those of the observed data. Our work demonstrates good fit for several RT sequences, indicating that the proposed mixture model can provide a solid building block for elaborating complex Bayesian hierarchies.

Keywords Bayesian models; reaction times; sequential dependencies; extreme observations; time series modeling.

1 Introduction

Response time (RT) is a ubiquitous measure of human performance. It is used to formulate theories of brain function and cognitive processing; to evaluate training regimens, user interface design, vehicle operation, and task design and to evaluate medical conditions, especially schizophrenia, learning disorders, and other psychological disorders. RT has been used as a window on psychological processes for almost two centuries and it forms the foundation for most work in cognitive psychology. Efforts to model RT, or, more fundamentally, the processes that give rise to certain patterns of RT, have been the focus of attention for much of experimental psychology for the last 50 years (see [Luce, 1986](#) for a comprehensive review).

There are several approaches to modeling RT data. Most commonly a researcher will derive a hypothesis about the behavior of a summary statistic (such as the mean) with changes in an experimental condition: for example, mean RT should decrease with increased practice ([Heathcote, Brown, and Mewhort 2000](#)). Fewer researchers formulate distributional hypotheses about the entire sample of RTs: for example, if response activation is controlled by neural events occurring according to a Poisson process, then the distribution of RTs should be gamma. Some common marginal distributions used to describe RTs are the gamma, Weibull, inverse normal, and the “ex-Gaussian,” a convolution of a normal and exponential. Other theoretical approaches are based on the minima of first passage time distributions, and involve Poisson, Wiener diffusion and Ornstein-Uhlenbeck processes ([Ratcliff and Smith 2004](#)).

These approaches implicitly assume that samples of RTs collapsed across similar conditions in an experiment are independent and identically distributed (*IID*). Other approaches, however, embrace the notion that sequences of RTs are neither independent nor identically distributed. Work investigating sequential effects, priming, inhibition of return, task-switching, and so forth is directed toward understanding how information about previous stimuli and responses influences processing of later stimuli and facilitates or inhibits responses to those stimuli (e.g., [Jones, Love, and Maddox; Meeter and Olivers; Stewart, Brown, and Chater, 2006, 2006, 2005](#)). Related work by [Gilden and others \(Gilden 1997; Gilden 2001; Pressing and Jolley-Rogers 1997; Thornton and Gilden 2005; Van Orden, Holden, and Turvey 2003\)](#) has emphasized long-range effects across sequences of trials, or $1/f$ (pink) noise, which has brought more attention to the autocorrelation structure of RT se-

quences (see also [Wagenmakers, Farrell, and Ratcliff, 2004](#) and [Farrell, Wagenmakers, and Ratcliff, 2005](#)).

Common to all analyses of RT data is a lack of uniformity in the treatment of extreme observations. Because RT data are positively skewed, some such observations are to be expected from the process that generated the data. However, other extreme observations are “clearly” due to other factors, such as unscheduled rests in the middle of a block of trials, lapses in attention and intrusions of the subject’s physiological system (sneezing, itching, etc.). Less easily detected are extreme observations on the lower tail of the distribution. Extremely fast RTs are commonly attributed to rapid guessing or error (twitches). How extremely slow and fast observations are to be teased from the data of interest is a question of longstanding interest in psychology and a focus of much research in statistics ([Barnett and Lewis 1994](#); [Ratcliff 1993](#); [Ulrich and Miller 1994](#); [Van Selst and Jolicoeur 1994](#)).

On occasion extreme observations have been modeled using mixture distributions. For example, ([Belin and Rubin](#)) model RTs among schizophrenics as a mixture of two normals and interpret an observation arising from the component with the larger mean as corresponding to the occurrence of an attention lapse. The most popular solution to the problem of extreme observations, however, is to trim or discard all observations above and below certain fixed values, determined by historical practice, quantiles, or standard deviations. Other practices involve transformations, such as the inverse or log, to minimize the effect of the extreme observations. When RT data are treated as a time series, trimming and deletion disturbs the autocorrelation structure of the sequence. Overall, the statistical methods currently employed in the social sciences do not attempt to describe simultaneously RT dependencies (autocorrelations) and the tail behavior of the marginal RT distribution.

Probably the most important recent development in RT analysis is the use of hierarchical Bayesian models for RT data ([Peruggia, Van Zandt, and Chen 2002](#); [Rouder, Lu, Speckman, Sun, and Jiang 2005](#); [Rouder, Sun, Speckman, Lu, and Zhou 2003](#)). The specification of the likelihood, the theoretically motivated “data generating” portion of the model, provides a solid foundation on which the Bayesian hierarchy can rest. Such specification is the main focus of this article. For a sequence of RT data collected on a given subject we wish to specify a probabilistic model that can capture *at once* the local dependencies and the tail behavior of the marginal distribution of the RTs. In the end, we would consider our modeling approach

successful if it can generate data sequences that exhibit the main features of observed data.

Some Bayesian models for RT data have been recently proposed. Such models are attractive because they afford the flexibility to specify RT distributions that depend on varying experimental conditions and manipulations, typically via hierarchical regression structures for certain model parameters assumed to capture the influence of those experimental conditions and manipulations. Examples of this approach are presented in (Peruggia, Van Zandt, and Chen) and (Rouder, Sun, Speckman, Lu, and Zhou). The models in (Rouder, Sun, Speckman, Lu, and Zhou) assume that the observed RTs are conditionally independent, ignoring sequential dependencies. The models in (Peruggia, Van Zandt, and Chen) attempt to capture sequential dependencies through an autoregressive structure for the logarithm of the scale parameter of the RT distribution. Ultimately, this approach seems to be more useful for uncovering lack-of-fit than for describing local dependencies (Peruggia 2005).

In this paper we propose that the suitably detrended sequence of RT logarithms can be described with a mixture likelihood. Detrending is a crucial modeling step, because it removes smooth changes in RT levels due to learning effects, fatigue, etc., distinguishing them from more localized dependencies. The three components of the proposed mixture account for additional features of the data. Specifically, a Gaussian autoregressive component captures local dependencies (i.e., dependencies that decay geometrically) and two exponential components model the two tails of the marginal log RT distribution. This likelihood provides a framework for performing analyses of RT time series that do not compromise model realism and autocorrelation structure. It can be used to build sensible Bayes hierarchical models which are not based entirely on consideration of convenience (namely, conditional conjugacy).

The organization of this paper is as follows: In Section 2 we carry out an exploratory analysis of a set of RT data that highlights those characteristics of the data that a good model must be able to accommodate. In Section 3 we propose a general mixture time series model for RT data. We apply the model to the data and summarize our inferences in Section 4 and we validate the model performance via Monte Carlo simulation in Section 5. We close with a brief discussion in Section 6.

2 Example response time data

(Wagenmakers, Farrell, and Ratcliff) published RT data collected with the intent to investigate autocorrelation structure across long sequences of trials.³ They recruited six subjects and considered three simple tasks to be performed in response to a fixed set of stimuli. The set of stimuli included all numerals from 1 to 9, except 5. The numerals were presented on a computer monitor and each task required the subject to make a keypress response on a computer keyboard. The first task was a simple detection: subjects had to respond as soon as any stimulus was presented. The second task was a choice response: subjects had to respond with one key when the numeral was greater than 5 and with a different key when the numeral was less than 5. The third task was a time estimation task: the subjects had to respond with a keypress one second after the numeral was presented, regardless of what the numeral was. Subjects were asked to perform each task in response to a sequence of stimuli presented in 1048 successive trials with no breaks and their RTs were recorded. The first 24 ‘practice’ responses of each sequence were discarded and the ensuing 1024 were retained for analysis. The experiment was designed so that the stimulus sequences were exactly the same regardless of which task a subject was to perform.

An important manipulation in the study was the interval between a response and the presentation of the subsequent stimulus, the response-stimulus interval or RSI. To prevent anticipatory responses (extremely fast RTs), RSIs were uniformly distributed between 550ms and 950ms in “short” conditions, and between 1150ms and 1550ms in “long” conditions. Each subject provided 6 RT sequences: one sequence for each task in each RSI condition.

The authors’ motivation for collecting these RT sequences was to specifically address claims by (Gilden) and others that RT variability over time can be described by a $1/f$ noise process. If confirmed, the presence of $1/f$ noise would suggest that the RT generating process is “fractal” in nature, possessing similar temporal characteristics regardless of the time scale of the measurement. Another way to conceive of $1/f$ noise is in terms of long-range dependencies, or serial correlations between responses that persist throughout the RT sequence. Gilden has suggested that this would relate to a subject’s formation of mental representations used to perform the task. The most important implication is that any realistic model of the RT generating process would have to produce,

³The data are available from <http://users.fmg.uva.nl/ewagenmakers/fnoise/noisedat.html>.

at least as a byproduct, $1/f$ noise.

