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 adding divisions (and thus increasing model complexity) yields 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 distribution (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

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

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

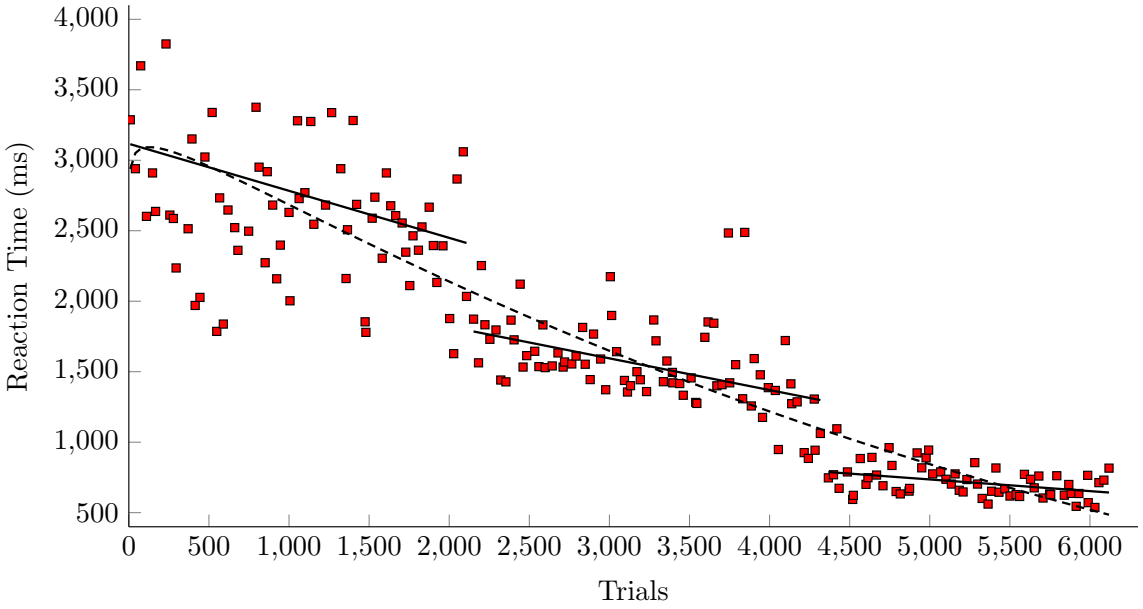

Figure 1. Reaction times from a single subject learning a psychophysical task, originally reported by Palmeri (1997). The dashed line corresponds to a four-parameter "learning curve," reported by Heathcote et al. (2000), while the solid lines interpret the same data as approximately linear, with two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

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

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

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

considered in more detail in a later section, but for now, they demonstrate an important issue in time-series analysis: Because different participants learn at different rates, and have insights at different times, group averaging conceals these moments and typically paints the erroneous picture of learning being gradual and continuous (a point argued in detail by Gallistel et al., 2004).

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

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

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

A Brief Review

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

Bayes' Theorem: A Machine For Priors

To answer the question, "Should these data be partitioned by a change-point?" the partitioning algorithm begins with Bayes' Theorem. Given a prior assumption that the

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

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

where

$$m(x, M) = \int_{\theta} f(x|\theta, M)\Pr(\theta, M)\, d\theta \tag{1}$$

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

Practical application of Bayes' Theorem requires a formal specification of prior assumptions. The shape of the prior distribution $\Pr(\theta, M)$ represents the existing evidence available to the analyst. Even when no empirical observations are available, this prior must be specified. A typical approach is to imagine a distribution derived from "hypothetical" observations. For example, if x consists of a series of die rolls, resulting in the numbers one through six, one might ask the question, "How likely is this die to roll a six?" If the die is fair[1], it should show six with a probability of $\frac{1}{6}$, so we might imagine five hypothetical failures and one hypothetical success. These hypothetical successes and failures are our *prior hyperparameters*: hypothetical failures are denoted by α_0 and hypothetical successes by α_1. Thus, we might begin by assuming $\alpha_0 = 5$ and $\alpha_1 = 1$.

This is an example of a 'subjective' prior, because the values for α_0 and α_1 are left to the analyst's discretion. Given high confidence in a fair die, a stronger prior might use $\alpha_0 = 25$ and $\alpha_1 = 5$, the equivalent of thirty hypothetical observations (which, again, are assumed before the die is ever actually rolled). Alternatively, a more agnostic (or "weak") prior might use $\alpha_0 = 1$ and $\alpha_1 = 1$. The weaker the prior, the less its assumptions influence the subsequent calculation. Because $\Pr(\theta, M)$ is a probability distribution, we do not have a 'known' value for θ. Instead, the relative weight of the possible values of θ are described by a beta distribution whose shape is defined by the prior hyperparameters, as depicted in Figure 2 (Left). As implied by the figure, the beta distribution supports the interval between 0 to 1, and represents our uncertainty about a simple probability. The thick red line represents the 'weak' prior, and it assigns even odds to any probability over the interval. The two stronger priors have converged somewhat towards the expected probability of a fair die, and these reflect greater confidence that $p(\text{six}) = \frac{1}{6}$.

[1]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.

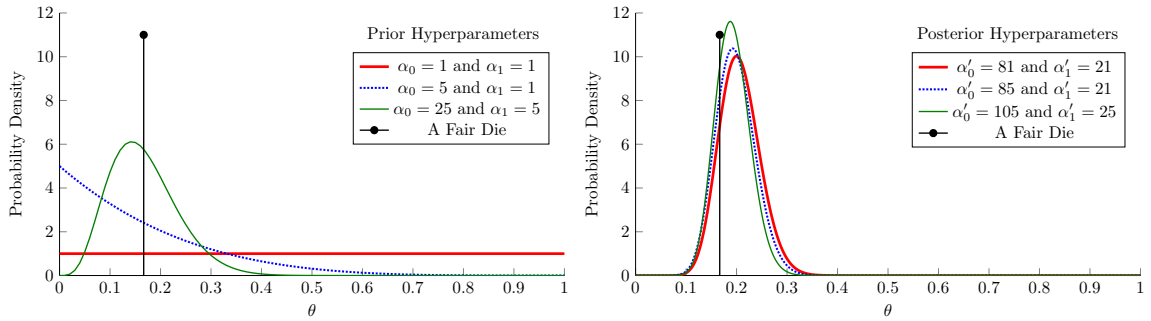

Figure 2. Prior hyperparameters α_0 and α_1 (left) and posterior hyperparameters α'_0 and α'_1 for the beta distribution given three different prior distributions and the same empirical observations. See Equation 2 and the Supplement for details.

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

On the one hand, this example demonstrates that the choice of the prior initially has a dramatic effect on the hypothesized distribution of θ before any data was collected (given how different the curves in Figure 2 (Left) appear). On the other hand, the influence of the prior diminishes towards infinitesimal as additional observations are made. When using a subjective Bayesian approach[2], many analysts prefer "weak" prior, such as $\alpha_0 = \alpha_1 = 1$, because such priors have an "anything could happen" flavor and do not have a determining effect on the shape of the posterior distribution.

