

# Closed-Form Estimation of Multiple Change-Point Models

Greg Jensen  
Columbia University

## Abstract

Identifying discontinuities (or *change-points*) in otherwise stationary time series is a powerful analytic tool. This paper outlines a general strategy for identifying an unknown number of change-points using elementary principles of Bayesian statistics. Using a strategy of *binary partitioning by marginal likelihood*, a time series is recursively subdivided on the basis of whether additional subdivisions (and thus increased model complexity) yield a justified improvement in the marginal model likelihood. When this approach is combined with the use of conjugate priors, it yields the *Conjugate Partitioned Recursion* (CPR) algorithm, which identifies change-points without computationally intensive numerical integration. Using the CPR algorithm, methods are described for specifying change-point models drawn from a host of familiar distributions, both discrete (binomial, geometric, Poisson) and continuous (exponential, Gaussian, uniform, and multiple linear regression), as well as multivariate distributions (multinomial, multivariate normal, and multivariate linear regression). Methods by which the CPR algorithm could be extended or modified are discussed, and several detailed applications to data published in psychology and biomedical engineering are described.

*Keywords:* bayesian statistics, change-point analysis, marginal likelihood, model selection, time series analysis

## Introduction

1

2 The analysis of time series data is essential to most scientific disciplines. Given the  
3 ability to measure the behavior of an agent, we often wish to know how the measured be-  
4 havior changes over time. This is true whether that agent is a single neuron firing action  
5 potentials, a human participant making choices, or a central bank reporting GDP. Some-  
6 times, conditions do not change, and observations appear consistent (and display consistent  
7 variability); in other cases, change happens gradually and continuously, in a manner befit-  
8 ting a fitted line or curve. Modeling phenomena in these terms is the bedrock of empirical  
9 statistics.

---

Correspondence should be directed to Greg Jensen by email at [greg.guichard.jensen@gmail.com](mailto:greg.guichard.jensen@gmail.com). Supported by NIMH grant (5-R01-MH068073-10) awarded to Peter Balsam.

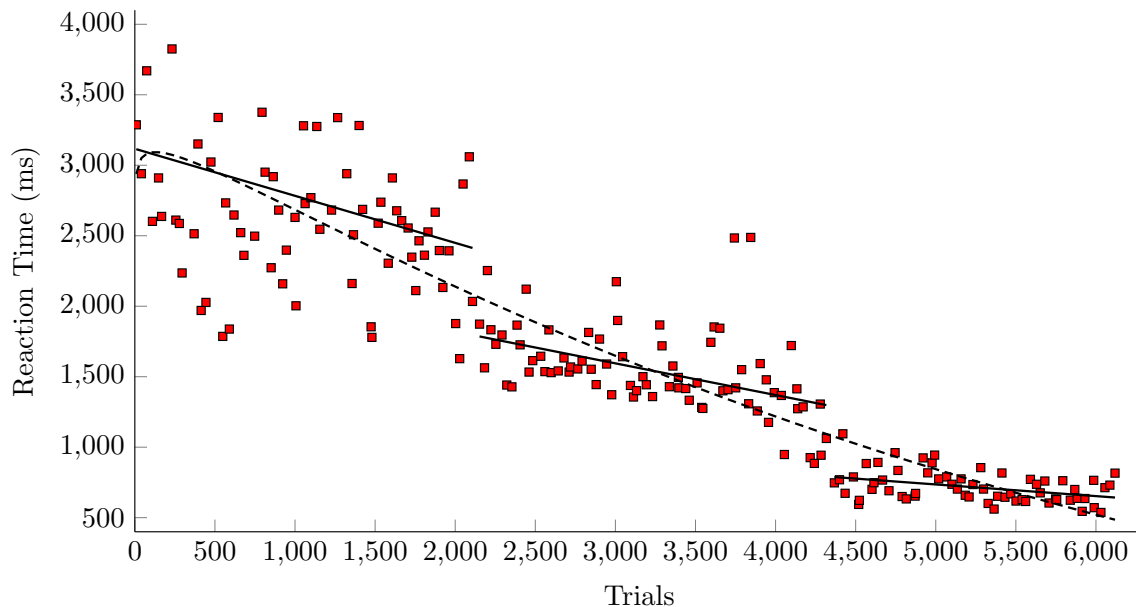


Figure 1. Reaction times from a single subject learning a psychophysical task, originally reported by Palmeri (1997). The dashed line corresponds to a four-parameter “learning curve,” reported by Heathcote et al. (2000), while the solid lines interpret the same data as approximately linear, with two change-points.

1 Time series often contain abrupt shifts. For example, most learning, when examined  
 2 on a trial-by-trial basis, is characterized by abrupt changes in behavior, rather than showing  
 3 gradual progress. Learning curves are often an averaging artifact, resulting from pooling  
 4 across trials (or across subjects). Building robust descriptive models of individual behavior  
 5 over time depends on identifying abrupt discontinuities, or *change-points*.

6 For example, consider the data in Figure 1, collected by Palmeri (1997). It depicts  
 7 reaction times from a single subject in a specific condition, as part of a cognitive study of  
 8 psychophysical learning. In a re-analysis of these and other data, Heathcote et al. (2000)  
 9 argued strongly in favor of a four-parameter pseudo-exponential function, whose best-fitting  
 10 form is represented by the dashed line. This curve is problematic for two reasons. Firstly,  
 11 a visual examination of the data suggests that two discontinuities (around trials 2,200  
 12 and 4,400) mark sudden drops in reaction times, which may be of experimental interest  
 13 but which the model is in principle incapable of identifying. Secondly, although the curve  
 14 follows the *overall* shape of the data reasonably well, it includes a slight reversal in direction  
 15 near the trial one, making the peculiar prediction that this participant is initially slowing  
 16 down.

17 If, on the other hand, the data are simply divided at the points algorithmically iden-  
 18 tified as change-points, the resulting segments appear comfortably linear (as depicted by  
 19 the solid lines). This is also reflected by the goodness of fit: the sum of squared residu-  
 20 als is smaller for each of the identified segments. Breaking the data into segments is not  
 21 merely convenient from a curve-fitting perspective: The discontinuities likely correspond to  
 22 theoretically important ‘eureka’ moments in the participant’s learning. These data will be

1 considered in more detail in a later section, but for now, they demonstrate an important is-  
2 sue in time-series analysis: Because different participants learn at different rates, and have  
3 insights at different times, group averaging conceals these moments and typically paints  
4 the erroneous picture of learning being gradual and continuous. The misleading nature of  
5 ‘learning curves’ is described in detail by Gallistel et al. (2004).

6 Evidence for change-points can be found in time series at every level of interest.  
7 In the analysis of high-level systems, ‘structural changes’ have long been of interest to  
8 economics (Hansen, 2001) and other social sciences that focus on historical data (Western  
9 & Kleykamp, 2004). However, the importance of identifying change-points on shorter time-  
10 scales is increasingly evident. In neuroscience, change-point analyses have been proposed  
11 for electrophysiology (Bélisle et al., 1998) and state-related fMRI (Lindquist et al., 2007).  
12 In applied clinical and public health domains, a wide variety of ‘turning points’ are crucial  
13 to understanding long-term outcomes (Cohen, 2008).

14 It is not appropriate, however, to pick out change-points by intuition alone, any  
15 more than it is to judge statistical significance by eyeball. Principled statistical tests are  
16 necessary to determine whether data require segmentation, where the change-points should  
17 be positioned, and how many divisions to make. Techniques of varying complexity have been  
18 developed for identifying change-points in a time series (Chen & Gupta, 2011), but change-  
19 point analysis as a domain is characterized both by its specialized focus and its technical  
20 complexity. Most published papers focus on a single facet of this general topic, doing so in  
21 great depth. As a result, change-point analysis is not typically presented in general terms,  
22 and digesting the change-point literature can be daunting for applied researchers.

23 This paper first provides a basic foundation for understanding the Bayesian logic on  
24 which it depends. The, it describes a straightforward approach called the *binary partition*  
25 *by marginal model likelihood* strategy, which combines divide-and-conquer design with el-  
26 ementary Bayesian reasoning. This approach is implemented as the *Conjugate Partitioned*  
27 *Recursion* (or “CPR”) algorithm, which relies on conjugate prior probability distributions.  
28 The CPR algorithm uses closed-form arithmetic to identify change-points rapidly using a  
29 minimal number of arbitrary parameters. Because it relies on neither bootstrapping nor  
30 numeric integration, it is able to process large, multivariate datasets rapidly.

## 31 A Brief Review

32 In order to use any statistical tool effectively, it is essential that the analyst understand  
33 how that tool works. Unfortunately, many readers will have only passing familiarity with  
34 some of the mathematical machinery that makes the CPR algorithm effective. In the interest  
35 of making the operations of the algorithm as clear as possible to as many readers as possible,  
36 the following review of core concepts is presented. While all readers are encouraged to  
37 reacquaint themselves with these topics, those who wish to proceed directly to the discussion  
38 of the CPR algorithm itself will find it beginning in the section entitled ‘The Conjugate  
39 Partitioned Recursion Algorithm.’

## 40 Bayes’ Theorem: A Machine For Priors

41 To answer the question, “Should these data be partitioned by a change-point?” the  
42 partitioning algorithm begins with Bayes’ Theorem. Given a prior assumption that the data

1 should conform to a model  $M$ , a distribution is specified describing the probability that the  
 2 data arise from parameters  $\theta$ . The purpose of the theorem is to update this probability  
 3 distribution over the parameters  $\theta$  as a result of having made the observations  $x$ . This  
 4 operation is commonly represented as follows:

$$\Pr(\theta|x, M) = \frac{f(x|\theta, M) \Pr(\theta, M)}{m(x, M)}$$

where (1)

$$m(x, M) = \int_{\theta} f(x|\theta, M) \Pr(\theta, M) d\theta$$

5 The *prior probability distribution* is denoted by  $\Pr(\theta, M)$ , and represents the distribution  
 6 associated with the parameters  $\theta$  at the outset, given model  $M$ . The function  $f(x|\theta, M)$   
 7 represents the *likelihood* of obtaining the observations  $x$  given model  $M$  with parameters  $\theta$ .  
 8 Convoluting these distributions does not typically result in a true probability distribution  
 9 (that is, the integral rarely equals 1.0), so the result must be divided by a *normalizing con-*  
 10 *stant*, represented by  $m(x, M)$ , in order to obtain the *posterior probability distribution*,  
 11 denoted by  $\Pr(\theta|x, M)$ . This posterior distribution represents the updated probability after  
 12 the observations are considered. More details on power of this approach are provided by  
 13 Chib (1998) and by Gelman et al. (2003).

14 Practical application of Bayes' Theorem requires a formal specification of prior as-  
 15 sumptions. The shape of the prior distribution  $\Pr(\theta, M)$  represents the existing evidence  
 16 available to the analyst. Even when no empirical observations are available, this prior must  
 17 be specified. A typical approach is to imagine a distribution derived from "hypothetical"  
 18 observations. For example, if  $x$  consists of a series of die rolls, resulting in the numbers one  
 19 through six, one might ask the question, "How likely is this die to roll a six?" If the die  
 20 is fair<sup>1</sup>, it should show six with a probability of  $\frac{1}{6}$ , so we might imagine five hypothetical  
 21 failures and one hypothetical success. These hypothetical successes and failures are our  
 22 *prior hyperparameters*: hypothetical failures are denoted by  $\alpha_0$  and hypothetical successes  
 23 by  $\alpha_1$ . Thus, we might begin by assuming  $\alpha_0 = 5$  and  $\alpha_1 = 1$ .

24 This is an example of a 'subjective' prior, because the values for  $\alpha_0$  and  $\alpha_1$  are left  
 25 to the analyst's discretion. Given high confidence in a fair die, a stronger prior might use  
 26  $\alpha_0 = 25$  and  $\alpha_1 = 5$ , the equivalent of thirty hypothetical observations (which, again, are  
 27 assumed before the die is ever actually rolled). Alternatively, a more agnostic (or "weak")  
 28 prior might use  $\alpha_0 = 1$  and  $\alpha_1 = 1$ . The weaker the prior, the less its assumptions influence  
 29 the subsequent calculation. Because  $\Pr(\theta, M)$  is a probability distribution, we do not have  
 30 a 'known' value for  $\theta$ . Instead, the relative weight of the possible values of  $\theta$  are described  
 31 by a beta distribution whose shape is defined by the prior hyperparameters, as depicted  
 32 in Figure 2 (Left). As implied by the figure, the beta distribution supports the interval  
 33 between 0 to 1, and represents our uncertainty about a simple probability. The thick red  
 34 line represents the 'weak' prior, and it assigns even odds to any probability over the interval.  
 35 The two stronger priors have converged somewhat towards the expected probability of a  
 36 fair die, and these reflect greater confidence that  $p(\text{six}) = \frac{1}{6}$ .

<sup>1</sup>This example reduces the die roll to the binary pairing "not 6" vs. "6" in the interest of simplicity. To ask whether the die is fair *in general*, the example may be extended to a multinomial case that models each of the six possible outcomes.

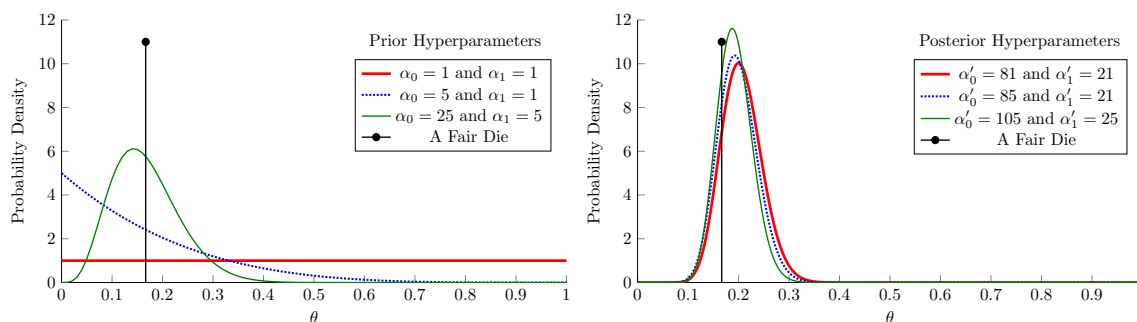


Figure 2. Prior hyperparameters  $\alpha_0$  and  $\alpha_1$  (left) and posterior hyperparameters  $\alpha'_0$  and  $\alpha'_1$  (right) for the beta distribution given three different prior distributions and the same empirical observations. See Equation 2 and the Supplement for details.

1 With priors in hand, the die is rolled 100 times, and the outcome is 20 sixes. Now  
 2 that observations have been collected, the prior may be updated to obtain the *posterior*  
 3 *hyperparameters*, which are distinguished by an apostrophe. These consist of the actual  
 4 observations plus the hypothetical ones. Thus, if we set the prior hyperparameters  $\alpha_0 = 5$   
 5 and  $\alpha_1 = 1$ , then our observations result in the posterior hyperparameters  $\alpha'_0 = 80 + \alpha_0 = 85$   
 6 and  $\alpha'_1 = 20 + \alpha_1 = 21$ . Using these updated values, the posterior probability of a parameter  
 7  $\theta$  can be calculated, again using the beta distribution, as depicted in Figure 2 (Right). The  
 8 process of updating the prior probability distribution using empirical evidence to obtain a  
 9 posterior probability distribution is the heart of Bayesian analysis (although, depending on  
 10 the model  $M$ , doing so is often more complex than adding up successes and failures).

11 On the one hand, this example demonstrates that the choice of the prior initially has  
 12 a dramatic effect on the hypothesized distribution of  $\theta$  before any data were collected (given  
 13 how different the curves in Figure 2 (Left) appear). On the other hand, the influence of the  
 14 prior diminishes towards infinitesimal as additional observations are made. When using a  
 15 subjective Bayesian approach<sup>2</sup>, many analysts prefer “weak” prior, such as  $\alpha_0 = \alpha_1 = 1$ ,  
 16 because such priors have an “anything could happen” flavor and do not have a determining  
 17 effect on the shape of the posterior distribution.

18 In an important way, however, even a very weak prior does not imply that *anything*  
 19 could happen. At one point in the satirical novel, *The Colour of Magic* (Pratchett, 1983),  
 20 the protagonist wishes to demonstrate that his companions are trapped in a strong magical  
 21 field where familiar physical laws are invalid. He offers to predict the outcome of a coin  
 22 toss, and while the coin is in mid-air, he calls, “Edge.” He succeeds in calling Edge four  
 23 times, but is incorrect on the fifth flip because the coin transforms into a caterpillar and  
 24 crawls away.

25 This whimsical example demonstrates how prior assumptions are constrained to those  
 26 outcomes permitted by the model  $M$ . A conventional model of coin-flipping allows no  
 27 possibility for the third outcome Edge or a fourth outcome Caterpillar (and rightly so).  
 28 Unfortunately, empirical observation almost never conforms precisely to a well-established  
 29 distribution, so an analyst must not only consider prior observations, but also justify their

<sup>2</sup>The alternative school of “objective Bayesian analysis” takes a different view, rejecting subjective priors, however weak, and instead favoring “non-informative priors” that minimally influence the posterior distribution. This approach is revisited in a later section.

