

DETECTING CHANGES IN MEAN OF CORRELATED DATA

Tze-San Lee
Western Illinois University
and

Lei Zhang
Hoechst Marion Roussel Pharm., Inc.

Abstract. Most tests for detecting changes in the mean usually assume that the errors in the data are independently distributed. However, in practical applications, the errors in the data are often correlated, e. g., time series data. Yet, little were discussed in the literature on detecting changes in the mean when the errors are, in fact, correlated. Here, proposed is a test to detect changes in the mean of the correlated data. A mean-change model with autocorrelated errors is formulated. An iterative scheme is constructed to estimate the change-point. The Illinois traffic data set is used to illustrate the proposed test. To facilitate the use of the proposed test, a SAS Macro which implements the proposed algorithm is also made available.

Key words: Autoregressive transformation; Markovian errors; Change-point model; Likelihood ratio test.

1. INTRODUCTION

Change-point problem has arisen originally from the quality control of industrial production, where the output of a production process is monitored sequentially and wants to detect any departure of the average output from some specified target value. When the observational errors are independently and identically distributed, sequential test for the detection of change in the mean was first proposed by Page ([9, 10]) and then quasi-Bayesian test by Chernoff and Zacks ([2]), Kander and Zacks ([6]), Gardner ([4]), Sen and Srivastava ([13, 14]). It was followed by likelihood ratio test (Hinkley [5]) and nonparametric test (Bhattacharyya-Johnson [1]; Pettitt [11]; Schechtman [12]; Wolfe-Schechtman [16]). However, in many practical applications, the data are often collected sequentially in time. In such a case, the errors in the data are likely to be serially correlated. Yet little work was done to detect change in the mean when the observational errors are correlated.

It is the aim of this note to propose an iterative algorithm based on a combination of Cochrane-Orcutt's autoregressive transformation and the maximum likelihood principle for an estimation

of the change-point when observational errors are assumed to follow a first-order autoregressive process. In Section 2, by conditioning on the known autocorrelation parameter, an iterative algorithm is proposed for the calculation of the change-point. Also, a test is given for testing the null hypothesis of no change-point versus the alternative hypothesis of having a single change-point. An application of the proposed test to the Illinois traffic data sets is presented in Section 3. Concluding remarks are given in Section 4.

2. A TEST

Assume that a sequence of the observations $x_1, x_2, \dots, x_n$ are collected over time and the observational error follows a first-order autoregressive (or Markovian) process. Suppose that at the unknown time point there is an abrupt change in the mean of the observations. Our goal is to estimate the unknown time at which the change in the mean has occurred.

It follows that the observations $x_1, x_2, \dots, x_n$ satisfy

$$\mu_1 + \epsilon_t, \quad t = 1, \dots, k-1, \quad (2.1a)$$

$$x_t = \begin{cases} \mu_1 + \epsilon_t, & t = 1, \dots, k-1, \\ \mu_2 + \epsilon_t, & t = k, \dots, n, \end{cases} \quad (2.1b)$$

and ϵ_t is Markovian error of the following form:

$$\epsilon_t = \rho \epsilon_{t-1} + \epsilon_t, \quad t = 2, \dots, n, \quad |\rho| < 1, \quad \rho \neq 0, \quad (2.1c)$$

where the ϵ_t 's are independent identically normal distributed with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma^2$, all the parameters μ_1, μ_2 ($\mu_1 \neq \mu_2$), ρ, σ , and k are unknown. In practical applications, k , called the change-point for the mean-change model (2.1a-c), is often unknown and has to be estimated from the data.

To estimate the parameters μ_1 and μ_2 by the principle of maximum likelihood, we need to first apply the autoregressive transformation (Cochrane and Orcutt [3]) given by

$$y_t = x_t - \rho x_{t-1}, \quad (2.2)$$

to (2.1a-b). It follows that (2.1a-b) can be rewritten as:

$$y_t = w_1 + \epsilon_t, \quad t = 2, \dots, k-1, \quad (2.3a)$$

$$y_k = \mu_2 - \rho \mu_1 + \epsilon_k, \quad (2.3b)$$

$$y_t = w_2 + \epsilon_t, \quad t = k+1, \dots, n, \quad (2.3c)$$

where w_1 and w_2 are defined by

$$w_1 = (1 - \rho)\mu_1, \quad (2.4a)$$

$$w_2 = (1 - \rho)\mu_2. \quad (2.4b)$$

Next the estimator for the unknown μ_1 and μ_2 can be obtained by minimizing the negative

logarithm of normal likelihood function given by

$$\text{Minimize } \sum_{t=2, k-1} (y_t - w_1)^2 + (y_k - \mu_2 + \rho \mu_1)^2 + \sum_{t=k+1, n} (y_t - w_2)^2 \quad (2.5)$$

