Closed-Form Estimation of Multiple Change-Points Models (Supplemental Material)

Greg Jensen
Columbia University

Abstract

Supplemental material to accompany "Closed-Form Estimation of Multiple Change-Points Models" by Greg Jensen. A variety of non-psychological datasets are analyzed to demonstrate broad applications of the *Conjugate Partitioned Recursion* (CPR) algorithm. A method for performing a sensitivity analysis is demonstrated. A weakness in the algorithm is identified in cases of large-scale stationary distributions, and two modifications to the standard procedure are proposed to circumvent that weakness: The 'dicing' operation (whose function is strictly exploratory) and the 'forward-retrospective' algorithm, which makes assessments sequentially rather than recursively. Finally, the mathematical basis for the conjugate priors invoked in the main article and the formulas for empirical Bayes 'rule-of-thumb' priors are provided.

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

Additional Examples

The examples provided in the main text showcase several experimental applications of the CPR algorithm. Although change-point algorithms remain uncommon in many experimental domains, they have a long history of use in actuarial, industrial, and econometric applications. In the interest of providing a bridge between that tradition, several further examples are provided here. These examples also demonstrate some of the probability distributions not discussed in the main text. In all cases, a decision criterion of $\tau = 10$ is used, and all analyses were performed on a dual-core 2.93GHz laptop.

British Coal-Mining Disasters

One of the signature datasets in the statistical analysis of change-points reports British coal-mining disasters (involving at least 10 men) over the period from March 15, 1851 to March 22, 1962. These data have been reported both as the interval between disasters in days (Jarrett, 1979) and as the number of disasters per year (Carlin et al., 1992).

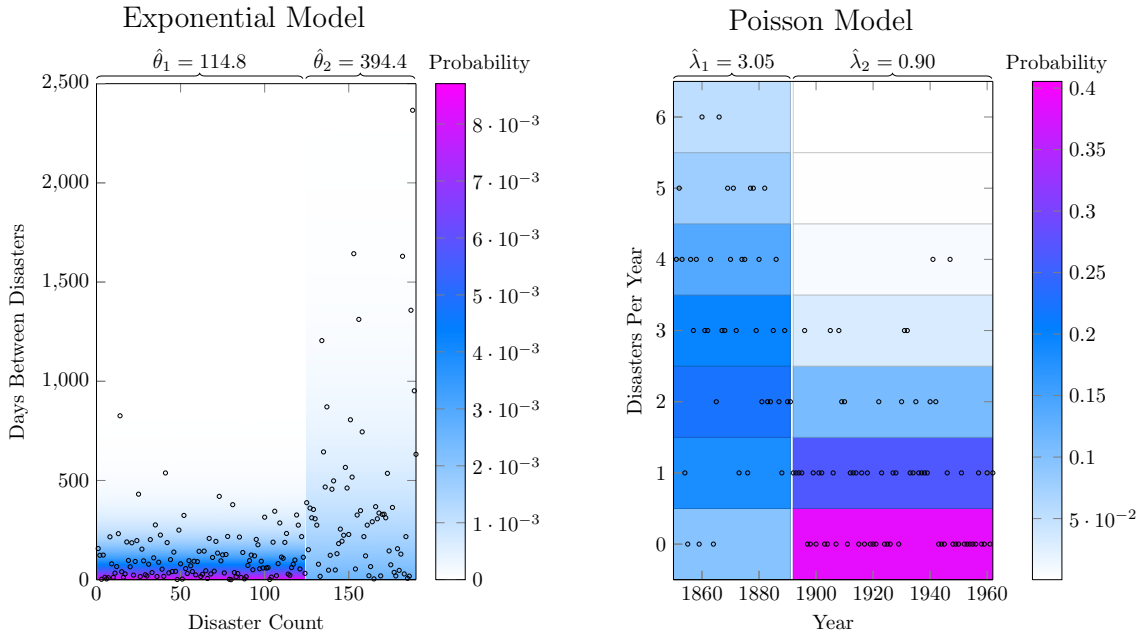

Figure 1. British coal-mining disasters between 1851 and 1962, represented as days between disasters (left) and number of disasters per year (right). In both cases, the CPR algorithm, using the appropriate rule-of-thumb empirical prior, finds a single change-point. Each plot shows its distribution's probability function as a color map, based on the posterior parameter estimates.

Because the probability of a major coal mining disaster is low, and because disasters at different mines are presumably independent of one another, the intervals between disasters follow an exponential distribution, while the number of disasters per year (or any other interval) is Poisson-distributed. This intimate relationship between the exponential distribution and the Poisson distribution is clear from their conjugate relationship: The MML for the exponential distribution is $\text{Gamma}\left(n, \sum x\right)$, while the MML for the Poisson distribution is $\text{Gamma}\left(\sum x, n\right)$.

Figure 1 shows the results of the CPR algorithm for the coal-mining disaster data, represented both as an exponential model of intervals measured in days (prior $= \Pr\left(\theta\right) = \text{Gamma}\left(1, 114\right)$) and as a Poisson model of disasters per year (prior $= \Pr\left(\lambda\right) = \text{Gamma}\left(1, 1\right)$). Here, the exponential parameter θ estimates the distribution of 'days between disasters' and the Poisson parameter λ estimates the distribution of 'disasters per year.' In both cases, a single change-point is identified. The exponential model places the change between 124th and 125th disasters, between March 10th, 1890 and April 2nd, 1891, whereas the Poisson model places the change between 1892 and 1893. These changes coincide with a period of regulatory reform in British mining, particularly the Coal Mines Act of 1887 (Anderson, 1911). Note that the parameters estimates are very similar, but not identical: $\left[\hat{\theta}_1 = 114.8 \approx \frac{365}{\hat{\lambda}_1} = 119.7\right]$; $\left[\hat{\theta}_2 = 394.4 \approx \frac{365}{\hat{\lambda}_2} = 405.6\right]$. Given the small size of the dataset (190 intervals over 112 years), computing the position of this change-point required less than a tenth of a second.

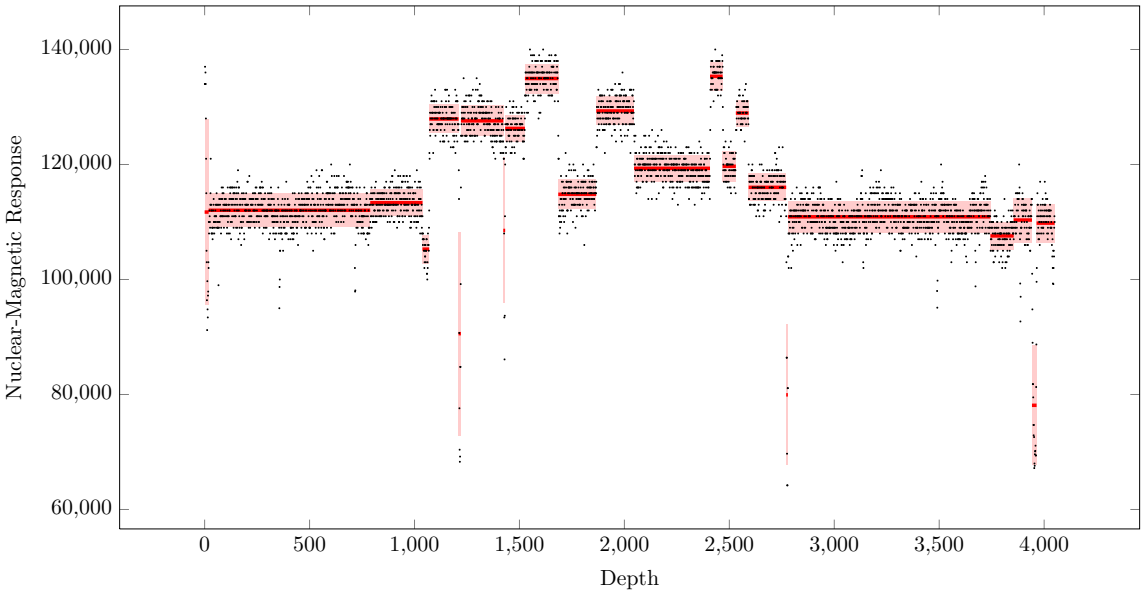

Figure 2. Nuclear-magnetic response of rock strata in an oil well bore-hole. Points correspond to individual observations. The red line indicates the estimated mean (with one standard deviation shown in light red) within segments identified by the CPR algorithm.