1 choice of which distribution describes the phenomenon under examination. For example,  
 2 Mandelbrot & Hudson (2006) argue that many of the “discontinuities” in the behavior of  
 3 the stock market arise from the mistaken assumption that market behavior should conform  
 4 to a Gaussian distribution when, in fact, the true distribution appears to have much heavier  
 5 tails.

6 Even rigorous model fitting can be dangerous without theoretical support. For ex-  
 7 ample, prior to the subprime crisis, risk models for mortgage-backed securities concluded  
 8 that default rates were Poisson distributed, rather than belonging to a more complex family  
 9 of models that allowed “contagion” (i.e. interdependence between outcomes) (Longstaff &  
 10 Rajan, 2008). These results depended on data collected during the US housing bubble of  
 11 the 1990s and 2000s, during which time housing prices steadily rose. When housing prices  
 12 began to fall in approximately 2007, rising contagion resulted in much higher rates of de-  
 13 fault than a Poisson model predicted were possible (Das et al., 2007; Silver, 2012). As such,  
 14 although the Poisson distribution was appropriate to the pattern of defaults during the  
 15 housing bubble, it grossly underestimated the *possible* rate of default under other economic  
 16 conditions, which resulted in a systematic failure to hedge against losses correctly. This  
 17 reflects an often under-appreciated aspect of statistical forecasting: The model  $M$  must  
 18 reflect the range of outcomes and model parameters that are theoretically plausible, rather  
 19 than being selected only because it best fits the current sample.

20 **Conjugate Priors**

21 Evaluating the posterior odds in Equation 1 is often prohibitive, even when the prob-  
 22 lem can be precisely specified. The most difficult operation is generally estimating the  
 23 value of  $m(x, M)$ . In many cases, even when the distributions are well-defined, the product  
 24  $f(x|\theta, M) \Pr(\theta, M)$  cannot be integrated across all values for  $\theta$ , particularly if  $\Pr(\theta, M)$  is  
 25 improper (that is, when the area under the curve is not finite). There is no guarantee of a  
 26 closed-form solution for the value  $m(x, M)$ .

27 A major advance in the practical use of Bayesian statistics was accomplished by  
 28 Raïffa & Schlaifer (1961), who observed that, in many cases, the posterior distribution  
 29  $\Pr(\theta|x, M)$  belonged to the same family as the prior distribution  $\Pr(\theta, M)$ . Furthermore,  
 30 these *conjugate priors* often possess closed-form solutions. Provided one is willing to select  
 31  $M$  from a particular set of distributions, then estimation of all functions in Equation 1  
 32 becomes straightforward.

33 Consider the die rolling example depicted in Figure 2. Given the prior hyperparame-  
 34 ters  $\alpha_0 = 1$  and  $\alpha_1 = 1$ , and the observation that a six was rolled 20 times out of 100 throws,  
 35 we would like to calculate the posterior probability that the die is fair, such that  $\theta = \frac{1}{6}$ .  
 36 Doing so requires computing the values of each of the elements in Equation 1. In this case,  
 37 the conjugate relationship identified by Raïffa & Schlaifer hinges on the Beta function:

$$\begin{aligned} \Pr(\theta, \text{binom}) &= \frac{\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}}{\text{Beta}(\alpha_0, \alpha_1)} \\ \Pr(\theta|x, \text{binom}) &= \frac{\theta^{\alpha'_1-1}(1-\theta)^{\alpha'_0-1}}{\text{Beta}(\alpha'_0, \alpha'_1)} \end{aligned} \quad \text{where } \text{Beta}(\alpha_0, \alpha_1) = \frac{(\alpha_0 - 1)! (\alpha_1 - 1)!}{(\alpha_0 + \alpha_1 - 1)!} \quad (2)$$

38 Here,  $\alpha'_0$  and  $\alpha'_1$  are the hyperparameters previously identified, based on combining observed  
 39 and hypothetical successes and failures. Because both depend only on the data and not on  
 40 the parameter  $\theta$ , it follows that  $m(x, \text{binom}) = \text{Beta}(\alpha'_0, \alpha'_1)$ , which is trivial to evaluate.

1 Since the publication of Raiffa & Schlaifer (1961), conjugate priors have been identified  
2 for a wide range of probability distributions. In cases where closed-form solutions exist for  
3  $m(x, M)$ , an analyst can use the sufficient statistics<sup>3</sup> describing the observations  $x$  and a  
4 set of prior hyperparameters  $\alpha_i$  to estimate the corresponding posterior hyperparameters  
5  $\alpha'_i$ ; those posterior hyperparameters can be used to evaluate  $m(x, M)$ .

6 Many statisticians on the cutting edge of Bayesian analysis see conjugacy as a crutch  
7 that can be set aside in favor of emerging computational techniques (Samaniego, 2012). The  
8 most notable of these is the family of techniques known collectively as Markov Chain Monte  
9 Carlo (MCMC) methods (Gamerman & Lopes, 2006). In practice, however, such techniques  
10 are computationally intensive. Many large datasets remain prohibitive to work with, and  
11 not all analysts have access to high-powered computers. Furthermore, the intricacies of  
12 implementing MCMC and other numerical integration methods (and the risk of applying  
13 them incorrectly) will dissuade many researchers from developing those skills. When non-  
14 stationary data render more familiar techniques inappropriate, the ‘clever trick’ of conjugacy  
15 is a crutch worth keeping. The CPR takes its name from this clever trick, as its proposed  
16 implementation takes advantage of conjugacy to compute results rapidly.

### 17 Subjective vs. Objective Bayes

18 Becoming acquainted with theoretical and applied Bayesian statistics can be intimi-  
19 dating for a variety of reasons. A major impediment for Bayesian neophytes is the ongoing  
20 debate over “subjective” vs. “objective” Bayesian methods (Goldstein, 2006; Berger, 2006,  
21 respectively).

22 As noted in Equation 1, the analyst must specify a prior, reflecting “past evidence”  
23 in some fashion. As shown in Figure 2, a change in the prior probability distribution  
24 necessarily influences the posterior probabilities, and critics are (often rightly) suspicious  
25 that this “subjectivity” creates an opportunity for abuse. Some alternatives are “objective”  
26 methods that, roughly, correspond to conditions in which prior assumptions are absolutely  
27 minimal. Many different varieties of minimally informative priors have been proposed (e.g.  
28 ‘Jeffreys priors,’ Jeffreys, 1961). Among the more recent are *reference priors* (Bernardo,  
29 2005). A reference prior requires that an analyst specify the model  $M$ , but does so without  
30 any ‘hypothetical’ observations<sup>4</sup>.

31 Unfortunately, working with objective priors is routinely challenging. Most such  
32 priors are improper, even when they result in proper posterior distributions. Additionally,  
33 most are not conjugate and must be computed using numerical integration techniques. The  
34 use of reference priors is especially problematic for distributions with more than one free  
35 parameter. A given model may possess multiple priors that each qualify as ‘objective’ in a  
36 mathematical sense, but nevertheless yield diverging outcomes (Klaunberg & Elster, 2012).  
37 The decision of which ‘objective’ prior to use in these cases remains subjective with respect

<sup>3</sup>The sufficient statistics required to update the hyperparameters differ from one distribution to the next. They are sometimes, but not always, “hypothetical observations” like those in the die-rolling example. Selection of a reasonable prior depends on understanding the relationship between the sufficient statistics and the hyperparameters, which are laid out in explicit detail in the Supplement.

<sup>4</sup>A more precise definition is that a reference prior should be the “maximally uninformative” distribution of shape  $M$ , measured in terms of how far its entropy diverges from those described by other parameters. Thus, a properly-specified reference prior is maximally divergent from all possible posterior distributions.

1 the analyst's preference. Put another way, a prior can never be truly 'uninformative,' only  
2 minimally so.

3 Objective priors may also find support for model parameters that are in practice  
4 impossible. For example, given the limits of a thermometer's precision, measurements  
5 taken at several consecutive times might appear identical. When presented with such data,  
6 a reference prior (being unaware of the *possibility* of a limit in the instrument's sensitivity)  
7 might yield the result that the best possible model consists of frequent changes, with many  
8 of the corresponding variance parameters equal to zero, suggesting infinite precision.

9 Several objective Bayesian methods that do not rely on improper priors have been  
10 developed, including "Bayes factor approximation" using information criteria (Wasserman,  
11 2000) and the use of "intrinsic priors" sampled systematically from the observed data  
12 (Berger & Pericchi, 1996). The intrinsic prior approach is applied to change-point analysis  
13 specifically by Girón et al. (2007). Both of these methods, in principle, permit closed-form  
14 approximation of non-subjective posterior distributions, albeit given considerable compu-  
15 tation. However, because they are approximation methods, they display instability when  
16 applied to small sample sizes, which can make them ill-suited to the demands of change-  
17 point analysis (which may need to divide data into small segments).

18 Although objective Bayesian methods may represent "best practices" when their use  
19 is reasonable (Wagenmakers, 2007), they also represent a departure from the fundamen-  
20 tally probabilistic character of orthodox Bayesian analysis (Samaniego, 2012). Put another  
21 way, it is very rare for an experiment to be performed without some expected constraints  
22 on the observations. Gelman (2006) argues that reference priors are best understood as  
23 "provisional" priors to be updated (or, indeed, abandoned) as observations are accrued.

24 Insisting on a strict adherence to an idealized standard of ignorance is unreasonable  
25 if doing so means entertaining obviously absurd hypotheses. Although the CPR algorithm  
26 may be implemented using objective priors, doing so is likely to be computationally in-  
27 tensive, or to yield nonsensical results, or both, particularly when considering very small  
28 subsets near the edges of the data. Most empirical data (being collected within practical  
29 and theoretical constraints) are better-served by a reasonable weakly informative subjective  
30 prior (Van Dongen, 2006). The challenge is to distinguish reasonable priors from unreason-  
31 able ones.

### 32 Subjective Prior Selection

33 In many cases, very weak priors yield results that are nearly indistinguishable from  
34 those based on objective priors, while also being conceptually straightforward and computa-  
35 tionally efficient. For example, using the function Beta (1, 1) as a prior for binary outcomes  
36 has the advantage that its integral is proper. One may also, however, use the function  
37 Beta (0.5, 0.5) as an even weaker prior whose integral is proper under very mild conditions  
38 (Beta (0.5, 0.5) is, in fact, the reference prior for the binomial model, a very rare case in  
39 which a reference prior is both proper and conjugate). For either of these priors, however,  
40 the biasing influence on the posterior distribution rapidly diminishes once observations have  
41 begun to be collected, as seen in Figure 2.

42 However, some distributions are much more powerfully influenced by their priors,  
43 particularly those whose prior hyperparameters are unbounded. For example, the conjugate



1 prior for a Gaussian distribution's  $\mu$  parameter is also Gaussian, with hyperparameters<sup>5</sup>  $\mu_\mu$   
 2 and  $\sigma_\mu$ . Because  $\mu$  can have real value, there is no "default" value that can be assigned to its  
 3 conjugate prior. In a case where observations fall in a range between 5000 and 6000, using  
 4  $\mu_\mu = 0.0$  as a prior hyperparameter has the effect of introducing a massive outlier to every  
 5 attempt to calculate the posterior hyperparameters. For these reasons, it is important to  
 6 select a subjective prior with reasonable hyperparameters. In some cases, reasonable priors  
 7 can be inferred from prior data; more controversially, they may be "elicited" from expert  
 8 opinion (Oakley & O'Hagan, 2007)

9 In the Supplement, each of the conjugate prior implementations includes a "rule-of-  
 10 thumb" subjective prior derived from the data being analyzed, which can be used as a default  
 11 value. This approach is an 'empirical Bayes method,' (Casella, 1985) and represents a  
 12 compromise between the standard logic of Bayesian calculation and the practical limitations  
 13 of experimentation. In the example above where observations  $x$  fall in the range  $5000 <$   
 14  $x < 6000$ , setting  $\mu_\mu = \text{median}(x)$  would be more appropriate than  $\mu_\mu = 0.0$ .

15 Empirical Bayes methods have become popular in recent years, as a result of more  
 16 robust (albeit computationally intensive) parameter estimation procedures (Carlin & Louis,  
 17 2000). Nevertheless, their use remains controversial, because a statistic derived from the  
 18 observations is used to validate those same observations (Gelman, 2008). The use of in-  
 19 dependently obtained priors avoids this sort of "double dipping." It would be acceptable,  
 20 for example, to use the rule-of-thumb prior calculated from pilot data in the analysis of a  
 21 subsequent experiment. Empirical priors may also be used reliably when the dataset from  
 22 which they are extracted is sufficiently large (Efron, 2010). Finally, rule-of-thumb priors  
 23 can be helpful in developing intuitions about the forms the prior might take. Even when  
 24 not used directly, an empirical prior distribution provides an idea of the form a reasonable  
 25 prior is likely to take. Regardless of its origin, an analyst must report which prior was used,  
 26 along with a justification for that prior.

## 27 Maximum Likelihood, Marginal Model Likelihood, and Bayes Factors

28 In Bayesian analysis, no single model has privileged status (there is not a canonical  
 29 "null hypothesis," for example, Gallistel, 2009). Instead, models are compared in terms  
 30 of their relative odds of being true given the evidence and the prior assumptions. This  
 31 is a substantial departure from the "frequentist" approach that is characterized by null-  
 32 hypothesis significance testing (Wagenmakers, 2007).

33 Many forms of parametric analysis are *maximum likelihood estimation* (or MLE)  
 34 procedures. Given data and a model, MLE procedures call for the selection of whichever  
 35 parameters have the highest likelihood. Often, these methods rely on theorems. For ex-  
 36 ample, the central limit theorem provides a proof that the 'maximum likelihood estimator'  
 37 for the population mean is the arithmetic average, and that the expected distribution of  
 38 sample means converges on Gaussian.

39 However, MLE has its shortcomings. One problem is that both data and density  
 40 functions of possible parameter values are often multi-modal: The maximum likelihood  
 41 estimator might be quite different from the second-best candidate, and may not approximate  
 42 any modal values in the data. Multi-modal distributions are more common in complex

<sup>5</sup>These hyperparameters correspond to the 'standard error of the mean,' such that  $\mu_\mu = \mu$  and  $\sigma_\mu = \frac{\sigma}{\sqrt{n}}$

1 models whose parameter space includes more than dimensions than can easily be visualized,  
 2 as well as when the independent measures are influenced by hidden variables.

3 MLE procedures are also vulnerable to overfitting because they are biased towards  
 4 higher model complexity (Myung & Pitt, 1997). All else being equal, the maximum likeli-  
 5 hood associated with a model with four free parameters will consistently be higher than one  
 6 with three parameters. As Enrico Fermi famously quipped, “With four parameters I can fit  
 7 an elephant, and with five I can make him wiggle his trunk” (Dyson, 2004), implying that  
 8 any model can appear superficially valid with a sufficiently large parameter space, regardless  
 9 of that model’s theoretical justification or later predictive strength. For example, Chen &  
 10 Gupta (2011) describe a variety of frequentist change-point methods, and consistently find  
 11 that the maximum likelihoods rise when additional change-points are added. This introduces  
 12 the additional difficulty of developing significance tests to determine whether the models  
 13 have improved more than expected by chance alone.

14 An alternative approach to model selection is to compare *marginal model likelihoods*  
 15 (or MML) (Wasserman, 2000). This approach judges the efficacy of a particular class of  
 16 model without considering any specific parameters for that model. Rather than favoring the  
 17 model with the tallest peak, an MML approach favors the model with the highest average  
 18 altitude. Put another way, the MML favors the model whose likelihood maximizes the  
 19 integrated volume of the likelihood function<sup>6</sup>.

20 Figure 3 conveys this intuition visually. Given 20 observed binary outcomes, it is  
 21 straightforward to calculate the likelihood of obtaining 10 successes and 10 failures, and to  
 22 estimate that the overall probability of success  $\theta$  is most likely to be 0.5 (Figure 3 Left).  
 23 However, splitting the data into the first ten observations vs. the second ten (perhaps to  
 24 test whether a change-point divides them) entails an increase in model complexity, because  
 25 two different parameters  $\theta_1$  and  $\theta_2$  are now needed. If splitting the data yields two subsets  
 26 that display a 6:4 ratio of success, then the split increases the maximum likelihood but  
 27 reduces the marginal model likelihood (Figure 3 Center). Under these circumstances, it  
 28 would be unjustified to split the data. However, if the split reveals an 8:2 ratio of success,  
 29 a dramatic increase in the marginal model likelihood is observed (Figure 3 Right). When  
 30 comparing these three scenarios, a two-parameter model is only justified when its volume  
 31 under the curve is larger than that of the one-parameter model, both of which are obtained  
 32 by integrating across possible values for the parameters.

33 The function  $m(x, M)$  in Equation 1 is precisely such an integral, and its value may  
 34 either be solved for (if a closed-form solution exists) or estimated numerically. If  $m(x, M)$   
 35 can be estimated, a *Bayes factor* (Kass & Raftery, 1995) can be computed. Bayes factors  
 36 can be interpreted as the ratio of support for  $M_1$  relative  $M_2$ , regardless of the MLE  
 37 parameters associated with either model. They may be calculated by comparing the MML  
 38 of two models:

$$\begin{aligned}
 K &= \frac{\int_{\Theta_1} f(x|\theta_1, M_1) \Pr(\theta_1, M_1) d\theta_1}{\int_{\Theta_2} f(x|\theta_2, M_2) \Pr(\theta_2, M_2) d\theta_2} \\
 &= \frac{m(x, M_1)}{m(x, M_2)}
 \end{aligned}
 \tag{3}$$

39 The MML is computed for two models,  $M_1$  and  $M_2$ , given the observations  $x$ . When  $M_1$

<sup>6</sup>Formally, the MML maximizes its Lebesgue measure with respect to the specified prior.

