A Bayesian Bernoulli Change-Point Problem: Derivation, Posterior Analysis, and Probability of Correct Detection

Guneesh Vats

February 5, 2025

Abstract

This report addresses an offline Bayesian change-point problem in a Bernoulli (0/1) sequence. We derive the posterior distribution for the change-point, discuss how to form a point estimate of that change-point, and show how to characterize the probability of correctly identifying the true change-point. References to classic work in Bayesian and frequentist change-point literature are provided, including [1], [2], and [3].

1 Introduction and Setup

We consider a sequence of n observations:

$$X_1, X_2, \ldots, X_n \quad \text{where each } X_i \in \{0, 1\}.$$

There is exactly *one* unknown change-point r^* in $\{1, \ldots, n\}$ such that:

$$X_i \sim \begin{cases} \text{Bernoulli}(p_1), & i = 1, \ldots, r^*, \\ \text{Bernoulli}(p_2), & i = r^* + 1, \ldots, n, \end{cases}$$

with $p_1 \neq p_2$. We call r^* the *true* change-point.

Goal. After observing the entire data set $(X_1, \ldots, X_n)$, we want:

1. A *posterior distribution* over all possible change-points $r \in \{1, \ldots, n\}$, given the data.
2. A *point estimate* $\hat{r}(X)$ of the change-point (often the posterior mode).
3. The *probability of correct detection*, i.e. $\mathbb{P}[\hat{r}(X) = r^*]$, under the true data generating process.

In the Bayesian approach, we assign:

- A *prior* $p_0(r)$ over r, satisfying $\sum_{r=1}^{n} p_0(r) = 1$.
- If p_1 or p_2 are unknown, we also specify priors for them (e.g., Beta distributions).

We then use Bayes' theorem to derive the posterior distribution $p(r \mid X_1, \ldots, X_n)$.

2 Bayesian Posterior Derivation

2.1 Likelihood for a Candidate Change-Point

Suppose first that p_1 and p_2 are both *known.* For a candidate change-point r, the data split as:

$$X_1, \ldots, X_r \quad \text{and} \quad X_{r+1}, \ldots, X_n.$$

Since each X_i is Bernoulli(p_1) in the first segment, and Bernoulli(p_2) in the second, the likelihood of observing a particular data sequence $x = (x_1, \ldots, x_n)$ is

$$L(r) \;=\; p(x \mid r, p_1, p_2) \;=\; \prod_{i=1}^{r} p_1^{x_i}(1-p_1)^{1-x_i} \;\times\; \prod_{i=r+1}^{n} p_2^{x_i}(1-p_2)^{1-x_i}. \quad (1)$$

Equivalently, in exponent form,

$$L(r) = p_1^{\sum_{i=1}^{r} x_i} \left(1 - p_1\right)^{r - \sum_{i=1}^{r} x_i} \;\times\; p_2^{\sum_{i=r+1}^{n} x_i} \left(1 - p_2\right)^{(n-r) - \sum_{i=r+1}^{n} x_i}.$$

2.2 Prior on the Change-Point

We write $p_0(r)$ for the prior probability that the change occurs at r. For instance:

$$p_0(r) = \begin{cases} \dfrac{1}{n}, & \text{(uniform)}, \\ \text{other shapes}, & \text{if prior info suggests early or late changes.} \end{cases}$$

We always have $\sum_{r=1}^{n} p_0(r) = 1$. If we interpret $r = n$ as "no change" (all from Bernoulli(p_1)), then $p_0(n)$ can capture our prior belief in "no change" vs. $r < n$.

2.3 Posterior for r

By Bayes' theorem, the posterior probability of r given the data x is:

$$p(r \mid x) \;=\; \frac{L(r)\, p_0(r)}{\sum_{s=1}^{n} \big[L(s)\, p_0(s)\big]}, \tag{2}$$

where $L(r)$ is as in (1). This yields a discrete distribution over $r = 1, \ldots, n$.

2.4 Estimating the Change-Point

Often, we pick a single best estimate

$$\hat{r}(x) \;=\; \arg\max_{r}\; p(r \mid x) \quad \text{(the posterior mode).}$$

If $p_0(r)$ is *uniform*, then maximizing $p(r \mid x)$ is equivalent to maximizing $L(r)$; hence $\hat{r}(x)$ coincides with the classical Maximum Likelihood Estimator (MLE). If $p_0(r)$ is non-uniform, it modifies the estimate toward values favored by the prior.