Given a choice between representing the data using either distribution, using the Poisson representation results in a loss of some information (dependent on the size of the intervals), so an exponential representation is preferred when precise timestamps are known for all events. Consequently, the exponential estimate is likely to be a better estimate. In many cases, however, counts within regular intervals are the only data available from the records.

Well-Log Data

Another well-studied dataset is a series of 4,050 nuclear-magnetic response measurements from a geological assay. Measurements were taken at regular intervals as a probe was lowered into a bore-hole, with changes in the response implying strata with different properties. Originally introduced by Ó Ruanaidh & Fitzgerald (1996), this dataset provides a robust test of change-point algorithms because it only weakly conforms to standard distributional assumptions, both in terms of outliers (which could reflect either distinctive strata or mere measurement noise) and substantial and inconsistent rounding error.

Assuming that the data followed a Gaussian distribution with unknown parameters, Figure 2 shows the change-points identified by the CPR algorithm, given a Normal-Gamma prior based on the median and median absolute deviation of the data ($a = 114000$, $b = 0.1$, $c = 1$, $d = 2965.2$). This prior remains relatively weak with respect to the mean (as $b = 0.1$ corresponds to only 1/10 of a hypothetical observation), but is somewhat stronger with respect to the variance (as $c = 1$ corresponds to one full hypothetical observation). The prior standard deviation d was estimated from the median absolute deviation of the the first differences (that is, the difference between consecutive observations), a robust econometric technique for assessing volatility in the presence of omitted variables (Wooldridge, 2002).

The resulting change-point model consists of 22 change-points, taking approximately 14 seconds to calculate.

One can get a sense of the CPR algorithm's rapidity by comparing it to the "posterior simulation" analyses performed on the dataset by Fearnhead (2006). These analyses use recursive sampling to effect a numerical integration of the posterior odds. When their analysis was performed on a 3.4GHz PC, a "piecewise-constant" posterior simulation algorithm performed 10,000 simulations that required 26 seconds to run, identifying between 45 and 60 change-points. Fearnhead also ran a more robust "random walk" model, which took approximately 19 minutes and only identified between 12 and 21 change-points; these models were subsequently compared to a brute-force Markov-chain Monte Carlo (MCMC) method.

This result is impressive, with a runtime that outperformed MCMC by about two orders of magnitude. Nevertheless, implementation requires that the analyst make substantial assumptions (such as consistent variance over time), and required that outliers be removed from the data prior to analysis. Additionally, because the method relies on random resampling, slightly different results are obtained with each simulation. The CPR algorithm compares very favorably, as it identifies close to the optimal number of change-points in less time than the faster algorithm reported by Fearnhead, produces identical results every time it is run, permits the variance to change from one segment to the next, and does not require the pruning of outliers.

Treasury Bill Rates

In econometrics, change-point analysis is often demonstrated using the nominal rate on three-month U.S. Treasury bills (or T-bills). Because these data are freely available to the public, and are regularly updated, it has been used as a test case in both retrospective (Bai, 1997) and prospective (Pesaran et al., 2006) change-point analysis.

Figure 3 shows two different implementation of the CPR algorithm to monthly T-bill data over the period from June 1947 to February 2013. The top plot shows the posterior estimates for a linear regression model. This approach identifies 28 change-points, and is highly sensitive to big shifts. The bottom plot shows the 'first differences' (Wooldridge, 2002), which emphasizes the volatility in the data over the current rate (Meligkotsidou & Dellaportas, 2011). Here, as in Figure 2, both the mean and one standard deviation are shown. This approach identifies only 15 change-points, because segments with similar volatility (e.g. throughout the 90s) are likely to be grouped together.

The contrasting benefits of these two methods depends on the purpose of the analysis. The regression analysis is more complex, devoting three parameters to each segment (slope, intercept, and the free cells of the estimated coefficient covariance matrix), and provides a much more accurate historical description. On the other hand, the first differences provide a more useful summary statistic for examining volatility. A comparable experimental case might be the study of motor impairment: Studying of the distance moved over each time interval is likely to be more informative than absolute position when tremors are a symptom of interest.

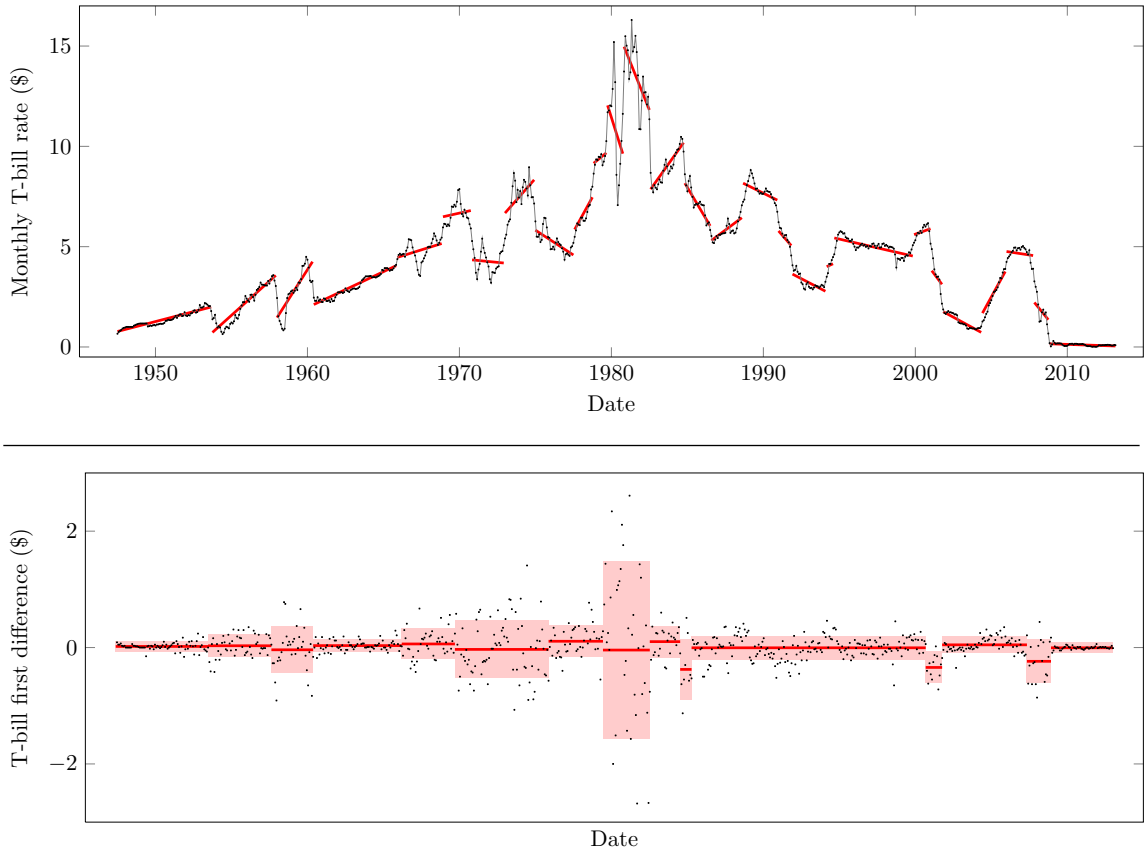

Figure 3. Rates on three-month U.S. Treasury bills, calculated monthly from June 1947 to February 2013. (Top) Nominal rates, with red lines displaying the model inferred by a linear regression change-point analysis. (Bottom) First differences of nominal rates ($x_i - x_{i-1}$), with the means (red) and one standard deviation (light red) from a Gaussian change-point analysis.

Sensitivity Analysis

An essential consideration in using the CPR algorithm is the selection of the decision criterion τ. In the examples presented, a criterion of $\tau = 10$ has been used because it provides a good balance between description and parsimony. 'Good' is a subjective term, however, and there is not always a good correspondence between results that are statistically vs. theoretically significant. This is further complicated by the relationship between a criterion's efficacy and subtle differences in the assumptions underlying different models.