Both Wagenmakers et al.’s and Gildea’s approaches are subject to some of the limitations that we noted in the introduction. They both used standard time-series techniques, such as those involved in fitting ARIMA-type models and estimating power spectra, that assume that the underlying noise distribution (the distribution of the innovations) is marginally normal. Wagenmakers et al. deleted extreme observations from the sequences, disturbing the autocorrelation structure and potentially introducing artifacts, although they report that the deletions did not affect their results. Gildea attempted to “detrend” his RT sequences by first subtracting the main effects from the individual observations in each experimental condition. In one analysis, he also attempted to eliminate short-term dependencies by subtracting main effects due to stimulus sequence. He then subtracted any linear trend from the sequence, and converted each sequence to z -scores to equate variability across different subjects.

While the general idea of detrending is sound, linear trends are not sufficient to capture all the features of the data. In the remainder of this section, we present the results from an initial exploratory analysis of the Wagenmakers et al. data, providing evidence for the presence of trend and both short and long-range dependencies. For our analyses we use only the data from the simple detection task under short and long RSIs.

2.1 Exploratory RT analysis

All RTs were transformed to log RTs (in base e). The logarithmic transformation enabled us to examine the RT sequences for potential extreme observations in both tails of the distribution and to assess the extent of serial dependence in each sequence, as we will demonstrate. Figure 1 displays time series plots of the log RTs for each of the six subjects. Note that, here and elsewhere, the “time” series is really a “trial” series. The abscissa of the plots is the trial number (from 1 to 1024) and the ordinate is the log RT for each trial. The left column shows the time series for the short RSIs and the right column shows the time series for the long RSIs. There are several common features in these data. The series show evidence of trends, and both short- and long-range serial dependence. Trend refers to the smooth changes of the level of the process over time. Trends occur for a variety of reasons, including learning (negative trends) or fatigue (positive trends). Fluctuations in attentional

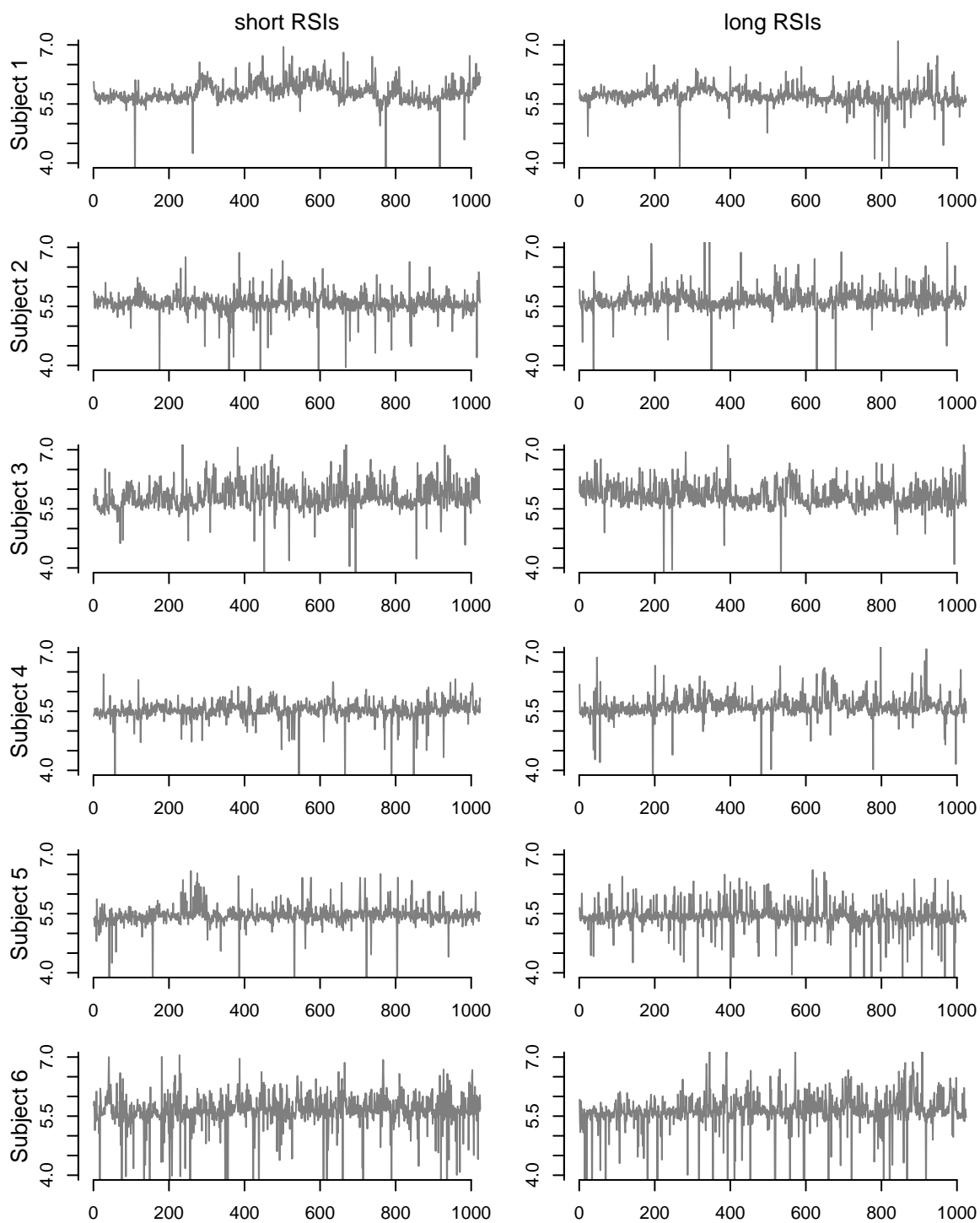


Figure 1: Time series plots of the log response times for the six subjects, for the short RSI experiments (left panels) and the long RSI experiments (right panels).

state or task readiness may result in a fixed shift in the mean of the process.

2.2 Detrending the log RT series

Before examining the other features of the data (serial dependence and extreme observations), we detrended each RT series using the following procedure. We first converted each series to normal scores. Let $q(\alpha)$ denote the α -th quantile of the standard normal distribution. For a given log RT sequence, $\{W_t\}_{t=1}^{1,024}$, let $\{R_t\}_{t=1}^{1,024}$ denote the associated ranks and let $\{W_{(t)}\}_{t=1}^{1,024}$ denote the sequence of order statistics. Then, the transformed data are given by $\{V_t = q((R_t - 1/2)/1,024)\}_{t=1}^{1,024}$ and we let $\{\hat{V}_{(t)}\}_{t=1}^{1,024}$ denote their order statistics.

After this conversion of the data to normal scores, we smoothed the data. Without special care, standard smoothing techniques, which depend on the assumption of independence, will result in series that are oversmoothed. A procedure that smooths dependent data was presented by (Wang). In this procedure, cubic smoothing splines are fit to Gaussian data and the dependence structure is modeled with an autoregressive process of order one. We applied this procedure to our normal scores series⁴, and obtained fitted trend values $\{\hat{V}_t\}_{t=1}^{1,024}$. For each t , we then determined the integer i such that $V_{(i)} < \hat{V}_t \leq V_{(i+1)}$ and estimated the trend \hat{W}_t on the original scale by linear interpolation between $W_{(i)}$ and $W_{(i+1)}$, using weights proportional to the distances of \hat{V}_t from $V_{(i)}$ and $V_{(i+1)}$.

Figure 2 displays the estimated trends for each subject and RSI condition. Subject 1’s trends are erratic for both conditions, while the other subjects have trends that are more steady. Most subjects show (oscillatory) increasing trend in both conditions, suggesting some tiring by the end of the trial sequence. Only Subject 3’s responses in the long RSI condition steadily and markedly decrease.

2.3 Dependence and extreme observations in the detrended log RT series

To illustrate the extent of serial dependence and extreme observations in these data, Figure 3 shows the long RSI, detrended log RT series for Subject 4 (the upper panel), along with common measures of dependence. This case is typical of the dependence and extreme observations found in the other RT sequences. Serial dependencies appear as “clustering” of data points in the series. For example,

⁴The R package ASSIST implements this procedure (<http://www.pstat.ucsb.edu/faculty/yuedong/software.html>).

