

On the Prediction of a Nonstationary Bernoulli Distribution based on Bayes Decision Theory

Daiki Koizumi ^a

Otaru University of Commerce, 3-5-21, Midori, Otaru-city, Hokkaido, 045-8501, Japan

Keywords: Probability Model, Bayes Decision Theory, Nonstationary Bernoulli Distribution, Hierarchical Bayesian Model.

Abstract: A class of nonstationary Bernoulli distribution is considered in terms of Bayes decision theory. In this nonstationary class, the Bernoulli distribution parameter follows a random walking rule. Even if this general class is assumed, it is proved that the posterior distribution of the parameter can be obtained analytically with a known hyper parameter. With this theorem, the Bayes optimal prediction algorithm is proposed assuming the 0-1 loss function. Using real binary data, the predictive performance of the proposed model is evaluated comparing to that of a stationary Bernoulli model.


1 INTRODUCTION

Binary data is popular subject for data analysis and is a topic of frequent research (Cox, 1970). From the perspective of Bayesian statistics, the stationary Bernoulli distribution and the stationary binomial distribution are frequently used to deal with binary data (Press, 2003) (Bernardo and Smith, 2000) (Berger, 1985). For Bayesian posterior parameter estimation under the stationary Bernoulli and binomial distributions, one of the most reasonable approaches is to assume the beta distribution as the prior of the parameter. This assumption drastically reduces the computational cost of obtaining the posterior distribution of the parameter using the Bayes theorem, and this prior is called the *natural conjugate* (Bernardo and Smith, 2000) (Berger, 1985).

In contrast, there have been many approaches to generalize the stationarity of the parameter by considering certain aspects of the nonstationarity of the parameter. In general, assuming the nonstationarity of parameters requires additional parameters compared to the stationary model. Furthermore, if the Bayesian approach is used, it is often difficult to save computational cost when obtaining the posterior of the parameter. This point depends on the class of nonstationarity of the parameter, and one important result is the *SPSM, Simple Power Steady Model* (Smith, 1979), to the best of author's knowledge. Under SPSM, it is

guaranteed that the posterior of the parameter can be obtained analytically. Similar aspects were discussed from the generalized perspective of the Kalman filter (Harvey, 1989). Some researchers have tried to apply this result to the discrete probability distributions and proposed predictive algorithms (Koizumi et al., 2009) (Koizumi, 2020) (Koizumi et al., 2012) (Yasuda et al., 2001). Koizumi et al. assumed a nonstationary Poisson distribution and proposed the Bayes optimal prediction algorithm under the known nonstationary hyper parameter (Koizumi et al., 2009). Koizumi recently generalized this prediction algorithm to the credible interval prediction (Koizumi, 2020). They obtained better predictive performance compared to a stationary Poisson distribution with real web traffic data. They also assumed a nonstationary Bernoulli distribution to predict SQL injection attacks in the field of network security (Koizumi et al., 2012). However, they defined an incorrect class of nonstationary parameters. Furthermore, they did not show any proof that the posterior parameter distribution was analytically obtained under their nonstationary model. Yasuda et al. assumed a similar nonstationary Bernoulli distribution and proposed the Bayes optimal prediction algorithm under the known nonstationary hyper parameter (Yasuda et al., 2001). However, they did not present any proof that the posterior parameter distribution can be obtained analytically under the nonstationary model again.

In this paper, a class of nonstationary Bernoulli distribution is proposed. This class has only one ad-

^a  <https://orcid.org/0000-0002-5302-5346>

ditional hyper parameter to express the nonstationarity of the Bernoulli parameter. Moreover, the prediction problem is considered under the proposed nonstationary Bernoulli distribution. Bayes decision theory (Weiss and Blackwell, 1961) (Berger, 1985) (Bernardo and Smith, 2000) is a powerful theoretical framework to define the prediction error. In terms of Bayes decision theory, the predictive estimator that minimizes the average predictive error is called the Bayes optimal prediction. Considering this point, this paper proposes the Bayes optimal prediction algorithm under a certain class of nonstationary Bernoulli distribution, if the nonstationary hyper parameter is known. The predictive performance of the proposed algorithm was evaluated with real binary data. When considering real data, the above-mentioned hyper parameter should be estimated. For this purpose, this study takes the empirical Bayesian approach, and the objective parameter is estimated by the approximate maximum likelihood estimation with numerical calculation.

The remainder of this paper is organized as follows. Section 2 provides the basic definitions of the nonstationary Bernoulli distribution, and some lemmas and corollaries in terms of the hierarchical Bayesian modeling approach. Section 3 begins with the basic definitions in terms of Bayes decision theory, then proves the main theorems of the proposed nonstationary Bernoulli distribution, discusses the hyper (nonstationary) parameter estimation, and proposes the Bayes optimal prediction algorithm. Section 4 gives some numerical examples with real binary data. Section 5 discusses the results. Section 6 concludes this paper.