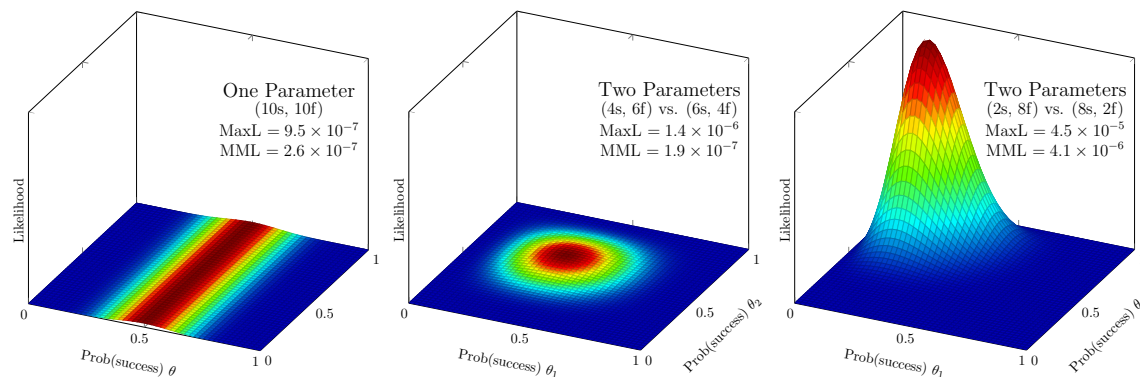


Figure 3. Likelihood as a function of the probability  $\theta$  for the binomial distribution, given 20 observations, consisting of 10 success and 10 failures. Darker areas indicate higher relative likelihoods. (Left Panel) When a single parameter is used to describe the data, the maximum likelihood estimator is  $\hat{\theta} = \frac{10}{20} = 0.5$ . (Center Panel) The first ten observations and the second ten observations are examined independently, and assigned their own MLE parameters  $\hat{\theta}_1 = \frac{4}{10} = 0.4$  and  $\hat{\theta}_2 = \frac{6}{10} = 0.6$ . This increase in model complexity improves the maximum likelihood, but also results in a smaller (and therefore less favorable) marginal model likelihood than that observed for the one-parameter model. (Right Panel) If splitting the observations into two groups instead reveals that  $\hat{\theta}_1 = \frac{2}{10} = 0.2$  and  $\hat{\theta}_2 = \frac{8}{10} = 0.8$ , this results in greater maximum and marginal model likelihoods. All estimations in this figure assume the prior hyperparameters  $\alpha_0 = \alpha_1 = 1$ . On the basis of these results, choosing a two-parameter model is justified in the scenario depicted in the right panel, but not for the center panel, despite both having higher maximum likelihoods than the one-parameter model in the left panel.

1 is more complex than  $M2$ , but also provides a better fit to the data, the Bayes factor  $K$   
 2 is an estimation of the relative odds of each model, given the evidence. For example, in  
 3 Figure 3, the Bayes factor favors the one-parameter model in the 6:4 case ( $K = 0.73$ ), while  
 4 the two-parameter model is favored in the 8:2 case ( $K = 15.8$ ).

5 A major advantage of the MML approach to model selection is that parsimony is  
 6 automatically factored into the calculation. This is because models with more free param-  
 7 eters must distribute their unit mass of prior probability over a larger number of dimensions  
 8 (Gallistel, 2009). All else being equal, each additional parameter lowers the MML by an  
 9 order of magnitude. This severely penalizes models with excessive complexity. MML meth-  
 10 ods favor models that balance the goodness of fit against the watering-down effects of its  
 11 complexity.

### 12 The Conjugate Partitioned Recursion Algorithm

13 Conjugate Partitioned Recursion uses conjugate priors to evaluate  $m(x, M)$  for mod-  
 14 els with and without a change-point, and recursively subdivides the data until none of  
 15 the resulting segments appear to possess further change-points. The Supplement lists con-  
 16 jugate priors and corresponding hyperparameters for four discrete distributions (binomial,  
 17 geometric, Poisson, and multinomial), four continuous distributions (exponential, Gaussian,  
 18 uniform, and multivariate Gaussian), and linear regression (single and multiple). These may

1 all be used as the basis for inferring the location of change-points using the CPR algorithm.

## 2 **Binary Partition by Marginal Model Likelihood**

3 The central problem in change-point analysis is parsimonious model selection. An  
4 algorithm that identifies too many change-points will slice a dataset into unusably small  
5 chunks with little predictive power, while an overly conservative algorithm will miss mean-  
6 ingful events. Effective change-point analysis must strike a balance between these extremes.

7 Unfortunately, it is functionally impossible to systematically examine every possible  
8 subdivision of the data. In a dataset with 100 observations, for example, there is 1 model  
9 with no change-points and 99 models with one change-point, a feasible set of possibilities.  
10 There are, however, 4851 models with two change-points and 156849 models with three  
11 change-points, a factorial progression. Because exhaustive testing of every combination is  
12 out of the question, attention must instead focus on models that seem sufficiently plausible  
13 to merit evaluation.

14 In most data, however, the number of points that are reasonable candidates for  
15 change-point status are a tiny subset, and any change-point identified by an algorithm  
16 seeking a single change is highly likely to also be selected in an analysis seeking two or  
17 more changes. This provides the grounds for a recursive process: Once a change-point is  
18 identified, the data are divided into two segments on either side of that change, and each  
19 of these can then searched for their own ‘best’ change-points, repeated until none of the re-  
20 sulting segments appear to possess any further changes. Divide-and-conquer methods were  
21 first proposed as a means partitioning data on the basis of change-points by Vostrikova  
22 (1981), who demonstrated that such a strategy was computationally efficient. This compu-  
23 tational efficiency makes it attractive to those designing algorithms that identify multiple  
24 change-points (Chen & Gupta, 2011).

25 In practice, the challenge is to statistically infer which point (if any) is most likely to  
26 be a change-point. Both the decision of whether to partition the data and where to make  
27 the split can be determined by estimating the MML for models with and without a change.

## 28 **Determining Whether To Partition The Data**

29 In order to determine whether to partition the data, a model comparison must pit the  
30 one-change model  $C_1$  against the no-change model  $C_0$ . In both cases, the data are assumed  
31 to arise from a distribution  $M$  that has unknown parameters  $\theta$  at times<sup>7</sup>  $(1 \dots n)$ . The  
32 marginal model likelihood of  $C_0$  follows from Equation 1:

$$m(x_{(1:n)}, C_0) = m(x_{(1:n)}, M_{(1:n)}) \quad (4)$$

33 In this and all subsequent equations, subscripts in parentheses refer to indices. For example,  
34  $x_{(1)}$  refers to the first datum in  $x$ , while  $x_{(i:j)}$  refers to the vector of all observations from  
35 datum  $i$  to datum  $j$ . Thus,  $x_{(1:n)}$  denotes the complete time series from observation 1 to  
36 observation  $n$ , and  $M_{(1:n)}$  denotes the distribution  $M$  over that data range.

37 The change-point model  $C_1$  presumes that a change-point splits the observations into  
38 two ranges,  $x_{(1:c-1)}$  (before the change) and  $x_{(c:n)}$  (after the change), each with its own

<sup>7</sup>Here, the intervals between observations are presumed to be uniform; the non-uniform case is discussed below

1 parameters for distribution  $M$ . The marginal model likelihood of  $C_1$  is the average<sup>8</sup>, given  
 2 every interval  $(c - 1 : c)$ , of the product between the model before the change and the  
 3 product after the change:

$$m(x_{(1:n)}, C_1) = \frac{1}{n-1} \sum_{c=2}^n m(x_{(1:c-1)}, M_{(1:c-1)}) m(x_{(c:n)}, M_{(c:n)}) \quad (5)$$

4 Given these values, the Bayes factor for whether the data favor including a change-point is  
 5 a simple ratio:

$$K = \frac{m(x, C_1)}{m(x, C_0)} \quad (6)$$

6 The Bayes factor shows the relative likelihood for two hypotheses: Either a single change-  
 7 point exists, or no change-points exist. The Bayes factor alone does not, however, indicate  
 8 the *posterior odds* that one model or the other is true. Consistent with Equation 1, the  
 9 likelihood ratio is just one part of the equation, and a prior probability must be specified  
 10 that indicates how likely a change is expected to be.

11 The prior probability of a change at any given time is denoted by  $p_c$ . Since there  
 12 are many possible locations at which a change might have occurred, two cumulative prior  
 13 probabilities must be calculated. The first,  $p_0$ , is the probability that no change-points are  
 14 observed across the entire data range; the second,  $p_1$  is the probability that exactly one  
 15 change-point is observed. These can be modeled by Poisson distributions, which describe  
 16 the probability of rare events with many opportunities to occur. The prior probability ratio  
 17 is then multiplied by the Bayes factor to calculate the posterior probabilities ratio between  
 18  $p'_0$  and  $p'_1$ :

$$\frac{p'_1}{p'_0} = K \cdot \frac{p_1}{p_0} = K \cdot \frac{\text{Pois}(1, p_c(n-1))}{\text{Pois}(0, p_c(n-1))} = K \cdot p_c(n-1) \quad (7)$$

19 If  $\frac{p'_1}{p'_0} > 1$ , then the evidence favors  $C_1$ ; otherwise, it favors  $C_0$ . When  $\frac{p'_1}{p'_0} \approx 1$ , the evidence  
 20 supporting either model is approximately equal. It is important to model the odds of  
 21 *exactly* one change-point (rather than, say, the odds of any number of changes) to match the  
 22 MML  $m(x, C_1)$ . Because  $m(x, C_{1,2,\dots})$  is computationally prohibitive, the binary partition  
 23 strategy only considers single change hypotheses at each stage of its recursion, and this  
 24 approach must be consistent when setting  $\frac{p_1}{p_0}$ .

25 In the absence of a strong theoretical case for a particular value for  $p_c$ , a good default  
 26 value is  $p_c = \frac{1}{n-1}$  because this indicates even odds. This is a relatively conservative prior,  
 27 however, and as change-points are discovered, its value should be relaxed, as described  
 28 below.

29 It is advisable to specify a *decision criterion*  $\tau > 1$ , and to partition the data only when  
 30  $\frac{p'_1}{p'_0} > \tau$ . Because binary partitioning is a recursive process, data with many change-points  
 31 must be divided into many small, noisy segments with some percentage of false positives.  
 32 This introduces a stopping problem: Each false positive drives further subdivision of the

<sup>8</sup>The act of averaging fulfills two functions: It converts the interval between  $t = 1$  and  $t = n$  to a unit length, and it then takes the sum of the discretized intervals in that space. This marginalizes the likelihood with respect to all possible positions of the change-point. The consequences of this interpretation are revisited in the section entitled “Biased MML Estimation in Small Samples.”

1 data, creating more opportunities for false positives. However, each false negative terminates  
 2 investigation of a particular segment.

3 The traditional interpretation of Bayes factors is that those in the range  $3 < K < 10$   
 4 are ‘substantial’ and  $10 < K < 30$  are ‘strong,’ whereas any value for  $K > 100$  is considered  
 5 ‘decisive’ (Jeffreys, 1961). The choice of a decision criterion depends on the main objective of  
 6 the analysis: A primarily descriptive model can entertain a criterion as low as 3 (introducing  
 7 some risk of overfitting), while a very strong theoretical test might use a criterion as high  
 8 as 100. Henceforth, this paper uses the decision criterion  $\tau = 10$  unless otherwise indicated.

9 When probability distributions are well-defined, an analyst seeking the optimal de-  
 10 cision criterion can perform an approximate sensitivity analysis by computing the log-  
 11 likelihood of all observations and comparing the change-point models identified using dif-  
 12 ferent decision criteria in terms of the Schwarz-Bayes Information Criterion (SBIC), which  
 13 provides a computationally straightforward estimate of the marginal model likelihood for  
 14 the entire time series (Schwarz, 1978). A demonstration sensitivity analysis is provided in  
 15 the Supplement. On the basis of that analysis,  $\tau = 10$  performs reasonably well in all cases.

16 **Determining Where To Partition The Data**

17 In the event that  $\frac{p'_1}{p_0} > \tau$ , the strength of the evidence supporting each possible  
 18 change-point must be compared. To do this, we may rewrite Equation 6 in the following  
 19 way:

$$\begin{aligned}
 K &= \sum_{c=2}^n \frac{m(x_{(1:c-1)}, C_0) m(x_{(c:n)}, C_0)}{(n-1) \cdot m(x_{(1:n)}, C_0)} \\
 &= \sum_{c=2}^n \frac{k_{(c)}}{n-1} \text{ where } k_{(c)} = \frac{m(x_{(1:c-1)}, C_0) m(x_{(c:n)}, C_0)}{m(x_{(1:n)}, C_0)}
 \end{aligned}
 \tag{8}$$

20 That is, the Bayes factor  $K$  is the average (for all possible change-points  $c$ ) of a series of  
 21 individual odds ratios  $(k_{(2)}, \dots, k_{(n)})$ . Provided the evidence favors identifying a change-  
 22 point, the best candidate available is given by:

$$\hat{c} = c \text{ when } k_{(c)} = \max(k_{(2)}, \dots, k_{(n)})
 \tag{9}$$

23 The estimated change-point  $\hat{c}$  represents the first trial  $c$  to take place after a change, and  
 24 it is determined by identifying the interval  $k_{(c)}$  whose odds ratio is the largest value in the  
 25 series  $(k_{(2)}, \dots, k_{(n)})$ .

26 Once a change-point has been identified, the process described above may be recur-  
 27 sively applied to the subsets of data on either side of the change. When the algorithm  
 28 concludes that there are no further change-points to be identified, all that remains is to  
 29 estimate the model parameters for each resulting segments of observations.

30 **Non-Uniform Event Times**

31 The computation of any MML, whether it be the no-change model  $m(x, C_0)$  or the  
 32 single-change model  $m(x, C_1)$ , consists of an integration across all possible models. This

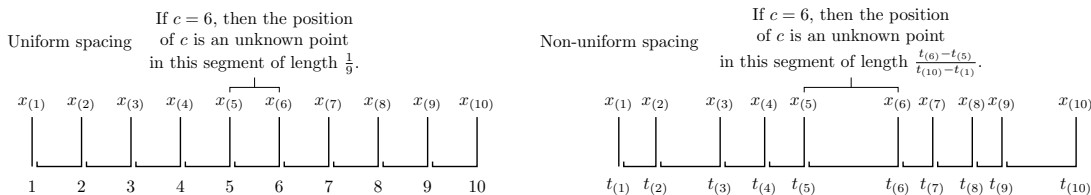


Figure 4. An example of uniform vs. non-uniform segmentation. (Left) In the uniform case, each segment of time is an equally strong candidate for a putative change-point. In this example, the nine segments each have a common weight of  $\frac{1}{9}$ . (Right) In the non-uniform case, a change-point is presumed to be located somewhere between  $t_{(1)}$  and  $t_{(10)}$ , and if any point along that interval is as likely as any other, then longer segments are correspondingly more likely to contain change-points.

1 includes the act of averaging the product of two models in Equation 5. Note that the  
 2 indices  $1 \dots n$  are not discrete points, but rather are regularly delimited intervals of time.  
 3 If a change-point is identified “at index  $c$ ,” that means, in practice, that a change-point is  
 4 likely to have occurred somewhere between index  $c - 1$  and index  $c$ . Without more fine-  
 5 grained data, the analysis offers no further insight about *when* the change occurred within  
 6 that interval. This is illustrated by Figure 4 (left), in which ten observations  $x_{(1)}$  to  $x_{(10)}$   
 7 appear at uniform intervals.

8 Figure 4 (right) presents a different case, in which each observation  $x_{(i)}$  took place  
 9 at a time  $t_{(i)}$ . Since the single change-point analysis treats the span from  $t_{(1)}$  to  $t_{(n)}$  as a  
 10 uniform interval within which a change might occur, the probability of a change-point being  
 11 located in segment  $c$  is equal to  $\frac{t_{(c)} - t_{(c-1)}}{t_{(n)} - t_{(1)}}$ , hereafter abbreviated by  $\mathcal{T}_{(c)}$ . With this mind,  
 12 we can update Equation 8 to accommodate the uneven intervals:

$$K = \sum_{c=2}^n k_{(c)} \cdot \mathcal{T}_{(c)} \quad \text{where } \mathcal{T}_{(c)} = \frac{t_{(c)} - t_{(c-1)}}{t_{(n)} - t_{(1)}} \tag{10}$$

13 Since the Bayes factor now assigns a different weight to each of the individual odds ratios,  
 14 this weight must be taken into consideration when selecting which segment of the time-series  
 15 is most likely to contain the change-point:

$$\hat{c} = c \text{ when } [k_{(c)} \cdot \mathcal{T}_{(c)}] = \max \left( [k_{(2)} \cdot \mathcal{T}_{(2)}], \dots, [k_{(n)} \cdot \mathcal{T}_{(n)}] \right) \tag{11}$$

16 This formulation is fully general given the assumption of uniform probability, with regular  
 17 inter-event intervals as a special case.