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

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

[2]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.

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

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

Conjugate Priors

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

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

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

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

Here, α_0' and α_1' are the hyperparameters previously identified, based on combining observed and hypothetical successes and failures. Because both depend only on the data and not on the parameter θ, it follows that $m(x, \text{binom}) = \text{Beta}(\alpha_0', \alpha_1')$, which is trivial to evaluate.

Since the publication of Raïffa & Schlaifer (1961), conjugate priors have been identified for a wide range of probability distributions. In cases where closed-form solutions exist for $m(x, M)$, an analyst can use the sufficient statistics[3] describing the observations x and a set of prior hyperparameters α_i to estimate the corresponding posterior hyperparameters α_i'; those posterior hyperparameters can be used to evaluate $m(x, M)$.

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

Subjective vs. Objective Bayes

Becoming acquainted with theoretical and applied Bayesian statistics can be intimidating for a variety of reasons. A major impediment for Bayesian neophytes is the ongoing debate over "subjective" vs. "objective" Bayesian methods (Goldstein, 2006; Berger, 2006, respectively).

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

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

[3]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 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.

[4]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.

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

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

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

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

Insisting on a strict adherence to an idealized standard of ignorance is often not a sensible position if doing so means entertaining obviously absurd hypotheses. Although the CPR algorithm may be implemented using objective priors, doing so is likely to be computationally intensive, or to yield nonsensical results, or both, particularly when considering very small subsets near the edges of the data. Most empirical data (being collected within practical and theoretical constraints) are better-served by a reasonable weakly informative subjective prior (Van Dongen, 2006). The challenge is to distinguish reasonable priors from unreasonable ones.

Subjective Prior Selection

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

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

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

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

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

Maximum Likelihood, Marginal Model Likelihood, and Bayes Factors

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

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

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

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

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

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

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

Figure 3 conveys this intuition visually. Given 20 observed binary outcomes, it is straightforward to calculate the likelihood of obtaining 10 successes and 10 failures, and to estimate that the overall probability of success θ is most likely to be 0.5 (Figure 3 Left). However, splitting the data into the first ten observations vs. the second ten (perhaps to test whether a change-point divides them) entails an increase in model complexity, because two different parameters θ_1 and θ_2 are now needed. If splitting the data results in two subsets of data that display a 6:4 ratio of success, then the split increases the maximum likelihood but reduces the marginal model likelihood (Figure 3 Center). Under these circumstances, it would be unjustified to split the data. However, if the split reveals an 8:2 ratio of success, a dramatic increase in the marginal model likelihood is observed (Figure 3 Right). When comparing these three scenarios, a two-parameter model is only justified when its volume under the curve is larger than that of the one-parameter model, both of which are obtained by integrating across possible values for the parameters.

The function $m(x, M)$ in Equation 1 is precisely such an integral, and its value may either be solved for (if a closed-form solution exists) or estimated numerically. If $m(x, M)$ can be estimated, a *Bayes factor* (Kass & Raftery, 1995) can be computed. Bayes factors can be interpreted as the ratio of support for $M1$ relative $M2$, regardless of the MLE parameters associated with either model. They may be calculated by comparing the MML of two models:

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

The MML is computed for two models, $M1$ and $M2$, given the observations x. When $M1$

[6]Formally, the MML maximizes its Lebesgue measure with respect to the specified prior.

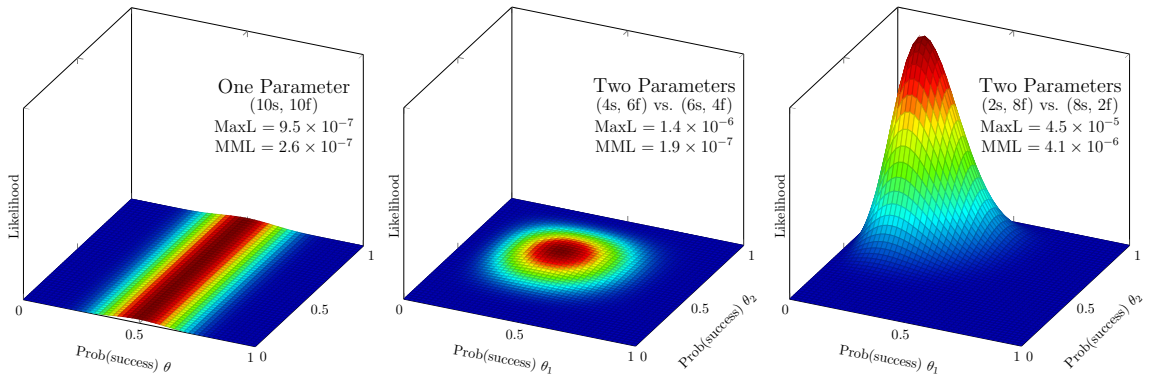

Figure 3. Likelihood as a function of the probability θ for the binomial distribution, given 20 observations, consisting of 10 success and 10 failures. Darker areas indicate higher relative likelihoods. (Left Panel) When a single parameter is used to describe the data, the maximum likelihood estimator 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.

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

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

The Conjugate Partitioned Recursion Algorithm

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

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

Binary Partition by Marginal Model Likelihood

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

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

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

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

Determining Whether To Partition The Data

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

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

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

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

[7] Here, the intervals between observations are presumed to be uniform; the non-uniform case is discussed below

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

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

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

$$K = \frac{m\left(x, C_1\right)}{m\left(x, C_0\right)} \tag{6}$$

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

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

$$\frac{p_1'}{p_0'} = K \cdot \frac{p_1}{p_0} = K \cdot \frac{\text{Poiss}\left(1, p_c\left(n-1\right)\right)}{\text{Poiss}\left(0, p_c\left(n-1\right)\right)} = K \cdot p_c\left(n-1\right) \tag{7}$$

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 supporting either model is approximately equal. It is important to model the odds of *exactly* one change-point (rather than, say, the odds of any number of changes) to match the MML $m\left(x, C_1\right)$. Because $m\left(x, C_{1,2,\ldots}\right)$ is computationally prohibitive, the binary partition strategy only considers single change hypotheses at each stage of its recursion, and this approach must be consistent when setting $\frac{p_1}{p_0}$.

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

It is advisable to specify a *decision criterion* $\tau > 1$, and to partition the data only when $\frac{p_1'}{p_0'} > \tau$. Because binary partitioning is a recursive process, data with many change-points must be divided into many small, noisy segments with some percentage of false positives. This introduces a stopping problem: Each false positive drives further subdivision of the

[8]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."

