BAYESIAN METHODS FOR THE ANALYSIS OF SERIAL DATA

December 1990

BY

NOBUHISA KASHIWAGI

ABSTRACT

This thesis investigates the use of the Bayesian approach in the analysis of serial data. Smoothing of serial Gaussian and non-Gaussian data is discussed. The detection of structural changes of the underlying distributions of serial data is discussed also.

In Chapter 1, a general formulation for smoothing is given, in which the smoothing problem is regarded as the simultaneous estimation problem of parameters depending on strata. This naturally leads us to Bayesian methods, and enables us to use the standard statistical theory in smoothing. Smoothing of serial Gaussian data and some related problems are discussed within the framework of this formulation.

In Chapter 2, a Bayesian method for smoothing serial count data is presented. Recursive formulas for estimating the trend and for evaluating the exact likelihood are developed. The exact likelihood yields the likelihood ratio test for homogeneity of the means. This method is constructed in accordance with the general formulation given in Chapter 1, and is versatile enough to permit various extensions including that for serial binomial data.

In Chapter 3, a Bayesian solution is given to the problem of making inferences about an unknown number of structural changes in serial data. Inferences are based on the posterior distribution of the number of change points and on the posterior probabilities of possible change points. Detailed analyses are given for serial binomial data and some regression problems. An approximation procedure to compute the posterior probabilities is also presented.

CONTENTS

CHAPTER 0

INTRODUCTION AND SUMMARY

The analysis of serial data is one of the most important subjects in statistical analysis. Much attention has been paid on this subject, and various methods have been developed. Smoothing methods are basic ones among those methods. Let $y_1, \cdots, y_T$ be T observations of a random variable Y taken at design points $x_1, \cdots, x_T$ $(x_1 \leq \cdots \leq x_T)$. The purpose of smoothing is to decompose each observation as

$$y_t = f(x_t) + \varepsilon_t$$

where $f(x_t)$ represents a systematic dependence of y_t on x_t and ε_t is a residual. To decompose it, smoothness is assumed for the dependence. Chapters 1 and 2 deal with the problem of estimating $f(x_t)$ from the Bayesian viewpoint.

There are three approaches available for estimating $f(x_t)$, i.e., the distribution-free approach, the likelihood approach and the Bayesian approach. The first approach has yielded several intuitive methods, e.g., moving average, local linear regression and kernel regression. Other examples of the first approach are robust methods such as moving midmean, moving median, and locally weighted regression. The smoothing spline is also a distribution-free method, though it can be derived by using the Wiener process. These distribution-free methods have the merit that they can be used easily. However, since likelihood is not defined in these methods, it is not easy to make detailed statistical inferences. Actually, no

general solution is given to the problems of model selection and tests, while the estimation of smoothing parameters which define the smoothness of $f(x_i)$ is possible by using intuitive criteria such as the risk, the prediction risk, C_p statistic, Cross-Validation and Generalized Cross-Validation.

On the contrary, the likelihood approach enables us to use the standard statistical theory, and consequently it is relatively easy to make detailed statistical inferences. As examples of the second approach, there are polynomial regression and the regression spline. The problems of the parameter estimation, model selection and tests have been studied in detail for these methods. However, in these methods, it is hard to vary the smoothness of the resulting estimate. Actually, polynomial cannot provide flexible curves, and it is not easy to determine the position of knots which define the smoothness of the regression spline.

The flexibility in varying the smoothness and the easiness in making statistical inferences are desirable properties for smoothing methods. These properties can be obtained by using the Bayesian approach. Actually, there are some Bayesian methods which have these properties. These methods are based on the difference constraint and the integrated likelihood. The difference constraint is an effective tool to provide flexible curves, and their smoothness can be varied easily by controlling the smoothing parameter which defines the weight of the constraint. The difference constraint was firstly used for smoothing by Whittaker (1923). Shiller (1973) used it again to estimate smooth lag curves, and introduced the term "smoothness prior". In their methods, however, the selection of the smoothing

parameter was left to the decision of the analyst, though it is a critical problem. A solution to this problem was given by Akaike (1980), who suggested the use of the integrated likelihood. Following Akaike, several Bayesian methods with the difference constraint and the integrated likelihood have been developed (e.g., Kashiwagi, 1982; Kashiwagi and Itani 1986). However, in these methods, the integrated likelihood is used only for estimating the smoothing parameter.

In Chapter 1, we give a formulation for smoothing, which is constructed from the viewpoint where the smoothing problem is regarded as the simultaneous estimation problem of parameters in a many strata model. The strata are assumed to be linearly ordered, and the neighboring strata are assumed to have densities close to each other. This viewpoint naturally leads us to Bayesian modeling, and enables us to embed smoothing methods in the standard statistical theory. Then we can construct estimators and test statistics by applying the likelihood inference. Our formulation is given in a general form, followed by the explicit description of the standard methods including the Stein problem, the one way design and useful smoothing methods. However, in this chapter based on Kashiwagi (1982), Kashiwagi and Itani (1986) and Yanagimoto and Kashiwagi (1990), we focus on the case where distributions are Gaussian.