A benefit of the CPR algorithm is that, given its reliance on closed-form solutions, it is trivial to perform a post-hoc calculation of the likelihood of the data once the parameters have been estimated. The overall efficacy of the entire model (across all segments) can then be compared to other such models using the Schwarz-Bayes Information Criterion (SBIC; Schwarz, 1978). As noted in the main text, SBIC arises from an approximation of the

marginal model likelihood $m(x, M)$:

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

The conventional practice is to convert this approximation into a score measured on a log scale:

$$SBIC(x) = -2 \cdot \log(f(x|\theta, M)) + p \cdot \log(n) + C_x \quad (2)$$

Here, given the observations x, a value is calculated that depends on twice the negative log of the likelihood function $f(x|\theta, M)$, on the number of free parameters p, and on the total number of observations n. For data x, there is also a constant C_x, which prevents the direct comparison of two datasets x and y (as they might have different constants). However, the SBIC can be used to compare how well different models fit the data x because the constant C_x will cancel out of such comparisons. The model with the *lowest* SBIC is considered the best, because its overall marginal model likelihood is approximately the highest (Yang, 2005).

Although the SBIC was originally intended to be used in cases where observations independent and identically distributed, its derivation has been generalized to non-identically distributed cases as well, having particular success in comparing time-series models (Cavanaugh & Neath, 1999). This makes it a suitable metric for performing a sensitivity analysis.

In order to examine the effects of the decision criterion τ on the resulting change-point model, a sensitivity analysis consists of running the analysis multiple times using different criteria but keeping other parameters constant. For each criterion that is tested, the resulting change-point model $\mathbf{M}$ is then used to calculate an SBIC using the following equation:

$$SBIC(x, \mathbf{M}) = -2\left[\sum_{i \in \mathbf{M}} \log(f(x_i|\theta_i, M))\right] + (\text{length}(\mathbf{M})(p_M + 1) - 1) \cdot \log(n) + C_x \quad (3)$$

Here, the log likelihood is summed in each of the segments of $\mathbf{M}$, and the free parameters consist of the one set of model parameters p_M for each segment, plus a free parameter for the position of each change-point (Yao, 1988). In the analyses performed below, the parameters θ_i were estimated post-hoc, using frequentist methods, to minimize the relationship between the prior assumptions and the subsequent sensitivity analysis.

Figure 4 demonstrates the results of such a sensitivity analysis for the datasets discussed in the main article and in the supplement. Although each dataset displays its own idiosyncrasies, several themes are evident from their comparison.

First, it is clear that models vary in their sensitivity to the decision criterion. For example, linear regression models are quite resistant to changes in τ. The regression model describing T-bill rates differs only by a few change-points over the range $1 \leq \tau \leq 100$ (Figure 4, lower left), and the reaction time data from Palmeri (1997) are not included in Figure 4 because they returned the same result regardless of the decision criterion used. On the other hand, the binomial data describing SimChain performance (Figure 4, upper left, Jensen et al., 2013), and the Gaussian distribution of T-bill rate first differences (Figure 4,

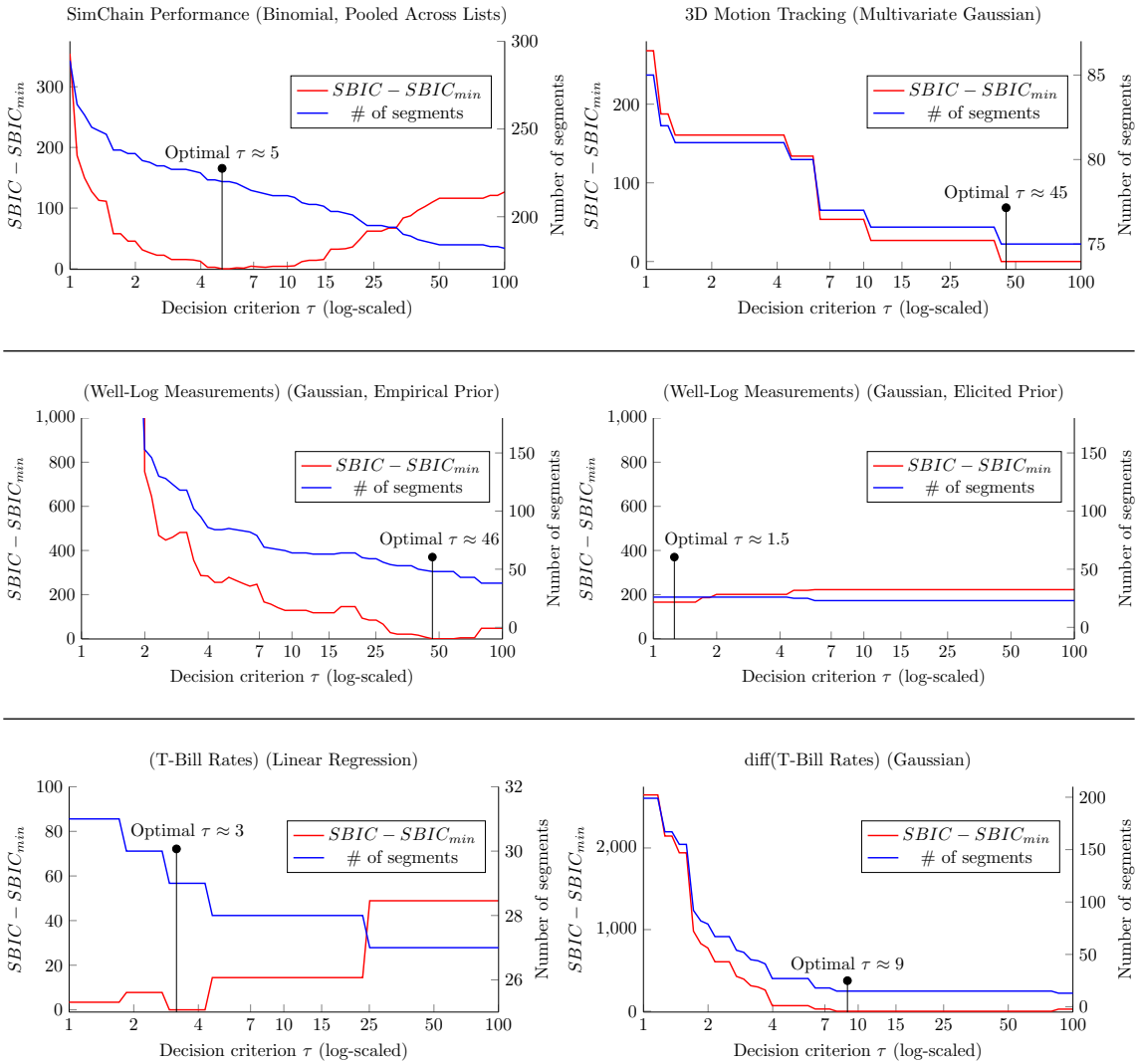

Figure 4. Sensitivity analysis performed on two datasets from the main body text and three from the supplement. The decision criterion τ was varied between 1 and 100, and the SBIC was calculated for each resulting model, using Equation 3. Relative SBIC among the models is plotted in red (with the best model set at zero), and the number of data segments is plotted in blue.

lower right) both demonstrate the risk of overfitting when the decision criterion is too lax. Additionally, the SimChain data show a risk of underfitting if the criterion is too harsh.

A different issue is raised by the sensitivity analysis performed on the well-log measurements. Figure 4 (center left) shows the results of the sensitivity analysis when the empirical 'rule-of-thumb' prior is used (hyperparameters $a = 114000$, $b = 0.1$, $c = 0.1$, $d = 5930.4$). However, these data do not conform well to the assumptions of a Gaussian distribution (due to device imprecision), and a different prior was used in the analysis described above (hyperparameters $a = 114000$, $b = 0.1$, $c = 1$, $d = 2965.2$). Figure 4 (center right) shows the results of using this 'elicited' prior (so-called because its selection depends on the analyst's discretion about the meaning of the data). When using the rule-of-thumb prior (which is weakly subjective), a great deal of sensitivity to the decision criterion was observed. Contrastingly, when a prior with a stronger prior assumption about the dispersion of the data was used, the influence of the decision criterion was much smaller. The sensitivity of posterior results to the prior is a well-established problem, and one of the central pillars of the argument in favor of objective Bayesian methods (Samaniego, 2012). Responsible use of subjective priors should include a sensitivity analysis with respect to the prior (Gelman et al., 2003).