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

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

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

Determining Where To Partition The Data

In the event that $\frac{p'_1}{p'_0} > \tau$, the strength of the evidence supporting each possible change-point must be compared. To do this, we may rewrite Equation 6 in the following way:

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

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

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

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

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

Non-Uniform Event Times

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

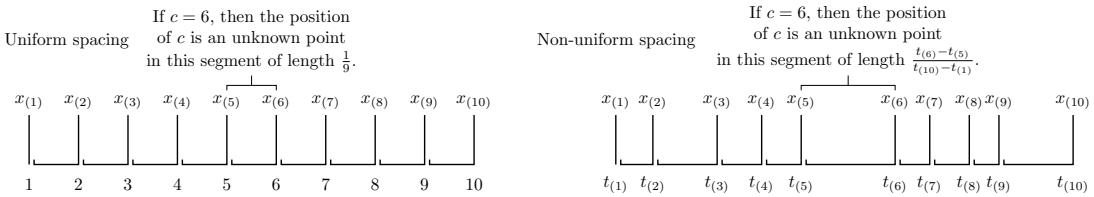

Figure 4. An example of uniform vs. non-uniform segmentation. (Left) In the uniform case, each segment of time is an equally strong candidate for a putative change-point. In this example, the nine segments each have a common weight of $\frac{1}{9}$. (Right) In the non-uniform case, a change-point is 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.

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

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

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

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

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

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

Biased MML Estimation in Small Samples

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

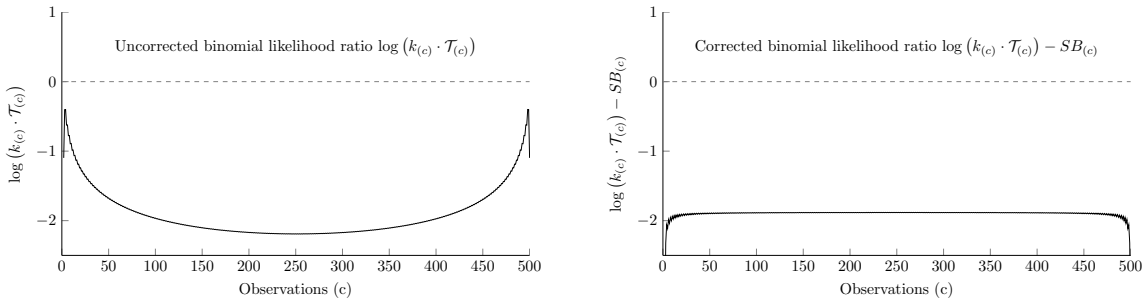

Figure 5. Values of $\log\left(k_{(c)} \cdot \mathcal{T}_{(c)}\right)$ given 100 binary observations $[0, 1, 0, 1, \cdots, 0, 1]$ both uncorrected (left) and corrected (right). The dashed line represents the boundary between evidence favoring a change at time (positive) vs. evidence against (negative)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

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

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

The degree of difference is approximated by the SBIC (Schwarz, 1978), which estimates the marginal likelihood:

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

From this approximation, it follows that:

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

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

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

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

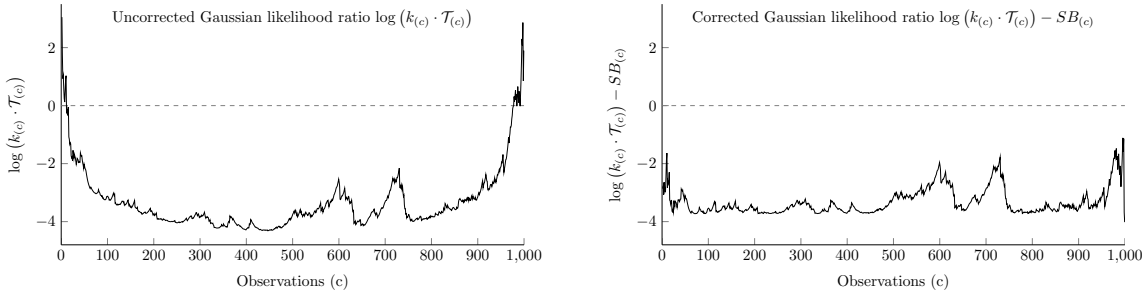

Figure 6. Values of $\log\left(k_{(c)} \cdot \mathcal{T}_{(c)}\right)$ given 1000 randomized Gaussian observations ($\mu = 0, \sigma = 1$), representing data in the range $\pm z = 3.3$ both uncorrected (left) and corrected (right). The dashed line represents the boundary between evidence favoring a change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

val:

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

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

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

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

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

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

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

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

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

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

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

Implementing Binary Partition by Marginal Model Likelihood

Given that the posterior odds ratio $\frac{p_1'}{p_0'}$ (Equation 7) indicates *whether* to partition the data and that the peak weighted odds ratio $\left[\frac{k_{(c)} \cdot \mathcal{T}_{(c)}}{\exp\left(SB_{(c)}\right)}\right]$ (Equation 11) indicates *where* to partition, the full description of the binary partitioning strategy is specified in Algorithm 1. The algorithm begins with an array $\mathbf{M}$ containing two indices, $\langle 0, n \rangle$; since these delimit the full span of the data, this array can be said to contain no change-points. Consequently, the number of change-points in the model is $(\text{length}\left(\mathbf{M}\right) - 2)$. The algorithm then tests whether to partition the data using Equation 7. If the posterior odds ratio exceeds the decision criterion τ, then the best available candidate is selected using Equation 11. This estimated change-point is added to $\mathbf{M}$, and the algorithm is then applied to each of the resulting segments in $\mathbf{M}$. This process iterates until no new change-points are identified.

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

[9] 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 criteron τ, 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* *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**
        $k_c \leftarrow \frac{m\left(x_{(i:c-1)}, C_0\right) m\left(x_{(c:j)}, C_0\right)}{m\left(x_{(i:j)}, C_0\right)}$ /* Eq. 8 */
        $\mathbf{K}_{(c)} \leftarrow \frac{k_c \cdot \mathcal{T}_{(c)}}{\exp\left(SB_{(c)}\right)}$ /* Eq. 15 */
      if *sum* $\left(\mathbf{K}_{(i+1:j)}\right) \cdot p_c \cdot (j - i) > \tau$ **then**
        $\hat{c} \leftarrow \text{index}(\max\left(\mathbf{K}_{(i+1:j)}\right))$-1 /* if Eq. 7 permits, find $\hat{c}$ */
        Push$(\mathbf{N}, \hat{c})$ /* insert $\hat{c}$ into N */
    Push$(\mathbf{M}, \mathbf{N})$; Sort$(\mathbf{M})$ /* merge N into M and sort */
  until *length*$(\mathbf{N}) = 0$
  for $s = 1$ *to* *length* $(\mathbf{M}) - 1$ **do**
    $\mathbf{P}_{(s)} \leftarrow \text{EstimateParameters}\left(x_{(\mathbf{M}_{(s)}+1:\mathbf{M}_{(s+1)})}\right)$;
  return $\mathbf{M}, \mathbf{P}$

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

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

Pitfalls & Considerations

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

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

The Law of the Instrument

In any analytic enterprise, it is important to recall (and resist) the law of the instrument:

> In addition to the social pressures from the scientific community there is also at work a very human trait of individual scientists. I call the *the law of the instrument*, and it may be formulated as follows: Give a small boy a hammer, and he will find that everything he encounters needs pounding.
> -Kaplan (1998), p. 28.