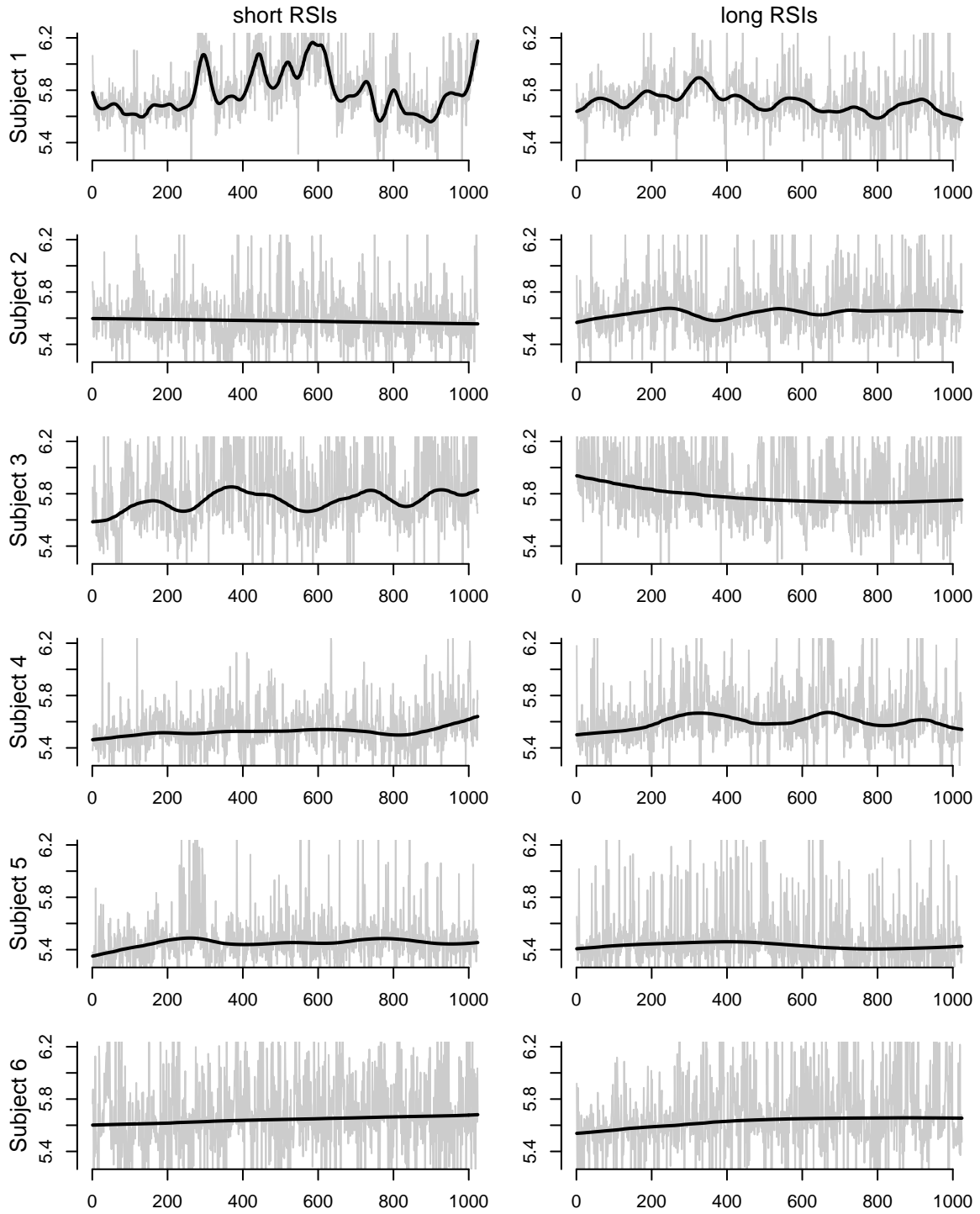


Figure 2: Zoomed-in plots of the estimated trends for the six subjects (dark lines) plotted over the original log RT values (grey lines), for the short RSI experiments (left panels) and the long RSI experiments (right panels).

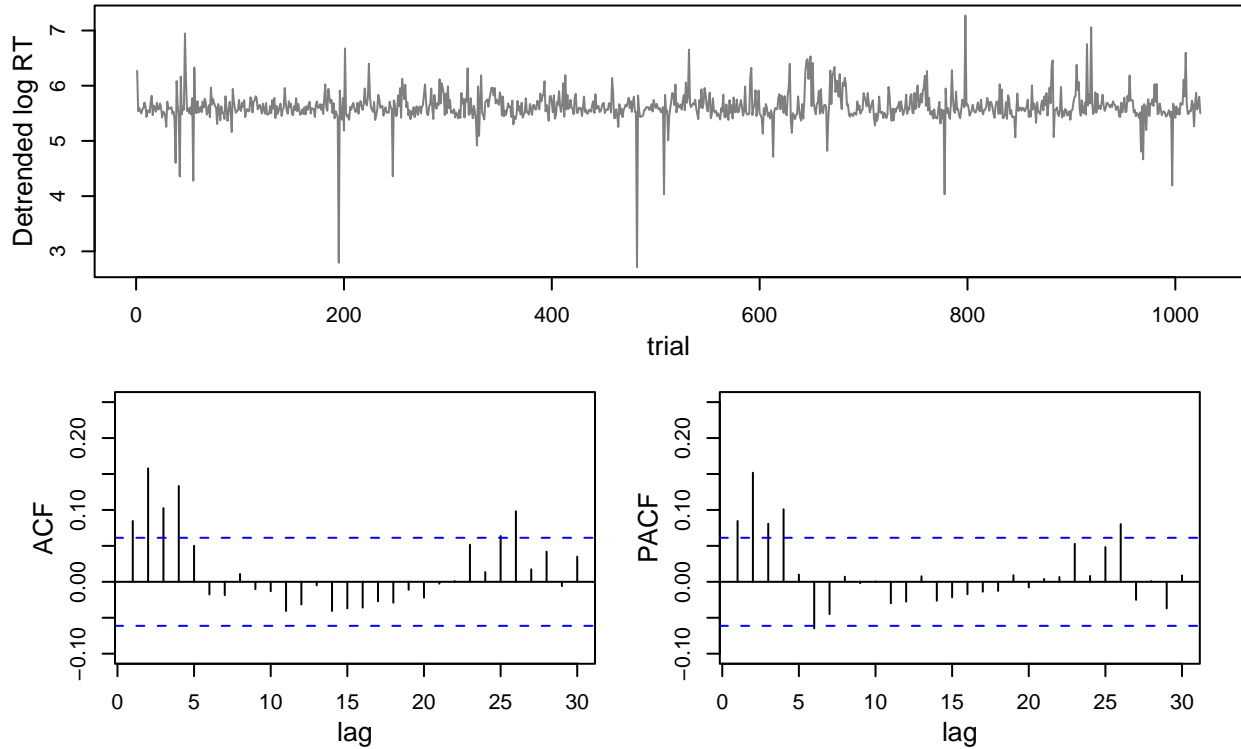


Figure 3: Long RSI, detrended log RT series for Subject 4. The top panel shows the time series plot of the data, the bottom left panel shows the sample autocorrelation function, and the bottom right panel shows the sample partial autocorrelation function.

there are two clusters of slow RTs around trial number 650 that look like “buttes” in the series. Another prominent feature of this sequence is the presence of extremely long and extremely short RTs, as can be seen in the time series plot. These observations would be considered outliers based on a model that specifies a Gaussian marginal distribution for the data. There are two types of extreme observations: slow (long RTs) which appear as upward spikes in the sequence, and fast (short RTs) that appear as downward spikes in the sequence.

Serial dependence may arise from both short-range and long-range effects, as we discussed earlier. Dependence over short time lags (short-range dependence) can be caused by carry-over effects from immediately preceding trials, either due to particular stimuli or to particular responses. When errors are committed in choice RT tasks, strong carry-over effects can be observed on subsequent trials. Short-range effects are quantified by autocorrelation and partial autocorrelations over short lags.

Given a nonnegative integer h , the lag h autocorrelation, $\rho(h)$, of a stationary sequence $\{X_t\}$ with finite second moments is defined as the correlation between elements in the sequence that are h time units apart, i.e., as $\rho(h) \equiv \text{corr}(X_t, X_{t+h})$. As h varies over the nonnegative integers, $\rho(h)$ defines a function called the autocorrelation function. For each h , the autocorrelation $\rho(h)$ can be estimated from the observed data via the Pearson product-moment correlation. This estimate is called the lag h sample autocorrelation. Figure 3 displays the sample autocorrelation function (i.e., the sample autocorrelations as a function of h) for Subject 4's data. The dashed horizontal lines on this plot indicate the pointwise upper and lower 95% confidence intervals for the autocorrelation function, assuming that there is no serial correlation (i.e., assuming a white noise sequence). Lags 1-4 (and perhaps lag 26) have autocorrelations significantly different from zero.

