

ON THE USE OF LOSS FUNCTIONS IN THE CHANGEPOINT PROBLEM

IRWIN GUTTMAN AND ULRICH MENZEFRICKE

(Received Sept. 30, 1981; revised Dec. 21, 1981)

Summary

We consider a sequence of independent random variables whose densities depend on a parameter which is subject to a change at an unknown time point. A Bayesian decision-theoretic approach is used to obtain an optimal choice of changepoint. The exponential and multivariate normal models are analyzed, and some numerical examples are given.

1. Introduction

Consider a sequence of n independent random variables y_t (possibly vector-valued) whose densities depend on a parameter (vector) θ . The value of this parameter or of some of its components changes at an unknown time point k so that the densities of y_t are $f(y_t|\theta_1)$ for $t=1, \dots, k$ and $f(y_t|\theta_2)$ for $t=k+1, \dots, n$. We are interested in the unknown changepoint $k \in \{1, 2, \dots, n-1\}$ and impose a loss structure on the problem, i.e., let $L(\hat{k}; k, \theta_1, \theta_2)$ be the loss incurred if it is decided that the change occurred at point \hat{k} when it actually occurred at point k . This loss depends on the magnitude of the change. A changepoint is then chosen to minimize the expected loss.

The changepoint problem has a long history. Most studies have been done about a change in the mean of a sequence of random variables, e.g., Srivastava and Sen [11], Smith [10], Lee and Heghinian [8] and Cobb [4]. Wichern, Miller and Hsu [12], and Menzefricke [9] examine changes in a variance. Hsu [5] examines a change in the scale parameter of a gamma distribution, and Chin Choy and Broemeling [3] examine a changing linear model.

In the next section we will briefly outline our general approach, followed by one section each for an exponential and a multivariate normal model. Some concluding remarks are offered in a final section.

Keywords: Changepoint problem; exponential; multivariate normal; loss function.

2. General model

We assume that prior information is available about the value of the parameter both before and after the change, expressed in a prior distribution $p(\theta_1, \theta_2)$. Prior information about the changepoint k is also available and is denoted by $p(k)$, $k=1, \dots, n-1$. We assume (θ_1, θ_2) and k to be independent. This prior information can be combined with the n observations, to yield a posterior distribution for θ_1 , θ_2 and k ,

$$(2.1) \quad p(k, \theta_1, \theta_2 | D) \propto p(k)p(\theta_1, \theta_2) \prod_{t=1}^k f(y_t | \theta_1) \prod_{t=k+1}^n f(y_t | \theta_2),$$

where $D=(y_1, \dots, y_n)$. Note that some of the parameters of the model may not change, i.e., θ_1 and θ_2 may have common components.

A decision regarding the changepoint can be made by minimizing the expected loss,

$$(2.2) \quad E L(\hat{k}) = \sum_{k=1}^{n-1} \int \int L(\hat{k}; k, \theta_1, \theta_2) p(k, \theta_1, \theta_2 | D) d\theta_2 d\theta_1.$$

The choice of loss function $L(\hat{k}; k, \theta_1, \theta_2)$ depends on the practical context of the problem and suggesting a particular form for all situations is clearly misguided. For purposes of illustrating our methods, we have chosen forms which are relatively simple to work with and lead to tractable results. Often $L(\hat{k}; k, \theta_1, \theta_2)$ will have special structure and simplification of (2.2) is then possible.

Let us consider a few special cases for which the loss function factors into two parts, one that depends only on \hat{k} and k , and another that depends only on θ_1 and θ_2 ,

$$L(\hat{k}; k, \theta_1, \theta_2) = L_1(\hat{k}, k) L_2(\theta_1, \theta_2).$$

Then (2.2) simplifies to

$$(2.3) \quad E L(\hat{k}) = \sum_{k=1}^{n-1} L_1(\hat{k}, k) E L_2(k, D) p(k | D),$$

where $E L_2(k, D) = \int \int L_2(\theta_1, \theta_2) p(\theta_1, \theta_2 | k, D) d\theta_1 d\theta_2$ and $p(\theta_1, \theta_2 | k, D)$ is the posterior distribution for θ_1 and θ_2 given the changepoint is at k . Within this framework, various simplifications are possible depending on the nature of L_1 and L_2 . We briefly list two such simplifications:

(i)

$$(2.4) \quad L_1(\hat{k}, k) = \begin{cases} 0 & \text{if } \hat{k} = k, \\ 1 & \text{if } \hat{k} \neq k. \end{cases}$$

Then, as is easily seen from (2.3), obtaining the optimal estimate of the changepoint is equivalent to finding the value of k which maximizes $R(k) = E L_2(k, D) p(k|D)$. If, in addition to (2.4), $L_2(\theta_1, \theta_2) = \text{constant}$, then $E L_2(k, D) = \text{constant}$ and the optimal choice of the estimate of the changepoint leads to finding the mode of the posterior distribution of k , $p(k|D)$.

$$(ii) \quad L_1(\hat{k}, k) = (\hat{k} - k)^2.$$

Then the minimizing value of \hat{k} in (2.3) is the integer k closest to

$$\left[\frac{\sum_{k=1}^{n-1} k E L_2(k, D) p(k|D)}{\sum_{k=1}^{n-1} E L_2(k, D) p(k|D)} \right].$$

If, in addition, $L_2(\theta_1, \theta_2) = \text{constant}$, then \hat{k} is simply the integer closest to the posterior mean of k .

3. Exponential case

3.1. Derivation of results

Let y_t , $t=1, \dots, n$, be a sequence of independently distributed exponential random variables with density $f(y_t|\theta_t) = \theta_t^{-1} \exp(-y_t/\theta_t)$, where $\theta_t = \theta_1$ for $1 \leq t \leq k$ and $\theta_t = \theta_2$ for $k+1 \leq t \leq n$. The prior distribution for θ_j is taken to be the inverse gamma with parameters t_j and s_j , $p(\theta_j) \propto \theta_j^{-t_j+1} \exp(-s_j/\theta_j)$, and the prior distribution for k is denoted by $p(k)$, $1 \leq k \leq n-1$. The posterior distribution for k , θ_1 and θ_2 is

$$(3.1) \quad p(k, \theta_1, \theta_2 | D) \propto \prod_{j=1}^2 \theta_j^{-t_{jk}+1} \exp(-s_{jk}/\theta_j) p(k),$$

$$\theta_1 > 0, \theta_2 > 0, \text{ and } k=1, 2, \dots, n-1,$$

where $t_{1k} = t_1 + k$, $t_{2k} = t_2 + n - k$, $s_{1k} = s_1 + \sum_{t=1}^k y_t$ and $s_{2k} = s_2 + \sum_{t=k+1}^n y_t$. The posterior distribution for the changepoint k is, for $k=1, \dots, n-1$,

$$(3.2) \quad p(k|D) \propto \Gamma(t_{1k}) s_{1k}^{-t_{1k}} \Gamma(t_{2k}) s_{2k}^{-t_{2k}} p(k),$$

and that for the magnitude $\zeta = \theta_2/\theta_1$ of the change given k is

$$(3.3) \quad p(\zeta|k, D) \propto \zeta^{t_{1k}-1} (s_{1k}\zeta + s_{2k})^{-(t_{1k}+t_{2k})},$$

i.e., $[(s_{1k}/t_{1k})/(s_{2k}/t_{2k})]\zeta$ has an F -distribution with $2t_{1k}$ and $2t_{2k}$ degrees of freedom.

We will next examine a loss function as in (2.4). A measure of change is $\zeta = \theta_2/\theta_1$, the ratio of the parameter values after and before the change. When ζ is very close to 1, failure to detect a change is not very serious. The loss $L_2(\theta_1, \theta_2)$ should thus increase as $\zeta = \theta_2/\theta_1$ moves away from 1. One convenient choice for $L_2(\theta_1, \theta_2)$ is $L_2(\theta_1, \theta_2) =$