Non-Gaussian smoothing is an important problem in practice. Especially, recent innovations of surveillance systems in health sciences and other fields have rapidly increased the need for smoothing of discrete data. It is still familiar in smoothing of

discrete data to assume the normality for the data. However, the characteristics of the normal and discrete distributions are quite different. It is desirable to assume discrete distributions for discrete data, since it is more realistic. Until recently, non-Gaussian smoothing was hard to execute because of its computational difficulties, but it has now become possible by the developments of the state-space approach.

In Chapter 2, we present a method for smoothing serial count data. The Poisson distributions are assumed for the data, and the difference constraints with the log normal distributions are assumed for the Poisson means. Under these assumptions, the Poisson means are estimated simultaneously. To estimate them, recursive formulas in the state-space approach are used, by which the computational difficulties in the simultaneous estimation are largely decreased. These formulas, at the same time, enable us to evaluate the exact likelihood, which yields the likelihood ratio test for homogeneity of the means. However, in the usual state-space approach, the initial state $f(x_1)$ is assumed to be a random variable, and a prior distribution of the initial state is given a priori. These assumptions cause an arbitrariness in making statistical inferences. To avoid such an arbitrariness, we assume the initial state is an unknown parameter, and give recursive formulas for estimating the initial state. The method presented here is constructed in accordance with the general formulation given in Chapter 1, and is versatile enough to permit various extensions including that for serial binomial data. This chapter is based on Kashiwagi and Yanagimoto (1990).

Smoothing is effective for observing a smooth trend. On the other hand, we are sometimes interested in finding structural changes of the underlying distributions of serial data. The residual analysis in smoothing can suggest the existence of structural changes to some extent. However, to detect them quantitatively, the method for that purpose is necessary. The problem of making inferences about structural changes is called the change point problem. This problem has been considered by many authors, and various methods including nonparametric, parametric and Bayesian methods have been developed. However, most studies are concerned with the single change case or the detection of multiple changes by using a stepwise procedure, and few studies are available on the problem of detecting multiple changes without using a stepwise procedure. Smith (1980) is one of such few studies, in which he suggested the usefulness of the Bayesian approach for the multiple change case.

In Chapter 3, we give a Bayesian solution to the problem of making inferences about an unknown number of changes. Inferences are based on the posterior distribution of the number of change points and on the posterior probabilities of possible change points. Any stepwise procedure is not used. These posterior probabilities are evaluated by using a combinatorial method, and consequently the large amount of computation is required. We present an approximation procedure to decrease the amount of computation. Detailed analyses are given for serial binomial data and some regression problems. This chapter is based on Kashiwagi (1990).

To illustrate the usefulness of the methods to be presented in each chapter, we provide some examples of application. In Chapters 1

and 2. smoothing methods are applied to epidemiological data sets. In Chapter 3. the Lindisfarne scribes problem and opinion poll data are analyzed.

CHAPTER 1

BAYESIAN METHODS FOR SMOOTHING DATA AND SIMULTANEOUS ESTIMATION OF MANY PARAMETERS

1. 1. INTRODUCTION

Consider a model with T strata having the density $g(y|\underline{\mu}_t, \underline{\theta})$ in the t-th stratum, where the parameter vector $\underline{\mu}_t$ depends on the stratum and $\underline{\theta}$ is common through the stratum. Suppose n_t observations $y_{t1}, \cdots, y_{tn_t}$ are obtained from the t-th stratum. We write $\underline{y}_t=(y_{t1}, \cdots, y_{tn_t})'$, $\underline{y}=(\underline{y}_1', \cdots, \underline{y}_T')'$ and $\underline{\mu}=(\underline{\mu}_1', \cdots, \underline{\mu}_T')'$. Our problems are:

(a) the estimation of $\underline{\mu}$

(b) the estimation of $\underline{\theta}$

(c) the test of the null hypothesis $\underline{\mu}\in M_0$.

We assume $\underline{\mu}$ is an outcome from a hyperpopulation having the density $h(\underline{\mu}|\underline{\delta})$ $\underline{\delta}\in D$, which is called a prior density in the Bayesian context. The parameter space D has a limiting point $\underline{\delta}_0$ such that $h(\underline{\mu}|\underline{\delta})$ tends to a degenerated measure; write it $h(\underline{\mu}|\underline{\delta}_0)$ for convenience. The null hypothesis in the test problem will be expressed as $\underline{\delta}=\underline{\delta}_0$.