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

Insensitivity Given Large-Scale Stationary Processes

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

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

Assumption-Free Change-Point Analysis

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

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

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

Specifying Prior Odds of a Change

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

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

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

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

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

Gradual Transitions

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

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

Multiple Time Series and Shared Parameters

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

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

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

Multiple Model Types

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

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

A multi-model CPR algorithm that can simultaneously evaluate these alternatives is beyond the scope of this paper. Such an approach also potentially invites 'fishing expeditions' in which different combinations of distributions are intermixed and only those that are favorable to the research are reported. However, given the building blocks provided in this manuscript, an analyst who has sufficient theoretical justification can assemble a custom variant suited to their empirical scenario. Such variants would need to be evaluated 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:

$$\begin{aligned} resp &= 00120101112224051532032355355552155555552 \\ b_1 &= 00110101111111011111011111111111111111111 \\ b_2 &= 00010000001111010111011111111111011111111 \\ b_3 &= 00000000000001010110010111111111001111110 \\ b_4 &= 00000000000001010100000011011111001111110 \\ b_5 &= 00000000000000010100000011011111001111110 \end{aligned}$$

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

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

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

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

Obtaining the marginal likelihood requires considering all possible change-points. With a prior $m\,(M) = \text{Beta}\,(1, 1)$ and the probability of a change $p_c = \frac{1}{39}$, the posterior odds ratio in support of a change was across all intervals was $\frac{p_1'}{p_0'} = \left[K \cdot \frac{39}{39}\right] = \sum \frac{k_{(c)}}{(n-1)\exp\left(SB_{(c)}\right)} =$

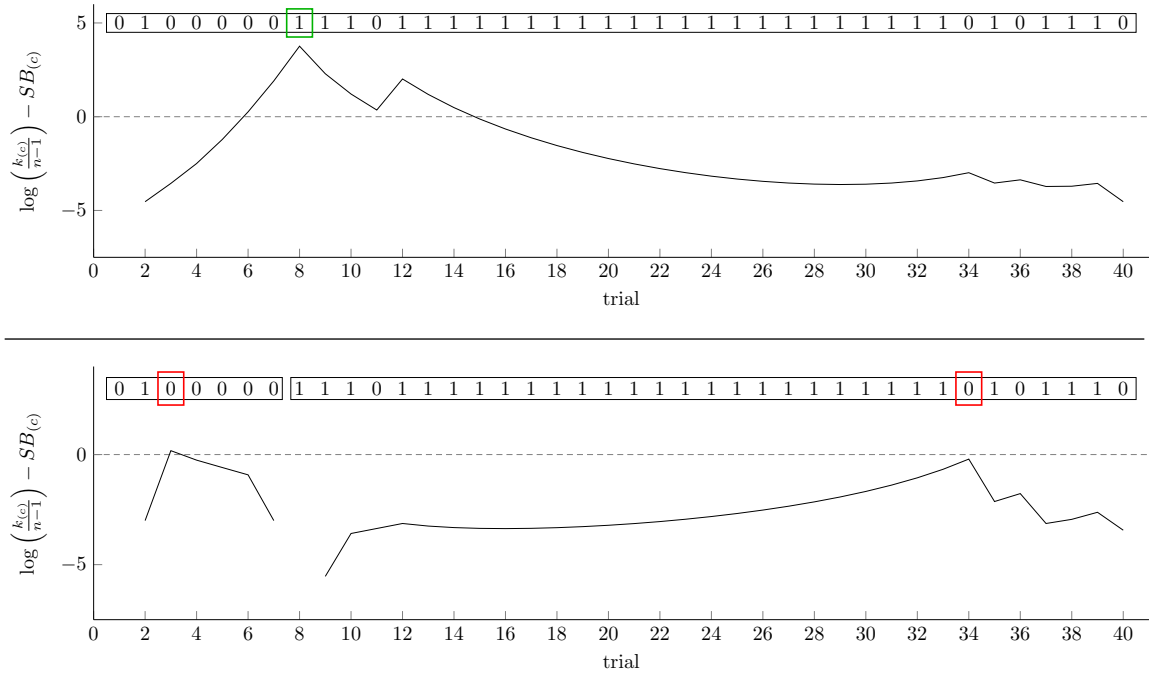

Figure 7. The CPR algorithm applied to a binary string. (Top) The individual Bayes factors $k_{(c)}$ (here presented on a log scale) are calculated for each point, and there is a clear peak when $c = 8$. (Bottom) The algorithm is then recursively applied to the two segment halves, with weak support found 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.

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

Figure 7 (Bottom) represents the second iteration of the algorithm. In the first 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 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 ratio slightly favoring a change, but falling well below the decision criterion of $\tau = 10$.

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

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

[10] 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.

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

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

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

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

Curve Fitting: Reaction Times

When performing an analysis of continuous data measurements, there is a powerful temptation to fit the data to a distribution and work with the resulting summary statistics. Indeed, the lion's share of null hypothesis significance testing is based on the tails of inferred distributions. To quote Wagenmakers (2007), "p values depend on data that were never observed[11]." A change-point analysis that can 'digest' continuous data without throwing out information can complement (or substitute) traditional analyses. A particularly relevant case is the practice of averaged curve-fitting.

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

[11] However, a counterpoint might be that many Bayes factors depend on priors invented from whole cloth.