$L_2(\zeta) = (\zeta - 1)^2$. The optimal changepoint estimate can then be found by choosing the value of k maximizing

$$(3.4) \quad R(k) = \left\{ \int (\zeta - 1)^2 p(\zeta | k, D) d\zeta \right\} p(k | D) \\ = \{ \text{Var}(\zeta | k, D) + [E(\zeta | k, D) - 1]^2 \} p(k | D),$$

where $E(\zeta | k, D) = (s_{2k}/(t_{2k} - 1)) / (s_{1k}/t_{1k})$, ($t_{2k} > 1$), and

$$(3.5) \quad \text{Var}(\zeta | k, D) = E^2(\zeta | k, D) \left[\frac{t_{1k} + t_{2k} - 1}{t_{1k}(t_{2k} - 2)} \right], \quad (t_{2k} > 2).$$

3.2. A numerical illustration

In this section we will present the result of a simulation study to illustrate the results for the exponential case. The simulation proceeded in the following way:

1. Generate n exponential variates x_t , $t=1, \dots, n$, from the exponential distribution with parameter 1.
2. Let q vary from 1 to $n-1$ in increments of 1. For each value of q , let $y_t = x_t$, $t=1, \dots, q$ and let $y_t = \zeta x_t$, $t=q+1, \dots, n$; find the mode k_q^p of the posterior distribution of the changepoint (3.2), and find the value of k , denoted by k_q^r , associated with the largest value of $R(k)$ from (3.4).
3. Repeat (1.) (2.) 300 times and find the average and standard deviation of the 300 values for k_q^p and k_q^r , denoted by $E(k_q^p)$, $E(k_q^r)$, $SD(k_q^p)$ and $SD(k_q^r)$.

In Table 1 are presented results for $n=20$, and $\zeta=3$ and 9. The prior distributions assumed are diffuse, i.e., $s_j = t_j = 0$, $j=1, 2$, in (3.1) to (3.4). An interesting point that emerges here is that this specification implies that $R(k)$ in (3.4) is not defined for $k=18$ and $k=19$, which may be seen on consulting (3.5).

Examining the results in Table 1, we find that use of k_q^r can lead to better (worse) decisions than k_q^p when q , the actual changepoint, is small (large). For example, consider the case where $\zeta=3$, i.e., the parameter of the exponential distribution after the change, θ_2 , is three times what it was before the change, θ_1 . When the actual changepoint occurred at $k=5$, the average changepoint selected based on the posterior mode is $E(k_q^p) = 7.6$, whereas that based on the loss function $L(\theta_1, \theta_2) = [(\theta_2/\theta_1) - 1]^2$, $E(k_q^r) = 5.4$, is closer to the correct value. In the latter case, however, the variability in the chosen values is slightly higher.

A different approach to this problem may be found in Hsu [5], who examined a related problem using the classical approach involving a change in the scale parameter of a sequence of gamma variates.

Table 1. Results of the simulation

q	$\zeta=3$				$\zeta=9$			
	$E(k_q^p)$	$SD(k_q^p)$	$E(k_q^r)$	$SD(k_q^r)$	$E(k_q^p)$	$SD(k_q^p)$	$E(k_q^r)$	$SD(k_q^r)$
1	8.9	5.8	4.8	5.8	6.8	6.0	1.8	2.9
2	7.8	5.6	4.5	5.1	4.5	4.6	2.2	2.0
3	7.3	5.0	4.7	4.7	4.5	3.3	2.9	1.4
4	7.3	4.6	5.0	4.5	4.9	2.5	3.8	1.5
5	7.6	4.2	5.4	4.5	5.6	2.0	4.9	1.8
6	7.9	3.7	6.1	4.5	6.4	1.4	5.8	2.0
7	8.6	3.7	6.7	4.5	7.5	1.5	6.9	1.9
8	9.1	3.5	7.5	4.7	8.3	1.3	7.9	2.0
9	9.7	3.3	8.3	4.8	9.2	1.3	8.7	2.1
10	10.4	3.5	9.0	4.9	10.2	1.3	9.7	2.2
11	11.1	3.5	9.6	5.0	11.1	1.4	10.8	2.4
12	11.7	3.7	10.3	5.2	12.1	1.5	11.8	2.4
13	12.3	3.9	10.8	5.4	13.0	1.7	12.7	2.6
14	12.8	4.2	11.5	5.8	14.0	1.8	13.8	2.9
15	13.1	4.6	12.3	5.8	14.9	1.9	14.5	3.2
16	13.5	4.8	12.8	6.1	15.7	2.2	15.5	3.1
17	13.1	5.3	12.5	6.5	16.3	2.8	15.8	3.8
18	12.5	5.8	11.6	6.8	16.2	4.2	15.0	4.7
19	11.3	6.0	10.3	6.9	14.6	5.6	12.9	6.2