The lag h partial autocorrelation of a stationary sequence $\{X_t\}$ is the partial correlation between X_t and X_{t+h} , given the values of the intervening elements of the sequence. Partial autocorrelations are useful for determining how many previous trials exert a direct influence on the present trial (the order of the autoregressive component of an ARMA process), because, given an autoregressive component of order j , all partial autocorrelations greater than j should theoretically be equal to 0. The estimated partial autocorrelations for Subject 4's data (i.e., the sample partial autocorrelation function) are shown in the lower right panel of Figure 3. Again, the dashed lines indicate the pointwise upper and lower 95% confidence limits assuming no serial dependence. The partial autocorrelations from lags 1-4 are significantly different from 0.

Longer range dependencies, those that may reflect the formation of mental representation, can also be examined using the sample autocorrelations and partial autocorrelations. Long-range dependencies and/or unmodeled trends are characterized by slowly decaying autocorrelation and partial autocorrelation sequences. The data shown in Figure 3 do not suggest long-range dependence, because the autocorrelations cut off quickly after four lags, and then are only just significantly different from zero at around 26 lags. These plots, then, do not provide strong evidence of long-range dependence.

Given the extreme observations observed in the detrended log RT series, some caution should be taken in interpreting the autocorrelation and partial correlation function. For example, consider the autocorrelation function. It has been established that the presence of a large additive outlier in an

ARMA proces will damp down the autocorrelation estimates (Peña, Pena, Tiao, and Tsay 2001). Similarly, for the mixture model that we will define in Section 3.2, the autocorrelation function for the process will be damped down by the presence of the independent additive exponential components.⁵ Also, because the data are not Gaussian, the variability and bias of the sample autocorrelation function increases as the extreme observations become more prominent. Trimming these observations is not a satisfactory solution, because this operation can also affect the estimation of autocorrelation.

To illustrate these problems we carried out a small Monte Carlo experiment. Let $\{U_t\}$ be a sequence of $N(0, 1)$ random variables, and let $\{X_t\}$ be the autoregressive process of order 1, AR(1), defined by

$$X_t = 0.6X_{t-1} + U_t,$$

for each t . A sample realization of this process is shown in the top left panel of Figure 4. An approximation to the sampling distribution of the autocorrelation function for this process based on 500 realizations of the AR(1) process is displayed in the middle left panel of the same figure. The circles denote the median value of the sampling distribution of the autocorrelation function at each lag, and the lines connect the 0.025 and 0.975 quantiles.

For each of the 500 realizations of $\{X_t\}$ we considered a mixture process, $\{W_t\}$, defined by

$$W_t = X_t + \delta_t Y_t,$$

where $\{\delta_t\}$ is an *IID* sequence of independent Bernoulli(0.2) random variables and $\{Y_t\}$ is an *IID* sequence of independent exponential random variables with mean two. We assume that the processes $\{X_t\}$, $\{\delta_t\}$, and $\{Y_t\}$ are mutually independent. An example of a realization of this mixture process based on the example realization of the AR(1) process is given in the top right panel of Figure 4. Note the presence of the additive outliers relative to the AR(1) process. An approximation to the sampling distribution of the autocorrelation function for the mixture process based on 500 realizations of the process is shown in the middle left panel of the same figure. Note that, as we described

⁵This follows by noting that we can rewrite the process as an AR(1) process plus an independent white noise process. The theoretical autocovariance function for the sum of these processes is the sum of the autocovariance functions for each process, and the result follows by noting that the autocovariance function for the white noise process is only nonzero at lag zero.

previously, the autocorrelation function is damped down relative to the autocorrelation function for the AR(1) process.

Two methods are commonly used to improve the estimate of the autocorrelation function: winsorizing and trimming. Both methods try to reduce the effects of the additive extremes. For the winsorized estimate of the autocorrelation function, we truncate values in the process that exceed some threshold before calculating the autocorrelation function. An approximation to the sampling distribution of the sample autocorrelation function estimated with winsorizing threshold set equal to the 95th percentile of the distribution is shown in the bottom left panel of Figure 4. In the trimmed estimate of the autocorrelation function, we first remove those trials that exceed a given threshold and then calculate the autocorrelation function of this reduced series. An obvious problem with this method is that it disturbs stationarity of the process, making interpretation of the autocorrelation function difficult. Ignoring the problem with stationarity for trimming, an approximation to the sampling distribution of the sample autocorrelation function estimated with trimming based the 95th percentile of the distribution for our simulated series is shown in the bottom right panel of Figure 4.

While trimming and winsorizing both slightly reduced the damping effect of the additive extreme values upon the sample autocorrelations, further experiments showed that changing the percentile cutoff value did not allow us to ever approach the autocorrelation function for the AR(1) process. We conclude that any method of analysis should model both the dependence and extreme observations jointly, rather than trimming or winsorizing.

3 A mixture model for the simple RT data

In the previous section we have identified three fundamental features of RT data: (1) long-range dependencies and/or trends; (2) short-range dependencies; and (3) extreme observations. With this in mind, we propose a “prototype” time series model for sequences of RT data. This model is intended to serve as a framework for the analysis of RT sequences in general contexts, especially in hierarchical Bayes models. The primary goal in this article is to evaluate the properties of the prototype model and we reserve construction of a Bayesian hierarchy for a subsequent article. Here,

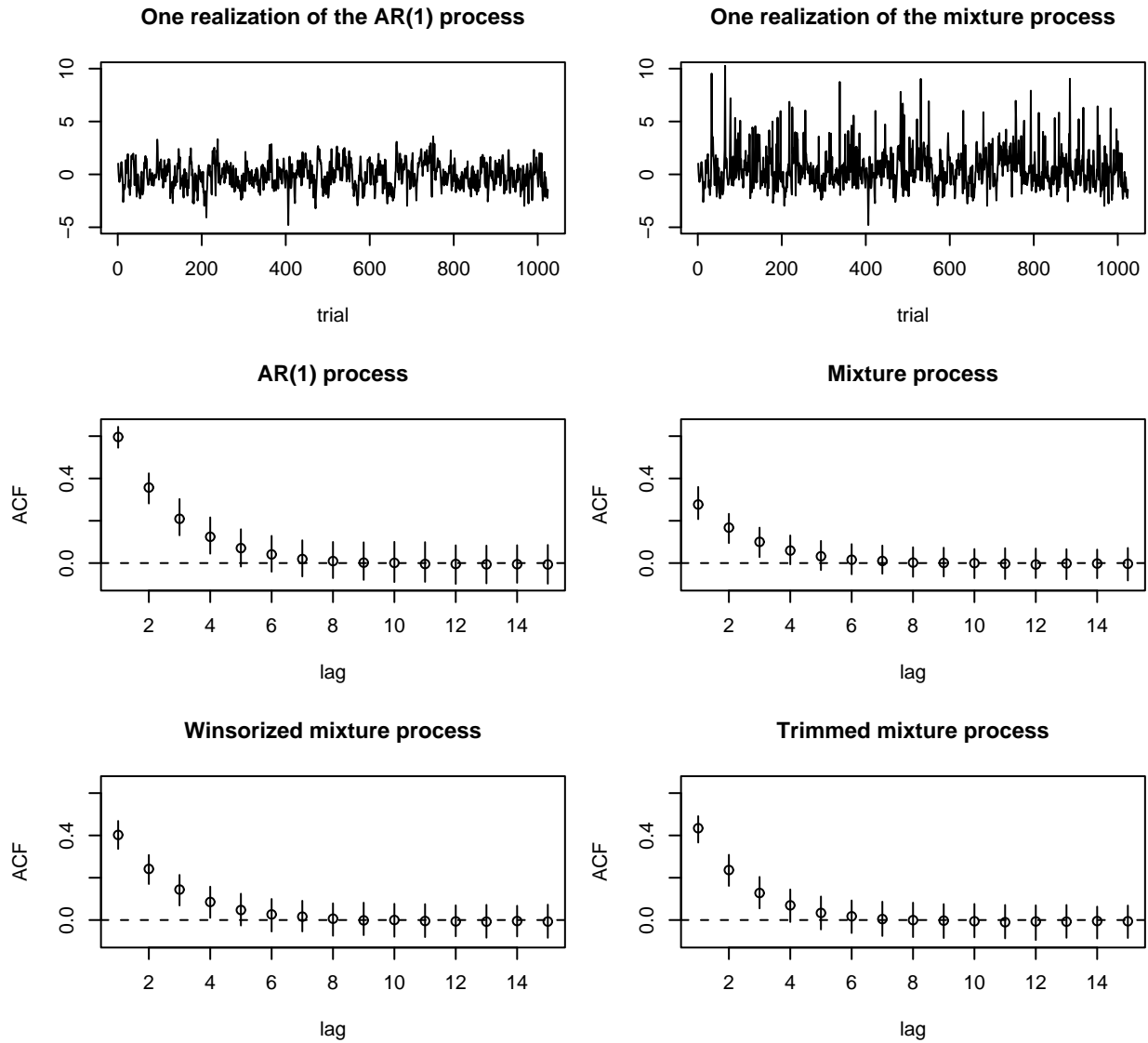


Figure 4: Illustration of the effect of additive extreme values on the sample autocorrelation function. See the text for more details.

then, we model the time series for each subject and type of RSI (short/long) separately. We assume that the the log RTs in each series, after a preliminary detrending, follow a mixture of Gaussian and ex-Gaussian distributions. The specified model accounts for sequential dependencies and allows for the possible occurrence of short and long responses.