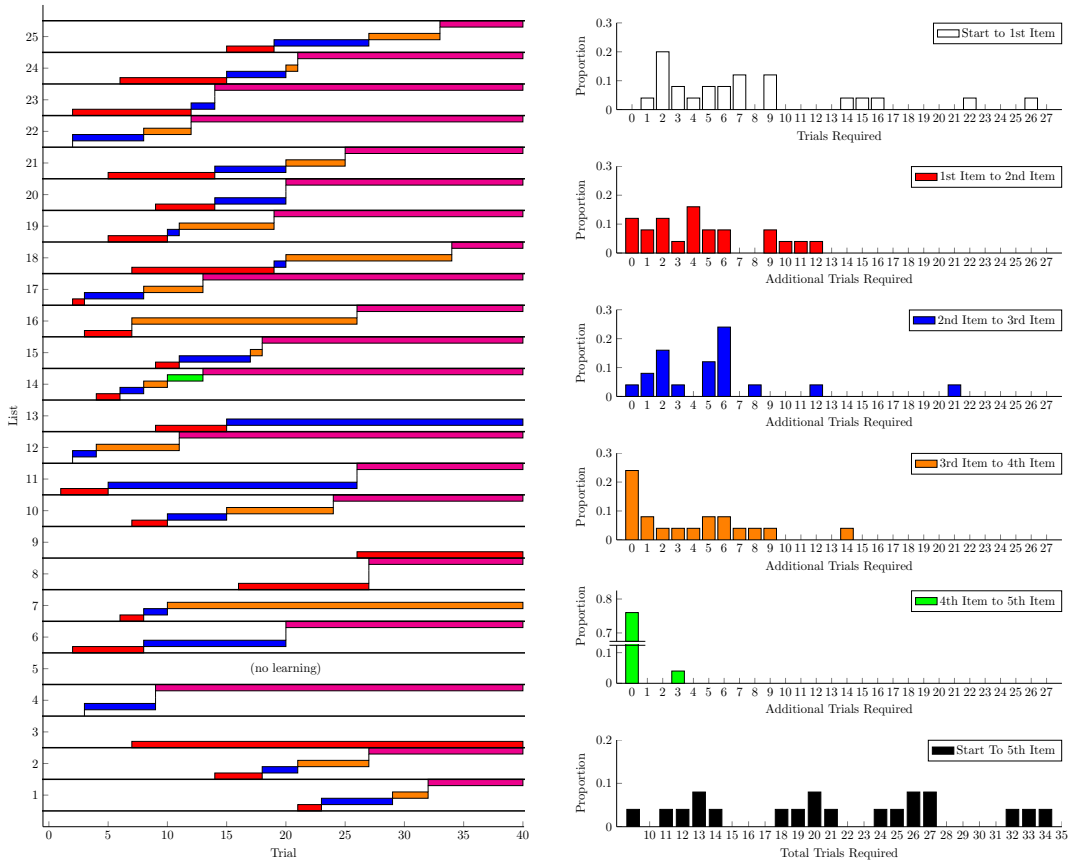

Figure 8. Plot of item acquisition in 25 different lists in Jensen et al. (2013), learned using the Simultaneous Chain procedure. (Left) Estimated time of acquisition for each list item is indicated both by elevation and by shade of gray. (Right) Histogram showing frequency of inter-acquisition periods, 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.

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

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

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

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

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

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

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

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

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

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

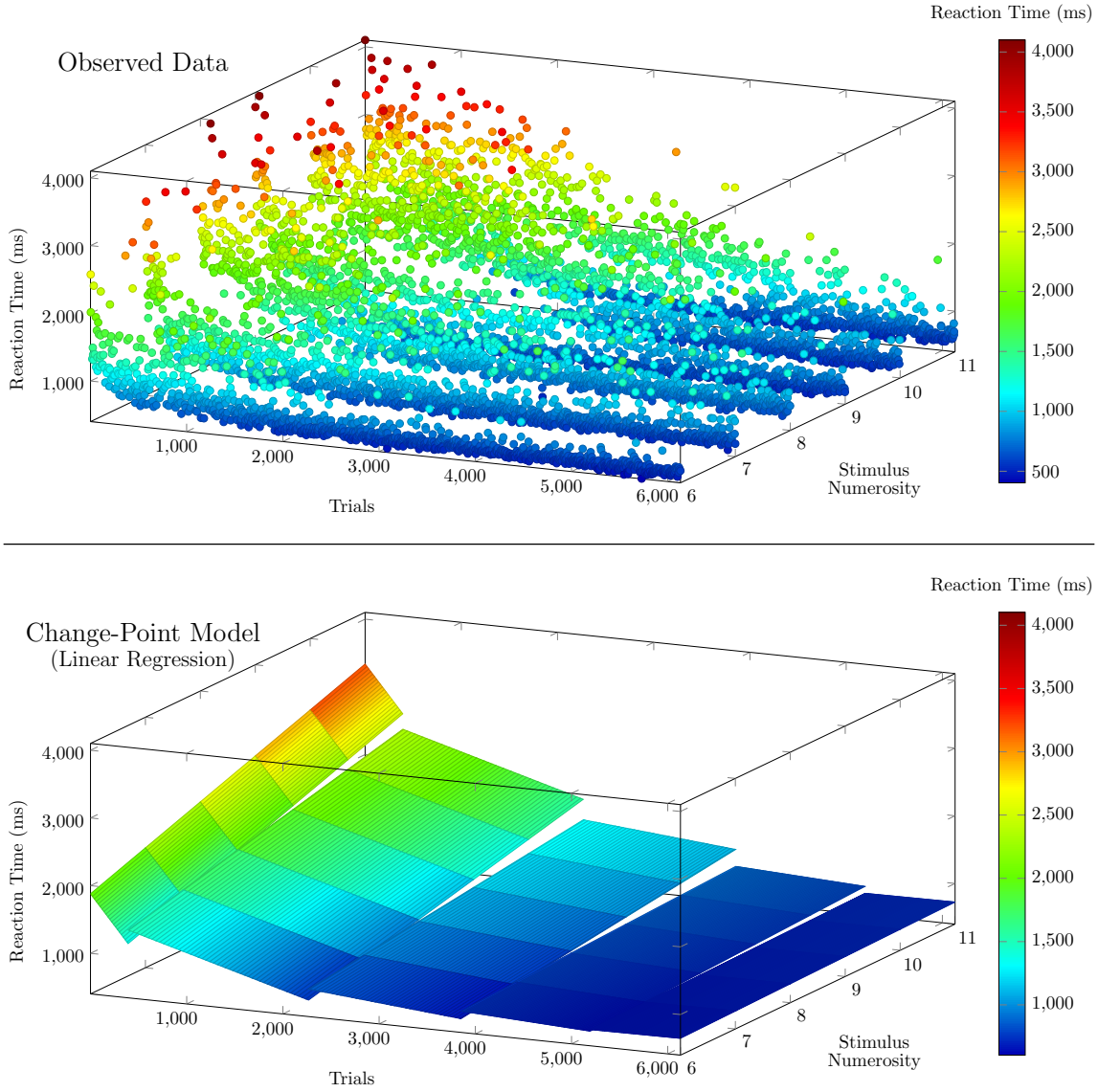

Figure 9. Reaction times as a function of trials and task difficulty in Palmeri (1997). (Top) Each of the 6132 reaction times, color-coded according to their speed and positioned with respect to trial and stimulus numerosity. (Bottom) The model fit resulting from a multiple regression, subdivided according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