4. Multivariate normal case

4.1. Derivation of results

Suppose the y_t are independent p -variate normal random variables with mean vector μ_t and precision matrix H . Here $\mu_t = \mu_1$ for $1 \leq t \leq k$ and $\mu_t = \mu_2$ for $k+1 \leq t \leq n$. The prior distribution for μ_j is taken to be normal with mean vector m_j and precision matrix $t_j H$, where t_j is a scalar; the prior distribution for H is Wishart with ν degrees of freedom and matrix parameter V , i.e., $p(H) \propto |H|^{(\nu-p-1)/2} \exp \{(-1/2) \text{tr} HV\}$; the prior distribution for the changepoint k is denoted by $p(k)$, $1 \leq k \leq n-1$. The posterior distribution for k , μ_1 , μ_2 and H can be written as $p(k, \mu_1, \mu_2, H|D) = p(\mu_1, \mu_2, H|k, D)p(k|D)$, where $p(k|D)$ is the posterior distribution of the changepoint k and $p(\mu_1, \mu_2, H|k, D)$ is the posterior distribution of μ_1, μ_2 , and H given k . It is well known that the joint distribution factors so that $p(\mu_1, \mu_2, H|k, D) = p(\mu_1|H, k, D)p(\mu_2|H, k, D) \cdot p(H|k, D)$, where $p(\mu_j|H, k, D)$ is a p -variate normal distribution with mean vector m_{jk} and precision matrix $t_{jk}H$, and $p(H|k, D)$ is a Wishart distribution with degrees of freedom $n+\nu$ and matrix parameter V_k . Note that

$$t_{1k} = t_1 + k, \quad t_{2k} = t_2 + n - k,$$

$$m_{1k} = (t_1 m_1 + k \bar{y}_{1k}) / t_{1k}, \quad m_{2k} = (t_2 m_2 + (n-k) \bar{y}_{2k}) / t_{2k},$$

$$\bar{y}_{1k} = \sum_{t=1}^k y_t / k, \quad \bar{y}_{2k} = \sum_{t=k+1}^n y_t / (n-k),$$

$$V_k = V + S_{1k} + S_{2k} + \frac{t_1 k}{t_{1k}} (m_1 - \bar{y}_{1k})(m_1 - \bar{y}_{1k})' + \frac{t_2(n-k)}{t_{2k}} (m_2 - \bar{y}_{2k})(m_2 - \bar{y}_{2k})'$$

$$S_{1k} = \sum_{t=1}^k (y_t - \bar{y}_{1k})(y_t - \bar{y}_{1k})' \quad \text{and} \quad S_{2k} = \sum_{t=k+1}^n (y_t - \bar{y}_{2k})(y_t - \bar{y}_{2k})'$$

Finally, the posterior distribution of the changepoint k is

$$(4.1) \quad p(k|D) \propto p(k)(t_{1k}t_{2k})^{-p/2} |V_k|^{-(n+\nu)/2} \quad k=1, 2, \dots, n-1.$$

We will next examine a loss function as in (2.4) which incorporates the magnitude of the change. A suitable measure of change is the squared Mahalanobis distance $(\mu_1 - \mu_2)'H(\mu_1 - \mu_2)$ and we take the loss $L_2(\theta_1, \theta_2)$ to be equal to the squared Mahalanobis distance. The optimal changepoint estimate \hat{k} can then be found by choosing the value of k corresponding to the largest value of

$$(4.2) \quad R(k) = \left\{ \int \int (\mu_1 - \mu_2)'H(\mu_1 - \mu_2)p(\mu_1, \mu_2, H|k, D)d\mu_1d\mu_2dH \right\} p(k|D) \\ = \{p(t_{1k}^{-1} + t_{2k}^{-1}) + (\nu + n)(m_{1k} - m_{2k})'V_k^{-1}(m_{1k} - m_{2k})\} p(k|D).$$

The second term of the expression in braces can be expected to be largest for the true value of the changepoint. When $t_1=t_2$, the first term in braces is largest when k equals 1 or $n-1$ and it is smallest when k equals $n/2$.