Across these six sensitivity analyses, $\tau = 10$ performs reasonably well, neither resulting in dramatic over- or under-fitting of the data. Although performing sensitivity analyses is an important element of statistical due diligence, $\tau = 10$ is a reasonable default value for the decision criterion.

A Failure Condition: CPR Insensitivity When Data Are Stationary

Although intended for large datasets, the sensitivity of the CPR algorithm to change-points is reduced in cases where changes occur frequently and values oscillate around some central value. Closed-form, conjugate solutions for $m\,(x, M)$ depend only on the prior hyperparameters and the data's sufficient statistics, both of which usually ignore the temporal structure of values within each segment (that is, they assume 'exchangeability'). Thus, segments containing many changes that oscillate around a stable central value are likely to be interpreted as a stationary process with an inflated variance.

This problem can be demonstrated using a simulated dataset. A "Gaussian iterated map" (or "mouse map") is a nonlinear iterated function with two parameters α and β:

$$x_{(i+1)} = \beta + \exp\left(-\alpha x_{(i)}^2\right) \tag{4}$$

This function's behavior is chaotic given certain parameters (Hilborn, 2001), making it useful for generating replicable datasets whose "true" parameters can be compared to those estimated. We defined two mouse maps:

$$x_{(i+1)} = -0.6 + \exp\left(-2\pi x_{(i)}^2\right); x_{(1)} = 0$$
$$y_{(i+1)} = -0.4 + \exp\left(-e^2 y_{(i)}^2\right); y_{(1)} = 1$$

These, in turn, were used to specify a series of parameters:

$$\mu_{(i)} = 4x_{(i)}$$
$$\sigma_{(i)} = \exp\left(y_{(i)}\right)$$

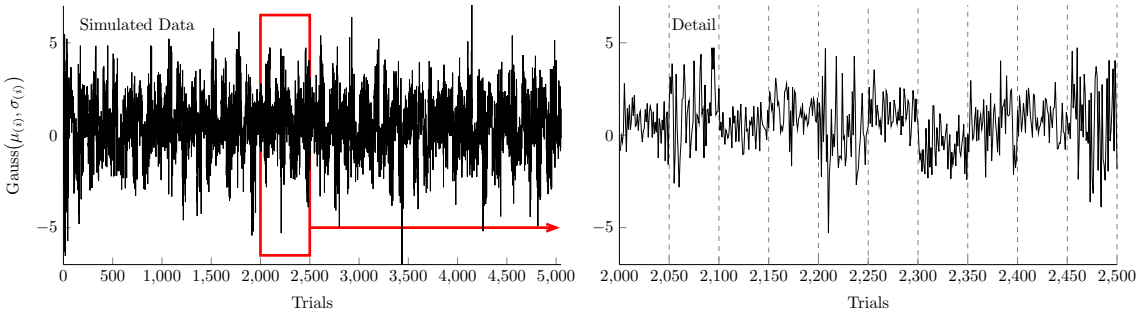

Figure 5. Simulated data generated using a chaotic distribution of Gaussian parameters derived from Equation 4. (Left) The full dataset D, consisting of 5050 simulated observations that were generated in blocks of 50. The default CPR algorithm will not identify most of these change-points. (Right) A subset of D, with the true change-points denoted by dashed gray lines. The CPR algorithm will reliably identify these change-points if it is run only on this subset, despite failing to detect them when run on the complete dataset.

The dataset D consisted of 101 segments, where the ith segment was generated using a Gaussian distribution with parameters $\left(\mu_{(i)}, \sigma_{(i)}\right)$. A characteristic example is presented in Figure 5, which presents both the full data (Left) and a detail view of a section of the data (Right). When the true change-points are clearly labeled, their subdivisions appear highly plausible, but the CPR algorithm is not omniscient, and these relatively subtle discontinuities are subsumed by the overall variance of the data when too many of them occur within a particular segment.

The Dicing Operation

A strictly exploratory solution to this problem is to introduce the preliminary step of 'dicing' the data into arbitrary segments and examining each segment for change-points, outlined in Algorithm 1.

When invoking the dicing operation, the analyst must specify some integer value $d > 1$ that specifies how many times to subdivide the data, doing so into equal parts. Each segment is checked for a single change-point, and any change-points that are identified are added to the initial model $\mathbf{M}$. After performing the dicing operation, the CPR Algorithm proceeds with the new model $\mathbf{M}$, rather than its default value of $\langle 0, n \rangle$. This effectively bypasses the first few partitioning operations (during which the sensitivity might be impaired) and applies the CPR algorithm to an already-subdivided dataset.

Figure 6 (top) demonstrates the efficacy of introducing the dicing operation to the data in Figure 5. Given one subdivision (the default for the CPR algorithm), no change-points are identified between $t = 150$ and $t = 5000$. However, with only three subdivisions, the number of change-points identified rises to 24. Given four or more subdivisions, the change-points identified were highly consistent, with difficulty only the range from 4400 to 5000. Figure 6 (center) shows the total number of times different points were identified over the thirty different dicing operations. Only 284 of the 5049 possible points were identified at least once. Of those, 27 points were specified at least twenty-five times, and 62 were specified at least twenty times, without even considering 'near misses.'

Algorithm 1: The dicing operation, which identifies the single best change-point in each arbitrary segment of data.

Data: events $x_{(1:n)}$, times $t_{(0:n)}$, model C_0, decision criterion τ, diced segment count d
Result: diced model $\mathbf{M}$
begin
  $\mathbf{S} \leftarrow 0, \frac{n}{d}, \frac{2n}{d}, \cdots, \frac{dn}{d}$ `/* get segment fenceposts */`
  $\mathbf{M} \leftarrow \langle 0, n \rangle ; p_c = \frac{1}{n-1}$ `/* empty "new CPs" array and prior odds of a change */`
  for $s = 1$ *to* $length(\mathbf{S}) - 1$ **do**
    $i \leftarrow \mathbf{S}(s) + 1$; $j \leftarrow \mathbf{S}(s+1)$; $|\mathbf{K}| \leftarrow j - i + 1$ `/* assign indices; allocate BF array */`
    for $c = i$ *to* j **do**
      $k_c \leftarrow \frac{m(x_{(i:c-1)}, C_0) m(x_{(c:j)}, C_0)}{m(x_{(i:j)}, C_0)}$; $\mathbf{K}_{(c)} \leftarrow \frac{k_c \cdot (t_{(c)} - t_{(c-1)})}{(t_{(j)} - t_{(i-1)}) \cdot \exp(SB_{(c)})}$ `/* Bayes factors */`
    if $sum(\mathbf{K}) \cdot p_c \cdot \frac{1}{j-i+1} > \tau$ **then**
      $\hat{c} \leftarrow \text{index}(\max(\mathbf{K}))$; $\text{Push}(\mathbf{M}, \hat{c})$ `/* insert` $\hat{c}$ `into` $\mathbf{M}$ `if` τ `permits */`
  Sort($\mathbf{M}$); **return** $\mathbf{M}$

Although the dicing operation has practical use when detecting stationary metastable periods within a more generally stable process, it should not be necessary in the vast majority of cases. Data should not be diced into segments so small that no more than one or two change-points can reasonably be expected to exist.

Forward-Retrospective Change-Point Detection

Another approach to the problem of oscillating data is to identify each change-point in chronological order. Rather than approaching the analysis as a batch process, change-point evaluations can take place within a more constrained frame.

Sequential change-point detection algorithms have a long history, with the work of Page (1954) being among the earliest (see also Venkatraman, 1992; Chen & Gopalakrishnan, 1998; Gallistel et al., 2004). Beginning with the indices $i = 1$ and $j = 2$, the data range $x_{(i:j)}$ is tested for a change-point repeatedly as j is incrementally increased. Once a change is detected at $\hat{c}$ (where $i < \hat{c} < j$), the data prior to $\hat{c}$ are removed from consideration and $i = \hat{c}$. This process continues until all observations have been considered.