2 HIERARCHICAL BAYESIAN MODELING WITH NONSTATIONARY BERNOULLI DISTRIBUTION

2.1 Preliminaries

Let $t = 1, 2, \dots$ be a discrete time index and $X_t = x_t$ be a discrete random variable at t . Assume that $x_t \in \{0, 1\}$ and $X_t \sim \text{Bernoulli}(\theta_t)$ where $0 \leq \theta_t \leq 1$ is a nonstationary parameter. Then the probability function of the nonstationary Bernoulli distribution $p(x_t | \theta_t)$ is defined as the following:

Definition 2.1. *Nonstationary Bernoulli Distribution*

$$p(x_t | \theta_t) = \theta_t^{x_t} (1 - \theta_t)^{1-x_t}, \quad (1)$$

where $0 \leq \theta_t \leq 1$. □

Definition 2.2. *Function for Θ_t, A_t and B_t*

Let $\Theta_t = \theta_t, A_t = a_t$, and $B_t = b_t$ be random variables where A_t and B_t are mutually independent, then a function for Θ_t is defined as,

$$\Theta_t = \frac{A_t}{A_t + B_t}, \quad (2)$$

where $0 < a_t, 0 < b_t$. □

Definition 2.3. *Nonstationarity of A_t, B_t*

Let $C_t = c_t, D_t = d_t$ be random variables, then the nonstationary functions for A_t and B_t are defined as,

$$A_{t+1} = C_t A_t, \quad (3)$$

$$B_{t+1} = D_t B_t, \quad (4)$$

where $0 < c_t < 1, 0 < d_t < 1$ and they are sampled from the following two types of Beta distributions:

$$C_t \sim \text{Beta}[k\alpha_t, (1-k)\alpha_t], \quad (5)$$

$$D_t \sim \text{Beta}[k\beta_t, (1-k)\beta_t], \quad (6)$$

where k is a real valued constant and $0 < k \leq 1$. □

Definition 2.4. *Conditional Independence for A_t, C_t (or B_t, D_t) under α_t (or β_t)*

$$p(a_t, c_t | \alpha_t) = p(a_t | \alpha_t) p(c_t | \alpha_t), \quad (7)$$

$$p(b_t, d_t | \beta_t) = p(b_t | \beta_t) p(d_t | \beta_t). \quad (8)$$

□

Definition 2.5. *Initial Distributions for A_1, B_1*

$$A_1 \sim \text{Gamma}(\alpha_1, 1), \quad (9)$$

$$B_1 \sim \text{Gamma}(\beta_1, 1), \quad (10)$$

where $0 < \alpha_1$ and $0 < \beta_1$. □

Definition 2.6. *Initial Distributions for C_1, D_1*

$$C_1 \sim \text{Beta}[k\alpha_1, (1-k)\alpha_1], \quad (11)$$

$$D_1 \sim \text{Beta}[k\beta_1, (1-k)\beta_1]. \quad (12)$$

□

Definition 2.7. *Gamma Distribution for q* Gamma distribution of $\text{Gamma}(r, s)$ is defined as,

$$p(q | r, s) = \frac{s^r}{\Gamma(r)} q^{r-1} \exp(-sq), \quad (13)$$

where $0 < q, 0 < r, 0 < s$, and $\Gamma(r)$ is the gamma function defined in Definition 2.9. □

Definition 2.8. *Beta Distribution for q*

Beta distribution of $\text{Beta}(r, s)$ is defined as,

$$p(q | r, s) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} q^{r-1} (1-q)^{s-1}, \quad (14)$$

where $0 < q < 1, 0 < r, 0 < s$. □

Definition 2.9. *Gamma Function for q*

$$\Gamma(q) = \int_0^{+\infty} y^{q-1} \exp(-y) dy, \quad (15)$$

where $0 < q$. □

2.2 Lemmas

Lemma 2.1. *Transformed Distribution for A_t*

For any $t \geq 1$, the transformed random variable $A_{t+1} = C_t A_t$ in Definition 2.3 follows the following Gamma distribution:

$$A_{t+1} \sim \text{Gamma}(k\alpha_t, 1). \quad (16)$$

□

Proof of Lemma 2.1.

See APPENDIX A. □

Lemma 2.2. *Transformed Distribution for B_t*

For any $t \geq 1$, the transformed random variable $B_{t+1} = D_t B_t$ in Definition 2.3 follows the following Gamma distribution:

$$B_{t+1} \sim \text{Gamma}(k\beta_t, 1). \quad (17)$$

□

Proof of Lemma 2.2.

The proof is exactly same as Lemma 2.1, replacing A_{t+1} by B_{t+1} , C_t by D_t , and α_t by β_t .

This completes the proof of Lemma 2.2. □

Lemma 2.3. *Transformed Distribution for Θ_t*

For any $t \geq 2$, the transformed random variable $\Theta_t = \frac{A_t}{A_t + B_t}$ in Definition 2.2 follows the following Beta distribution:

$$\Theta_t \sim \text{Beta}(k\alpha_{t-1}, k\beta_{t-1}). \quad (18)$$

□

Proof of Lemma 2.3.

See APPENDIX B. □

Corollary 2.1. *Transformed Initial Distribution for Θ_1*

The transformed random variable $\Theta_1 = \frac{A_1}{A_1 + B_1}$ in Definition 2.2 follows the following Beta distribution:

$$\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1). \quad (19)$$

□

Proof of Corollary 2.1.