18 **Biased MML Estimation in Small Samples**

19 Although Equation 10 gives the appearance of weighing each point in time equally,  
 20 this is not the case in practice. Figure 5 (left) displays the values for  $\log(k_{(c)} \cdot \mathcal{T}_{(c)})$  at  
 21 each observation  $c$  in the sequence  $[0, 1, 0, 1, \dots, 0, 1]$ , as computed using Equation 10 and  
 22 assuming a binomial distribution. Although there is no signal in the data, the values of

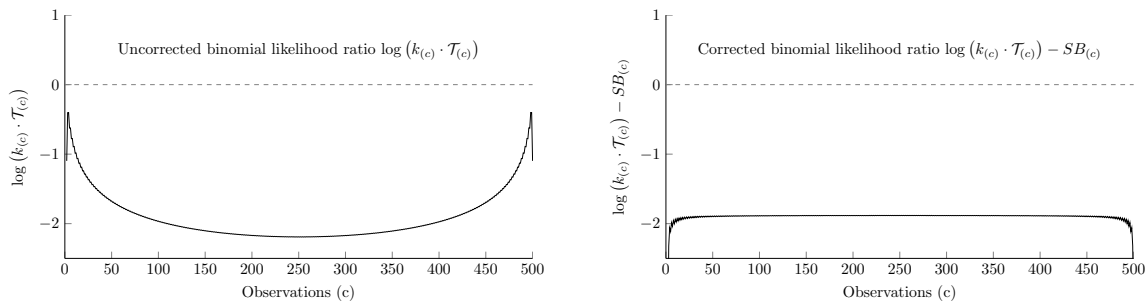


Figure 5. Values of  $\log(k_{(c)} \cdot \mathcal{T}_{(c)})$  given 100 binary observations  $[0, 1, 0, 1, \dots, 0, 1]$  both uncorrected (left) and corrected (right). The dashed line represents the boundary between evidence favoring a change at time (positive) vs. evidence against (negative) at event  $c$ .

1  $\log(k_{(c)} \cdot \mathcal{T}_{(c)})$  are nevertheless closer to 0.0 at the edges of the sequence than they are in  
 2 the center.

3 The inflated ratios near the edges of a segment indicate a distortion in the estimate of  
 4 model complexity for the one-change model  $C_1$  relative to the no-change model  $C_0$ . Since  
 5 the model  $C_1$  consists of two distributions (each with  $p$  unknown parameters), it should be  
 6 lower by a consistent amount *unless* a discontinuity is present in the data.

7 The degree of difference is approximated by the SBIC (Schwarz, 1978), which esti-  
 8 mates the marginal likelihood:

$$m(x, M) \tilde{\propto} f(x|\theta, M) \left(\frac{1}{n}\right)^{p/2} \tag{12}$$

9 From this approximation, it follows that:

$$k_{(c)} \tilde{\propto} \frac{f(x_{(1:c-1)}|\theta, C_0) \cdot f(x_{(c:n)}|\theta, C_0)}{f(x_{(1:n)}|\theta, C_0)} \left(\frac{n}{(c-1) \cdot (n-c+1)}\right)^{p/2} \tag{13}$$

10 The distortion predicted by formulation matches the shape of the bias observed in Figure 5  
 11 (left) precisely.

12 Since introducing a second distribution effectively adds  $p$  parameters to the model,  
 13 without changing the number of observations, any given ratio  $k_{(c)}$  should be expected to  
 14 have a value of  $\left(\frac{1}{n}\right)^{p/2}$  if the data contain no change. Without a correction, however, values  
 15 for  $k_{(c)}$  close to the edges of a segment get much closer to 1.0, which introduces a substantial  
 16 risk of false positives.

17 With these considerations in mind, a ‘Schwarz-Bayes correction’ for the value of  
 18  $\log(k_{(c)})$  is denoted by  $SB_{(c)}$ , and is obtained from an integral over the relevant inter-



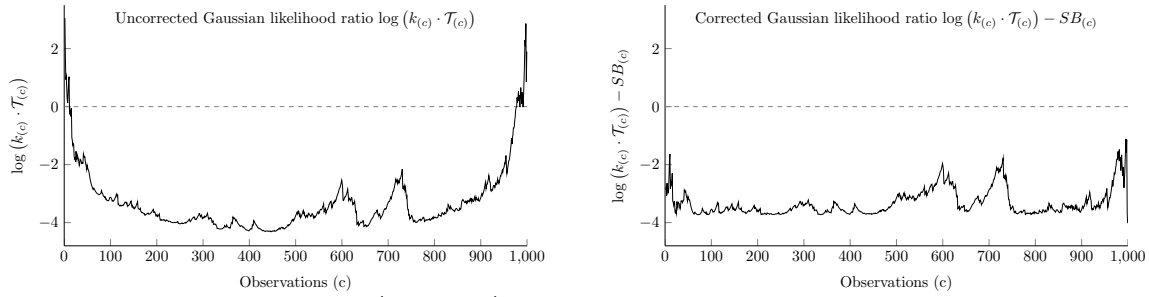


Figure 6. Values of  $\log(k_{(c)} \cdot \mathcal{T}_{(c)})$  given 1000 randomized Gaussian observations ( $\mu = 0, \sigma = 1$ ), representing data in the range  $\pm z = 3.3$  both uncorrected (left) and corrected (right). The dashed line represents the boundary between evidence favoring a change at time (positive) vs. evidence against (negative) at event  $c$ .

1 val:

$$\begin{aligned}
 &\text{Given that } r_{(c)} = \frac{t_{c-1}}{t_n} \\
 &\text{and } \int_0^i \log\left(\frac{1}{x(1-x)}\right) = i(2 - \log(i - i^2)) + \log(1 - i) - 2i \\
 &\quad \dots \\
 &SB_{(c)} = \frac{pn}{2} \int_{r_{(c-1)}}^{r_{(c)}} \log\left(\frac{1}{x(1-x)}\right) \\
 &= \frac{pn}{2} \left[ \int_0^{r_{(c)}} \log\left(\frac{1}{x(1-x)}\right) - \int_0^{r_{(c-1)}} \log\left(\frac{1}{x(1-x)}\right) \right]
 \end{aligned} \tag{14}$$

2 Given that  $SB_{(c)}$  is a modification to the log-likelihood, Equation 10 should be further  
 3 updated to the following form:

$$K = \sum_{c=2}^n \frac{k_{(c)} \cdot \mathcal{T}_{(c)}}{\exp(SB_{(c)})} \tag{15}$$

4 And this, in turn, suggests the following criterion for selecting the best candidate for a  
 5 change-point:

$$\hat{c} = c \text{ when } \left[ \frac{k_{(c)} \cdot \mathcal{T}_{(c)}}{\exp(SB_{(c)})} \right] = \max \left( \left[ \frac{k_{(2)} \cdot \mathcal{T}_{(2)}}{\exp(SB_{(2)})} \right], \dots, \left[ \frac{k_{(n)} \cdot \mathcal{T}_{(n)}}{\exp(SB_{(n)})} \right] \right) \tag{16}$$

6 Figure 5 (right) shows the values for  $k_{(c)}$  in the binary example after the correction have been  
 7 applied. Although estimates become noisier close to the edges, they no longer show a dra-  
 8 matic bias near the edges. Given that this sequence has 100 observations and one parameters  
 9 per distribution, the default odds ratio should be approximately  $\sqrt{\frac{1}{100}} = \exp(-2.302) = .1$ ,  
 10 close to the corrected value across all observations.

11 Although this correction may appear arbitrary, or may seem to violate the logic of  
 12 the Bayes factor, it is actually a sensible modification of the prior distribution associated

1 with the location of the change-point. If we conceptualize the interval from  $t_{(1)}$  to  $t_{(n)}$  as  
 2 a unit interval (such that  $0 \leq t_{(c)} \leq 1$ ), then a uniform prior is equivalent to the prior  
 3 distribution Beta (1, 1). As originally shown by Jeffreys (1946), however, this prior is *not*  
 4 unbiased, and the appropriate unbiased prior distribution should resemble the reference  
 5 prior, Beta (0.5, 0.5). The depth of the correction must reflect the correct number of free  
 6 parameters, here given by  $p$ :

$$SB_{(c)} \propto \int_{r_{(c-1)}}^{r_{(c)}} p \cdot \log(\text{Beta}(0.5, 0.5)) + C \quad (17)$$

7 Here,  $C$  is a constant that keeps  $\Sigma(SB_{(c)}) = 0$ , such that the overall prior odds of a change  
 8 remain equal to  $p_c$ . The link to the reference prior for the binomial is not coincidental: as  
 9 pointed out in the discussion of Equation 5, the position of a change-point may be conceived  
 10 of as a point on a unit line between  $t_{(1)}$  and  $t_{(n)}$ .

11 Because the depth of the correction depends on the number of free parameters, its  
 12 necessity grows as a function of model complexity. The distorting effects of the small sample  
 13 bias are mild enough for the binomial model that the correction may appear unnecessary,  
 14 but can produce aberrant model estimates for more complex models. For example, Figure 6  
 15 (left) shows the values for  $\log(k_{(c)})$  calculated from 1000 observations in a uniform fashion<sup>9</sup>  
 16 from a standard Gaussian distribution ( $\mu = 0, \sigma = 1$ ) and subsequently randomized. The  
 17 naïve calculation of  $k_{(c)}$  in the left figure suggests that although the overall pattern of  
 18 evidence is inclined against a change-point, the curvature is more extreme because of the  
 19 additional free parameter, and the weight of the evidence favoring a change exceeds 0 at  
 20 the edges of the distribution. These are simply a result of the small sample bias, however,  
 21 and when the correction is applied in Figure 6 (right), the individual marginal likelihoods  
 22 reveal that the evidence is consistently inclined against a change-point across all intervals.

### 23 Implementing Binary Partition by Marginal Model Likelihood

24 Given that the posterior odds ratio  $\frac{p'_1}{p_0}$  (Equation 7) indicates *whether* to partition the  
 25 data and that the peak weighted odds ratio  $\left[ \frac{k_{(c)} \cdot \mathcal{T}_{(c)}}{\exp(SB_{(c)})} \right]$  (Equation 11) indicates *where* to  
 26 partition, the full description of the binary partitioning strategy is specified in Algorithm 1.  
 27 The algorithm begins with an array  $\mathbf{M}$  containing two indices,  $\langle 0, n \rangle$ ; since these delimit  
 28 the full span of the data, this array can be said to contain no change-points. Consequently,  
 29 the number of change-points in the model is  $(\text{length}(\mathbf{M}) - 2)$ . The algorithm then tests  
 30 whether to partition the data using Equation 7. If the posterior odds ratio exceeds the  
 31 decision criterion  $\tau$ , then the best available candidate is selected using Equation 11. This  
 32 estimated change-point is added to  $\mathbf{M}$ , and the algorithm is then applied to each of the  
 33 resulting segments in  $\mathbf{M}$ . This process iterates until no new change-points are identified.

34 The prior odds of a change begin with a value of  $p_c = \frac{1}{n-1}$ . This corresponds to even  
 35 odds that a single change-point is present in the data, as noted earlier. As further change-  
 36 points are identified, this prior is updated to reflect the newly discovered change-points,

<sup>9</sup>1000 evenly spaced values from .0005 to .9995 were converted to  $z$ -scores using an inverse cumulative normal distribution, generating simulated observations in the range  $z \pm 3.3$ . These were then ordered randomly.

---

**Algorithm 1:** The binary partition by marginal model likelihood strategy, allowing for non-uniform time stamps.

---

**Data:** events  $x_{(1:n)}$ , times  $t_{(1:n)}$ , model  $C_0$ , decision criterion  $\tau$ , initial model

$\mathbf{M} = \langle 0, n \rangle$

**Result:** updated model  $\mathbf{M}$ , model parameters  $\mathbf{P}$

**begin**

**repeat**

$\mathbf{N} \leftarrow \langle \rangle; |\mathbf{K}| \leftarrow n$  /\* set up arrays \*/

$p_c = \frac{\max(1, \text{length}(\mathbf{M}) - 2)}{n - 1}$  /\* set odds of a change \*/

**for**  $s = 1$  **to**  $\text{length}(\mathbf{M}) - 1$  **do**

$i \leftarrow \mathbf{M}_{(s)} + 1; j \leftarrow \mathbf{M}_{(s+1)}$  /\* assign indices \*/

**for**  $c = i + 1$  **to**  $j$  **do**

$k_c \leftarrow \frac{m(x_{(i:c-1)}, C_0) m(x_{(c:j)}, C_0)}{m(x_{(i:j)}, C_0)}$  /\* Eq. 8 \*/

$\mathbf{K}_{(c)} \leftarrow \frac{k_c \cdot \mathcal{T}_{(c)}}{\exp(SB_{(c)})}$  /\* Eq. 15 \*/

**if**  $\text{sum}(\mathbf{K}_{(i+1:j)}) \cdot p_c \cdot (j - i) > \tau$  **then**

$\hat{c} \leftarrow \text{index}(\max(\mathbf{K}_{(i+1:j)})) - 1$  /\* if Eq. 7 permits, find  $\hat{c}$  \*/

                Push( $\mathbf{N}, \hat{c}$ ) /\* insert  $\hat{c}$  into  $\mathbf{N}$  \*/

        Push( $\mathbf{M}, \mathbf{N}$ ); Sort( $\mathbf{M}$ ) /\* merge  $\mathbf{N}$  into  $\mathbf{M}$  and sort \*/

**until**  $\text{length}(\mathbf{N}) = 0$

**for**  $s = 1$  **to**  $\text{length}(\mathbf{M}) - 1$  **do**

$\mathbf{P}_{(s)} \leftarrow \text{EstimateParameters}(x_{(\mathbf{M}_{(s)} + 1 : \mathbf{M}_{(s+1)})});$

**return**  $\mathbf{M}, \mathbf{P}$

---

1 such that  $p_c = \frac{\text{length}(\mathbf{M}) - 2}{n - 1}$  on any given iteration of the algorithm. Although this appears  
2 to bias the algorithm slightly in favor of finding a change on the first iteration, it must be  
3 emphasized that, per the logic of Equation 7, the only models being compared in any single  
4 computation are those with *exactly* zero changes and *exactly* one change. Consequently,  
5  $p_c = \frac{1}{n - 1}$  on the first iteration corresponds to even odds for either outcome.

6 Algorithm 1 has several desirable qualities. In principle, its generality allows it to  
7 be applied regardless of which distribution  $M$  is specified for  $C_0$ , and regardless of the  
8 approach used to compute the value of  $m(x, C_0)$ . Furthermore, because the evaluation of  
9  $m(x, C_0)$  is the runtime's primary limiting factor, closed-form solutions for this integral  
10 permit Algorithm 1 to run rapidly even on large datasets.

### 11 Pitfalls & Considerations

12 The CPR algorithm, as described above and encapsulated in Algorithm 1, is a pow-  
13 erful tool for asking a specific kind of question. If an analyst believes a time series under  
14 examination is well-described by a model whose MML has a closed form, is willing to treat  
15 model changes as being discontinuities, and wishes to treat the resulting segments indepen-

1 dently, the CPR algorithm will identify likely change-points rapidly without requiring the  
2 fine-tuning of a host of operational parameters. If any of these assumptions are unreason-  
3 able, however, the CPR algorithm may not be ideal.

#### 4 **The Law of the Instrument**

5 In any analytic enterprise, it is important to recall (and resist) the law of the instru-  
6 ment:

7 In addition to the social pressures from the scientific community there is  
8 also at work a very human trait of individual scientists. I call the *the law of the*  
9 *instrument*, and it may be formulated as follows: Give a small boy a hammer,  
10 and he will find that everything he encounters needs pounding.

11 -Kaplan (1998), p. 28.

12 The temptation is to use statistical methods that are familiar or that require minimal  
13 effort. Although the CPR algorithm is designed to be easy to understand, easy to use,  
14 and computationally efficient, it is not intended to supplant all change-point analyses, as  
15 it can entertain only certain hypotheses. The limitations of the CPR algorithm should be  
16 understood clearly by anyone wishing to make use of it. In many cases, these limitations can  
17 be mitigated or bypassed. Above all, to paraphrase the counsel of Wilkinson et al. (1999),  
18 “Analysts should never report statistics whose operations they do not understand.” While  
19 the CPR algorithm is less sophisticated than many other change-point analyses currently  
20 available, its generality and computational simplicity will hopefully permit a wider range  
21 of applied researchers to understand its operations.

#### 22 **Insensitivity Given Large-Scale Stationary Processes**

23 Although the CPR algorithm is highly effective with a wide range of data, it is entirely  
24 ineffective with data in which a large-scale stationary processes conceals many small sub-  
25 processes. Because the marginal likelihoods are calculated for only a single change-point  
26 at a time, the posterior hyperparameters are estimated from wide swaths of data for early  
27 change-point evaluations and the resulting summary statistics may mistake many brief  
28 segments for a single continuous segment with high variability.