3 Probability of Correct Detection

Let r^* be the *true* change-point, and the data be generated by

$$\begin{cases} X_i \sim \text{Bernoulli}(p_1), & i = 1, \ldots, r^*, \\ X_i \sim \text{Bernoulli}(p_2), & i = r^* + 1, \ldots, n. \end{cases}$$

We want to know

$$\mathbb{P}\big[\hat{r}(X) = r^*\big]$$

where $X = (X_1, \dots, X_n)$ is drawn according to the true process. Formally:

$$\mathbb{P}\big[\hat{r}(X) = r^*\big] \;=\; \sum_{x \in \{0,1\}^n} \mathbf{1}\{\hat{r}(x) = r^*\}\; \underbrace{p\big(x \mid r^*, p_1, p_2\big)}_{\text{true distribution}}. \tag{3}$$

Here, $\mathbf{1}\{\hat{r}(x) = r^*\}$ is the indicator that our estimate on data x matches the true change-point r^*.

3.1 No Simple Closed-Form for Finite n

Because $\hat{r}(x)$ is defined via the posterior $\arg\max_r p(r \mid x)$, and $p(x \mid r^*)$ is a Bernoulli mixture, summation over all 2^n possible data sequences x does not reduce to a neat closed-form expression.

Monte Carlo Approximation. A typical approach is to *simulate* many datasets $\{X^{(m)}\}_{m=1}^M$ from the true distribution and compute

$$\widehat{\mathbb{P}}(\hat{r} = r^*) \;\approx\; \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\hat{r}(X^{(m)}) = r^*\}.$$

As $M \to \infty$, this converges to (3).

3.2 Asymptotic Consistency

Classical results ([2], [3], and also [1]) show that as $n \to \infty$, if $p_1 \neq p_2$ and certain regularity conditions hold, the maximum-likelihood or Bayesian estimator $\hat{r}(X)$ is *consistent*. In other words:

$$\mathbb{P}\big[\hat{r}(X) = r^*\big] \;\longrightarrow\; 1 \quad \text{as } n \to \infty.$$

Hence for large sample sizes, the probability of exact detection of r^* goes to 1 (assuming a single abrupt change and non-degenerate p_1, p_2).

4 Extension to Unknown p_1, p_2

If p_1 or p_2 are unknown, we place *priors* on them—commonly Beta distributions for Bernoulli data. For instance:

$$p_1 \sim \mathrm{Beta}(a_1, b_1), \quad p_2 \sim \mathrm{Beta}(a_2, b_2), \quad \text{independently.}$$

Then for each candidate r, the *marginal likelihood* becomes

$$\int_0^1 \Big[\prod_{i=1}^{r} p_1^{x_i}(1-p_1)^{1-x_i}\Big]\,\pi(p_1)\,dp_1 \;\times\; \int_0^1 \Big[\prod_{i=r+1}^{n} p_2^{x_i}(1-p_2)^{1-x_i}\Big]\,\pi(p_2)\,dp_2,$$

where $\pi(p_1), \pi(p_2)$ denote the Beta prior densities. Conjugacy ensures each integral is a Beta function ratio. The posterior on r becomes

$$p(r \mid x) \;=\; \frac{\Big[\text{marginal-likelihood for segment } 1 \times \text{marginal-likelihood for segment } 2\Big] \times p_0(r)}{\sum_{s=1}^{n}\Big[\dots\Big] \times p_0(s)},$$

exactly analogous to (2), but replacing $L(r)$ with integrated expressions. The estimation and correct-detection probability is then defined the same way, only now

$$p\big(x \mid r^*, p_1, p_2\big)$$

could reflect the actual (true) p_1, p_2 even though in the model we are integrating them out as unknown. Again, no closed-form for $\mathbb{P}[\hat{r} = r^*]$ is typically available, so simulation-based or asymptotic reasoning is used.

5 Conclusion

We have shown that the Bayesian method for detecting a single change-point r in a Bernoulli sequence involves:

1. Specifying a prior $p_0(r)$ on the change index r.
2. Computing the (marginal) likelihood for each r.
3. Deriving the posterior $p(r \mid x)$ via Bayes' rule, (2).
4. Picking $\hat{r}(x) = \arg\max_r p(r \mid x)$ or another suitable posterior-based estimate.

The *probability of correct detection*, $\mathbb{P}[\hat{r}(X) = r^*]$, sums (or integrates) over all data sequences weighted by the true distribution. For large samples and distinct Bernoulli parameters $p_1 \neq p_2$, the estimator is *consistent*—the probability of exact detection approaches 1. For finite n, direct computation is often done via Monte Carlo.

Acknowledgments. This derivation is based on the framework introduced by Smith [1] and earlier foundational works on change-point estimation [2, 3].

References

[1] A. F. M. Smith (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2): 407–416.

[2] H. Chernoff and S. Zacks (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Annals of Mathematical Statistics*, 35: 999–1018.

[3] A. Kander and S. Zacks (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *Annals of Mathematical Statistics*, 37: 1196–1210.