From Definition 2.5,

$$\begin{aligned} A_1 &\sim \text{Gamma}(\alpha_1, 1), \\ B_1 &\sim \text{Gamma}(\beta_1, 1). \end{aligned}$$

If Lemma 2.3 is applied to the above A_1 and B_1 , then the following holds.

$$\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1). \quad (20)$$

This completes the proof of Corollary 2.1. □

3 PREDICTION ALGORITHM BASED ON BAYES DECISION THEORY

3.1 Preliminaries

Definition 3.1. *Loss Function*

$$L(\hat{x}_{t+1}, x_{t+1}) = \begin{cases} 0 & \text{if } \hat{x}_{t+1} = x_{t+1}; \\ 1 & \text{if } \hat{x}_{t+1} \neq x_{t+1}. \end{cases} \quad (21)$$

□

Definition 3.2. *Risk Function*

$$\begin{aligned} R(\hat{x}_{t+1}, \theta_{t+1}) \\ = \sum_{x_{t+1}=0}^1 L(\hat{x}_{t+1}, x_{t+1}) p(x_{t+1} | \theta_{t+1}), \end{aligned} \quad (22)$$

where $p(x_{t+1} | \theta_{t+1})$ is from Definition 2.1. □

Definition 3.3. *Bayes Risk Function*

$$\begin{aligned} BR(\hat{x}_{t+1}) \\ = \int_0^1 R(\hat{x}_{t+1}, \theta_{t+1}) p(\theta_{t+1} | \mathbf{x}^t) d\theta_{t+1}. \end{aligned} \quad (23)$$

□

Definition 3.4. *Bayes Optimal Prediction*

The Bayes optimal prediction \hat{x}_{t+1}^* is obtained by,

$$\hat{x}_{t+1}^* = \arg \min_{\hat{x}_{t+1}} BR(\hat{x}_{t+1}). \quad (24)$$

□

3.2 Main Theorems

Theorem 3.1. *Posterior Distribution for θ_t*

Let the prior distribution of parameter θ_1 of the nonstationary Bernoulli distribution in Definition 2.1 be $\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$. For any $t \geq 2$, let $\mathbf{x}^{t-1} = (x_1, x_2, \dots, x_{t-1})$ be the observed data sequence. Then, the posterior distribution of $\Theta_t | \mathbf{x}^{t-1}$ can be obtained as the following closed form:

$$\Theta_t | \mathbf{x}^{t-1} \sim \text{Beta}(\alpha_t, \beta_t), \quad (25)$$

where the parameters α_t, β_t are given as,

$$\begin{cases} \alpha_t = k^{t-1} \alpha_1 + \sum_{i=1}^{t-1} k^{t-i} x_i; \\ \beta_t = k^{t-1} \beta_1 + \sum_{i=1}^{t-1} k^{t-i} (1 - x_i). \end{cases} \quad (26)$$

□

Proof of Theorem 3.1.

For any $t \geq 2$, the posterior of parameter distribution $p(\theta_t | \mathbf{x}^{t-1})$ remains in the closed form $\Theta_t \sim \text{Beta}(\alpha_t, \beta_t)$ if $X_t \sim \text{Bernoulli}(\theta_t)$ in Definition 2.1 and $\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$ in Corollary 2.1 according to the nature of *conjugate families* (Bernardo and Smith, 2000, 5.2, p.265) (Berger, 1985, 4.2.2, p.130).

Furthermore, assuming that x_{t-1} is the observed data,

$$\begin{cases} \alpha_t = \alpha_{t-1} + x_{t-1}; \\ \beta_t = \beta_{t-1} + 1 - x_{t-1}, \end{cases} \quad (27)$$

holds by conjugate analysis (Bernardo and Smith, 2000, Example 5.4, p.271). This is the proof of Eq. (25).

In this paper, nonstationary parameter model is assumed. Therefore, if both Lemma 2.1, and Lemma 2.2 are recursively applied to Eq. (27), then,

$$\begin{cases} \alpha_t = k(\alpha_{t-1} + x_{t-1}); \\ \beta_t = k(\beta_{t-1} + 1 - x_{t-1}), \end{cases} \quad (28)$$

holds.

Finally, Eq. (26) is obtained if Eq. (28) is recursively applied until the initial conditions α_1, β_1 from both Definition 2.5 and Corollary 2.1 appear.

This completes the proof of Theorem 3.1. \square

Remark 3.1.

For the second terms of the right hand sides of Eq. (26), each observed data x_i is exponentially weighted by k^{t-i} where $i = 1, 2, \dots, t-1$. This structure is called the *EWMA, Exponentially Weighted Moving Average* (Harvey, 1989, 6.6, p.350).

Theorem 3.2. Predictive Distribution

$$p(x_{t+1} | \mathbf{x}^t) = \begin{cases} \frac{\beta_{t+1}}{\alpha_{t+1} + \beta_{t+1}} & \text{if } x_{t+1} = 0; \\ \frac{\alpha_{t+1}}{\alpha_{t+1} + \beta_{t+1}} & \text{if } x_{t+1} = 1, \end{cases} \quad (29)$$

where α_{t+1} and β_{t+1} are in Eq. (26). \square

Proof of Theorem 3.2.