29 When data show large-scale uniformity that conceals small-scale distributional shifts,  
30 it is often more appropriate to rely on other forms of time-series analysis, such as wavelet-  
31 based methods (Lio & Vannucci, 2000). However, two modifications can be made to the  
32 CPR algorithm that are also effective at overcoming this problem. The first approach is  
33 to arbitrarily subdivide (or to ‘dice’) the data into small segments and run preliminary  
34 analyses. If these reveal evidence for changes, the detected changes may be retained and  
35 used as the prior model of changes  $\mathbf{M}$ . The second approach, which is more principled, is  
36 to examine the data sequentially, using a widening window to restrict analysis to a local  
37 region. Both the dicing operation and a form of sequential analysis called the ‘forward-  
38 retrospective’ strategy, originally proposed by Gallistel et al. (in press), are described in  
39 detail in the Supplement.

## 1 Assumption-Free Change-Point Analysis

2 It is important to reiterate that many Bayesian techniques rely on distributions that  
3 do not have conjugate priors (for example, most that use reference priors). Furthermore,  
4 although closed-form solutions exist for the marginal model likelihoods of many conjugate  
5 priors, there are notable exceptions, such as the gamma distribution and the negative bi-  
6 nomial distribution.

7 As noted above, the binary partitioning strategy is not limited to conjugate analysis,  
8 although it becomes much more computationally intensive if integrals must be approximated  
9 numerically (e.g. using MCMC, Carlin & Chib, 1995; Bauwens & Rombouts, 2012). In some  
10 cases, arithmetic approximations may also be available (e.g. Raftery, 1996; Fearnhead, 2006;  
11 Hannart & Naveau, 2012). Such methods should be used, however, under those conditions  
12 in which conjugate methods are inappropriate for the observed data or are theoretically  
13 incoherent.

14 In practice, all analyses rely on some assumptions. However, methods are available  
15 with far more relaxed assumptions than those employed by the CPR algorithm. The most  
16 famous of these is reversible jump MCMC (Green, 1995), which not only numerically ap-  
17 proximates an unknown model, but does so with an initially unknown number of model  
18 dimensions. These attributes have contributed to its widespread application in technical  
19 circles (Sisson, 2005). However, both implementing such methods and interpreting their re-  
20 sults can be challenging, however, and remain beyond the reach of many applied researchers  
21 (Han & Carlin, 2001).

## 22 Specifying Prior Odds of a Change

23 The CPR algorithm, as described above, identifies a single change-point under the  
24 assumption that it is equally likely at every moment in time. In other words, the distribution  
25 associated with the change is uniform. Koop & Potter (2009) demonstrate that the hazard  
26 function for the uniform function is not flat, and depends on the analyst's prior assumptions  
27 about the likely number of changes and the maximum period between changes. Fortunately,  
28 as Koop & Potter note, the uniform assumption is appropriate when seeking to identify a  
29 single change-point. Binary partitioning only identifies the strongest change-point in each  
30 segment, which prevents its uniform assumption from creating a bias.

31 There are other circumstances, however, in which the odds of a change are themselves  
32 a function of the passage of time. In principle, if this relationship is known, it can be  
33 integrated into Algorithm 1. However, making such a modification would likely have such a  
34 powerful impact on the posterior odds that doing so severely undermines the objectivity of  
35 the test. In effect, this changes the prior probability distribution for a change (with respect  
36 to time), and carries with it all of the concerns articulated above regarding the possibility  
37 for abuse. Because substantive changes are likely to be detected even when the odds of a  
38 change are assumed to be uniform, imposing additional assumptions about when changes  
39 are likely to occur should have very substantial theoretical support.

40 The CPR algorithm also assumes that the probability of a change  $p_c$  (Equation 7) does  
41 not display discontinuities. It may be the case, however, that the probability of observing  
42 a change is itself subject to regular changes. A very successful paradigm for Bayesian  
43 change-point analysis hinges on representing the data in terms of hidden Markov models

1 in which the number of states is either known (Chib, 1998) or unknown (Meligkotsidou &  
2 Dellaportas, 2011), wherein each state has its own value for  $p_c$ . These more complex models  
3 are best approached using robust numerical methodologies.

4 One scenario that may be dealt with simply is the case of the ‘impossible change-  
5 point.’ Under some circumstances, an analyst can identify conditions under which change-  
6 points are impossible. For example, in a learning paradigm, it is reasonable to assume  
7 that an organism could not know the identity of a never-before-seen stimulus prior to being  
8 exposed to it, since there is no persuasive evidence for precognitive or oracular abilities  
9 (Wagenmakers et al., 2011). A simple adjustment to Algorithm 1 is to fix the value of  
10  $k_{(c)}$  at zero for any interval  $c$  during which a change is impossible *a priori*, leaving all  
11 other values unchanged. This correction is inherently conservative, as it always lowers the  
12 resulting value of  $K$ , and protects against edge conditions in which measurement error  
13 accidentally suggests a causally impossible result. An example of this is provided in the  
14 Simultaneous Chain example below.

### 15 Gradual Transitions

16 The CPR algorithm assumes that the change-points under consideration are either  
17 genuine discontinuities or transitions that are sufficiently rapid that they occur between  
18 time points  $t_{(c-1)}$  and  $t_{(c)}$ . In some contexts, this is theoretically assumed: Economists,  
19 for example, often refer to change-points as ‘structural changes’ because they often arise  
20 in economic data as a result of substantial changes in leadership or policy. Furthermore,  
21 abrupt and discontinuous change is the norm rather than the exception in much of behavior  
22 analysis, a fact that has been obscured by over-reliance on averaging (Gallistel et al., 2004).  
23 Often, however, gradual transitions are an appropriate assumption (see, for example, the  
24 description of ‘turning points’ by Cohen, 2008), and these are not ideally suited to the CPR  
25 algorithm as it is implemented.

26 In some cases, mixture distributions may be used for Bayesian modeling of gradual  
27 transitions (e.g. Kheifets & Gallistel, 2012) or growth curves (e.g. Zhang et al., 2007); in  
28 these cases, the corresponding marginal likelihoods do not typically have closed-form solu-  
29 tions, and must be modeled numerically. However, several simpler methods may be used  
30 in concert with the CPR algorithm. For example, conjugate priors are relatively straight-  
31 forward to implement for simple and multiple linear regression, and the CPR algorithm  
32 performs well at fitting approximately linear transitions. Linear fits of this kind are pro-  
33 vided in the reaction time and 3D position examples below, and in additional examples  
34 presented in the supplement.

### 35 Multiple Time Series and Shared Parameters

36 Another major limitation of the CPR algorithm is that it applies to a single (poten-  
37 tially multivariate) time series, as opposed to a series of inter-related time series that may  
38 share parameters (such as a set of individuals). Furthermore, each segment of data assessed  
39 by the CPR algorithm is assumed to be independent of every other segment. While such  
40 limitations are appropriate for unambiguous measurements under controlled experimen-  
41 tal conditions, they pose considerable difficulties when dealing with opportunity samples,  
42 loosely operationalized measures, or contexts with substantial nuisance variables.

1 For the most part, these limitations arise from the objective of making a compu-  
2 tationally efficient and conceptually straightforward change-point analysis. However, it is  
3 not automatically appropriate to favor more sophisticated methods. For example, although  
4 mixed-effect models are a powerful way to characterize both population parameters and  
5 individual deviations, they nevertheless impose strong assumptions about the qualitative  
6 similarity across individuals (Schielzeth & Forstmeier, 2009). Although other methods, such  
7 as those based on ‘hidden Markov models’ (Robert et al., 2000), require only weak depen-  
8 dence between modeled phenomena, they nevertheless also imposes structural assumptions  
9 on the resulting segments.

10 In this regard, the CPR algorithm should be seen as having limited power. Because  
11 it relies on isolating segments of data for analysis, it is necessarily less powerful than a test  
12 that can make use of all of the data to evaluate every point. At the same time, however  
13 the power of comprehensive tests stem from assumptions about how well different regions  
14 of the data can inform one another.

## 15 Multiple Model Types

16 As it is currently implemented, the CPR algorithm assumes that all segments arise  
17 from models with identical forms (e.g. all are Poisson distributed, or all are normally  
18 distributed). There is no reason *a priori*, however, why it could not consider a wider  
19 range of possible models in parallel, particularly if there are additional covariates involved.  
20 For example, in a continuous univariate dataset, each segment could not only compare  
21 whether or not a change-point is appropriate, but also whether each segment is drawn from  
22 a normal, exponential, or uniform distribution. In principle, marginal likelihoods can be  
23 used to approach this problem, and so long as each segment can be considered in isolation,  
24 the closed form solutions made available for conjugate priors may also be used. However,  
25 in this scenario, the posterior odds of any given model is influenced of how many other  
26 functions are being considered, as described by Kass & Raftery (1995).

27 Alternatively, if a specific kind of model comparison arises from theoretical consider-  
28 ations, a custom analysis can potentially be constructed to accommodate that comparison.  
29 For example, the contrast presented in Figure 1 cannot be directly tested using the CPR  
30 algorithm as written, because the equation specified by Heathcote et al. (2000) has not been  
31 formally studied. However, numerical methods could be used to assess its likelihood func-  
32 tion and corresponding marginal likelihoods (which would supplant the uniform no-change  
33 hypothesis in the CPR algorithm’s logic), while conjugate formulas could be used for the  
34 linear segments.

35 A multi-model CPR algorithm that can simultaneously evaluate these alternatives is  
36 beyond the scope of this paper. Such an approach also potentially invites ‘fishing expedi-  
37 tions’ in which different combinations of distributions are intermixed and only those that  
38 are favorable to the research are reported. However, given the building blocks provided  
39 in this manuscript, an analyst who has sufficient theoretical justification can assemble a  
40 custom variant suited to their empirical scenario. Such variants would need to be evaluated  
41 on a case-by-case basis.

## Example Implementations

Because the CPR algorithm is highly general with respect to the Bayesian models it is equipped to consider, it can be applied to a wide range of data, provided those data reasonably conform to distributional assumptions. Here, three examples are considered, each relying on a different statistical model. Additional examples from other empirical disciplines are provided in the Supplement.

In the first example, task performance in an animal cognition experiment is modeled in terms of success or failure using the binomial distribution. Doing so not only permits general statements to be made about how well the subject learned, but additionally permits identification of the trials during which learning occurred on a session-by-session basis. This reveals dynamics of learning glossed over by learning curves.

In the second example, the CPR algorithm is used to examine changes in human reaction time as a result of practice over consecutive trials. Not only does this permit a within-subject multiple regression to compare the effects of practice and task difficulty on reaction time, but also allows intermixed sub-tasks to be teased apart and individually examined. This analysis reveals a variety of previously concealed discontinuities in the response times.

In the third example, a motion-tracking device produced multivariate data, signaling position in 3D space at rapid but irregular time intervals. The CPR algorithm permits this highly complex dataset to be reduced to a tractable summary of position and motion.

### Task Acquisition: Simultaneous Chains

An elementary question in psychology is “when did learning occur?” Despite being a prerequisite to a host of questions (including *how* learning occurs), identifying and describing these events is routinely difficult using traditional statistics. Learning is often discontinuous, a ‘eureka’ moment marked by an abrupt shift in a behavior. Despite this, the “learning curve” is often invoked, despite being an average across many trials, or worse, many subjects (Gallistel et al., 2004). Change-point analyses provide a different way of thinking about learning, since change-point functions are chiefly concerned with identifying discontinuities rather than smoothing them over.

Jensen et al. (2013) trained rhesus macaques to learn the ordering of lists of otherwise arbitrary stimuli using the simultaneous chaining paradigm (Terrace, 2005). In each of a session’s 40 trials, five photographs were simultaneously displayed on a touchscreen, and remained until subjects either touched all five in the correct order (obtaining a food pellet) or made any mistakes (leading to a time-out, followed by a new trial). Subjects learned by trial and error. First, they were required to identify Item One, and only then could they proceed to identify Item Two. In this fashion, subjects learned 25 novel lists, each composed of stimuli never seen prior to that session.

A straightforward way to assess a subject’s progress in a novel list is to decompose



that progress into a series of binary strings, as follows:

```

resp = 0012010111222405153203235535555215555552
b1 = 001101011111110111110111111111111111111
b2 = 000100000011110101110111111111110111111
b3 = 000000000000010101100101111111001111110
b4 = 0000000000000101010000001101111001111110
b5 = 0000000000000001010000001101111001111110
    
```

1 Here, *resp* corresponds to the performance of one subject (Coltrane), where each value  
 2 represents number of correct presses made by the subject on a given trial (with food pellets  
 3 only delivered on a ‘5’). As such, if *resp* indicates a ‘3’ at a given position, that means  
 4 that the subject correctly selected the first, second, and third items, but did not select the  
 5 correct 4th item. Thus, item one is correctly selected on the third trial, while items one  
 6 and two are correctly selected on the fourth trial. In each binary string, the trial  $b_j(i) = 1$   
 7 if  $resp(i) \geq j$ . The first time an item is correctly selected is not necessarily the point at  
 8 which that item’s position is learned. For example, although Item Two is first successfully  
 9 chosen on trial 4, there is a dramatic shift in responding at trial 11.

10 A change-point analysis was performed treating ‘probability of a correct response’ as  
 11 a Bernoulli process. This entails calculating the value of  $m(x, M)$  in Equation 1, namely,  
 12  $m(b_j, binomial)$ . The Supplement lists the closed-form solutions for  $m(x, M)$  given variety  
 13 of distributions, including the binomial; in this case, the closed-form estimate is obtained  
 14 by computing Beta (successes, failures).

15 The first step to applying the CPR algorithm is the specification of a prior. As noted  
 16 previously, the MML of the reference prior for binomial data is  $Beta(0.5, 0.5) = 3.142$ , which  
 17 corresponds to half a success and half a failure. Thus, given the binary string presented in  
 18 Figure 7, the value of  $m(x_{(1:40)}, C_0)$  (from Equation 4) depends on the observed successes  
 19 and failures, plus the prior. The result is also computed using the Beta function, thanks  
 20 to conjugacy:  $m(x_{(1:40)}, C_0) = Beta(30.5, 10.5) = 0.00000000007$ . This is marginal model  
 21 likelihood for a no-change-point model. Additionally, a decision criterion must be selected;  
 22 this analysis uses the default value of  $\tau = 10$ .

23 Figure 7 (Top) shows how the CPR algorithm uses the series of Bayes factors  $k_{(c)}$   
 24 are used to identify the first change-point. The *x*-axis shows individual trials, whereas the  
 25 *y*-axis shows each element’s value of  $k_{(c)}$  (plotted on a log scale to emphasize its shape). For  
 26 example,  $k_{(8)}$  requires that we know three values:  $m(x_{(1:40)}, C_0)$  (which was just calculated)  
 27 and the MML for each of the two segments,  $m(x_{(1:7)}, C_0)$  and  $m(x_{(8:40)}, C_0)$ . These are  
 28 obtained using the Beta function, just as before:  $m(x_{(1:7)}, C_0) = Beta(1.5, 6.5) = 0.0506$ ,  
 29 and  $m(x_{(8:40)}, C_0) = Beta(29.5, 4.5) = 0.0000022$ . Filling in the values, we determine that  
 30  $k_{(8)} = \frac{0.0506 \times 0.0000022}{0.00000000007} = 1654.9$ .

31 Obtaining the marginal likelihood requires considering all possible change-points.  
 32 With a prior  $m(M) = Beta(1, 1)$  and the probability of a change  $p_c = \frac{1}{39}$ , the posterior odds  
 33 ratio in support of a change was across all intervals was  $\frac{p'_1}{p'_0} = \left[ K \cdot \frac{39}{39} \right] = \sum \frac{k_{(c)}}{(n-1) \exp(SB_{(c)})} =$

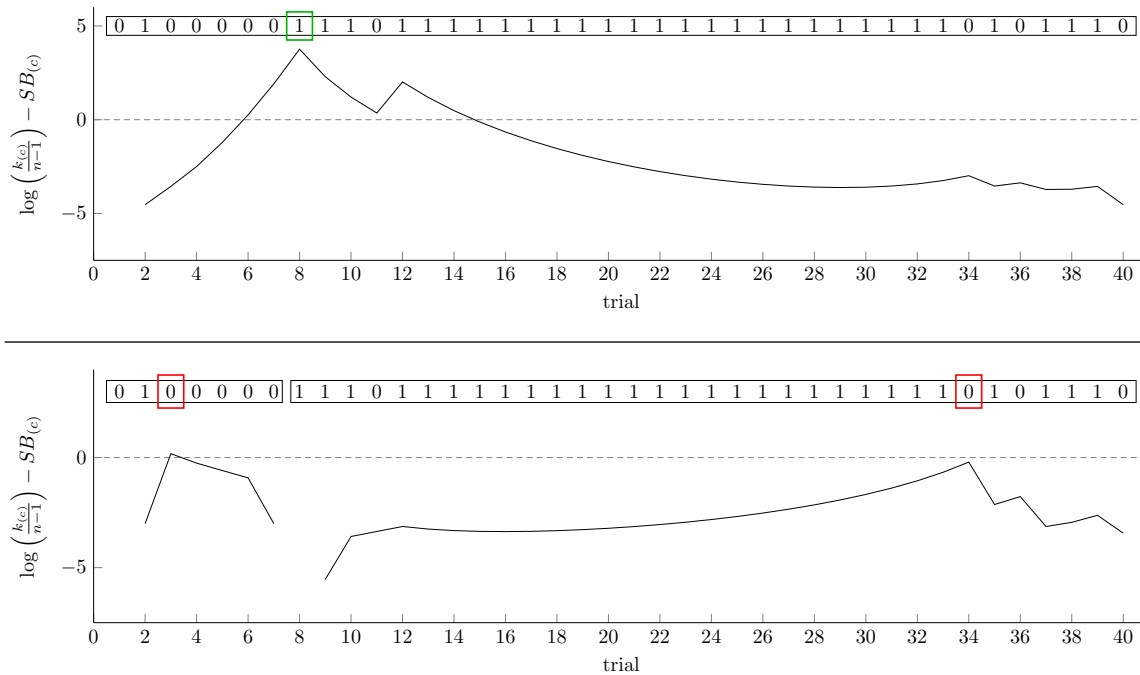


Figure 7. The CPR algorithm applied to a binary string. (Top) The individual Bayes factors  $k_{(c)}$  (here presented on a log scale) are calculated for each point, and there is a clear peak when  $c = 8$ . (Bottom) The algorithm is then recursively applied to the two segment halves, with weak support found when  $c = 34$ . Because this support is too weak to satisfy the decision criterion, it is not selected. The dashed line represents the boundary between evidence favoring a change at time (positive) vs. evidence against (negative) at event  $c$ ; however, note that a change-point is only added to the model if the posterior odds  $\frac{p'_1}{p'_0}$  for a segment favors a change, which is not the case for either segment in the bottom panel.