where w_1 and w_2 are given, respectively, by (2.4a-b). By conditioning on that $k = K$ and $\rho = r$ are given, let μ_1 and μ_2 be the solution to the minimization problem of (2.5). Thus, μ_1 and μ_2 satisfy the following equations which are obtained by setting the partial derivatives of the objective function in (2.5) with respect to μ_1 and μ_2 equal to zero:

$$[(K-2)(1-r)^2 + r^2]\mu_1 - r\mu_2 = (1-r)\left(\sum_{t=2, K-1} y_t\right) - ry_K \quad (2.6a)$$

$$-r\mu_1 + [(n-K)(1-r)^2 + 1]\mu_2 = (1-r)\left(\sum_{t=K+1, n} y_t\right) + y_K \quad (2.6b)$$

By using the method of determinant in solving system of linear equations (Strang [15]), a closed form solution for (2.6a-b) is given by

$$\hat{\mu}_1 = \{[(n-K)(1-r)^2 + 1]\left(\sum_{t=2, K-1} y_t\right) - r(1-r)(n-K)y_K + r\sum_{t=K+1, n} y_t\}/D, \quad (2.7a)$$

and

$$\hat{\mu}_2 = \{[(K-2)(1-r)^2 + r^2]\left(\sum_{t=K+1, n} y_t\right) + (K-2)(1-r)y_K + r\sum_{t=2, K-1} y_t\}/D, \quad (2.7b)$$

where D is given by

$$D = (1-r)\{(K-1)(n-K)r^2 - 2(K-2)(n-K)r + (K-2)(n-K-1)\}. \quad (2.7c)$$

Let $S(K) \equiv \text{RSS}(K | \hat{\mu}_1, \hat{\mu}_2, r)$ be the corresponding residual sum of squares of (2.5) and the optimal estimate of change-point, $\hat{k}$, is obtained from

$$S(\hat{k}) = \min_{2 \leq k \leq n-1} \{S(K)\}. \quad (2.8)$$

Once the value of $\hat{k}$ is found, the mean square error associated with (2.1a-c) is calculated as

$$\text{MSE} = S(\hat{k})/(n-4), \quad (2.9)$$

where the degrees of freedom is $n-4$ since two linear parameters, μ_1 and μ_2 , and two nonlinear parameters, k and ρ , have been estimated. Let $S(n)$ be the residual sum of squares of (2.5) having no change-point. Then, a testing statistic F given by

$$F = (S(n) - S(\hat{k}))/\text{MSE}, \quad (2.10)$$

is used to test the null hypothesis of no change-point versus the alternative hypothesis of having a single change-point, where MSE is given by (2.9). Note that the sampling distribution of F of (2.10) is approximately distributed as a Fisher's F -distribution with degrees of freedom 1 and $n-4$, respectively.

An iterative algorithm to implement the above method is described as follows:

Step 1. Set $r = 0$. Let the optimal solutions to (2.5) be denoted by $\hat{\mu}_1$ and $\hat{\mu}_2$ corresponding to $\hat{k}$ which is obtained from (2.8). In this initial case, note that, instead of using (2.7a-b), $\hat{\mu}_1$ and $\hat{\mu}_2$ are given, respectively, by

$$\hat{\mu}_1 = \sum_{t=1, k-1} x_t/(\hat{k}-1), \quad (2.11a)$$

and

$$\hat{\mu}_2 = \sum_{t=k, n} x_t/(n-\hat{k}). \quad (2.11b)$$

Step 2. Compute the residual $\hat{\epsilon}_t$ defined by

$$y_t - (1-r)\hat{\mu}_1, \quad t = 1, \dots, \hat{k}-1, \quad (2.12a)$$

$$\hat{\epsilon}_t = \{y_t - \hat{\mu}_2 + r\hat{\mu}_1, \quad t = \hat{k}, \quad (2.12b)$$

$$y_t - (1-r)\hat{\mu}_2, \quad t = \hat{k}+1, \dots, n, \quad (2.12c)$$

where y_t 's are defined by (2.3a-c) with $\rho = r$.

Step 3. To determine whether the residuals $\hat{\epsilon}_t$ of (2.12a-c) are independent or not plot the residuals against time and judge from the plot. If the independence assumption of (2.12a-c) is satisfied, stop. Otherwise, go to Step 4.

Step 4. Compute the maximum likelihood estimator for ρ defined by

$$r = \sum_{t=2, n} \hat{\epsilon}_t \hat{\epsilon}_{t-1} / \left(\sum_{t=2, n} \hat{\epsilon}_t^2\right). \quad (2.13)$$

Step 5. Set $\rho = r$. Let the optimal solutions to (2.5) in which $\hat{k}$ is obtained from (2.8) be denoted by $\hat{\mu}_1$ and $\hat{\mu}_2$. Go to Step 2.