See APPENDIX C. \square

Theorem 3.3. Bayes Optimal Prediction

$$\hat{x}_{t+1}^* = \begin{cases} 0 & \text{if } \alpha_{t+1} < \beta_{t+1}; \\ 1 & \text{if } \alpha_{t+1} > \beta_{t+1}, \end{cases} \quad (30)$$

\square

Proof of Theorem 3.3.

In terms of Bayes decision theory (Weiss and Blackwell, 1961) (Berger, 1985) (Bernardo and Smith, 2000), the Bayes optimal prediction $\hat{x}_{t+1} = \hat{x}_{t+1}^*$ maximizes the predictive distribution $p(x_{t+1} | \mathbf{x}^t)$ if 0-1 loss function in Definition 3.1 is

defined. Since $\hat{x}_{t+1}^* \in \{0, 1\}$ and Theorem 3.2 holds, this maximization can be done by comparing just two cases. Therefore,

$$\begin{aligned} \hat{x}_{t+1}^* &= \arg \max_{x_{t+1}} p(x_{t+1} | \mathbf{x}^t) \\ &= \begin{cases} 0 & \text{if } \alpha_{t+1} < \beta_{t+1}; \\ 1 & \text{if } \alpha_{t+1} > \beta_{t+1}, \end{cases} \end{aligned}$$

This completes the proof of Theorem 3.3. \square

3.3 Hyper Parameter Estimation with Empirical Bayes Method

Since a hyper parameter $0 < k \leq 1$ in Definition 2.3 is assumed to be known, it should be estimated in practice. One of estimation methods can be the maximum likelihood estimation with numerical approximation in terms of empirical Bayes approach. This is,

$$\hat{k} = \arg \max_k L(k), \quad (31)$$

where $0 < k \leq 1$ and,

$$\begin{aligned} L(k) &= p(x_1 | \theta_1, k) p(\theta_1) \prod_{i=2}^t p(x_i | \mathbf{x}^{i-1}, k) \\ &= \prod_{i=1}^t \left(\frac{\beta_i}{\alpha_i + \beta_i} \right)^{1-x_i} \left(\frac{\alpha_i}{\alpha_i + \beta_i} \right)^{x_i}. \end{aligned} \quad (32)$$

Note that Eq. (32) is obtained by applying Theorem 3.2.

Therefore, its log-likelihood function $\log L(k)$ is,

$$\begin{aligned} \log L(k) &= \sum_{i=1}^t \{ (1-x_i) [\log \beta_i - \log(\alpha_i + \beta_i)] \\ &\quad + x_i [\log \alpha_i - \log(\alpha_i + \beta_i)] \}. \end{aligned} \quad (33)$$

Eqs. (31) and (33) can not be solved analytically and then the approximate numerical method should be applied.

3.4 Proposed Bayes Optimal Prediction Algorithm

Based on main Theorems in Subsection 3.2, the following Bayes optimal prediction algorithm is proposed.

Algorithm 3.1. Proposed Bayes Optimal Algorithm

1. Estimate hyper parameter k from training data by approximate maximum likelihood estimation with Eqs. (31) and (33).

2. Set $t = 1$ and define $\alpha_1 > 0, \beta_1 > 0$ in Definition 2.5 in order to set the initial prior of parameter distribution $\Theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$ in Corollary 2.1.
3. With test data sequence \mathbf{x}^t , update the posterior of nonstationary parameter distribution $p(\theta_{t+1} | \alpha_{t+1}, \beta_{t+1}, \mathbf{x}^t)$ with Eq. (26) in Theorem 3.1.
4. Calculate the predictive distribution $p(x_{t+1} | \mathbf{x}^t)$ in Theorem 3.2.
5. Obtain the Bayes optimal prediction \hat{x}_{t+1}^* in Theorem 3.3.
6. If $t < t_{max}$, then update $t \rightarrow t + 1$ and back to 3.
7. If $t = t_{max}$, then terminate the algorithm.

□

4 NUMERICAL EXAMPLES

This section shows numerical examples to evaluate the performance of Algorithm 3.1. Subsection 4.1 explains both the training and test data specifications. Training data is applied to estimate the hyper parameters: k in Definition 2.3 and α_1, β_1 in Definition 2.5, where the latter is used for the prior of parameter Θ_1 to predict test data. Test data were applied to evaluate the predictive performances of the proposed algorithm.

4.1 Binary Data Specifications

Table 1 and 2 show the training and test data specifications, respectively. These binary data were obtained from the daily rainfall data in Tokyo from January 1, 2018 to December 31, 2019 (Japan Meteorological Agency, 2020). Note that the threshold of binary data is defined by the following rule: i th daily rainfall: $x_i = 1$ if its amount is greater than 0.5 mm, otherwise $x_i = 0$.

Table 1: Training Data Specifications.

Items	Values
From:	January 1, 2018
To:	December 31, 2018
Total Days:	365

Table 2: Test Data Specifications.

Items	Values
From:	January 1, 2019
To:	December 31, 2019
Total Days:	365

4.2 Evaluations for Bayes Optimal Predictions

This subsection mainly evaluates two aspects of the Bayes optimal predictions from both the proposed nonstationary and conventional stationary Bernoulli distribution models. The first is the predictive performance between two models with non-informative priors. The second is that with informative priors.