1 81.40, supporting a change-point because it exceeds the decision criterion of  $\tau = 10$ . Given  
 2 the peak value of  $k_{(c)}$ , the most likely position for a change-point therefore lay between  $x_{(7)}$   
 3 and  $x_{(8)}$ .

4 Figure 7 (Bottom) represents the second iteration of the algorithm. In the first  
 5 segment,  $K = 3.52$  and  $\frac{p'_1}{p'_0} = \left[ K \cdot \frac{6}{39} \right] = 0.54$ , indicating an 11:5 odds ratio against a new  
 6 change-point. In the second segment,  $K = 3.72$  and  $\frac{p'_1}{p'_0} = \left[ K \cdot \frac{32}{39} \right] = 3.05$ , a posterior odds  
 7 ratio slightly favoring a change, but falling well below the decision criterion of  $\tau = 10$ .

8 With no further change-points expected, we formally define the change-point model  
 9  $M_{(0;2)} = [0, 7, 40]$ , dividing the data into segments 1:7 and 8:40. Only at this stage are the  
 10 rate parameters estimated for each segment. Given our prior Beta(0.5, 0.5), the Bayesian<sup>10</sup>  
 11 posterior parameter estimates are  $P_{(1;2)} = \left[ \frac{1.5}{8}, \frac{28.5}{34} \right] = [0.19, 0.84]$ .

12 In the Simultaneous Chain paradigm, subjects must work through the list incremen-  
 13 tally, solving each subsequent item by trial and error. In rare cases, the CPR algorithm

<sup>10</sup>Although these parameter estimates continue the Bayesian logic of this analysis by incorporating the prior hyperparameters, the model parameters for each segment may also be estimated using frequentist methods. The only parameters estimated by the CPR algorithm are the positions of change-points.



1 reported that subjects ‘learned’ the identity of an item a trial or two before ever receiving  
2 feedback confirming its position. In order to prevent these cases of the algorithm making  
3 a slightly early prediction, values for  $k_{(c)}$  for intervals prior to the first successful press to  
4 an item were fixed at 0.0 in advance, in keeping with the advice for avoiding impossible  
5 conclusions described in the ‘Specifying Prior Odds of a Change’ section.

6 The CPR algorithm provides a precise account of the behavior for each session in-  
7 dependently. For each of Coltrane’s 25 novel lists, Figure 8 (Left) presents a plot of every  
8 change-point, corresponding to learning the identity of each item. Progress was coded both  
9 in terms of position along the  $y$ -axis and with increasingly dark shades of gray. On the  
10 0th trial, no progress is assumed. The discovery of the first item (as determined by the  
11 change-point analysis) is marked by the first gray bar, with each additional item denoted  
12 by another step. Figure 8 (Right) presents histograms graphing the distribution of intervals  
13 to “learning the next item,” as well as the total number of trials needed to learn the list.

14 Learning varies quite a bit from session to session, and few sessions resemble “average  
15 learning.” Acquisition is sometimes very rapid (lists 7 and 14), sometimes gradual and  
16 incremental (list 10 and 25), and sometimes a mix of the two (lists 8 and 23). Another  
17 problem with averaging is that once a subject correctly identifies the fourth item, the fifth  
18 and final item is also usually identified by process of elimination. As such, a statement  
19 that “on average, four items were learned by time  $t$ ” almost certainly refers to a mix of  
20 3-item-learning and 5-item-learning.

21 Using binomial change-points to identify learning events (e.g. Gallistel et al., 2004)  
22 is one of the more straightforward applications of change-point analysis to psychological  
23 data because binomial data do not possess ‘outliers’ as such. In scenarios where outcomes  
24 can reasonably be represented as binomial, the ease of calculating the Beta function makes  
25 this approach appealing. However, in many scenarios, representing events in binary terms  
26 discards much of the information that might be relevant. Fortunately, because conjugate  
27 prior analysis is possible for a wide range of distributions, the CPR algorithm is not limited  
28 to binary data.

### 29 **Curve Fitting: Reaction Times**

30 When performing an analysis of continuous data measurements, there is a powerful  
31 temptation to fit the data to a distribution and work with the resulting summary statistics.  
32 Indeed, the lion’s share of null hypothesis significance testing is based on the tails of inferred  
33 distributions. To quote Wagenmakers (2007), “ $p$  values depend on data that were never  
34 observed<sup>11</sup>.” A change-point analysis that can ‘digest’ continuous data without throwing out  
35 information can complement (or substitute) traditional analyses. A particularly relevant  
36 case is the practice of averaged curve-fitting.

37 For example, Palmeri (1997) performed a series of experiments contrasting cognitive  
38 and memory strategies in processing visual stimuli. The training phase presented one of  
39 thirty pre-generated stimuli consisting of between 6 and 11 black dots arrayed on a white  
40 field (with five stimuli per number of dots). Participants responded as quickly as possible,  
41 indicating how many dots were on screen. Because the stimuli were pre-generated, par-  
42 ticipants gradually transitioned from explicit cognitive strategies (such as counting dots)

<sup>11</sup>However, a counterpoint might be that many Bayes factors depend on priors invented from whole cloth.

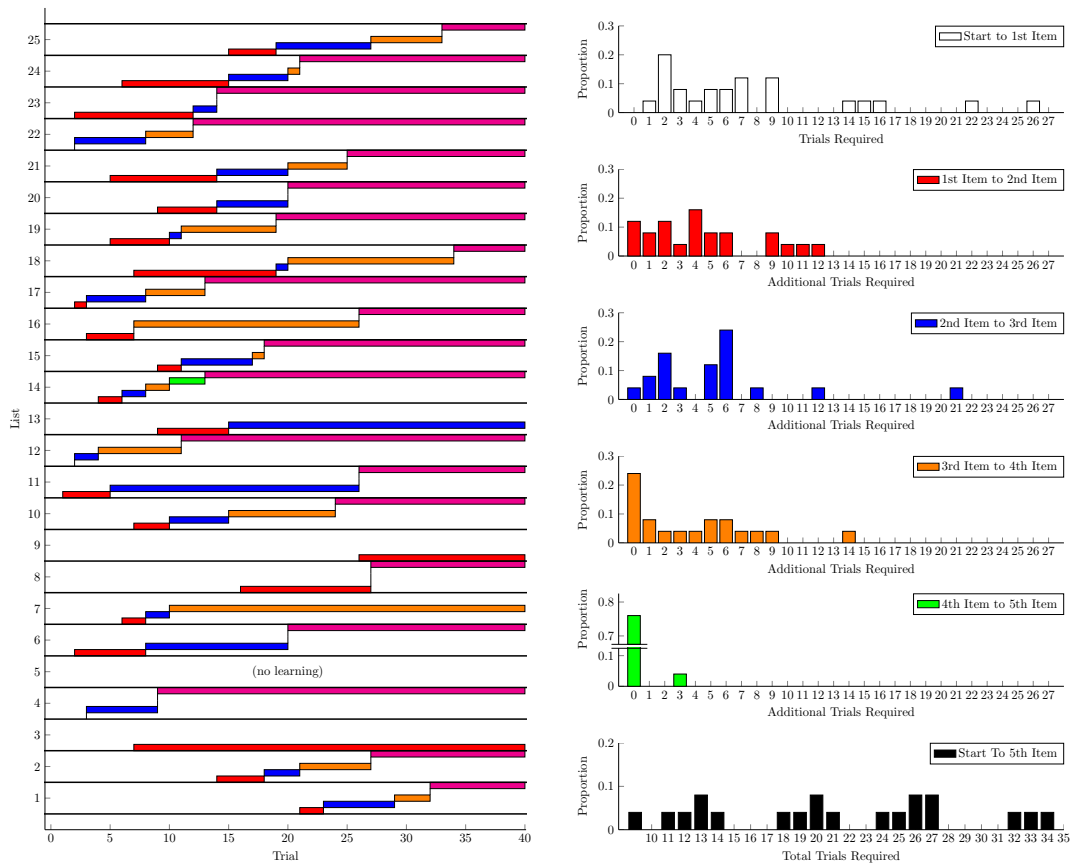


Figure 8. Plot of item acquisition in 25 different lists in Jensen et al. (2013), learned using the Simultaneous Chain procedure. (Left) Estimated time of acquisition for each list item is indicated both by elevation and by shade of gray. (Right) Histogram showing frequency of inter-acquisition periods, in trials, for each item, as well as the number of trials overall needed to acquire all list items. Bars do not sum to 100%, because acquisition did not occur in all lists.

1 to a memory strategy in which they recognized the stimulus and recalled their previous  
 2 answer. These data were subsequently published<sup>12</sup> by Heathcote et al. (2000) as part of a  
 3 meta-analysis of reaction time data.

4 A visual examination of Palmeri’s data suggests that stimuli containing more dots  
 5 elicited longer reaction times, and that responses generally became faster over consecutive  
 6 training trials. Fitting a curve to these data results in an approximately exponential di-  
 7 minishing returns function (Heathcote et al., 2000). However, an explicit component of  
 8 Palmeri’s hypothesis is that two processes (explicit cognition vs. memory) contribute to  
 9 reaction times. If the transition between strategies occurs abruptly, and these transitions  
 10 occur at a variety of times, discontinuities are likely to appear in the data, which a standard

<sup>12</sup>Data are available for download at the Newcastle Cognition Lab Data Repository, <http://www.newcl.org/?q=node/7>, retrieved March 6, 2013.

PeerJ PrePrints

1 curve-fitting approach will be unable to properly detect or quantify.

2 A change-point analysis was performed using a linear regression model. Thus, instead  
3 of dividing the data into segments representing flat rates (as in the simultaneous chain  
4 example above), the linear approach split the data into linear segments with  $\beta$  parameters  
5 corresponding to model's covariates. The model intercept is hereafter denoted by  $\beta_{\text{const}}$ , the  
6 slope with respect to trials is denoted by  $\beta_{\text{trials}}$ , and the slope with respect to numerosity  
7 (i.e. the number of dots in the stimulus) is denoted by  $\beta_{\text{num}}$ .

8 As in the binomial case, setting a prior is an important prerequisite to performing  
9 the Bayesian analysis. In addition to observations  $y$  and a matrix of explanatory variables  
10  $\mathbf{X}$ , four prior hyperparameters go into the analysis to describe the residuals, presumed to  
11 conform to a multivariate normal distribution. Two,  $\mathbf{m}$  and  $b$ , correspond to the intercept  
12 of the function, while the other two,  $c$  and  $\mathbf{\Lambda}$  correspond to the precision matrix (i.e. the  
13 inverse of the covariance matrix) of the residuals. Without access to pilot data, an empirical  
14 Bayes method was used to estimate a 'rule-of-thumb' prior (described in greater detail in  
15 the Supplement).

16 In Figure 9 (Top), 6132 consecutive reaction times from a single participant, are  
17 plotted with respect to trials and numerosity (darker points correspond to slower reaction  
18 times). A handful of outliers are omitted from the plot but were included in the analysis. At  
19 a glance, it is clear that the larger numerosities (10 and 11) initially elicit much longer reac-  
20 tion times than the shorter ones (6 and 7). Additionally, it is clear that reaction times late  
21 in training (after around 4,000 trials) depend very little on the stimulus numerosity. Fig-  
22 ure 9 (Bottom) shows how the CPR algorithm, using a linear regression model, subdivides  
23 the data.

24 In an important sense, this 'kitchen-sink' change-point analysis is merely a different  
25 flavor of the familiar shortcomings of curve fitting. For example, this linear model overes-  
26 timates the reaction time for 6-item numerosities: While performance is consistently below  
27 1 s by trial 1,000, the model does not predict performance on 6-item stimuli reaching this  
28 speed until around trial 2,000. More importantly, however, is the experimental detail that  
29 there were thirty stimuli, with five stimuli belonging to each numerosity. If faster reaction  
30 times signal a transition to a memorial strategy based on stimulus recognition, then each  
31 stimulus was presumably learned at a different point in time. Thus, a per-stimulus analysis  
32 would be provide a more compelling description of learning.

33 Figure 10 contrasts the fastest and slowest learning for the 6-dot stimuli (Left) and the  
34 11-dot stimuli (right), drawn from the same dataset that was used in Figure 9. Once again,  
35 change-point analyses were performed, this time using only  $\beta_{\text{const}}$  and  $\beta_{\text{trials}}$ . Equations 10  
36 and 11 were used because the order of stimulus presentation was randomized, resulting in  
37 non-uniform intervals between events.

38 As expected from Figure 9, participants initially had a much higher reaction time  
39 when presented with 11-dot stimuli than with 6-dot stimuli. There was also considerable  
40 overlap between times associated with each stimulus very early in responding. It is also  
41 clear that reactions times were largely independent of stimulus complexity late in training  
42 (when all responding is quite rapid). This consistency falls apart in mid-range responding,  
43 however, with the fastest 11-dot stimulus quite reliably acquired faster than the slowest 6-dot  
44 stimulus was. Furthermore, although some of the distributions resemble traditional learning  
45 curves, others displayed dramatic and irregular discontinuities. Interestingly, despite having

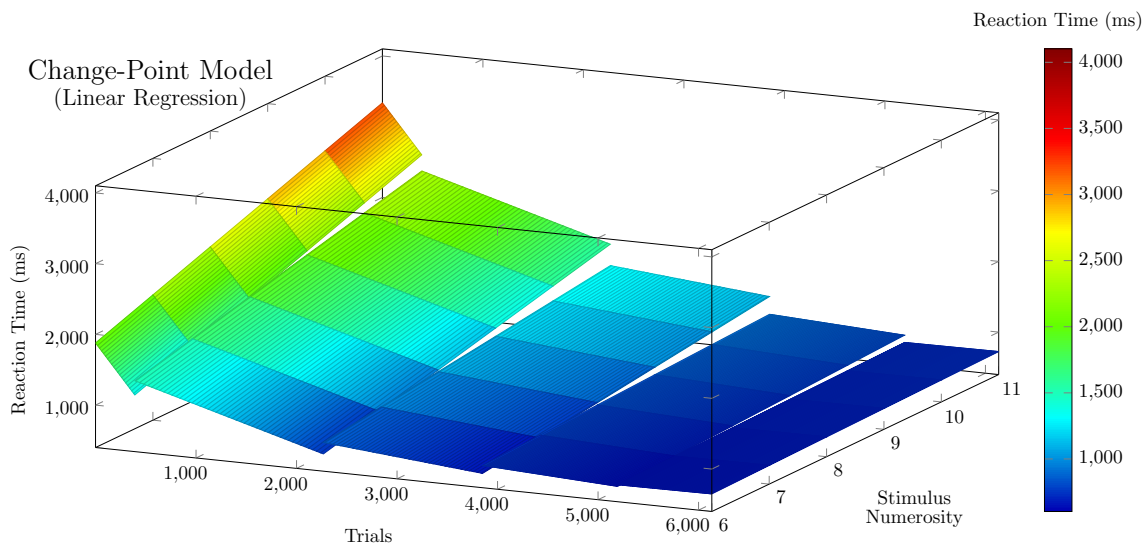
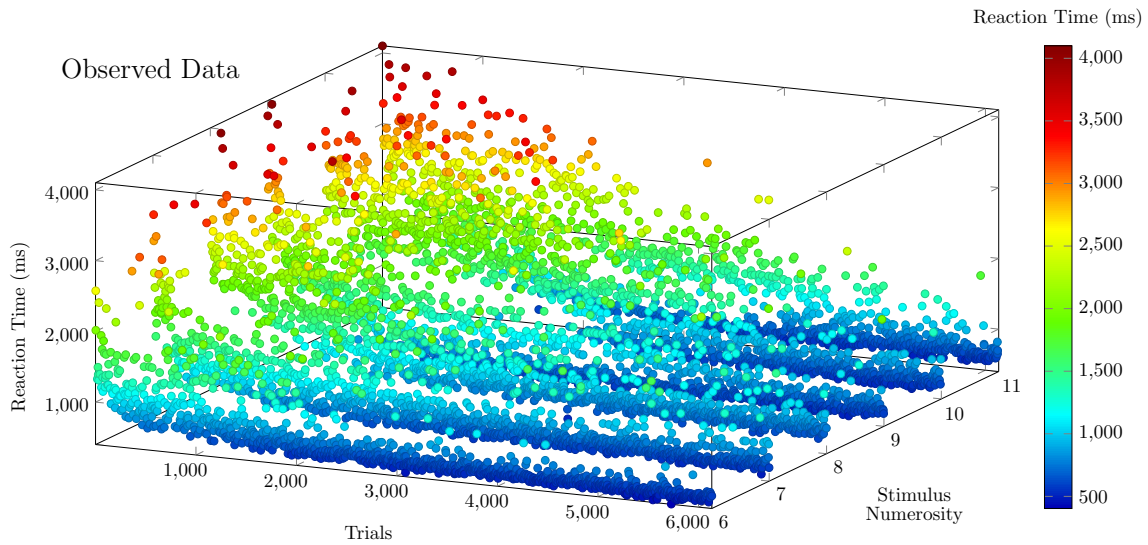


Figure 9. Reaction times as a function of trials and task difficulty in Palmeri (1997). (Top) Each of the 6132 reaction times, color-coded according to their speed and positioned with respect to trial and stimulus numerosity. (Bottom) The model fit resulting from a multiple regression, subdivided according to the CPR algorithm.