Remark. In Step 3, if the sample size is too small, it is difficult to judge the validity of independence assumption from the residual plot. In this case, it is probably better to apply the run test to the residuals (Madansky[7]).

A SAS Macro program to implement the above algorithm will be available upon request..

3. APPLICATION

To illustrate the proposed test, the Illinois traffic data sets including deaths, injuries, accidents, and death rates, taken from Sen and Srivastava ([14]), is used as an example. A plot of the data points clearly shows that the mean of all four data sets appears to have changed (Figures are not shown here). By using the algorithm described in Section 2, the computation stops after one iteration for all four data sets since the residual plot has already satisfied the independence assumption. The estimated change-points $\hat{k}$ and the autocorrelation coefficients r for all four data sets are given in Table 1. The 95% confidence interval (C.I.) for $\hat{k}$ is calculated by choosing those K such that the corresponding value of $(S(K) - S(\hat{k}))/\text{MSE}$ is less than the critical value of $F_{1, n-4}$ (Table 1). All the estimated change-points are significant at the level of significance less than 0.01, namely, the null hypothesis of having no change-point is rejected for all four traffic data sets. It is interesting to note that the estimated autocorrelation coefficients for all four data sets are zero. However, should the change-point model not be used, the estimates of autocorrelation coefficient for the data of deaths, injuries, accidents and death rates are 0.6, 0.65, 0.73 and 0.67, respectively, which are clearly spurious.

Incidentally, our results here are different from the results of Sen and Srivastava. This is because what Sen and Srivastava considered is the rate of

change in the data rather than the original data.

Table 1. The estimated change-points and autocorrelation coefficients

Data set	k (95% C.I.)	r	F-value of (2,10)
Deaths (hundreds)	1965 (1963-65)	0.0	17.4
Injuries (thousands)	1964 (1963-65)	0.0	26.1
Accidents (thousands)	1964 (1963-68)	0.0	14.5
Death rate (per 10 ⁴ vehicle miles)	1969 (1967-69)	0.0	25.8

4. CONCLUDING REMARKS

A test is proposed for detecting a change in the mean of the autocorrelated data. It seems to work quite well on the Illinois traffic data sets. Should the change-point model not be applied to fit the Illinois traffic data sets, the spurious estimate for the autocorrelation coefficient would occur as a result. Although the formulation in Section 2 only consider a single change-point, it can, in principle, be used in handling more than one change-point case. All we have to do is to use the proposed test to find the most significant change-point first, and then use the obtained change-point to divide the data set into two separate subsets. Then, apply the proposed test separately to each of the subsets. By repeating this process as long as necessary, we will be able to find all change-points contained in the data. Finally, although only autoregressive process of order one, AR(1), is considered, other time series processes like ARIMA(p,d,q) can also be used to model the correlated errors (Mills [8]), which is, however, beyond the scope of this paper. Nevertheless, AR(1) is the simplest plausible scheme to model the errors of autocorrelated observations, and can usually serve as a good approximation to other more complex ARIMA processes.

References

[1] Bhattacharyya, G.K., and Johnson, R.A. (1968). Nonparametric tests for shift at an unknown time point. *Annals of Mathematical Statistics*, 39, 1731-

1743.

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

[3] Cochran, D., and Orcutt, G.H. (1949). Application of least squares regression to relationship concerning autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32-61.

[4] Gardner, L.A. (1969). On detecting changes in the means of normal variates. *Annals of Mathematical Statistics*, 40, 116-126.

[5] Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57, 1-17.

[6] Kander, Z., and Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distribution occurring at unknown time points. *Annals of Mathematical Statistics*, 37, 1196-1210.

[7] Madansky, A. (1988). *Prescriptions for Working Statisticians*. Springer-Verlag Inc., New York.

[8] Mills, T.C. (1990). *Time Series Techniques for Economists*. Cambridge University Press, Cambridge, England.

[9] Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100-115.

[10] Page, E.S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42, 523-527.

[11] Pettitt, A.N. (1979). A nonparametric approach to the changepoint problem. *Applied Statistics*, 28, 126-135.

[12] Schechtman, E. (1982). A nonparametric test for detecting changes in location. *Communications in Statistics*, A11, 1475-1482.

[13] Sen, A., and Srivastava, M.S. (1975a). On tests for detecting change in mean. *Annals of Statistics*, 3, 98-108.

- [14] Sen, A., and Srivastava, M.S. (1975b). Some one-sided tests for change in levels. *Technometrics*, 17, 61-64.
- [15] Strang, G. (1976). *Linear Algebra and Its Applications*. Academic Press, New York.
- [16] Wolfe, D.A., and Schechtman, E. (1984). Nonparametric statistical procedures for the changepoint problem. *Journal of Statistical Planning and Inference*, 9, 389-396.