4.2.1 Prediction Results with Non-informative Priors

Before evaluating the predictive performance, the hyper parameter \hat{k} is estimated using Eq. (31) from training data. This is the approximate maximum likelihood estimation with numerical calculation. The results are shown in Table 3.

Table 3: Estimated Hyper Parameter from Training Data.

Item	Value
\hat{k}	0.971

In this evaluation, the hyper parameters α_1 and β_1 of the prior distribution $p(\theta_1 | \alpha_1, \beta_1)$ are assumed to be non-informative. This initial prior should be a uniform distribution. The defined values of the hyper parameters are shown in Table 4.

Table 4: Defined Hyper Parameters for Non-informative Priors of Test Data.

α_1	β_1
1.000	1.000

Using \hat{k}, α_1 , and β_1 from Tables 3 and 4, the predictive errors for the proposed and stationary Bernoulli models $\sum_{i=1}^{365} L(\hat{x}_i, x_i)$ are calculated with test data. The results are shown in Table 5.

Table 5: Predictive Errors with Test Data for Proposed and Stationary Models with Non-informative Priors.

Items	Proposed	Stationary
$\sum_{i=1}^{365} L(\hat{x}_i, x_i)$	173	187

4.2.2 Prediction Results with Informative Priors

In this evaluation, the hyper parameters α_1 and β_1 of the prior distribution $p(\theta_1 | \alpha_1, \beta_1)$ are assumed to be informative. In this case, the empirical Bayesian approach is adopted. Both α_1 and β_1 are obtained from

the posterior distribution of $p(\theta_t | \mathbf{x}^t, \alpha_t, \beta_t)$ from the training data, and these are used as the initial prior of $p(\theta_1 | \alpha_1, \beta_1)$ to predict the test data. These values are listed in Table 6.

Table 6: Defined Hyper Parameters for Informative Priors of Test Data.

α_1	β_1
16.429	34.612

Using \hat{k}, α_1 , and β_1 from Tables 3 and 6, the predictive errors are calculated for both models with the test data. The results are shown in Table 7.

Table 7: Predictive Errors with Test Data for Proposed and Stationary Models with Informative Priors.

Items	Proposed	Stationary
$\sum_{i=1}^{365} L(\hat{x}_i, x_i)$	178	179

5 DISCUSSIONS

Table 5 shows that the total loss of the proposed nonstationary Bernoulli model is smaller than that of the stationary model, with accuracies of 52.6% and 48.8%, respectively. Moreover, the time series of the posterior probability $p(\theta_t = 1 | \mathbf{x}^t)$ ¹ is calculated and plotted in Figure 1. In Figure 1, the vertical axis is the posterior probability, the horizontal axis shows the indices of days, the red line is the time series of the posterior probabilities from the proposed model, and the blue line is that from the stationary model. From Figure 1, it can be observed that the posterior from the proposed model drifts more drastically than that of the stationary model. Thus, the extra hyper parameter k in the proposed model must work relatively well with a non-informative prior.

However, if the AIC, Akaike Information Criterion (Akaike, 1973) values for both models are calculated with test data, the values in Table 8 are obtained. From the perspective of model selection theory, the smaller the AIC value, the more appropriate the model is under the observed data. Table 8 indicates that the stationary model is more appropriate than the proposed model with test data. However, as mentioned above, the proposed model is superior to the stationary model in terms of predictive performance. Thus, the result of the first evaluation with a non-informative prior cannot be explained by AIC

¹Each value is the daily probability of rainfall.

with the specific test data in this paper.

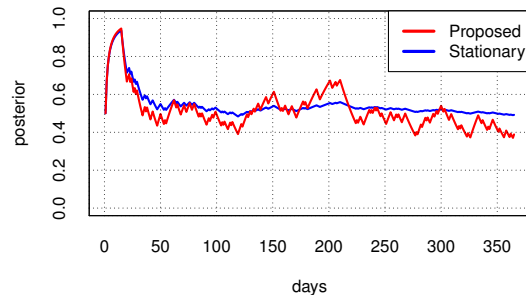


Figure 1: Posterior Probability Plot of $p(\theta_t = 1 | \mathbf{x}^t)$ with Non-informative Priors.

Table 8: AIC values for Proposed and Stationary Models with Non-informative Priors.

Items	Proposed	Stationary
AIC	-500.476	-505.316

In contrast, according to Table 7, the difference in the predictive performance for both models becomes smaller than that of the first evaluation. In fact, the result is almost a draw, with accuracies of 51.2% and 51.0%, respectively. Moreover, Figure 2 shows the time series of the posterior rainfall probability for both models. Note that an informative prior is assumed in this evaluation. From Figure 2, the first 50 points of the time series of the proposed model (red line) are more stable than those of the proposed model in Figure 2. This difference can be interpreted as the effect of informative priors. However, the predictive performance becomes worse in the proposed model. In this case, it can be considered that the setting of the informative prior weakens the effect of the estimated nonstationary hyper parameter \hat{k} . From Figure 2, the entire blue plot of the stationary model becomes more stable than that of the stationary model in Figure 1. In this case, the posterior of the stationary model almost converges, and its predictive performance is improved effectively as shown by the comparison of the results from Tables 5 and 7.