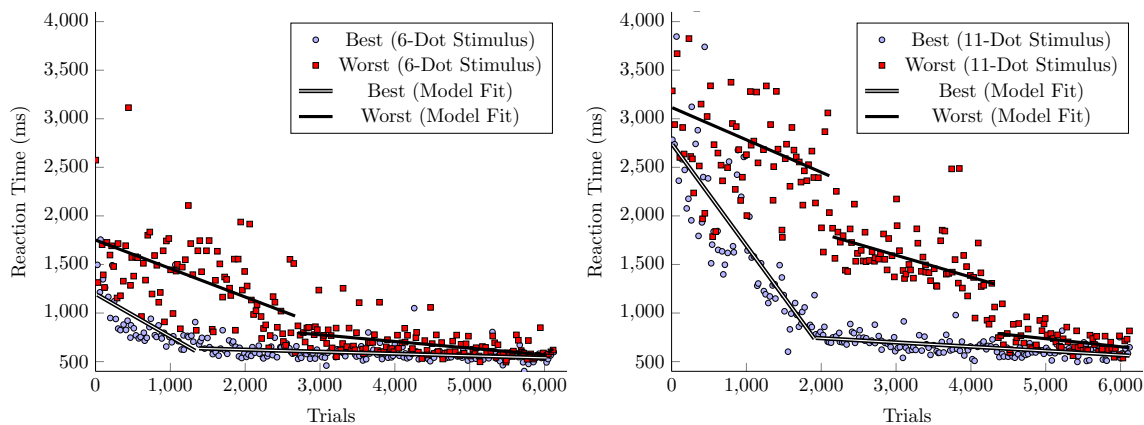


Figure 10. Reaction times as a function of trials in Palmeri (1997) for specific stimuli in a single participant. Lines represent Bayesian regression estimates whose subdivisions were determined using the change-point algorithm. (Left) 6-dot stimulus learning for most rapidly acquired stimulus (blue) and the least rapid (red). (Right) 11-dot stimulus learning for most rapidly acquired stimulus (blue) and the least rapid (red).

1 a reputation for being distributed in a non-normal fashion, the residuals of these individual  
 2 regressions were reasonably close to normal<sup>13</sup> with respect to both skewness ( $\mu = 1.15 \pm 0.55$ )  
 3 and kurtosis ( $\mu = 4.88 \pm 1.78$ ). Although these display moderate departures from normality,  
 4 they nevertheless fall well below the rule-of-thumb guidelines for regression specified by  
 5 Kline (1998) that skew  $< 3$  and kurtosis  $< 10$ .

6 Although the analysis of reaction times is ubiquitous across many domains of psy-  
 7 chology, its time-series character and its often non-normal distributions raise the concern  
 8 that it is often analyzed incorrectly (Whelan, 2008). Given the variety of patterns displayed  
 9 by a single participant in Figures 9 and 10, the practice of fitting curves as a form of bulk  
 10 averaging will need to give way to analyses that are more sensitive to individual learning his-  
 11 tories and discrete changes in behavior. While it is not a panacea, a change-point approach  
 12 like the CPR algorithm can nevertheless contribute to a more nuanced understanding of  
 13 reaction time data.

14 **Multivariate Data: 3D Position Tracking**

15 As the impacts of Big Data continue to be felt, and data-gathering technologies be-  
 16 come less expensive, there is an underexploited opportunity to ask psychological questions  
 17 on a larger and more multivariate scale. The challenge for many, however, is that traditional  
 18 methods of analysis are not adequate to fully exploit such datasets because of their non-  
 19 stationary characteristics. A computationally efficient change-point algorithm can distill  
 20 otherwise daunting datasets into practical chunks, as well as provide important large-scale  
 21 parametric measures.

<sup>13</sup>Skewness and kurtosis were calculated for the residuals from each of the segments, omitting outliers falling over 4 standard deviations from the mean. These 37 censored outliers constituted only 0.6% of the data. The resulting means are reported  $\pm 1$  standard deviation.

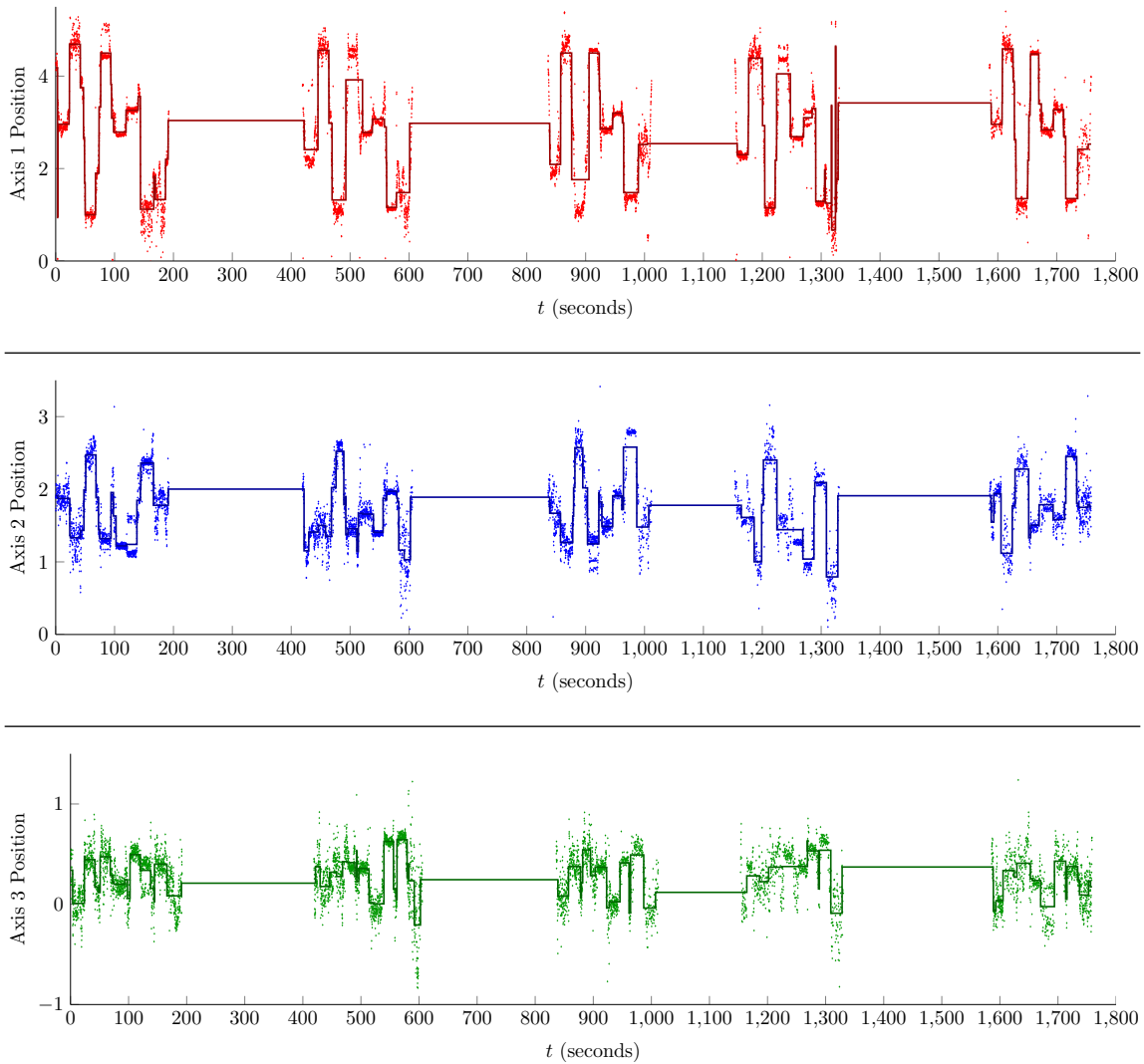


Figure 11. 3D position data collected by Kaluža et al. (2010). Individual points are multivariate, such that an observation an  $x$ -coordinate (in the top panel), a  $y$ -coordinate (in the center panel), and a  $z$ -coordinate (in the bottom panel). Change-points were identified using the CPR algorithm assuming a multivariate normal distribution.



1 Kaluža et al. (2010) demonstrated a proof of concept for a movement tracker designed  
2 to be worn by elderly individuals. Each sensor continuously transmitted X-Y-Z coordinates.  
3 By monitoring the patterns of movement of several sensors at different places on the body  
4 in parallel, a participant's behavior be characterized, and emergency situations (such as  
5 sudden falls) could be detected immediately.

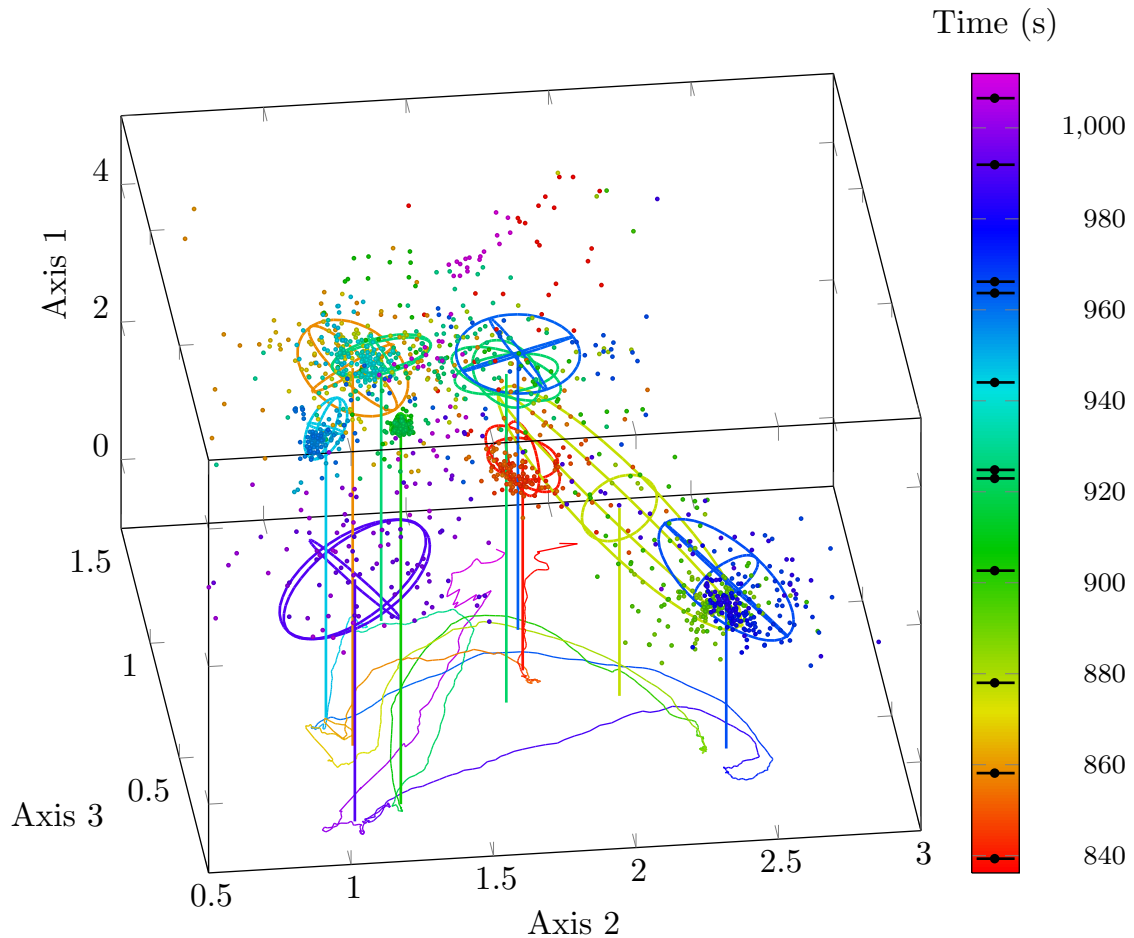
6 There is considerable experimental potential for this style of multivariate data. For  
7 example, a clinical trial of psychiatric medication could measure both small- and large-scale  
8 movement over days and weeks in order to study effects of chronic administration on motor  
9 coordination (when measured on the order of inches) or social isolation (when measured on  
10 the order of miles). In a laboratory context, the movement and position of animal subjects  
11 is often a covariate of interest, and could be determined more precisely than measuring  
12 "time spent in each quadrant."

13 Figure 11 depicts the data collected from one sensor worn by one participant over  
14 a period of approximately 30 minutes. Because their objective was to demonstrate the  
15 efficacy of a machine-learning algorithm, the data consisted of participants performing the  
16 same series of physical movements five times. In addition to some inter-temporal variability  
17 in the times at which coordinates were measured, there were four large gaps in the data  
18 during which the experimenters reset the conditions to permit the action script to be re-  
19 peated. The colored lines corresponds to the means of a representative multivariate normal  
20 distribution, whose change-points and parameters were estimates using the CPR algorithm.  
21 Although possessing a closed form, the arithmetic solution for  $m([X, Y, Z], MVnormal)$   
22 is intimidating, and is presented in the Supplement. As in the linear regression example  
23 described above, the analyst must specify four prior hyperparameters (corresponding to the  
24 mean and covariance). A rule-of-thumb empirical prior was used, based on robust mean  
25 and covariance estimates (Campbell, 1980).

26 Figure 12 shows, in three-dimensional terms, the correspondence between individual  
27 observations in time and the corresponding segments specified by the change-point model  
28 over a subset of the data. Color is used to indicate the passage of time. Every point  
29 corresponds to a discrete observation from the 3D sensor, color-coded according to event  
30 time. Additionally, a moving average of 50 responses is plotted on the marginal horizontal  
31 plane.

32 In addition to raw sensor readings, Kaluža et al. also reported 'activity labels' in-  
33 dicating the behavior being performed by the participant. These labels alternate between  
34 steady-state behaviors (walking, sitting) and transitional behaviors (falling, rising). Concep-  
35 tually, the change-points detected by the CPR algorithm should fall within or near these  
36 transitional periods, dividing behavior into segments with distinct properties. Figure 13  
37 shows the congruence between the detected change-points (marked with dashed lines) and  
38 the reported transition periods (indicated as gray zones) over the same interval as depicted  
39 in Figure 12, with a corresponding color bar indicating the passage of time. The algorithm  
40 detected most of transitions, although it was not generally able to detect several transitions  
41 occurring in quick succession.

42 Figure 13 also showcases a limitation of the multivariate Gaussian distribution used  
43 in this example. Because the model presumed abrupt transitions between otherwise uni-  
44 form states, gradual movement through space (e.g. between 875s and 905s) was not very  
45 effectively modeled. However, the algorithm performed admirably in periods of stability,



*Figure 12.* 3D depiction of the a subset of the movement data from Kaluža et al. (2010) and its associated change-point model. Individual points represent discrete observations and are color-coded continuously with respect to time, as noted on the color bar. The points on the color bar denote the times at which change-points were detected by the CPR algorithm. Ellipses represent the means and covariance associated with particular segments of data, estimated post-hoc using the robust method described by Campbell (1980). The thin colored line, which is drawn along the horizontal plane, represents a moving average of 50 points.