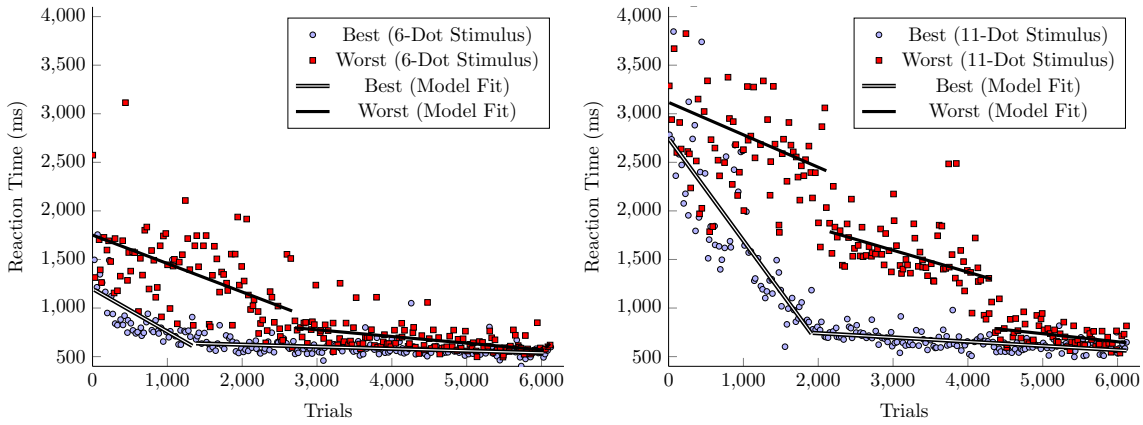

Figure 10. Reaction times as a function of trials in Palmeri (1997) for specific stimuli in a single participant. Lines represent Bayesian regression estimates whose subdivisions were determined using to the change-point algorithm. (Left) 6-dot stimulus learning for most rapidly acquired stimulus (blue)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

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

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

Multivariate Data: 3D Position Tracking

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

[13]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 ± 1 standard deviation.

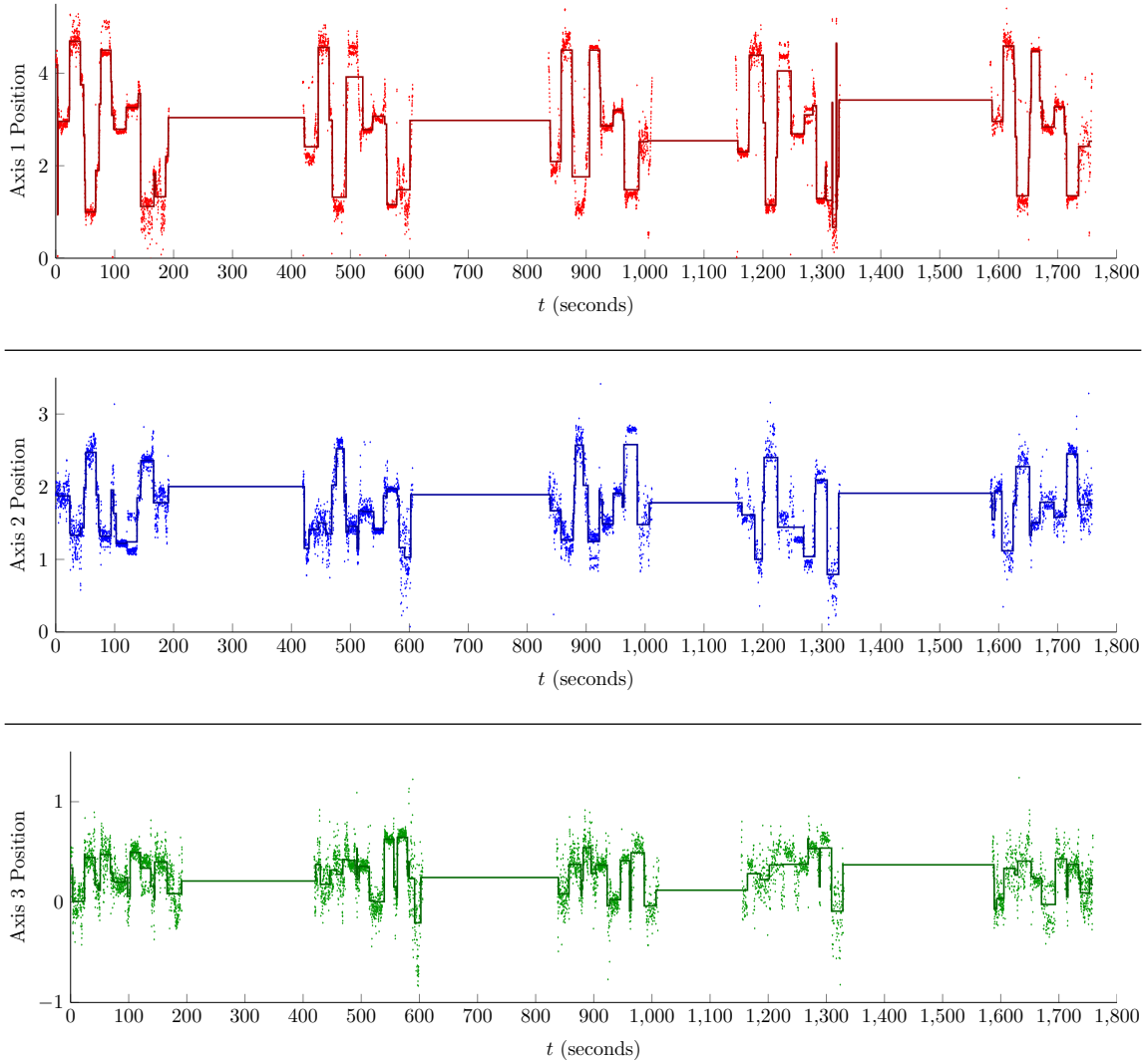

Figure 11. 3D position data collected by Kaluža et al. (2010). Individual points are multivariate, such that an observation an x-coordinate (in the top panel), a y-coordinate (in the center panel), and a z-coordinate (in the bottom panel). Change-points were identified using the CPR algorithm assuming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