Table 9 shows the AIC values for both models. From the perspective of AIC, the value of the proposed model with the informative prior become slightly smaller than that of the proposed model with the non-informative prior. For the stationary model, this difference becomes larger. Thus, the theory of AIC explains the predictive performance of the stationary model well. However, the same situation does not hold true for the proposed model.

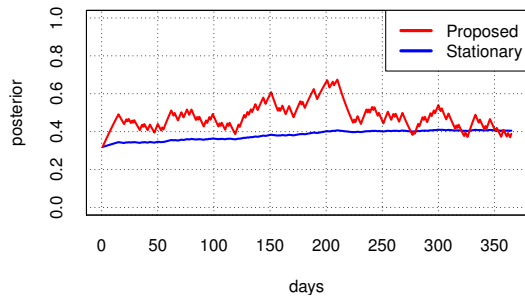


Figure 2: Posterior Probability Plot of $p(\theta_t = 1 | \mathbf{x}^t)$ with Informative Priors.

Table 9: AIC values for Proposed and Stationary Models with Informative Priors.

Items	Proposed	Stationary
AIC	-505.776	-522.896

6 CONCLUSIONS

This paper proposes a class of nonstationary Bernoulli distribution and the Bayes optimal prediction algorithm under the known nonstationary hyper parameter. The proposed class has only one extra hyper parameter compared to the stationary Bernoulli distribution, and it is proved that the posterior distribution of the Bernoulli parameter is obtained analytically. Furthermore, the predictive performance of the proposed algorithm is evaluated using real binary data. As a result, a certain advantage for predictive performance is discovered by comparing the results to those of the stationary Bernoulli model; however, this point cannot be explained in terms of model selection theory.

As important factor in the abovementioned advantage is the additional nonstationary hyper parameter in the proposed model. Because the empirical Bayesian approach is used in this study and the additional hyper parameter is estimated by the approximate maximum likelihood estimation, the objective likelihood function should be analyzed in detail. This point will be left for future work.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, pages 267–281.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian Theory*. John Wiley & Sons, Chichester.

Cox, D. R. (1970). *The Analysis of Binary Data*. Chapman and Hall, London.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Marsa, Malta.

Japan Meteorological Agency (2020). ClimatView (in Japanese). <https://www.data.jma.go.jp/gmd/cpd/monitor/dailyview/>. Browsing Date: Nov. 27, 2020.

Koizumi, D. (2020). Credible interval prediction of a non-stationary poisson distribution based on bayes decision theory. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 995–1002, Valletta, Malta. INSTICC, SciTePress.

Koizumi, D., Matsuda, T., and Sonoda, M. (2012). On the automatic detection algorithm of cross site scripting (xss) with the non-stationary bernoulli distribution. In *The 5th International Conference on Communications, Computers and Applications (MIC-CCA2012)*, pages 131–135, Istanbul, Turkey. IEEE.

Koizumi, D., Matsushima, T., and Hirasawa, S. (2009). Bayesian forecasting of www traffic on the time varying poisson model. In *Proceeding of The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'09)*, volume II, pages 683–689, Las Vegas, NV, USA. CSREA Press.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, Hoboken.

Smith, J. Q. (1979). A generalization of the bayesian steady forecasting model. *Journal of the Royal Statistical Society - Series B*, 41:375–387.

Weiss, L. and Blackwell, D. (1961). *Statistical Decision Theory*. McGraw-Hill, New York.

Yasuda, G., Nomura, R., and Matsushima, T. (2001). A study of coding for sources with nonstationary parameter (in Japanese). *Technical Report of IEICE (IT2001-15)*, 101(177):25–30.

APPENDIX

A: Proof of Lemma 2.1

Suppose $t = 1$, $A_1 = a_1$ and $C_1 = c_1$ are defined as,

$$A_1 \sim \text{Gamma}(\alpha_1, 1), \quad (34)$$

$$C_1 \sim \text{Beta}[k\alpha_1, (1-k)\alpha_1], \quad (35)$$

according to Definition 2.5 and Definition 2.6, respectively.

Since $A_2 = C_1 A_1$ from Definition 2.3, and A_t and C_t are conditional independent from Definition 2.4,

the joint distribution of $p(c_1, a_1)$ becomes,

$$\begin{aligned}
 p(c_1, a_1) &= p[c_1 | k\alpha_1, (1-k)\alpha_1] p(a_1 | \alpha_1, 1) \\
 &= \frac{\Gamma(\alpha_1)}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} c_1^{k\alpha_1-1} (1-c_1)^{(1-k)\alpha_1-1} \\
 &\quad \cdot \frac{a_1^{\alpha_1-1}}{\Gamma(\alpha_1)} \exp(-a_1) \\
 &= \frac{c_1^{k\alpha_1-1} (1-c_1)^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} a_1^{\alpha_1-1} \exp(-a_1).
 \end{aligned}$$

Now, denote the two transformation as,

$$\begin{cases} v = a_1 c_1; \\ w = a_1 (1 - c_1), \end{cases} \tag{36}$$

where $0 < v, 0 < w$.