Thus, when a change is detected, the range of data subsequently considered is truncated. Although a deliberate feature in the interests of reducing the algorithm's computational load, this creates problems when 'false alarms' are observed. Over a very long sequence of Gaussian observations, for example, a handful of sizable outliers are reasonable; when only a small segment of data is being observed, however, estimates of variance are likely to be conservative and a random outlier may be confused for a true change in the distribution. This problem is particularly salient in data that suffer from some degree of

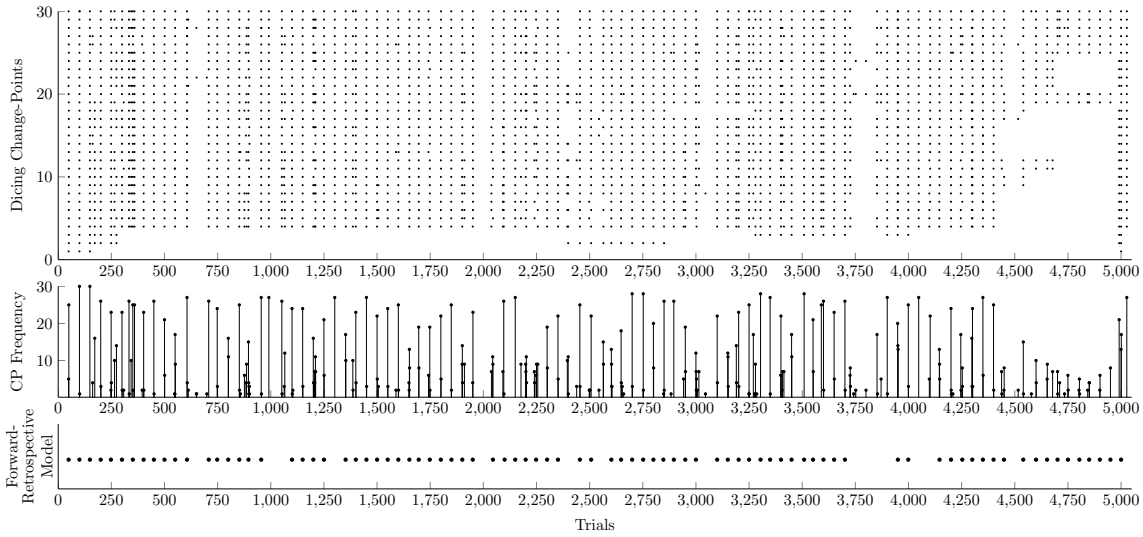

Figure 6. Change-points identified in simulated data. (Top) Trials for which change-points were identified, as a function of the number of subdivisions examined by the dicing operation. (Center) A histogram of how often, across 30 values for the dicing parameter, a precise change-point was identified. (Bottom) Trials for which change-points were identified using the retrospective sequential algorithm.

out-of-sample 'contamination.' Furthermore, even in cases where the data are well-behaved, sequential analyses are vulnerable to mistaken inference because of multiple comparisons. When a true change is identified, there is a risk that its position will be estimated prematurely.

In order to combat this, a substantial improvement to traditional sequential analyses is proposed by Gallistel et al. (Submitted). After the first change-point is identified, each additional change-point's identification is immediately followed by a retrospective test of the preceding change-point. For example, once $\hat{c}_{(3)}$ is identified, the interval $x_{\hat{c}(1):\hat{c}(3)}$ is reanalyzed. If $\hat{c}_{(2)}$ was the result of a false alarm, it will probably not be detected again, and it can be removed from the model. If, on the other hand, the evidence still supports a change between $\hat{c}_{(1)}$ and $\hat{c}_{(3)}$, this re-analysis can be used to update the position of $\hat{c}_{(2)}$.

The forward-retrospective change-point analysis proposed by Gallistel et al. (Submitted) is combined with the CPR algorithm in Algorithm 2. Here, rather than making use of recursive binary partitioning, the CPR algorithm is only invoked to find a single change-point at a time.

Figure 6 (bottom) shows the change-point model reported by the retrospective sequential algorithm when analyzing the simulated data from Figure 5. Whereas the CPR algorithm displays considerable insensitivity in the absence of the dicing operation, the forward-retrospective analysis identifies nearly all changes, doing so close to their true times of occurrence.

Although an effective alternative in many circumstances, the forward-retrospective approach also has shortcomings. Both the standard CPR algorithm and its sequential counterpart have expected runtimes that scale as a function both of the number of data

Algorithm 2: The forward-retrospective change-point algorithm, which scans forward through the data in order to identify each change-point in chronological order

Data: events $x_{(1:n)}$, parameters $\mathbf{P}$
Result: change-point model $\mathbf{M}$
begin
  $\mathbf{M} \leftarrow \langle 0, n \rangle ; s \leftarrow 1$ `/* empty "new CPs" array, segment index */`
  for $i = 2$ *to* n **do**
    $\mathbf{N} \leftarrow \text{CPRBayes}\left(x_{(\mathbf{M}(s)+1:i)}, \mathbf{P}\right)$ `/* check for single cp */`
    if *length* $(\mathbf{N}) > 2$ **then**
      Push $\left(\mathbf{M}, \mathbf{M}_{(s)} + \mathbf{N}_{(2)}\right)$ `/* insert cp into M */`
      Sort$(\mathbf{M}) ; s = s + 1$ `/* sort M, update s */`
      if $s > 2$ **then**
        $\mathbf{O} \leftarrow \text{CPRBayes}\left(x_{(\mathbf{M}(s-2)+1:\mathbf{M}(s))}, \mathbf{P}\right)$ `/* check previous cp */`
        if *length* $(\mathbf{O}) > 2$ **then**
          $\mathbf{M}_{(s-1)} = \mathbf{M}_{(s-2)} + \mathbf{O}_{(2)}$ `/* refine previous cp */`
        else
          Pull $\left(\mathbf{M}_{(s-1)}\right) ; s = s - 1$ `/* remove previous cp */`
  return M

segments s and of the number of observations n. When using a binary partitioning strategy, the CPR algorithm's runtime is expected to fall between $O(n \cdot s)$ and $O(n \cdot \log s)$ (depending on the order in which the points are identified). However, the sequential alternative has a runtime between $O\left((n-s)^2\right)$ and $O\left(\frac{n^2}{s}\right)$, depending on how evenly the change-points are spaced. When changes are infrequent, the forward-retrospective algorithm is potentially much slower.

Table 1 gives a sense of these contrasting relationships. When change-points occur relatively frequently (as in the 3D motion tracking or the T-Bill rate datasets), the sequential approach only performs somewhat less well. However, when there are very few changes relative to the number of observations (as is particularly the case for the reaction times data), the increase in runtime can be dramatic. The quadratic runtime renders the sequential approach unusable for large datasets with large gaps between change-points.

One way to mitigate the quadratic runtime is to increment the index i by more than one observation at a time (Chen & Gopalakrishnan, 1998). If change-points are consistently widely spaced, then the value of i could conceivably increase by tens or hundreds of observations with each step. In practice, however, consistently wide spacing renders the sequential approach unnecessary, while inconsistent spacing (in which some change-points are spaced much closer together than others) introduces the risk that large increments of i will overlook some changes.

Another difficulty with the forward-retrospective strategy is its vulnerability to the stopping problem. In principle, this is mitigated to some extent by the retrospective oper-

Dataset	Binary Partitioned	Forward-Retrospective
SimChain	2.84 s	24.80 s
Reaction Time	21.23 s	3756.04 s
3D Motion Tracking	317.77 s	527.57 s
Mining Disasters (Exponential)	0.35 s	4.79 s
Mining Disasters (Poisson)	0.14 s	1.62 s
Well-Log Measurements	15.08 s	169.43 s
T-Bill Rates	4.45 s	15.70 s
diff(T-Bill Rates)	1.78 s	9.25 s
Simulation Data	(diced) 23.15 s	71.72 s

Table 1
Processing time for the CPR algorithm in each of the datasets described in the main text and supplement, performed using a dual-core 2.93GHz laptop.

ation, but the forward-retrospective strategy will, all things being equal, identify a larger number of change-points than the CPR strategy, because it is more likely to be impacted by small subsets of the data. As a counterweight against this, the forward-retrospective algorithm uses a default criterion of $\tau = 20$, twice the default recommended for the CPR algorithm.

Finally, the forward-retrospective algorithm yields different results examining the data forward and backward in time. This lack of symmetry arises from several factors, but the most substantive is the process of updating p_c as additional change-points are detected. This results in a growing propensity to identify change-points over time.

On the basis of these limitations, there are two scenarios in which the sequential approach should consistently be favored. The first is when changes are expected to occur relatively frequently, particularly if they occur in the oscillating manner characterized by the simulated data in Figure 5. The second is when temporal order is of primary importance, such as when the earliest change can be said to have occurred (e.g. the first trial in which an animal engaged in conditioned responding to a stimulus).