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

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

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

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

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

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

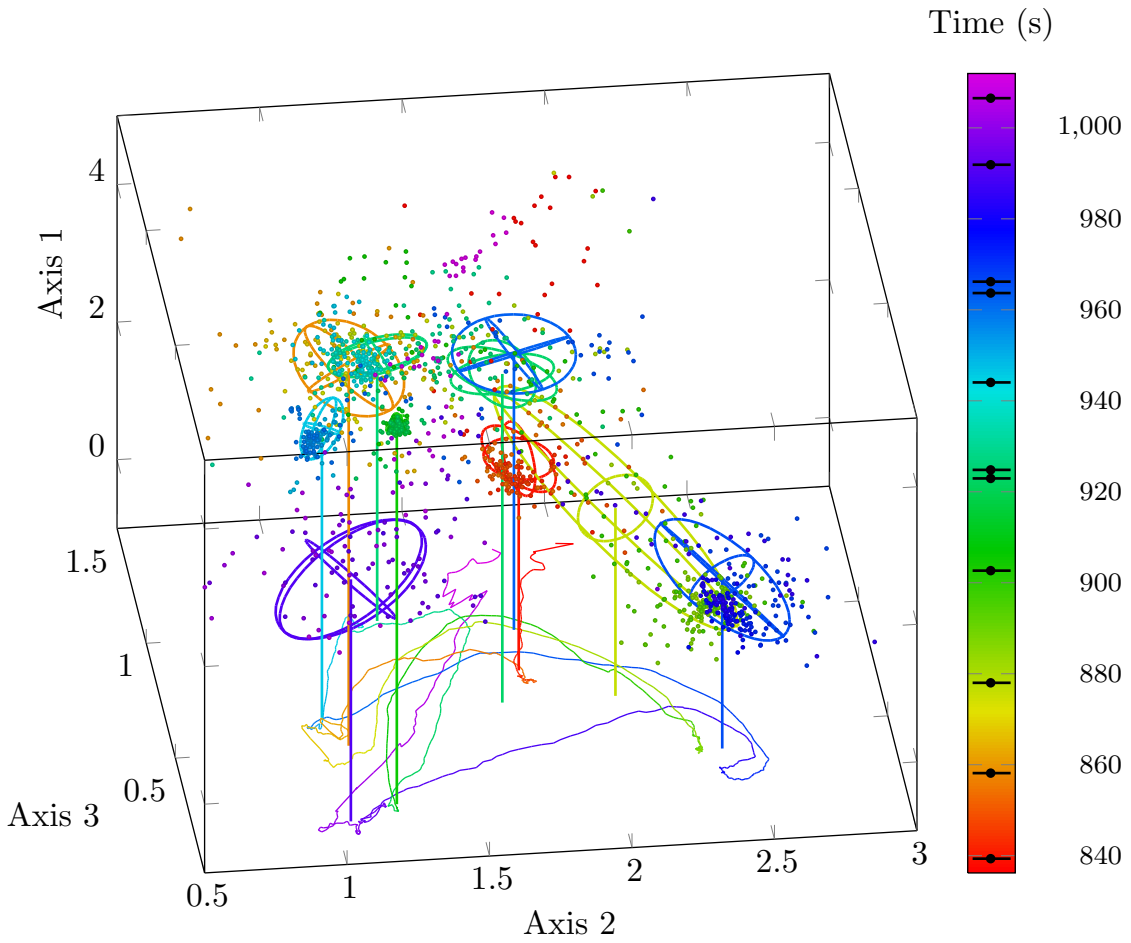

Figure 12. 3D depiction of the a subset of the movement data from Kaluža et al. (2010) and its associated change-point model. Individual points represent discrete observations and are color-coded continuously with respect to time, as noted on the color bar. The points on the color bar denote the times 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.

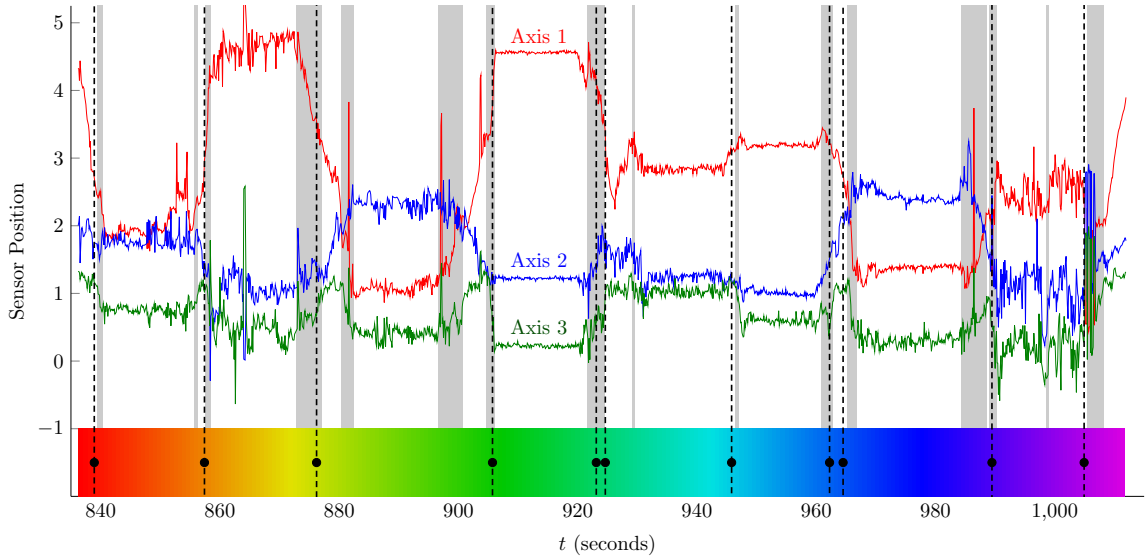

Figure 13. Detailed plot of the data presented in Figure 11, represented in terms of recorded sensor values alone each axis. Gray zones are identified by Kaluža et al. as 'transitional periods' (such as falling or sitting down), while dashed lines indicated change-points detected by the CPR algorithm,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

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

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

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

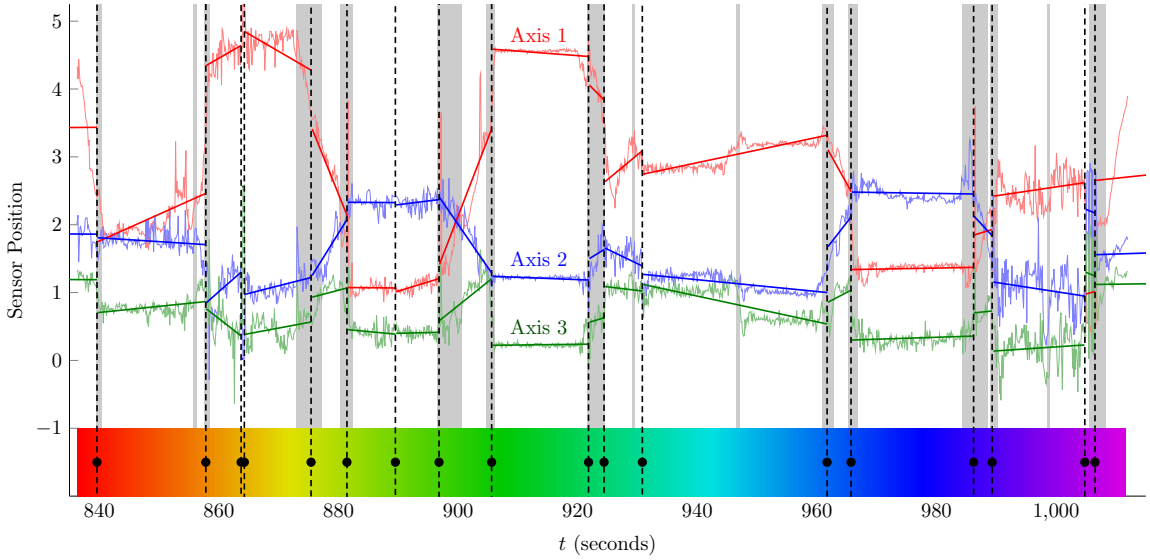

Figure 14. Detailed plot of the data presented in Figure 11, re-analyzed using a multiple linear regression model. The thin, pale lines represent the raw data, while the thick lines correspond to the best-fitting linear fits. Gray zones are identified by Kaluža et al. as 'transitional periods' (such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

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

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

Conclusions

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

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

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

Acknowledgements

The author wished to thank Peter Balsam, Alina Bica-Huiu, Niall Bolger, Sy-Miin Chow, Daniel Fürth, and Charles Gallistel for their feedback and commentary.