4.2. Illinois traffic data

To illustrate these results we will use some Illinois traffic data adapted from Srivastava and Sen [11] and given in Table 2. Let x_{i1} and x_{i2} , $i=1962, \dots, 1971$, be the number of deaths per 10^8 miles and the number of injuries per 10^7 miles. We will apply the results of

Table 2. Illinois traffic data*

Year	Deaths per 10^8 miles	Injuries per 10^7 miles
1962	4.9	28.2
1963	5.1	30.1
1964	5.2	31.6
1965	5.1	32.9
1966	5.3	31.3
1967	5.1	30.6
1968	4.9	29.2
1969	4.7	29.2
1970	4.2	28.6
1971	4.2	26.1

Table 3. Illinois traffic data application of Subsection 4.1

Year, k	$p(k D)$	$R(k)$
1963	.01	.02
1964	.02	.09
1965	.06	.38
1966	.74	10.42
1967	.04	.21
1968	.02	.06
1969	.06	.36
1970	.05	.41

* From Srivastava and Sen [11]

Subsection 4.2 to first differences $y_{ij} = x_{ij} - x_{i-1,j}$; $i = 1963, \dots, 1971$; $j = 1, 2$. Note that $n = 9$. We will use diffuse prior distributions for the parameters, i.e., we let $\nu = -2$, $V = 0$, and $t_1 = t_2 = 0$ in Subsection 4.2. In Table 3 are given the values of $p(k|D)$ and $R(k)$, computed from (4.1) and (4.2), respectively. The mode of the posterior distribution of the changepoint is at $k = 1966$, indicating that a change in the first differences occurred after 1965. A similar result is obtained when examining the values of $R(k)$. The magnitude of the change could be found by a straightforward extension of the results in Subsection 4.1. Such an analysis suggests that, before the changepoint, rates increased whereas they decreased after the changepoint.

5. Conclusion

In this paper we have shown how loss functions can be incorporated into the changepoint problem. In our discussion we tacitly assumed that a change must have occurred, by setting $p(k=n) = 0$. When proper prior distributions are available for the parameters, this restriction can be relaxed without difficulty. In the case of improper prior distributions some difficulties arise when $p(k=n) \neq 0$ which are related to the fact that the changepoint problem is related to the model choice problem. When $k = n$, we have a model with parameter θ_1 only but when $k < n$, we have a model with an additional parameter θ_2 . The choice between two models with different numbers of parameters can lead to difficulties when improper prior distributions are used, e.g., see Atkinson [1] or Bernardo [2].

UNIVERSITY OF TORONTO

REFERENCES

- [1] Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model, *Biometrika*, **65**, 39-48.
- [2] Bernardo, J. M. (1980). A Bayesian analysis for classical hypothesis testing, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), Ch. 14, University Press, Valencia, Spain.
- [3] Chin Choy, J. H. and Broemeling, L. D. (1980). Some Bayesian inferences for a changing linear model, *Technometrics*, **22**, 71-78.
- [4] Cobb, G. W. (1978). The problem of the Nile: Conditional solution to a changepoint problem, *Biometrika*, **65**, 243-251.
- [5] Hsu, D. A. (1979). Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis, *J. Amer. Statist. Ass.*, **74**, 31-40.
- [6] Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, II, Houghton Mifflin Company, Boston.
- [7] Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*, J. Wiley & Sons, New York.
- [8] Lee, A. F. S. and Heghinian, S. M. (1977). A shift of the mean level in a sequence

- of independent normal random variables—A Bayesian approach, *Technometrics*, **19**, 503-506.
- [9] Menzefricke, U. (1981). A Bayesian analysis of a change in the precision of a sequence of independent normal random variables at an unknown time point, *Appl. Statist.*, **30**, 141-146.
- [10] Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables, *Biometrika*, **62**, 407-416.
- [11] Srivastava, M. S. and Sen, A. (1975). Some one-sided tests for change in level, *Technometrics*, **17**, 61-64.
- [12] Wichern, D. W., Miller, R. B. and Hsu, D. A. (1976). Changes of variance in first-order autoregressive time series models—with an application, *Appl. Statist.*, **25**, 248-256.