Closed Form Solutions for Conjugate Priors

This section provides closed-form solutions for the conjugate priors and marginal model likelihoods of various distributions that are likely to be of interest. These are used to compute the values of $m(x, M)$ that appear in the change-point algorithm. This list is not meant to be exhaustive. Unless otherwise noted, the conjugacy of these distributions is demonstrated by DeGroot (2005).

As detailed in the body text, the prior probability distribution (and the corresponding hyperparameters that describe its shape) can have a determining effect in Bayesian analysis, and must be selected in a principled manner. Whenever possible, the hyperparameters should reflect an analyst's principled assumptions, and they should be explicitly reported along with a justification for their use. For example, if data were collected using equipment with a known degree of imprecision, this should be reflected by the prior. For each of the distributions below, the interpretation of the hyperparameters is provided to facilitate this process.

In some cases, an analyst's intuitions must arise from the data themselves. In such exploratory analyses, setting an inappropriate prior will not only result in misleading marginal likelihoods, but can also severely impair the sensitivity of the CPR algorithm. To the extent possible, the "rule-of-thumb" priors provided for each distribution are intended to facilitate the identification of change-points in keeping with an empirical Bayes approach (Casella, 1985; Carlin & Louis, 2000). Insofar as their use deviates from standard Bayesian inference, an analyst choosing to use them should explicitly note cases where the prior hyperparameters are derived from the data.

By preference, the rule-of-thumb priors make use of robust estimation techniques (Huber & Ronchetti, 2009). Time series with substantial discontinuities will, almost by definition, appear to contain 'outliers' when viewed without regard for time, so robust estimators (such as the median and median absolute deviation) are less likely to suggest parameters that are not characteristic of the preponderance of the data.

Notation	Formula	Interpretation
$\Gamma(x)$	$(x-1)!$	Gamma Function
$\text{Beta}(x,y)$	$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$	Beta Function
$\text{SumSq}(x,y)$	$(x-y)(x-y)^\top$	Sum of squared differences
$\text{MAD}(x)$	$\text{median}(\lvert x-\text{median}(x)\rvert)$	Median absolute deviation

Discrete Data (Univariate)

The Binomial Distribution.

- Support:

$$\{x_1,\ldots,x_n | x_i = 0 \text{ or } 1\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad S_x = \sum_{i=1}^{n} x_i$$

- With unknown probability of success p:

Prior Hyperparameters

$$\begin{array}{ll} \{a|a>0\} & \text{Number of successes}-1 \\ \{b|b>0\} & \text{Number of failures}-1 \end{array}$$

Conjugate Prior

$$\Pr(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{\text{Beta}(a,b)}$$

Normalizing Constant

$$m(a,b) = \text{Beta}(a,b) \tag{5}$$

Posterior Hyperparameters

$$a' = a + S_x \qquad b' = b + n - S_x$$

Posterior Predictive

$$\hat{p} = \frac{a'}{a'+b'}$$

'Rule-of-Thumb' Empirical Prior

$$a = 0.5 \qquad b = 0.5$$

The Geometric Distribution.

- Support:

$$\{x_1, \ldots, x_n | x_i \in (0, 1, 2, 3, \cdots)\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad S_x = \sum_{i=1}^{n} x_i$$

- With unknown probability of success p:

Prior Hyperparameters

$\{a \mid a > 0\}$	Number of observations
$\{b \mid b > 0\}$	Sum of observations

Conjugate Prior

$$\Pr(p|a, b) = \frac{p^{a-1}(1-p)^{b-1}}{\text{Beta}(a, b)}$$

Normalizing Constant

$$m(a, b) = \text{Beta}(a, b) \tag{6}$$

Posterior Hyperparameters

$$a' = a + n \qquad b' = b + S_x$$

Posterior Predictive

$$\hat{p} = \frac{a' + n}{a' + b' + n + S_x}$$

'Rule-of-Thumb' Empirical Prior

$$a = 0.5 \qquad b = 0.5$$

The Poisson Distribution.

- Support:

$$\{x_1, \ldots, x_n | x_i \in (0, 1, 2, 3, \cdots)\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad S_x = \sum_{i=1}^{n} x_i$$

- With unknown rate λ:

Prior Hyperparameters

$\{a \mid a > 0\}$	Total occurrences
$\{b \mid b > 0\}$	Number of intervals

Conjugate Prior

$$\Pr(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\frac{\lambda^{a-1}}{\exp(b\lambda)}$$

Normalizing Constant

$$m(a,b) = \frac{\Gamma(a)}{b^a} \tag{7}$$

Posterior Hyperparameters

$$a' = a + S_x \qquad b' = b + n$$

Posterior Predictive

$$\hat{\lambda} = \frac{a'}{b'}$$

'Rule-of-Thumb' Empirical Prior

$$a = \text{median}(x) \qquad b = 1$$

Continuous Data (Univariate)

The Exponential Distribution.

- Support:

$$\left\{x_1, \ldots, x_n | x_i \in \mathbb{R}^{>0}\right\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad S_x = \sum_{i=1}^{n} x_i$$

- With unknown θ:

Prior Hyperparameters

$\{a\|a > 0\}$	Number of observations
$\{b\|b > 0\}$	Sum of observations

Conjugate Prior

$$\Pr(\theta|a,b) = \frac{b^a}{\Gamma(a)}\frac{\theta^{a-1}}{\exp(b\theta)}$$

Normalizing Constant

$$m(a,b) = \frac{\Gamma(a)}{b^a} \tag{8}$$

Posterior Hyperparameters

$$a' = a + n \qquad b' = b + S_x$$

Posterior Predictive

$$\hat{\theta} = \frac{b'}{a'}$$

'Rule-of-Thumb' Empirical Prior

$$a = 1 \qquad b = \text{median}(x)$$

The Gaussian Distribution.

The conjugate analysis in this section, originally outlined by DeGroot (2005), is based on the excellent treatment of the subject by Murphy (2007).

- Support:

$$\{x_1, \ldots, x_n | x_i \in \mathbb{R}\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad S_x = \sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{S_x}{n}$$

- With unknown mean μ and known precision λ:

  Prior Hyperparameters

$\{a \mid a \in \mathbb{R}\}$	Prior mean
$\{b \mid b > 0\}$	Prior total precision

  Conjugate Prior

$$\Pr(\mu | a, b) = \sqrt{\frac{b}{2\pi}} \exp\left(-\frac{b}{2}(\mu - a)^2\right)$$

  Normalizing Constant

$$m(a, b) = \sqrt{\frac{2\pi}{b}} \exp\left(\frac{a^2 b}{2}\right) \tag{9}$$

  Posterior Hyperparameters

$$a' = \frac{ab + \lambda S_x}{b + \lambda n} \qquad b' = b + \lambda n$$

  Posterior Predictive

$$\hat{\mu} = a'$$

  'Rule-of-Thumb' Empirical Prior

$$a = \text{median}(x) \qquad b = 0.1$$

- With known mean μ and unknown precision λ:

  Prior Hyperparameters

$\{a \mid a > 0\}$	(Number of observations)/2
$\{b \mid b > 0\}$	(Sum of squared deviations from μ)/2

  Conjugate Prior

$$\Pr(\lambda | a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$

  Normalizing Constant

$$m(a, b) = \frac{\Gamma(a)}{b^a} \tag{10}$$

Posterior Hyperparameters

$$a' = a + \frac{n}{2} \qquad b' = b + \frac{\text{SumSq}\,(x, \mu)}{2}$$

Posterior Predictive

$$\hat{\lambda} = \sqrt{\frac{b'}{a'}}$$

'Rule-of-Thumb' Empirical Prior

$$a = 0.1 \qquad b = \text{MAD}\,(x) \cdot 1.4826$$

- With unknown mean μ and unknown precision λ:

Prior Hyperparameters

$\{a \mid a \in \mathbb{R}\}$	Prior mean
$\{b \mid b > 0\}$	Observations supporting the prior mean
$\{c \mid c > 0\}$	(Observations supporting the prior precision)/2
$\{d \mid d > 0\}$	(Sum of squared deviations)/2