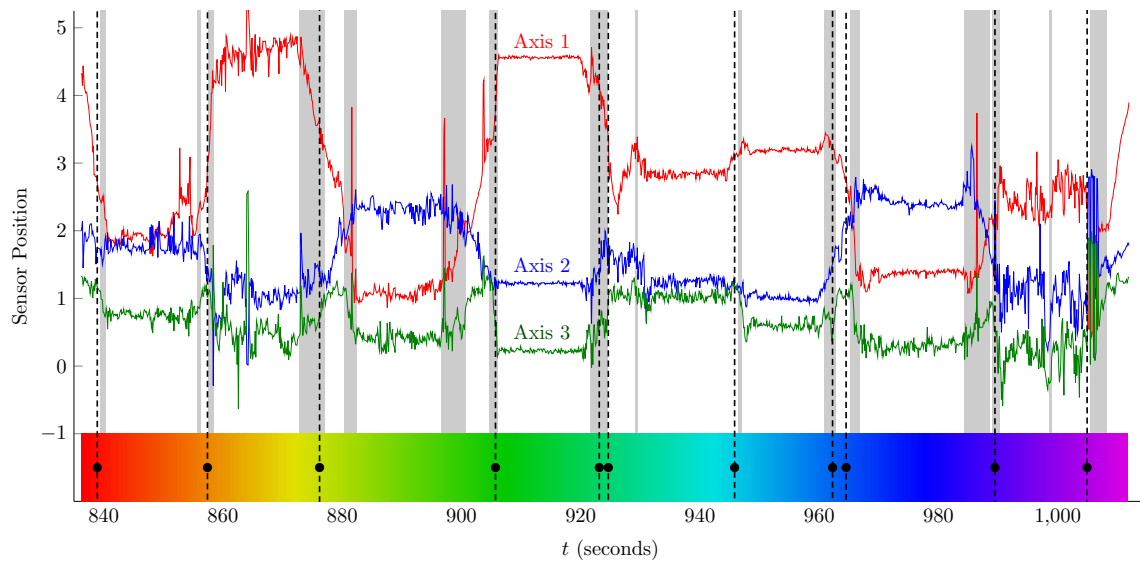


Figure 13. Detailed plot of the data presented in Figure 11, represented in terms of recorded sensor values along each axis. Gray zones are identified by Kaluža et al. as ‘transitional periods’ (such as falling or sitting down), while dashed lines indicated change-points detected by the CPR algorithm, given a multivariate normal model.

1 despite segments each possessing distinct patterns of covariance. To precisely model sensor  
 2 data in cases of incremental movement, a multivariate regression approach would be more  
 3 appropriate. The decision of which approach to use depends largely on whether movements  
 4 are more commonly expected to be abrupt or gradual. Because the multivariate Gaussian  
 5 distribution is simpler (having no slope parameter to consider), it is more sensitive to abrupt  
 6 changes.

7 An alternative analysis is presented in Figure 14, this time using a multiple linear re-  
 8 gression model (also described in the Supplement). Rather than emphasize static positions,  
 9 this analysis instead captures overall drift through space. Again, the dashed lines indicate  
 10 detected change-points, which overlap even more closely with the gray ‘transition’ periods  
 11 identified by Kaluža et al.. However, some transitions (such as the small shift at approx-  
 12 imately 945 seconds) are missed by this analysis. In general, because regression models  
 13 have more free parameters than multivariate step functions, they can be expected to be less  
 14 sensitive to small changes, as the MML’s penalties for model complexity may overwhelm  
 15 real but subtle discontinuities. Although the multiple linear regression performs well, the  
 16 simpler multivariate normal model may be better suited to the goals laid out by Kaluža et  
 17 al., particularly with respect to detecting abrupt falls.

18 The above analyses are based on the reported movements of a single sensor, but  
 19 Kaluža et al. report findings for participants wearing four sensors simultaneously on different  
 20 parts of the body. On the one hand, more data should permit an analyst to better distinguish  
 21 true movements from sensor noise. On the other hand, however, integrating inputs across  
 22 sensors that are not perfectly synchronized requires additional processing. Although a full  
 23 treatment of this problem is beyond the scope of this example, it is important to note  
 24 that dramatically expanding the dimensionality of the sample space makes detecting subtle

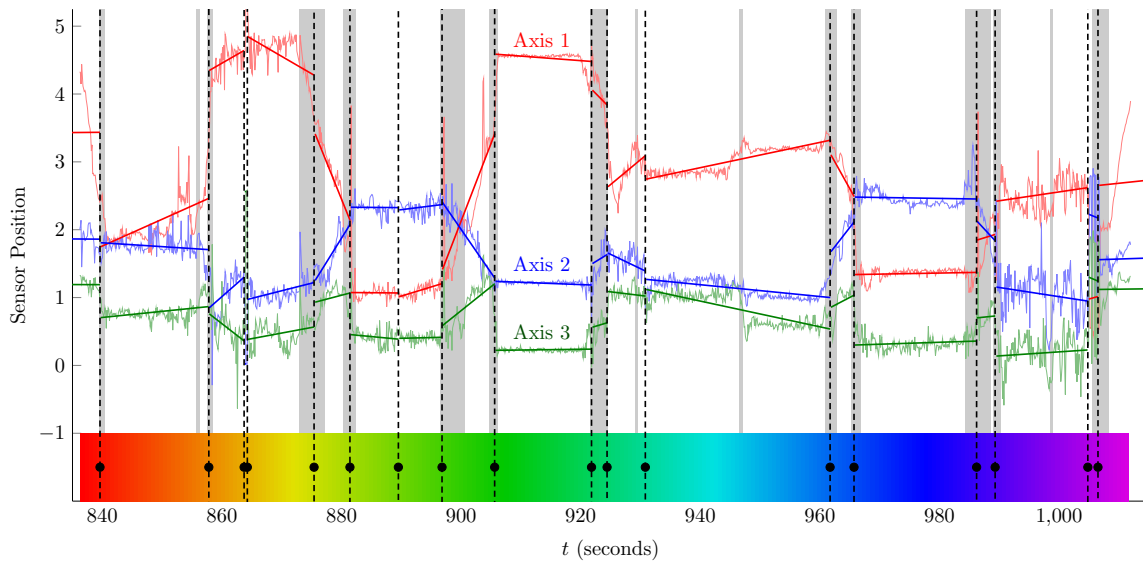


Figure 14. Detailed plot of the data presented in Figure 11, re-analyzed using a multiple linear regression model. The thin, pale lines represent the raw data, while the thick lines correspond to the best-fitting linear fits. Gray zones are identified by Kaluža et al. as ‘transitional periods’ (such as falling or sitting down), while dashed lines indicated change-points detected by the CPR algorithm, given a multiple regression model.

1 changes more difficult. Provided that synchronized estimates be obtained (for example by  
 2 computing means for each parameter within regular windows of time), an effective solution is  
 3 to use Principal Component Analysis (PCA) to reduce the sample space. The data reported  
 4 by Kaluža et al. could be interpreted as having 12 dimensions (3 spatial dimensions for each  
 5 of four sensors), but much of the information from the sensors was redundant, such that  
 6 the first three components of the PCA were consistently able to explain over 90% of the  
 7 variance, and were largely indistinguishable from the single-sensor analysis reported here.

8 Although tracking positions in space over time is an obvious application of multivari-  
 9 ate time-series analysis, many other forms of data can be illuminated using this approach.  
 10 Other measures, such as acceleration, could be used to distinguish between types of behav-  
 11 ior (fixation vs. saccade in eye-tracking, for example). Biological measures such as heart  
 12 rate and blood pressure could be examined, either continuously or in a longitudinal fash-  
 13 ion. Because these basic forms of multivariate analysis are straightforward to implement,  
 14 change-point analysis opens behavior analysis up to tasks and measures previously limited  
 15 to fields with a stronger engineering focus, such as machine learning and computer vision.

16 **Conclusions**

17 Despite the best efforts of experimentalists to build simple theories, empirical data  
 18 remain complicated and discontinuous. Precise experimental control in laboratory exper-  
 19 iments remains crucial, but it is often the processes being studied themselves that are a  
 20 source of frustrating inconsistency. The traditional approach of averaging across subjects  
 21 and across situations is perhaps most problematic for asking “when” questions, such as  
 22 “When did learning occur?” Different participants often learn at different rates, or expe-

1 rience epiphanies at different times, and an researcher interested in the characteristics of  
2 those moments of learning is not well-served by the smearing effect of an averaged learning  
3 curve. Averaging over time becomes more obviously absurd when asking “When did the  
4 man fall down?” Change-point analysis is an important framework for addressing these  
5 questions, and for moving theory away from indiscriminate averaging.

6 The Conjugate Partitions Recursion algorithm for change-point analysis, and the  
7 broader Bayesian strategy of binary partitioning by marginal model likelihood, provide  
8 tools that make non-stationary time-series analysis practical for use by applied researchers.

### 9 Acknowledgements

10 The author wished to thank Peter Balsam, Alina Bica-Huiu, Niall Bolger, Sy-Miin  
11 Chow, Daniel Fürth, Charles Gallistel, and Dobromir Rahnev for their feedback and com-  
12 mentary.

### 13 References

- 14 Bauwens, L., & Rombouts, J. V. K. (2012). On marginal likelihood computations in change-point  
15 models. *Computational Statistics and Data Analysis*, *56*, 3415–3429.
- 16 Bélisle, P., Joseph, L., MacGibbon, B., Wolfson, D. B., & du Berger, R. (1998). Change-point  
17 analysis of neuron spike train data. *Biometrics*, *54*, 113–123.
- 18 Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, *1*, 385–402.
- 19 Berger, J., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction.  
20 *Journal of the American Statistical Association*, *91*, 109–122.
- 21 Bernardo, J. M. (2005). Reference analysis. In D. K. Dey & C. R. Rao (Eds.), *Handbook of statistics*  
22 *25: Bayesian thinking, modeling, and computation* (pp. 17–90). Elsevier.
- 23 Campbell, N. A. (1980). Robust procedures in multivariate analysis i: Robust covariance estimation.  
24 *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *29*, 231–237.
- 25 Carlin, B. P., & Chib, S. (1995). Bayesian model choice vis markov chain monte carlo methods.  
26 *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 473–484.
- 27 Carlin, B. P., & Louis, T. A. (2000). Empirical bayes: Past, present, and future. *Journal of the*  
28 *American Statistical Association*, *95*, 1286–1289.
- 29 Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*,  
30 *39*, 83–87.
- 31 Chen, J., & Gupta, A. K. (Eds.). (2011). *Parametric statistical change point analysis*. Birkhäuser  
32 Boston.
- 33 Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econo-*  
34 *metrics*, *86*, 221–241.
- 35 Cohen, P. (Ed.). (2008). *Applied data analytic techniques for turning points research*. Taylor &  
36 Francis Group.
- 37 Das, S. R., Duffie, D., Kapadia, N., & Saita, L. (2007). Common failings: How corporate defaults  
38 are correlated. *Journal of Finance*, *62*, 93–117.

- 1 Dyson, F. (2004). A meeting with enrico fermi. *Nature*, *427*, 297.
- 2 Efron, B. (2010). *Large-scale inference*. Cambridge University Press.
- 3 Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems.  
4 *Statistics and Computing*, *16*, 203–213.
- 5 Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.
- 6 Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative  
7 analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*,  
8 13124–13131.
- 9 Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (in press). The perception of  
10 probability. *Psychological Review*.
- 11 Gamerman, D., & Lopes, H. F. (2006). *Markov chain monte carlo: Stochastic simulation for bayesian*  
12 *inference* (2nd ed.). Chapman & Hall/CRC.
- 13 Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian*  
14 *Analysis*, *1*, 515–533.
- 15 Gelman, A. (2008). Objections to bayesian statistics. *Bayesian Analysis*, *3*, 445–450.
- 16 Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman  
17 & Hall/CRC.
- 18 Girón, F. J., Moreno, E., & Casella, G. (2007). Objective bayesian analysis of multiple changepoints  
19 for linear models. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, & A. P. David (Eds.), *Bayesian*  
20 *statistics 8* (pp. 1–27). Oxford University Press.
- 21 Goldstein, M. (2006). Subjective bayesian analysis: Principles and practice. *Bayesian analysis*, *1*,  
22 403–420.
- 23 Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model  
24 determination. *Biometrika*, *82*, 711–732.
- 25 Han, C., & Carlin, B. P. (2001). Markov chain monte carlo methods for computing bayes factors.  
26 *Journal of the American Statistical Association*, *96*, 1122–1132.
- 27 Hannart, A., & Naveau, P. (2012). An improved bayesian information criterion for multiple change-  
28 point models. *Technometrics*, *54*, 256–268.
- 29 Hansen, B. E. (2001). The new econometrics of structural change: Dating breaks in u.s. labor  
30 productivity. *Journal of Economic Perspectives*, *15*, 117–128.
- 31 Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an  
32 exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.
- 33 Jeffreys, H. (1946). An invariant form for the prior probability in estimating problems. *Proceedings*  
34 *of the Royal Society of London, Series A, Mathematical and Physical Sciences*, *186*, 453–461.
- 35 Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Clarendon Press.
- 36 Jensen, G., Altschul, D., Danly, E., & Terrace, H. S. (2013). Transfer of a spatial representation of  
37 two distinct serial tasks by rhesus macaques. *PLOS ONE*, *8*, e70825.

- 1 Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M., & Gams, M. (2010). An agent-based ap-  
2 proach to care in independent living. In B. de Ruyter et al. (Eds.), *Ambient intelligence: First*  
3 *international joint conference, 2010* (pp. 177–186). Springer Berlin Heidelberg.
- 4 Kaplan, A. (1998). *The conduct of inquiry: Methodology for behavioral science*. Transaction Pub-  
5 lishers.
- 6 Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*,  
7 *90*, 773–795.
- 8 Kheifets, A., & Gallistel, C. R. (2012). Mice take calculated risks. *Proceedings of the National*  
9 *Academy of Sciences of the United States of America*, *109*, 8776–8779.
- 10 Klauenberg, K., & Elster, C. (2012). The multivariate normal mean - sensitivity of its objective  
11 bayesian estimates. *Metrologia*, *49*, 395–400.
- 12 Kline, R. B. (1998). *Principles and practice of structural equation modeling*. Guilford.
- 13 Koop, G., & Potter, S. M. (2009). Prior elicitation in multiple change-point models. *International*  
14 *Economic Review*, *50*, 751–772.
- 15 Lindquist, M. A., Waugh, C., & Wager, T. D. (2007). Modeling state-related fmri activity using  
16 change-point theory. *NeuroImage*, *35*, 1125–1141.
- 17 Lio, P., & Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins.  
18 *Bioinformatics*, *16*, 376–382.
- 19 Longstaff, F. A., & Rajan, A. (2008). An empirical analysis of pricing in collateralized debt obliga-  
20 tions. *Journal of Finance*, *63*, 529–563.
- 21 Mandelbrot, B. B., & Hudson, R. L. (2006). *The (mis)behavior of markets*. Basic Books.
- 22 Meligkotsidou, L., & Dellaportas, P. (2011). Forecasting with non-homogeneous hidden markov  
23 models. *Statistics and Computing*, *21*, 439–449.
- 24 Myung, J., & Pitt, M. A. (1997). Applying occam’s razor in modeling cognition: A bayesian  
25 approach. *Psychonomic Bulletin & Review*, *4*, 79–95.
- 26 Oakley, J. E., & O’Hagan, A. (2007). Uncertainty in prior elicitation: A nonparametric approach.  
27 *Biometrika*, *94*, 427–441.
- 28 Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of*  
29 *Experimental Psychology: Learning, Memory, and Cognition*, *23*, 324–354.
- 30 Pratchett, T. (1983). *The colour of magic*. Colin Smythe.
- 31 Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalised  
32 linear models. *Biometrika*, *82*, 251–266.
- 33 Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate  
34 School of Business Administration, Harvard University.
- 35 Robert, C. P., Rydén, T., & Titterton, D. M. (2000). Bayesian inference in hidden markov models  
36 through the reversible jump markov chain monte carlo method. *Journal of the Royal Statistical*  
37 *Society, Series B*, *62*, 57–75.

- 1 Samaniego, F. J. (2012). *A comparison of the bayesian and frequentist approaches to estimation*.  
2 Springer New York.
- 3 Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in  
4 mixed models. *Behavioral Ecology*, *20*, 416–420.
- 5 Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
- 6 Silver, N. (2012). *The signal and the noise*. Penguin Press.
- 7 Sisson, S. A. (2005). Transdimensional markoc chains. *Journal of the American Statistical Associ-*  
8 *ation*, *100*, 1077-1089.
- 9 Terrace, H. S. (2005). The simultaneous chain: A new approach to serial learning. *Trends in*  
10 *Cognitive Sciences*, *9*, 202–210.
- 11 Van Dongen, S. (2006). Prior specification in bayesian statistics: Three cautionary tales. *Journal*  
12 *of Theoretical Biology*, *242*, 90–100.
- 13 Vostrikova, L. J. (1981). Detecting ‘disorder’ in multidimensional random processes. *Soviet Mathe-*  
14 *matics Doklady*, *24*, 55–59.
- 15 Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic*  
16 *Bulletin & Review*, *14*, 779–804.
- 17 Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists  
18 must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal*  
19 *of Personality and Social Psychology*, *100*, 426–432.
- 20 Wasserman, L. (2000). Bayesian model election and model averaging. *Journal of Mathematical*  
21 *Psychology*, *44*, 92–107.
- 22 Western, B., & Kleykamp, M. (2004). A bayesian change point model for historical time series  
23 analysis. *Political Analysis*, *12*, 354–374.
- 24 Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482.
- 25 Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology  
26 journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.
- 27 Zhang, Z., Hamagami, F., Wang, L. L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis  
28 of longitudinal data using growth curve models. *International Journal of Behavioral Development*,  
29 *31*, 374–383.