3.1 The Ex-Gaussian distribution

Suppose that X is a normally distributed random variable, with mean μ and variance σ^2 , and let Y denote an exponential random variable with rate parameter λ . Assume that X and Y are independent of one another. Then the sum $W = X + Y$ is said to have an *ex-Gaussian* distribution with parameters μ , σ^2 and λ . The density for W is thus a convolution of an exponential and a Gaussian density. We can think of W as arising either from a shifted exponential distribution with a normally distributed shift, or from a normal distribution with an exponentially distributed mean. Also, since X is supported on the entire real line, the distribution of W puts positive mass on the negative values. Thus, the ex-Gaussian model is not by default a model for positively valued data, unless some transformation of the data is entertained or the weight given to negative values can be considered negligible.

The original motivation for the ex-Gaussian distribution is due to [Hohle \(1965\)](#). In his paper he reasoned that the Gaussian component models the peripheral processing time, and the exponential component models the cognitive processing time. This is no longer an accepted theory in the response time literature. However, the ex-Gaussian model sees wide use as a description of RT data because it is very flexible and can adequately describe extremely large observations ([Andrews and Heathcote 2001](#); [Gottlob 2004](#); [Heathcote, Popiel, and Mewhort 1991](#); [Penner-Wilger, Leth-Steensen, and LeFevre 2002](#); [Ratcliff and Murdock 1976](#); [Reber, Alvarez, and Squire 1997](#)). In the next section we extend the ex-Gaussian model to allow for extreme observations in the negative direction as well.

3.2 The Bayesian mixture model

Let $\{W_t : t = 1, \dots, 1024\}$ denote the detrended log-transformed values of the RT sequence for a given subject, with either short or long RSIs. The prototype model for W_t is then

$$W_t = \begin{cases} X_t & \text{with prob. } p_X \\ X_t + Y_t & \text{with prob. } p_Y \\ X_t - Z_t & \text{with prob. } p_Z. \end{cases} \quad (1)$$

Here $\{X_t\}$ is a hidden AR(1) process defined by

$$\begin{aligned} X_1 - \mu &= U_1, \\ X_t - \mu &= \phi(X_{t-1} - \mu) + U_t, \quad t = 2, \dots, 1024, \end{aligned}$$

where U_1 is a $N(0, \sigma_1^2)$ random variable and $\{U_t\}_{t=2}^{1024}$ is an *IID* sequence of $N(0, \sigma^2)$ random variables. This AR(1) process models the serial dependence in the response times. The possible occurrence of extreme observations is modeled by the two independent sequences of exponentially-distributed random variables $\{Y_t\}$ and $\{Z_t\}$, with means $1/\lambda_Y$ and $1/\lambda_Z$, respectively. We complete the specification of the Bayesian model by assuming that $\phi \sim N(\mu_\phi, \sigma_\phi^2)$, $\mu \sim N(\eta, \sigma_\mu^2)$, $1/\sigma_1^2 \sim \text{Gamma}(\alpha_{\sigma_1^2}, \beta_{\sigma_1^2})$, $1/\sigma^2 \sim \text{Gamma}(\alpha_{\sigma^2}, \beta_{\sigma^2})$, $\lambda_Y \sim \text{Gamma}(\alpha_Y, \beta_Y)$, $\lambda_Z \sim \text{Gamma}(\alpha_Z, \beta_Z)$, $\mathbf{p} = (p_X, p_Y, p_Z) \sim \text{Dirichlet}(\gamma_X, \gamma_Y, \gamma_Z)$, all independently. The Gamma distributions is parameterized so that a $\text{Gamma}(\alpha, \beta)$ random variable has mean α/λ . The Dirichlet distribution has density proportional to $p_X^{\gamma_X-1} p_Y^{\gamma_Y-1} p_Z^{\gamma_Z-1}$ over the unit simplex. All the constants appearing in these specifications are assumed to be known.

4 Results

We used the program WinBUGS (Spiegelhalter, Thomas, Best, and Lunn 2003) (<http://www.mrc-bsu.cam.ac.uk/bugs/>) to fit the models by Markov Chain Monte Carlo (MCMC) simulation, postprocessing the output in R. We used the convenient interface provided by the package RBUGS for communication between WinBUGS and R.

We fixed the constants in the prior specification as follows. In the distribution of the autoregressive parameter ϕ we set $\mu_\phi = 0$ and $\sigma_\phi^2 = 1$. In the distributions for the error variances, σ_1^2 and σ^2 , we

set $\alpha_{\sigma_1^2} = 3$, $\beta_{\sigma_1^2} = 0.3$, $\alpha_{\sigma^2} = 0.1$, and $\beta_{\sigma^2} = 0.1$. The mean of ϕ was set to zero, a value reflecting serial independence. The model does not enforce conditional stationarity of the autoregressive component of the mixture. (Aside from any modeling considerations, this choice has the advantage of simplifying considerably the model fitting steps in WinBUGS.) However, setting $\sigma_\phi^2 = 1$ ensures that the interval $(-1, 1)$, corresponding to values of ϕ yielding stationary AR(1) processes, receives considerable prior weight. Since we do not enforce stationarity, we untied the variance of the first error term, U_1 , from the variance of the subsequent error terms, U_2, \dots, U_{1024} . The specification of the constants in the gamma distribution for σ^2 is intended to make the data drive the posterior inferences. The constants in the gamma distribution of σ_1^2 were set so as to yield a more informative distribution, because a single error term, in and of itself, cannot provide much information. (A separate sensitivity investigation, not reported in this article, using a much less informative prior showed that the inferences are quite stable, regardless of what the specification is.) The constants for the gamma distributions of λ_Y and λ_Z were all set equal to 0.001, yielding rather diffuse priors and the parameters of the Dirichlet distribution for $\mathbf{p} = (p_X, p_Y, p_Z)$ were all set equal to 1, yielding a uniform distribution on the unit simplex.

For each subject and either the long or short RSIs experiment, we carried out a burn-in of 12,500 iterations, discarded the output, and then ran each chain for a further 125,000 iterations. We thinned out the chains by subsampling every 25th iteration, thus retaining the draws from 5,000 equally spaced iterations for analysis. We assessed convergence of the model using trace plots and autocorrelation plots.

Figure 5 displays posterior summaries of the parameters μ , ϕ , and σ which characterize the process modeling the serial dependence in the data. In each of the three panels, the posterior medians for one of the three parameters are plotted against subject index for the short RSI experiments (circles) and for the long RSI experiments (boxes). The lines extend from the 2.5 to the 97.5 empirical percentiles of the posterior samples.

One interesting feature of the results is that, within subjects, there appears to be a substantial amount of agreement between the posterior parameter estimates for the long and short RSI experiments. This suggests that the nature of the serial correlation exhibited by the RTs is more closely related to subject specific characteristics than to the experimental manipulations. It can also be

Variables	RSI	Subject					
		1	2	3	4	5	6
μ and ϕ	Short	-0.15	0.02	-0.09	-0.30	-0.07	-0.32
μ and ϕ	Long	-0.04	-0.01	-0.26	-0.34	0.01	-0.02
μ and μ_y	Short	0.48	0.53	0.22	0.52	0.29	0.20
μ and μ_y	Long	0.33	0.54	-0.09	0.51	0.21	0.30
μ and μ_z	Short	-0.02	-0.01	-0.10	-0.01	-0.09	-0.10
μ and μ_z	Long	-0.13	0.03	0.03	0.05	-0.14	-0.03

Table 1: Spearman’s rank correlations between the posterior draws of μ and ϕ , μ and μ_Y , and μ and μ_Z , for each subject by RSI combination. Bold numbers indicate those correlations that are significantly different from zero. (These tests are not adjusted for multiplicity of comparisons).

seen that the posterior distributions of the standard deviation parameters for the innovations of the AR(1) model, σ , are similar for the majority of the subjects. The amount of autocorrelation in the model, as judged by the posterior distributions of ϕ , ranges from weak for subjects 1, 2 and 5 to stronger for subjects 3, 4, and 6. These differing autocorrelations give some indication that carry-over effects are larger for some subjects than for others.