Then, the inverse transformation of Eq. (36) becomes,

$$\begin{cases} a_1 = v + w; \\ c_1 = \frac{v}{v+w}, \end{cases} \tag{37}$$

The Jacobian J_1 of Eq. (37) is,

$$\begin{aligned}
 J_1 &= \begin{vmatrix} \frac{\partial a_1}{\partial v} & \frac{\partial a_1}{\partial w} \\ \frac{\partial c_1}{\partial v} & \frac{\partial c_1}{\partial w} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ \frac{w}{(v+w)^2} & -\frac{v}{(v+w)^2} \end{vmatrix} \\
 &= -\frac{1}{v+w} = -\frac{1}{a_1} \neq 0.
 \end{aligned}$$

Then, the transformed joint distribution $p(v, w)$ is obtained by the product of $p(c_1, a_1)$ and the absolute value of J_1 .

$$\begin{aligned}
 p(v, w) &= p(c_1, a_1) \left| -\frac{1}{a_1} \right| \\
 &= \frac{\left(\frac{v}{v+w}\right)^{k\alpha_1-1} \left(\frac{w}{v+w}\right)^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \\
 &\quad \cdot (v+w)^{\alpha_1-1} \exp[-(v+w)] \cdot \frac{1}{v+w} \\
 &= \frac{v^{k\alpha_1-1} w^{(1-k)\alpha_1-1}}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \exp[-(v+w)]. \tag{38}
 \end{aligned}$$

Then, $p(v)$ is obtained by marginalizing Eq. (38) with respect to w ,

$$\begin{aligned}
 p(v) &= \int_0^\infty p(v, w) dw \\
 &= \frac{v^{k\alpha_1-1} \exp(-v)}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \\
 &\quad \cdot \int_0^\infty w^{(1-k)\alpha_1-1} \exp(-w) dw \\
 &= \frac{v^{k\alpha_1-1} \exp(-v)}{\Gamma(k\alpha_1)\Gamma[(1-k)\alpha_1]} \cdot \Gamma[(1-k)\alpha_1] \\
 &= \frac{1}{\Gamma(k\alpha_1)} v^{k\alpha_1-1} \exp(-v). \tag{39}
 \end{aligned}$$

Eq. (39) exactly corresponds to $Gamma(k\alpha_1, 1)$ according to Definition 2.7. Recalling $v = a_1 c_1$ from Eq. (36) and $A_2 = C_1 A_1$ from Definition 2.3,

$$A_2 \sim Gamma(k\alpha_1, 1).$$

Thus if $t = 1$, then $A_{t+1} \sim Gamma(k\alpha_t, 1)$ holds.

For $t \geq 2$, by substituting $\alpha_t = k\alpha_{t-1}$, $A_t = a_t$ and $C_t = c_t$ are defined as,

$$A_t \sim Gamma(\alpha_t, 1), \tag{40}$$

$$C_t \sim Beta[k\alpha_t, (1-k)\alpha_t]. \tag{41}$$

Eqs. (40) and (41) correspond to Eqs. (34) and (35), respectively. Therefore the same proof can be applied for the case of $t \geq 2$ and it can be proved that,

$$\forall t, A_{t+1} \sim Gamma(k\alpha_t, 1).$$

This completes the proof of Lemma 2.1. □

B: Proof of Lemma 2.3

From Lemma 2.1 and 2.2,

$$\forall t \geq 2, A_t \sim Gamma(k\alpha_{t-1}, 1),$$

$$\forall t \geq 2, B_t \sim Gamma(k\beta_{t-1}, 1).$$

According to Definition 2.2, two random variables A_t and B_t are independent. Therefore, the joint distribution pf $p(a_t, b_t)$ becomes,

$$\begin{aligned}
 p(a_t, b_t) &= p(a_t | k\alpha_{t-1}, 1) p(b_t | k\beta_{t-1}, 1) \\
 &= \left[\frac{a_t^{k\alpha_{t-1}-1} \exp(-a_t)}{\Gamma(k\alpha_{t-1})} \right] \cdot \left[\frac{b_t^{k\beta_{t-1}-1} \exp(-b_t)}{\Gamma(k\beta_{t-1})} \right] \\
 &= \frac{a_t^{k\alpha_{t-1}-1} b_t^{k\beta_{t-1}-1}}{\Gamma(k\alpha_{t-1})\Gamma(k\beta_{t-1})} \exp[-(a_t + b_t)].
 \end{aligned}$$

Denoting the two transformations,

$$\begin{cases} \lambda = a_t + b_t; \\ \mu = \frac{a_t}{a_t + b_t}, \end{cases} \tag{42}$$

where $0 < \lambda, 0 < \mu$.