Conjugate Prior

$$\Pr\left(\mu, \lambda | a, b, c, d\right) = \frac{d^c \lambda^{c-1}}{\Gamma\left(c\right) \exp\left(\lambda d\right)} \cdot \sqrt{\frac{\lambda b}{2\pi}} \cdot \exp\left(-\frac{\lambda b}{2}\left(\mu - a\right)^2\right)$$

Normalizing Constant

$$m\left(a, b, c, d\right) = \frac{\Gamma\left(c\right)}{d^c}\sqrt{\frac{2\pi}{b}} \tag{11}$$

Posterior Hyperparameters

$$a' = \frac{ab + S_x}{b + n} \qquad b' = b + n \qquad c' = c + \frac{n}{2}$$

$$d' = d + \frac{\text{SumSq}\,(x, \bar{x})}{2} + \frac{bn\left(\bar{x} - a\right)^2}{2\left(b + n\right)}$$

Posterior Predictive

$$\hat{\mu} = a'\hat{\lambda} = \sqrt{\frac{d' \cdot (b' + 1)}{b' \cdot c'}}$$

'Rule-of-Thumb' Empirical Prior

$$a = \text{median}(x) \qquad b = 0.1$$

$$c = 0.1 \qquad d = \text{MAD}\,(x) \cdot 1.4826$$

Basic Linear Regression.

Linear regression presents an analyst with a hybrid problem, because the dependent observations y are being summarized in terms of a multivariate vector of regression coefficients $\boldsymbol{\beta}$ and a univariate variance parameter σ^2. As with the Gaussian distribution, these parameters can each be understood in terms of their respective priors distributions. The implementation below uses the multivariate Gaussian distribution as the conjugate prior for the coefficients $\boldsymbol{\beta}$, and the inverse gamma distribution as the conjugate prior for the variance.

This implementation is also distinctive in that the most arithmetically tractable form for the normalizing constant incorporates both the prior and posterior hyperparameters.

- Support:

$$\{y_1, \ldots, y_n | y_i \in \mathbb{R}\}\, X = \begin{bmatrix} X_{1,1} & \ldots & X_{1,k} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \ldots & X_{n,k} \end{bmatrix}, X_{j,i} \in \mathbb{R}$$

- With unknown regression coefficients $\boldsymbol{\beta}$ and unknown variance σ^2:

Prior Hyperparameters

$\text{model} = m\left(\mathbf{m}, a, b, \boldsymbol{\Lambda}\right)$	Gaussian error model
$\mathbf{m} = \{\mathrm{m}_1, \ldots, \mathrm{m}_k | \mathrm{m}_i \in \mathbb{R}\}$	Prior regression coefficients
$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$	Prior coefficient precision matrix
$\{a | a > 0\}$	(Observations supporting the error)/2
$\{b | b > 1\}$	(Sum of squared deviations)/2

Conjugate Prior

$$\Pr\left(\boldsymbol{\beta}, \sigma^2 | y, X, \text{model}\right) = \frac{\Pr\left(\boldsymbol{\beta}, \sigma^2 | \text{model}\right) \cdot \Pr\left(y | X, \boldsymbol{\beta}, \sigma^2, \text{model}\right)}{\Pr\left(y | \text{model}\right)}$$

$$\text{Given the model, } \Pr\left(\boldsymbol{\beta}, \sigma^2 | y, X\right) \propto \Pr\left(\boldsymbol{\beta} | \sigma^2, y, X\right) \cdot \Pr\left(\sigma^2 | y, X\right)$$

$$\Pr\left(\boldsymbol{\beta} | \sigma^2, y, X\right) = \mathcal{N}\left(\mathbf{m}, \sigma^2 \boldsymbol{\Lambda}^{-1}\right) \Pr\left(\sigma^2 | y, X\right) = \text{InvGamma}\left(a, b\right)$$

Normalizing Constant

$$\Pr\left(y | \text{model}\right) = \frac{1}{(2\pi)^{n/2}} \cdot \sqrt{\frac{|\boldsymbol{\Lambda}|}{|\boldsymbol{\Lambda'}|}} \cdot \frac{b^a}{b'^{a'}} \cdot \frac{\Gamma\left(a'\right)}{\Gamma\left(a\right)} \tag{12}$$

Posterior Hyperparameters

$$\mathbf{m}' = \left(X^\top X + \boldsymbol{\Lambda}\right)^{-1} \cdot \left(\boldsymbol{\Lambda}\mathbf{m} + X^\top y\right) \qquad a' = a + \frac{n}{2}$$

$$\boldsymbol{\Lambda}' = X^\top X + \boldsymbol{\Lambda} \qquad b' = b + \frac{1}{2}\left(y^\top y + \mathbf{m}^\top \boldsymbol{\Lambda} \mathbf{m} - \mathbf{m}'^\top \boldsymbol{\Lambda}' \mathbf{m}'\right)$$

Posterior Predictive

$$\hat{\boldsymbol{\beta}} = \mathbf{m}' \qquad \hat{y} = X^\top \hat{\boldsymbol{\beta}}$$

$$\hat{\sigma}^2 = \frac{\text{SumSq}\,(y, \hat{y})}{n}$$

'Rule-of-Thumb' Empirical Prior

$$a = regress(x) \qquad b = 1$$

$$c = 1 \qquad d = \frac{X^\top \cdot X}{n^2}$$

The Uniform Distribution.

Applying the CPR algorithm to the continuous uniform distribution is a difficult because of how it responds to virtual observations. The sufficient statistics for a uniform distribution are the number of observations n, the minimum x_{mn}, and the maximum x_{mx}. Given n observations from a uniform distribution, each additional observation x_{n+1} always increases n, but otherwise it *only* impacts the shape of the distribution when it falls outside the range of $[x_{\text{mn}}, x_{\text{mx}}]$ and changes the position of one of the endpoints. Thus, a "virtual" observation provided by a prior has very little effect on the distribution when $x_{\text{mn}} < x_{n+1} < x_{\text{mx}}$ but has a very powerful effect otherwise. This confounds the principle that one should use a 'weak' prior, because setting prior minimum and maximum values for the data either has a negligible effect, or a massive one, with no intermediary scenario.

When both the minimum λ_α and maximum λ_ω are unknown, setting a prior for the corresponding hyperparameters b_1 and b_2 has a determining effect on the viable range of values. This is reasonable in special cases where the system being modeled is very well defined, but in almost all empirical scenarios, the role played by the prior is to make an inference about the *support* of the distribution, which corresponds to a uniform step whose width is $(\lambda_\omega - \lambda_\alpha)$ and whose position is arbitrary. As such, an alternate parameterization of the conjugate prior only requires that the analyst specify prior observations a and the minimum allowable range d, which describes minimum variance (i.e. the maximum precision) of the measurement. For example, if observations are rounded to the nearest tens place, then $d = 10$ is an appropriate value.

In the more straightforward case that the minimum is known *a priori* to be 0.0, the prior maximum b and the allowable range d are identical. The likelihoods in any case when either the minimum or maximum is known can be reduced to this simple case by repositioning the origin to the known parameter and (in the case of a known maximum) reversing the resulting signs.

- Support:

$$\{x_1, \ldots, x_n | x_i \in \mathbb{R}\}$$

- Sufficient statistics:

$$n \text{ observations} \qquad x_{\text{mn}} = \min(x) \qquad x_{\text{mx}} = \max(x)$$

- With known minimum 0.0 and unknown maximum λ_ω:

Prior Hyperparameters

$\{a \mid a > 0\}$	Number of observations
$\{b \mid b > 0\}$	Minimum width of the interval

Conjugate Prior

$$\Pr\left(\lambda_\omega | a, b\right) = \frac{ab^a}{\left(\lambda_\omega\right)^{a+1}}$$

Normalizing Constant

$$m\left(a, b\right) = \frac{1}{ab^a} \tag{13}$$

Posterior Hyperparameters

$$a' = a + n \qquad b' = \max\left(b, x_{\text{mx}}\right)$$

Posterior Predictive

$$\hat{\lambda_\omega} = \max\left(b, x_{\text{mx}}\right)$$

'Rule-of-Thumb' Empirical Prior

$$a = 1 \qquad b = \text{minimum nonzero value of diff}\left(\text{sort}\left(x\right)\right)$$

- With unknown minimum λ_α and unknown maximum λ_ω:

Prior Hyperparameters

$\{a \mid a > 0\}$	Number of observations
$\{b_1, b_2 \mid b_1, b_2 \in \mathbb{R}\}$	Minimum and maximum values

OR

$\{a \mid a > 0\}$	Number of observations
$\{d \mid d > 0\}$	Minimum non-zero value for $\lambda_\omega - \lambda_\alpha$

Conjugate Prior

$$\Pr\left(\lambda_\alpha, \lambda_\omega | a, b_1, b_2\right) = \frac{a\left(a+1\right)\left(b_2 - b_1\right)^a}{\left(\lambda_\omega - \lambda_\alpha\right)^{a+2}} = \frac{a\left(a+1\right)\left(d\right)^a}{\left(\lambda_\omega - \lambda_\alpha\right)^{a+2}}$$

Normalizing Constant

$$m\left(a, b_1, b_2\right) = \frac{1}{a\left(a+1\right)\left(b_2 - b_1\right)^a} = \frac{1}{a\left(a+1\right)\left(d\right)^a} \tag{14}$$

Posterior Hyperparameters

$$a' = a + n \qquad b_1' = \min\left(b_1, x_{\text{mn}}\right) \qquad b_2' = \max\left(b_2, x_{\text{mx}}\right)$$

$$d' = \max\left(d, x_{\text{mx}} - x_{\text{mn}}\right)$$

Posterior Predictive

$$\hat{\lambda_\alpha} = b_1' \qquad \hat{\lambda_\omega} = b_2'$$

$$\hat{\lambda_\alpha} = \min\left(\left(\frac{x_{\text{mn}} + x_{\text{mx}} - d'}{2}\right), x_{\text{mn}}\right) \qquad \hat{\lambda_\omega} = \max\left(\left(\frac{x_{\text{mn}} + x_{\text{mx}} + d'}{2}\right), x_{\text{mx}}\right)$$

'Rule-of-Thumb' Empirical Prior

$$a = 1 \qquad d = \text{minimum nonzero value of diff}\left(\text{sort}\left(x\right)\right)$$

$$b_1 = N/A \qquad b_2 = N/A$$

Discrete Data (Multivariate)

The Multinomial Distribution.

• Support:

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{bmatrix}, x_{j,i} = 0 \text{ or } 1$$

• With unknown vector $\mathbf{w}$:

Prior Hyperparameters

$$\mathbf{a} = \{a_1, \dots, a_k | a_i > 0\} \quad \text{Occurences per category} - 1$$

Conjugate Prior

$$\Pr(\mathbf{w}|\mathbf{a}) = \frac{\Gamma(\sum a_i)}{\prod(\Gamma(a_i))} \prod w_i^{a_i - 1}$$

Normalizing Constant

$$m(\mathbf{a}) = \frac{\prod(\Gamma(a_i))}{\Gamma(\sum a_i)} \tag{15}$$

Posterior Hyperparameters

$$\mathbf{a}' = \{a'_1, \dots, a'_k\} \text{ given that } a'_i = a_i + \sum_{j=1}^{n} x_{j,i}$$

Posterior Predictive

$$\hat{w}_i = \frac{a'_i}{\sum \mathbf{a}'}$$

'Rule-of-Thumb' Empirical Prior

$$a_i = 1 \text{ for all } i$$

Continuous Data (Multivariate)

The Multivariate Normal Distribution.

As in the univariate case above, the support for this section is based on the demonstrations provided by by DeGroot (2005) and Murphy (2007).

Because covariance matrices are highly sensitive to outliers, robust covariance estimation is highly complex (Huber & Ronchetti, 2009). The rule-of-thumb prior provided in this section is based on the estimates for the mean and covariance provided by Campbell (1980), which was selected because its implementation is straightforward and closed-form. However, this method's robustness against outliers depends on the dimensionality of the data, tolerating at most $\frac{1}{k+1}$ outliers across the data's k dimensions (Rousseeuw & van Zomeren, 1990). Consequently, this estimator should be used with caution when the data display very high dimensionality.

• Support:

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{bmatrix}, x_{j,i} \in \mathbb{R}$$

- Sufficient statistics:

$$n \text{ observations} \qquad \bar{\mathbf{x}} = \left\{ \frac{1}{n}\sum_{j=1}^{n} x_{j,1}, \cdots, \frac{1}{n}\sum_{j=1}^{n} x_{j,k} \right\}$$

$$I_n = [n \times n] \text{ identity matrix} \qquad J_n = [n \times n] \text{ matrix of ones}$$

- With unknown vector $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$:

  Prior Hyperparameters

$\mathbf{m} = \{\mathrm{m}_1, \ldots, \mathrm{m}_k \vert \mathrm{m}_i \in \mathbb{R}\}$	Prior mean vector
$\{p \vert p > 0\}$	Observations supporting the prior mean
$\{a \vert a > k-1\}$	Observations supporting the prior precision
$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$	Prior precision matrix

  Conjugate Prior

$$\Pr\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \vert \mathbf{m}, p, a, \boldsymbol{\Lambda}\right) = \frac{p^{k/2}\left\vert\boldsymbol{\Lambda}\right\vert^{a/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\right) - \frac{p}{2}\left(\boldsymbol{\mu}-\mathbf{m}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\mathbf{m}\right)\right)}{\left(2\pi\right)^{k/2}\left\vert\boldsymbol{\Sigma}\right\vert^{(a+k)/2+1} 2^{ak/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{a+1-j}{2}\right)}$$

  Normalizing Constant

$$m\left(\mathbf{m}, p, a, \boldsymbol{\Lambda}\right) = \left(\frac{2\pi}{p}\right)^{k/2} \cdot \frac{2^{ak/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{a+1-j}{2}\right)}{\left\vert\boldsymbol{\Lambda}\right\vert^{a/2}} \tag{16}$$

  Posterior Hyperparameters

$$\mathbf{m}' = \frac{p}{p+n}\mathbf{m} + \frac{n}{p+n}\bar{\mathbf{x}} \qquad p' = p+n \qquad a' = a+n$$

$$\boldsymbol{\Lambda}' = \boldsymbol{\Lambda} + \mathbf{x}^{\top}\left(I_n - \frac{1}{n}J_n\right)\mathbf{x} + \frac{pn}{p+n}\mathrm{SumSq}\left(\bar{\mathbf{x}}, \mathbf{m}\right)$$

  Posterior Predictive

$$\hat{\boldsymbol{\mu}} = \mathbf{m}' \qquad \hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Lambda}'}{a'-k-1}$$

  'Rule-of-Thumb' Empirical Prior

$$a = robustmean(x) \qquad b = 1$$

$$c = 1 \qquad d = inv(robustcov(x))$$

References

Anderson, A. (1911). Labour legislation. In H. Chisholm (Ed.), *Encyclopædia britannica* (11th ed., Vol. 16, pp. 7–28). Encyclopedia Britannica Company.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, *79*, 551–563.

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., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical bayesian analysis of changepoint problems. *Applied Statistics*, *41*, 389–405.

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.

Cavanaugh, J. E., & Neath, A. A. (1999). Generalizing the derivation of the schwarz information criterion. *Communications in Statistics - Theory and Methods*, *28*, 49–66.

Chen, S. S., & Gopalakrishnan, P. S. (1998). Speaker, environment, and channel change detection and clustering via the bayesian information criterion. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 8-13.

DeGroot, M. H. (2005). *Optimal statistical decisions*. Wiley-Interscience.

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

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.

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

Hilborn, R. C. (2001). *Chaos and nonlinear dynamics* (2nd ed.). Oxford University Press.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). John Wiley & Sons.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, *66*, 191–193.

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.

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

Murphy, K. P. (2007). *Conjugate bayesian analysis of the gaussian distribution* (Tech. Rep.). University of British Columbia.

Ó Ruanaidh, J. J. K., & Fitzgerald, W. J. (1996). *Numerical bayesian methods applied to signal processing*. Springer.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*, 100–115.

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

Pesaran, M. H., Pettenuzzo, D., & Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, *73*, 1057–1084.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633–639.

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

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

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems* (Tech. Rep. No. 24). Department of Statistics, Stanford University.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* MIT Press.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, *92*, 937–950.

Yao, Y. C. (1988). Estimating the number of change-points vis schwarz' criterion. *Statistics and Probability Letters*, *6*, 181–189.