Figure 6 displays analogous posterior summaries for the parameters that characterize the mixture components describing the occurrence of extreme observations ($p_Y, p_Z, \mu_Y = 1/\lambda_Y$, and $\mu_Z = 1/\lambda_Z$). As with the time series parameters, there is good agreement within subjects for long and short RSI experiments. There is, however, a great deal of between subject variability for the probability of a fast extreme component (p_y), with estimates that range from close to zero to close to one. The estimated probabilities of a slow extreme component are much less variable and typically smaller in size, ranging between zero and 0.15 for all subjects. The posterior estimates of μ_Y suggest that, within subject, the expected values of the fast components are generally larger for the long RSI experiments than for the short RSI experiments. The slow extreme components, while occurring less frequently, have expected values that are typically larger than those of the fast extreme components.

We also examined the bivariate associations between the mean of the AR(1) component, μ , and the AR(1) coefficient, ϕ , and between μ and the means of the slow and fast components (μ_Y and μ_Z). We used Spearman’s rank correlation as a measure of association, thus avoiding the need to make

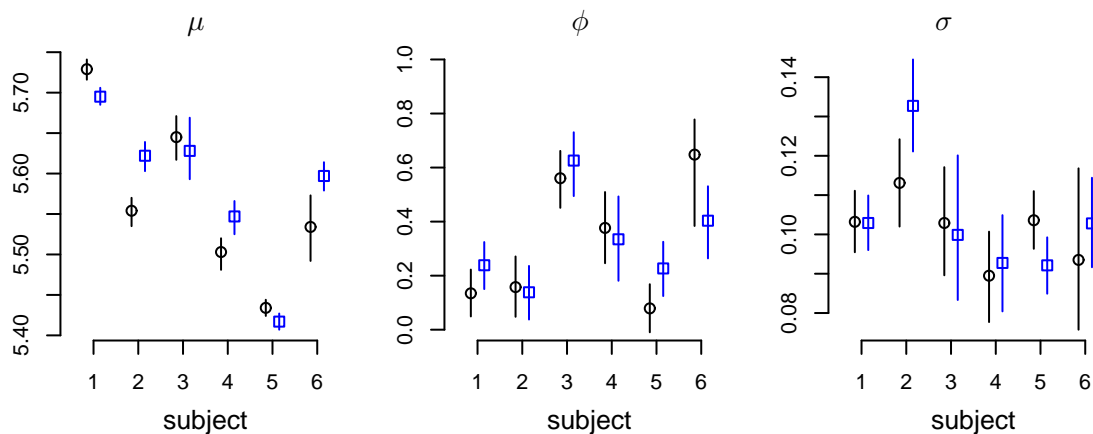


Figure 5: Posterior summaries of the parameters of the AR mixture components for each of the six subjects. The circles denote the posterior medians for the short RSI experiments and the boxes denote the posterior medians for the long RSI experiments. For both experiments the vertical lines connect the 2.5 and 97.5 percentiles of the empirical posterior distributions.

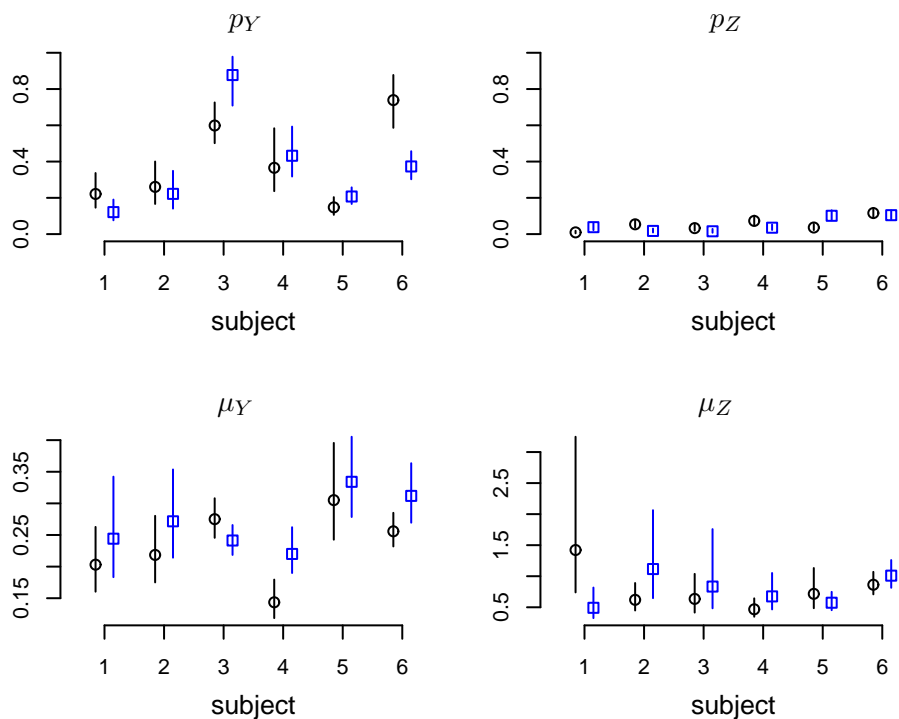


Figure 6: Posterior summaries of the parameters of the mixture components describing the extreme observations for each of the six subjects. The circles denote the posterior medians for the short RSI experiments and the boxes denote the posterior medians for the long RSI experiments. For both experiments the vertical lines connect the 2.5 and 97.5 percentiles of the empirical posterior distributions.

distributional assumptions about the posterior draws.

The calculated correlations are reported in Table 1 for each of the six subjects at the two RSIs. Bold type is used, mainly as an exploratory data analysis device, to denote those correlations that are significantly different from zero (without multiplicity adjustment). Although there is a great deal of variability across subjects, and to some extent across long/short RSIs, we observe a general tendency for the correlation between the mean and autoregressive parameter of the AR(1) component to be negative. This would reflect a direct link between a tendency to respond more slowly and the weakening of carry over effects from one trial to the next. The mean of the AR(1) component is, with one exception, positively correlated with the mean of the fast component, and, generally, negatively correlated with the mean of the slow component, but interpretation of these association is far from straightforward because the posterior probabilities of the various components also play a role.

The mixture model of Equation (1) specifies that, at any given time t , the detrended log RT process can be in one of three states: state X denoting the fact that $W_t = X_t$, state Y denoting the fact that $W_t = X_t + Y_t$, and state Z denoting the fact that $W_t = X_t - Z_t$. The model makes an implicit independence assumption about the states in which the process can be at different times, an assumption whose validity we wanted to investigate. In particular we wanted to determine if there is evidence of the existence of a first order Markovian association in the transitions between the three states. In the implementation of the MCMC sampling we used 0-1 indicator variables to code the state of the mixture at any given time t . Thus, for each subject by RSI experimental condition, we were able to estimate easily the first order transition probabilities by tabulating the frequencies with which the various transitions occurred in the MCMC simulations. The estimated probabilities are plotted against subject index in Figure 7, using circles for short RSIs and squares for long RSIs. The panels are arranged as in a transition probability matrix, with row 1 summarizing the transition probabilities from state X to states X , Y , and Z , row 2 summarizing the transition probabilities from state Y to states X , Y , and Z , and row 3 summarizing the transition probabilities from state Z to states X , Y , and Z . The figure shows no evidence of dependence of the transition probabilities into a given state based on the state of origin and reveals that there is close agreement between the estimated values of these conditional probabilities and the corresponding unconditional estimates presented in Figure 6. This finding rules out the need for a more complicated model.