The inverse transformation of Eq. (42) becomes,

$$\begin{cases} a_t = \lambda\mu; \\ b_t = \lambda(1-\mu). \end{cases} \tag{43}$$

Then, the Jacobian J_2 of Eq. (43) is,

$$\begin{aligned}
 J_2 &= \begin{vmatrix} \frac{\partial a_t}{\partial \lambda} & \frac{\partial a_t}{\partial \mu} \\ \frac{\partial b_t}{\partial \lambda} & \frac{\partial b_t}{\partial \mu} \end{vmatrix} = \begin{vmatrix} \mu & \lambda \\ 1-\mu & -\lambda \end{vmatrix} \\
 &= -\lambda = -(a_t + b_t).
 \end{aligned}$$

Then, the transformed joint distribution $p(\lambda, \mu)$ is obtained by the product of $p(a_t, b_t)$ and the absolute value of J_2 as the following,

$$\begin{aligned}
 p(\lambda, \mu) &= p(a_t, b_t) \cdot |-(a_t + b_t)| \\
 &= \frac{(\lambda\mu)^{k\alpha_t-1} [\lambda(1-\mu)]^{k\beta_t-1}}{\Gamma(k\alpha_t)\Gamma(k\beta_t)} \exp(-\lambda) \cdot \lambda \\
 &= \frac{\mu^{k\alpha_t-1} (1-\mu)^{k\beta_t-1}}{\Gamma(k\alpha_t)\Gamma(k\beta_t)} \lambda^{k\alpha_t+k\beta_t-1} \exp(-\lambda).
 \end{aligned} \tag{44}$$

Then, $p(\mu)$ is obtained by marginalizing Eq. (44) with respect to λ ,

$$\begin{aligned}
 p(\mu) &= \int_0^\infty p(\lambda, \mu) d\lambda \\
 &= \frac{\mu^{k\alpha_t-1} (1-\mu)^{k\beta_t-1}}{\Gamma(k\alpha_t)\Gamma(k\beta_t)} \cdot \int_0^\infty \lambda^{k\alpha_t+k\beta_t-1} \exp(-\lambda) d\lambda \\
 &= \frac{\mu^{k\alpha_t-1} (1-\mu)^{k\beta_t-1}}{\Gamma(k\alpha_t)\Gamma(k\beta_t)} \cdot \Gamma(k\alpha_t + k\beta_t) \\
 &= \frac{\Gamma(k\alpha_t + k\beta_t)}{\Gamma(k\alpha_t)\Gamma(k\beta_t)} \mu^{k\alpha_t-1} (1-\mu)^{k\beta_t-1}.
 \end{aligned} \tag{45}$$

Eq. (45) exactly corresponds to $Beta(k\alpha_t, k\beta_t)$ according to Definition 2.8.

Recalling $\mu = \frac{a_t}{a_t+b_t}$ from Eq. (42) and $\Theta_t = \frac{A_t}{A_t+B_t}$ from Definition 2.2,

$$\forall t \geq 2, \Theta_t \sim Beta(k\alpha_{t-1}, k\beta_{t-1}),$$

holds.

This completes the proof of Lemma 2.3. \square

C: Proof of Theorem 3.2

Since the predictive distribution is Binomial-Beta distribution (Bernardo and Smith, 2000, p.117), $p(x_{t+1} | \mathbf{x}^t)$ becomes,

$$\begin{aligned}
 p(x_{t+1} | \mathbf{x}^t) &= \int_0^1 p(x_{t+1} | \theta_{t+1}) p(\theta_{t+1} | \mathbf{x}^t) d\theta_{t+1} \\
 &= c \cdot \Gamma(\alpha_{t+1} + x_{t+1}) \Gamma(\beta_{t+1} + 1 - x_{t+1}),
 \end{aligned}$$

where $c = \frac{\Gamma(\alpha_{t+1} + \beta_{t+1})}{\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1})\Gamma(\alpha_{t+1} + \beta_{t+1} + 1)}$.

If $x_{t+1} = 0$, then,

$$\begin{aligned}
 p(x_{t+1} | \mathbf{x}^t) &= \frac{\Gamma(\alpha_{t+1} + \beta_{t+1})}{\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1})\Gamma(\alpha_{t+1} + \beta_{t+1} + 1)} \\
 &\quad \cdot \Gamma(\alpha_{t+1})\Gamma(\beta_{t+1} + 1) \\
 &= \frac{\Gamma(\alpha_{t+1} + \beta_{t+1})}{\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1})(\alpha_{t+1} + \beta_{t+1})\Gamma(\alpha_{t+1} + \beta_{t+1})} \\
 &\quad \cdot \Gamma(\alpha_{t+1})\beta_{t+1}\Gamma(\beta_{t+1}) \\
 &= \frac{\beta_{t+1}}{\alpha_{t+1} + \beta_{t+1}}.
 \end{aligned} \tag{46}$$

Note that Eq. (46) is obtained by applying the following property of Gamma function: $\Gamma(q + 1) = q\Gamma(q)$.

If $x_{t+1} = 1$, then,

$$\begin{aligned}
 p(x_{t+1} | \mathbf{x}^t) &= \frac{\Gamma(\alpha_{t+1} + \beta_{t+1})}{\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1})\Gamma(\alpha_{t+1} + \beta_{t+1} + 1)} \\
 &\quad \cdot \Gamma(\alpha_{t+1} + 1)\Gamma(\beta_{t+1}) \\
 &= \frac{\Gamma(\alpha_{t+1} + \beta_{t+1})}{\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1})(\alpha_{t+1} + \beta_{t+1})\Gamma(\alpha_{t+1} + \beta_{t+1})} \\
 &\quad \cdot \alpha_{t+1}\Gamma(\alpha_{t+1})\Gamma(\beta_{t+1}) \\
 &= \frac{\alpha_{t+1}}{\alpha_{t+1} + \beta_{t+1}}.
 \end{aligned}$$

This completes the proof of Theorem 3.2. \square