References

Bauwens, L., & Rombouts, J. V. K. (2012). On marginal likelihood computations in change-point models. *Computational Statistics and Data Analysis*, *56*, 3415–3429.

Bélisle, P., Joseph, L., MacGibbon, B., Wolfson, D. B., & du Berger, R. (1998). Change-point analysis of neuron spike train data. *Biometrics*, *54*, 113–123.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, *1*, 385–402.

Berger, J., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Bernardo, J. M. (2005). Reference analysis. In D. K. Dey & C. R. Rao (Eds.), *Handbook of statistics 25: Bayesian thinking, modeling, and computation* (pp. 17–90). Elsevier.

Campbell, N. A. (1980). Robust procedures in multivariate analysis i: Robust covariance estimation. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *29*, 231–237.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice vis markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 473–484.

Carlin, B. P., & Louis, T. A. (2000). Empirical bayes: Past, present, and future. *Journal of the American Statistical Association*, *95*, 1286–1289.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, *39*, 83–87.

Chen, J., & Gupta, A. K. (Eds.). (2011). *Parametric statistical change point analysis*. Birkhäuser Boston.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*, 221–241.

Cohen, P. (Ed.). (2008). *Applied data analytic techniques for turning points research*. Taylor & Francis Group.

Das, S. R., Duffie, D., Kapadia, N., & Saita, L. (2007). Common failings: How corporate defaults are correlated. *Journal of Finance*, *62*, 93–117.

Dyson, F. (2004). A meeting with enrico fermi. *Nature*, *427*, 297.

Efron, B. (2010). *Large-scale inference*. Cambridge University Press.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, *16*, 203–213.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 13124–13131.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (Submitted). The perception of probability.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain monte carlo: Stochastic simulation for bayesian inference* (2nd ed.). Chapman & Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533.

Gelman, A. (2008). Objections to bayesian statistics. *Bayesian Analysis*, *3*, 445–450.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman & Hall/CRC.

Girón, F. J., Moreno, E., & Casella, G. (2007). Objective bayesian analysis of multiple changepoints for linear models. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, & A. P. David (Eds.), *Bayesian statistics 8* (pp. 1–27). Oxford University Press.

Goldstein, M. (2006). Subjective bayesian analysis: Principles and practice. *Bayesian analysis*, *1*, 403-420.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, *82*, 711-732.

Han, C., & Carlin, B. P. (2001). Markov chain monte carlo methods for computing bayes factors. *Journal of the American Statistical Association*, *96*, 1122-1132.

Hannart, A., & Naveau, P. (2012). An improved bayesian information criterion for multiple change-point models. *Technometrics*, *54*, 256–268.

Hansen, B. E. (2001). The new econometrics of structural change: Dating breaks in u.s. labor productivity. *Journal of Economic Perspectives*, *15*, 117–128.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Jeffreys, H. (1946). An invariant form for the prior probability in estimating problems. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, *186*, 453–461.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Clarendon Press.

Jensen, G., Altschul, D., Danly, E., & Terrace, H. S. (2013). Transfer of a spatial representation of two distinct serial tasks by rhesus macaques. *PLOS ONE*, *8*, e70825.

Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M., & Gams, M. (2010). An agent-based approach to care in independent living. In B. de Ruyter et al. (Eds.), *Ambient intelligence: First international joint conference, 2010* (pp. 177–186). Springer Berling Heidelberg.

Kaplan, A. (1998). *The conduct of inquiry: Methodology for behavioral science*. Transaction Publishers.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kheifets, A., & Gallistel, C. R. (2012). Mice take calculated risks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 8776–8779.

Klauenberg, K., & Elster, C. (2012). The multivariate normal mean - sensitivity of its objective bayesian estimates. *Metrologia*, *49*, 395–400.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. Guilford.

Koop, G., & Potter, S. M. (2009). Prior elicitation in multiple change-point models. *International Economic Review*, *50*, 751–772.

Lindquist, M. A., Waugh, C., & Wager, T. D. (2007). Modeling state-related fmri activity using change-point theory. *NeuroImage*, *35*, 1125–1141.

Lio, P., & Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, *16*, 376–382.

Longstaff, F. A., & Rajan, A. (2008). An empirical analysis of pricing in collateralized debt obligations. *Journal of Finance*, *63*, 529–563.

Mandelbrot, B. B., & Hudson, R. L. (2006). *The (mis)behavior of markets*. Basic Books.

Meligkotsidou, L., & Dellaportas, P. (2011). Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, *21*, 439–449.

Myung, J., & Pitt, M. A. (1997). Applying occam's razor in modeling cognition: A bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Oakley, J. E., & O'Hagan, A. (2007). Uncertainty in prior elicitation: A nonparametric approach. *Biometrika*, *94*, 427–441.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 324–354.

Pratchett, T. (1983). *The colour of magic*. Colin Smythe.

Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, *82*, 251–266.

Raïffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Robert, C. P., Rydén, T., & Titterington, D. M. (2000). Bayesian inference in hidden markov models through the reversible jump markov chain monte carlo method. *Journal of the Royal Statistical Society, Series B*, *62*, 57–75.

Samaniego, F. J. (2012). *A comparison of the bayesian and frequentist approaches to estimation*. Springer New York.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Silver, N. (2012). *The signal and the noise*. Penguin Press.

Sisson, S. A. (2005). Transdimensional markoc chains. *Journal of the American Statistical Association*, *100*, 1077-1089.

Terrace, H. S. (2005). The simultaneous chain: A new approach to serial learning. *Trends in Cognitive Sciences*, *9*, 202–210.

Van Dongen, S. (2006). Prior specification in bayesian statistics: Three cautionary tales. *Journal of Theoretical Biology*, *242*, 90–100.

Vostrikova, L. J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, *24*, 55–59.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Wasserman, L. (2000). Bayesian model election and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Western, B., & Kleykamp, M. (2004). A bayesian change point model for historical time series analysis. *Political Analysis*, *12*, 354–374.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Zhang, Z., Hamagami, F., Wang, L. L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*, 374–383.