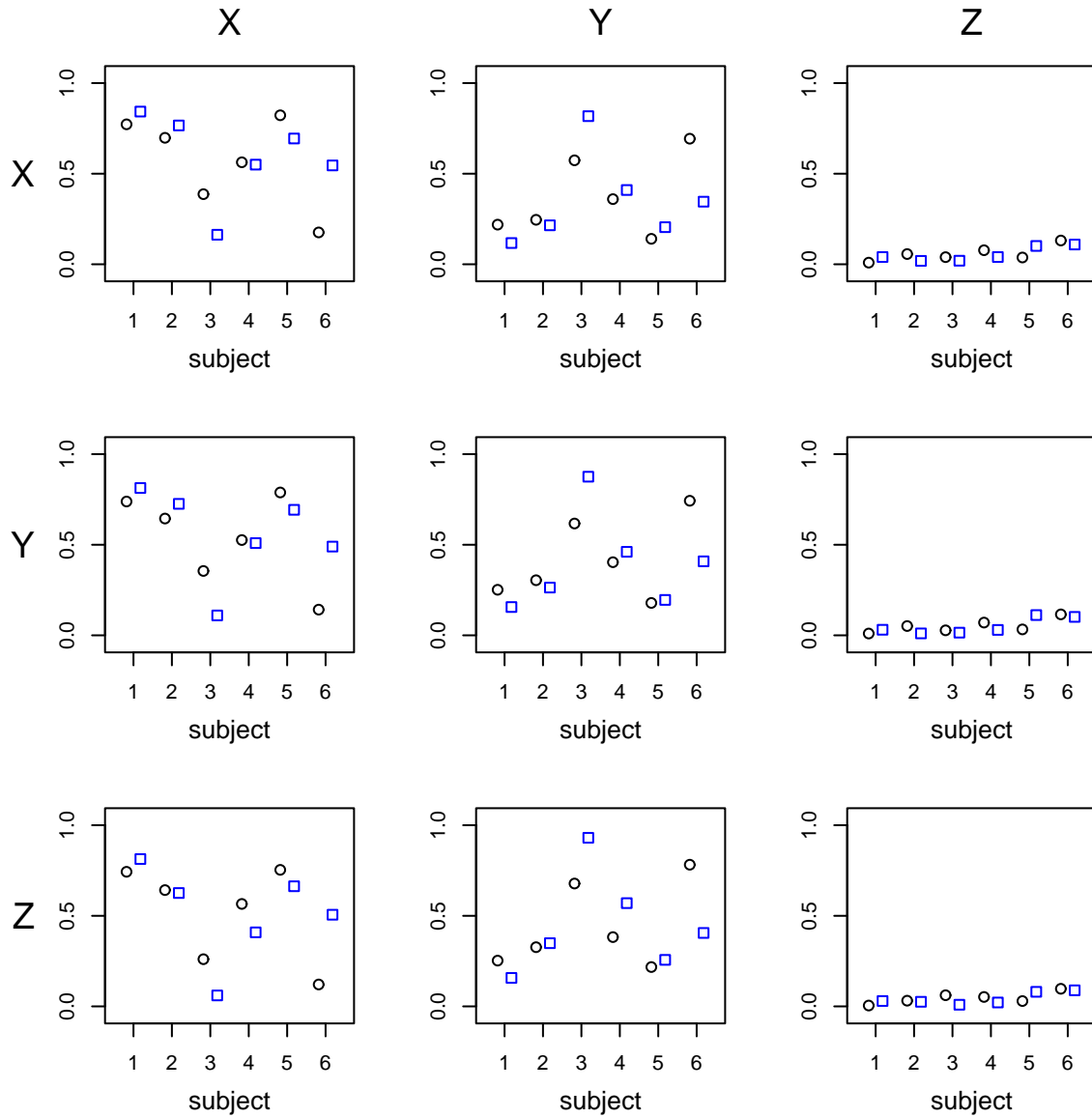


Figure 7: Estimated conditional probabilities of one-step transitions between the different states of the mixture. (X stands for $W_t = X_t$, Y for $W_t = X_t + Y_t$ and Z for $W_t = X_t - Z_t$.) The circles denote the posterior medians for the short RSI experiments, and the boxes denote the posterior medians for the long RSI experiments.

5 Model validation via Monte Carlo simulation

In this section we present the result of a simulation experiment to evaluate the performance of the prototype mixture model. We do so by assessing whether the behavior of RT series simulated from the posterior predictive distributions of the fitted models is compatible with the statistical properties of the original RT sequences. In particular, we investigate the agreement of marginal properties and first order serial dependence properties.

The basic steps of the simulation are as follows. Consider the RT sequence of a given subject with either short or long RSI and the corresponding 5,000 sets of parameter draws from the posterior distribution of the prototype model that were used to compute the estimates presented in Section 4. Conditional on each of these sets of parameter draws, we simulated a sequence of detrended log RT values of length 1024 from the model specified in Section 3.2. (These simulated log RT sequences are thus realizations from the posterior predictive distribution for that subject and experimental condition (long or short RSI).) Finally, we added the trend estimated from the original logged data and transformed the resulting sequence back to the original scale. Repeating this operation for each set of parameter draws from the posterior distribution gave us a collection of 5,000 simulated posterior predictive RT sequences for each subject by experimental condition combination.

To investigate agreement between marginal distributions, we calculated the 2.5, 25, 50, 75 and 97.5 percentiles of each simulated RT sequence. The first five panels of Figure 8 describe the empirical distributions of the simulated percentiles for each of the the (six) subject by (two) RSI combinations. For example, the top left panel summarizes the distributions of the 2.5 percentiles. For each subject, the circle denotes the median of the 2.5 empirical percentiles of the 5,000 simulated sequences for the short RSI, and the box denotes the median of the 2.5 empirical percentiles of the 5,000 simulated sequences for the long RSI. In all cases, the vertical lines show 95% equal-tailed intervals for the 2.5 percentiles of the simulated realizations. The crosses identify the observed values of the corresponding percentiles for the actual data. There is an overall evidence of a very good marginal agreement between the data and synthetic RT sequences generated from the posterior predictive distributions, across all the percentiles tabulated, for all subjects and RSI condition, with all observed percentiles falling within the corresponding 95% intervals.

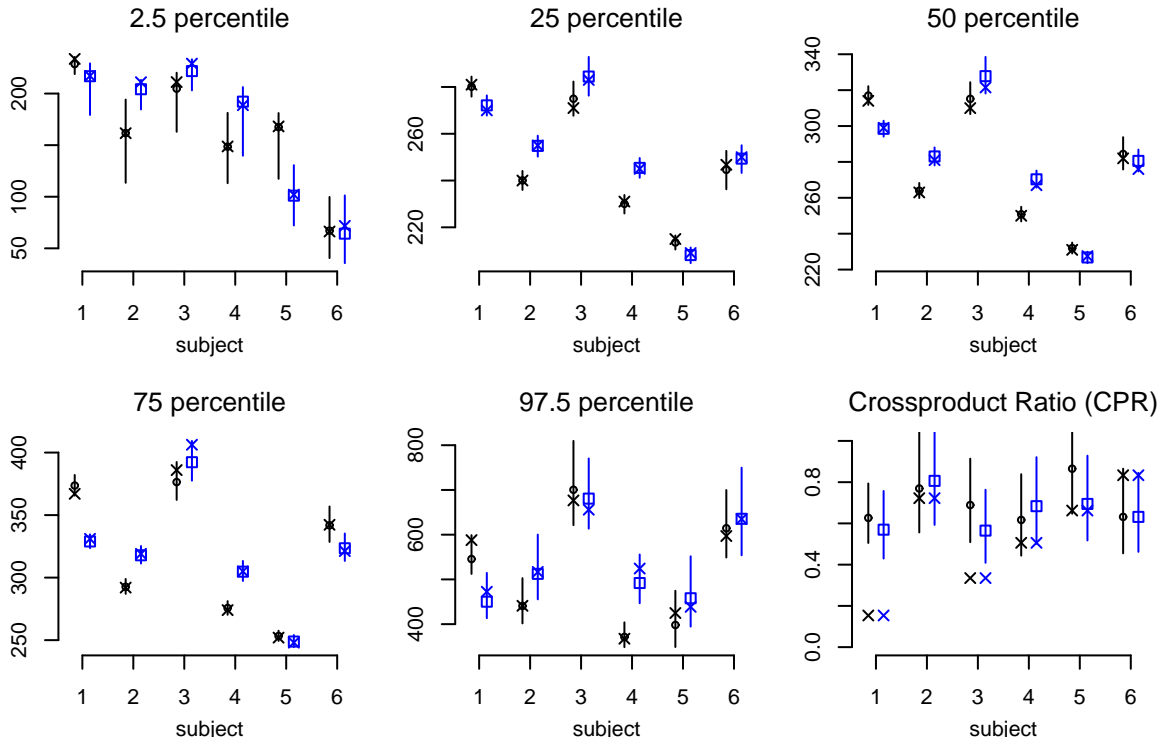


Figure 8: Marginal (percentiles) and first-order dependence (CPR) summaries of the posterior predictive distributions for each subject by RSI combination. The circles denote the posterior medians of the displayed quantities for the short RSI experiments, and the boxes denote the posterior medians for the long RSI experiments. For both experiments the vertical lines denote 95% equal-tailed intervals. The crosses identify the observed values of the corresponding quantities for the actual data.

Next we assess the adequacy of the first-order dependence properties of the simulated sequences. Because the models are not multivariate Gaussian, we do not rely on lag-one autocovariances for this evaluation but we calculate cross-product ratios for the lagged series instead. Given a time series of logged RT values, $\{w_t\}$, let M denote the median value of the series. Consider plotting w_t versus w_{t+1} and counting the number of points that fall in each of the four quadrants centered at the point of coordinates (M, M) . Specifically, let n_{11} denote the number of points in the lower left quadrant (i.e., the number of values for which $w_t < M$ and $w_{t+1} < M$), let n_{12} denote the number of points in the upper left quadrant (i.e., the number of values for which $w_t < M$ and $w_{t+1} \geq M$), let n_{21} denote the number of points in the lower right quadrant (i.e., the number of values for which

$w_t \geq M$ and $w_{t+1} < M$), and let n_{22} denote the number of points in the upper right quadrant (i.e., the number of values for which $w_t \geq M$ and $w_{t+1} \geq M$). The cross-product ratio is then calculated according to the formula:

$$\text{CPR} = \frac{n_{12} n_{21}}{n_{11} n_{22}}. \quad (2)$$

Stronger direct association between successive trials would result in more points falling in the upper right and lower left quadrant and produce lower CPR values.

The bottom right panel of Figure 8 shows the median and 95% intervals for the CPRs calculated from the posterior predictive simulated sequences for each subject by RSI combination. The crosses identify the corresponding CPR values computed from the original data. For most of the subjects there is satisfactory agreement between the observed and the simulated CPR values. Subject 1 and, to a lesser extent, subject 3 are exceptions, having observed CPR values that are lower than the CPR values predicted by the models.

6 Discussion

In the beginning of this article, we discussed the typical assumptions that are made about RT data to simplify the modeling enterprise, most importantly the assumption that sequences of observations are independent. We also discussed the models of RT distributions used by psychologists for descriptive purposes. In many cases, this approach has been fruitfully applied to RT data and resulted in elegant and powerful explanations of human choice behavior (Ratcliff 1978). As more theoretical interest develops in dependent sequences of RTs, we currently have no equivalent modeling tools for exploring effects that unfold over trials.

In this article we have introduced a mixture time series models for RT data that is simple to understand and captures a number of typical features of RT data, including serial dependencies and both fast and slow extreme observations. This comprehensive modeling approach distinguishes our work from more *ad hoc*, traditional treatments of RT data, in which, for example, extreme observations might be trimmed before an analysis is carried out. The model can be conveniently fit by MCMC methods, using publicly available software (WinBUGS, R, and the RBUGS package), with no need for specialized programming.

Using Monte Carlo simulations, we were able to verify that synthetic data generated from the posterior predictive distributions of the fitted models behaved, to a large extent, like the observed data with regard to both their marginal distributional properties and the nature of their first order serial dependencies. In this respect, our proposed approach is preferable to methods that deal with extreme observations by trimming them or by otherwise modifying their nature, even in those cases in which such methods are model based. In fact, these sorts of data manipulations performed prior to estimation will typically result in estimated models that cannot generate data that capture some of the important features of RT data. If the extreme observations are eliminated from the analysis, the information about the nature of the tail behavior of the marginal distribution of the data will be lost. Furthermore, such manipulations can also have an adverse impact on the estimation of the serial dependencies of the data.

In practice, this model can be used as a starting point for the systematic investigation of effects of experimental factors that influence both the marginal RT distributions and dependencies in a sequence of trials. For example, in an ongoing extension of this line of work, we are modeling trends directly within a hierarchical Bayes structure. More broadly, the hierarchical approach gives a coherent way to specify regression structures for the parameters that depend on the varying experimental conditions (fixed effects) and that include subject specific terms (random effects reflecting individual differences), ensuring pooling of information across RT sequences. In this way, this model could be used to test hypothesis about experimental effects in any RT experiment without neglecting the effects of serial dependencies.

We have begun with the simplest possible structure, but the basic prototype model could be enhanced in several ways. The AR(1) mixture component offers a simple device to model serial dependencies. This device appears to perform satisfactorily for the data that we analyzed, but more complex time series models (for example higher order ARMA models) could be entertained. More flexible parametric families of distributions (Gamma, Weibull, etc.) could be employed to model the two mixture components dealing with extreme observations. The data that we analyzed showed no evidence of a dependence of the probability of transitioning into a given state of the mixture at time $t + 1$ (AR, AR + fast observation, AR - slow observation) on the state of the mixture at time t . However, if needed for other data, the introduction of a first order Markovian structure governing the evolution of the state of the mixture over time would be fairly straightforward. Of course, all of

these changes would come at the expense of an increased complexity of the model.

In this article we did not model trends. Rather, we executed a preliminary, nonparametric detrending of the data. We were careful not to oversmooth the data, and to use a procedure that would not disturb the underlying dependence structure of the data. Undetected trend can at times be mistaken for serial correlation and the distinction between one and the other is often blurry. Researchers should be concerned that failures to adequately deal with trend, and to distinguish between trend and serial correlation, may reduce the impact of findings of long-term serial correlations ($1/f$ noise) in RT data. These failures also make it difficult, if not impossible, to systematically evaluate dependencies and trends across subjects and conditions. The fact that we applied a precisely prespecified detrending procedure to all the RT sequences that we analyzed ensures that the inferential conclusions are comparable across subjects and experimental conditions.

References

- Andrews, S. and A. Heathcote (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory and Cognition* 27, 514–544.
- Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data* (3rd ed.). New York: John Wiley & Sons.
- Belin, T. R. and D. B. Rubin (1995). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine* 14, 747–768.
- Farrell, S., E.-J. Wagenmakers, and R. Ratcliff (2005). ARFIMA time series modeling of serial correlations in human performance. Submitted for publication.
- Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science* 8, 296–301.
- Gilden, D. L. (2001). Cognitive emissions of $1/f$ noise. *Psychological Review* 108, 33–56.
- Gottlob, L. R. (2004). Location cuing and response time distributions in visual attention. *Perception and Psychophysics* 66, 1293–1302.
- Heathcote, A., S. Brown, and D. J. K. Mewhort (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review* 7, 185–207.
- Heathcote, A., S. J. Popiel, and D. J. Mewhort (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin* 109, 340–347.
- Hohle, R. H. (1965). Inferred components of reaction time as a function of foreperiod duration. *Journal of Experimental Psychology* 69, 382–386.
- Jones, M., B. C. Love, and W. T. Maddox (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, 316–332.
- Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York: Oxford University Press.
- Meeter, M. and C. N. L. Olivers (2006). Intertrial priming stemming from ambiguity: A new account of priming in visual search. *Visual Cognition* 13, 202–222.

- Peña, D., D. Pena, G. C. Tiao, and R. S. Tsay (2001). *A Course in Time Series Analysis*. John Wiley & Sons.
- Penner-Wilger, M., C. Leth-Steensen, and J.-A. LeFevre (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory and Cognition* 30, 1160–1167.
- Peruggia, M. (2005). Bayesian model diagnostics based on artificial autoregressive errors. Technical Report 758, The Ohio State University.
- Peruggia, P., T. Van Zandt, and M. Chen (2002). Was it a car or a cat I saw? An analysis of response times for word recognition. In *Case Studies in Bayesian Statistics*, Volume 6, pp. 319–334. New York: Springer-Verlag.
- Pressing, J. and G. Jolley-Rogers (1997). Spectral properties of human cognition and skill. *Biological Cybernetics* 76, 339–347.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review* 85, 59–108.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin* 114, 510–532.
- Ratcliff, R. and B. B. Murdock, Jr. (1976). Retrieval processes in recognition memory. *Psychological Review* 83, 190–214.
- Ratcliff, R. and P. L. Smith (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review* 111, 333–367.
- Reber, P. J., P. Alvarez, and L. R. Squire (1997). Reaction time distributions across normal forgetting: Searching for markers of memory. *Learning and Memory* 4, 284–290.
- Rouder, J., D. Sun, P. Speckman, J. Lu, and D. Zhou (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika* 68, 589–606.
- Rouder, J. N., J. Lu, P. Speckman, D. Sun, and Y. Jiang (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review* 12, 195–223.
- Spiegelhalter, D. J., A. Thomas, N. G. Best, and D. Lunn (2003). *WinBUGS User Manual, Version 1.4*. Cambridge, UK: MRC Biostatistics Unit.

- Stewart, N., G. D. A. Brown, and N. Chater (2005). Absolute identification by relative judgment. *Psychological Review* 112, 881–911.
- Thornton, T. L. and D. L. Gilden (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin and Review* 12, 409–441.
- Ulrich, R. and J. Miller (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General* 123, 34–80.
- Van Orden, G. C., J. G. Holden, and M. T. Turvey (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General* 132, 331–350.
- Van Selst, M. and P. Jolicoeur (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology* 47, 631–650.
- Wagenmakers, E.-J., S. Farrell, and R. Ratcliff (2004). Estimation and interpretation of $1/f$ noise in human cognition. *Psychonomic Bulletin & Review*.
- Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* 93, 341–348.