Most smoothing methods have been developed separately from the standard statistical theory. However, the smoothing problem can be regarded as the simultaneous estimation of the parameters in a many strata model under the assumptions that the strata are linearly ordered and the neighboring strata have densities close to each other. This viewpoint naturally leads us to Bayesian modeling and

enables us to embed smoothing methods in the standard statistical theory. Then we can construct estimators and test statistics by applying the likelihood inference. In this chapter, we discuss smoothing methods in relation to the standard statistical methods.

In Section 1.2, a procedure to solve the problems (a)~(c) is proposed. In Section 1.3, the Stein problem, the one way design and several Bayesian smoothing methods are discussed within the framework of our formulation. In Section 1.4, our experiences in analyzing epidemiological data sets in terms of the smoothing methods are presented.

1.2. PROPOSED PROCEDURE

To construct the procedure to solve the problems (a)~(c), we introduce the following likelihoods.

Definition 1.1. We define the overall likelihood by

$$L(\underline{\mu}, \underline{\theta}, \underline{\delta}) = \left\{ \prod_{t=1}^{T} \prod_{i=1}^{n_t} g(y_{ti} | \underline{\mu}_t, \underline{\theta}) \right\} \cdot h(\underline{\mu} | \underline{\delta}).$$

Let M be the support of $h(\underline{\mu} | \underline{\delta})$.

$$IL(\underline{\theta}, \underline{\delta}) = \int_M L(\underline{\mu}, \underline{\theta}, \underline{\delta}) \, d\underline{\mu}$$

is called the integrated likelihood. ■

Using the integrated likelihood, the likelihood ratio test statistic is defined.

Definition 1.2. Let $(\hat{\underline{\theta}}, \hat{\underline{\delta}})$ and $(\tilde{\underline{\theta}}, \tilde{\underline{\delta}}_0)$ be the solutions which attain the maximums of $IL(\underline{\theta}, \underline{\delta})$ and $IL(\underline{\theta}, \underline{\delta}_0)$, respectively. We define the

test statistic for the null hypothesis $\underline{\mu} \in M_0$ by

$$S = 2 \cdot \log\{IL(\hat{\underline{\theta}}, \hat{\underline{\delta}}) / IL(\tilde{\underline{\theta}}, \tilde{\underline{\delta}}_0)\}. \blacksquare$$

Our procedure is constructed as:

(1) Estimate $\hat{\underline{\theta}}$ and $\hat{\underline{\delta}}$ by maximizing $IL(\underline{\theta}, \underline{\delta})$

(2) Estimate $\hat{\underline{\mu}}$ by maximizing $L(\underline{\mu}, \hat{\underline{\theta}}, \hat{\underline{\delta}})$

(3) Reject the null hypothesis when $S > c_\alpha$, where c_α is a critical value with the level α.

Remark 1.1. The overall likelihood is proportional to the posterior density of $\underline{\mu}$, $p(\underline{\mu} | \underline{y}, \underline{\theta}, \underline{\delta}) = L(\underline{\mu}, \underline{\theta}, \underline{\delta}) / IL(\underline{\theta}, \underline{\delta})$, and consequently maximization of $L(\underline{\mu}, \hat{\underline{\theta}}, \hat{\underline{\delta}})$ with respect to $\underline{\mu}$ corresponds with that of the empirical posterior density $p(\underline{\mu} | \underline{y}, \hat{\underline{\theta}}, \hat{\underline{\delta}})$. Therefore, our estimate $\hat{\underline{\mu}}$ coincides with the mode of the empirical posterior distribution, and also it coincides with the empirical posterior mean when the posterior distribution is Gaussian.

1.3. APPLICABLE MODELS

Selecting densities $g(y | \underline{\mu}, \underline{\theta})$ and $h(\underline{\mu} | \underline{\delta})$ suitably, we can give a variety of methods.

1.3.1. Stein problem

First, we derive a Stein type estimator with a likelihood ratio test. Let Y_t be a random sample of size 1 from the t-th normal population $N(\mu_t, 1)$. Suppose $\underline{\mu}$ is a random sample of size T from a normal hyperpopulation $N(0, \delta)$. In this problem, $\underline{\theta}$ does not appear,

and the null hypothesis is expressed as $\delta=0$. When $\delta=0$, it follows that $\mu_1=\cdots=\mu_T=0$.

Theorem 1. 1.

$$\hat{\mu}_t = [\|\underline{y}\|^2-T]^+ y_t/\|\underline{y}\|^2$$

$$\hat{\delta} = [\|\underline{y}\|^2-T]^+/T$$

$$S = \begin{cases} 0 & \|\underline{y}\|^2 \leq T \\ \|\underline{y}\|^2-T\cdot\log(\|\underline{y}\|^2/T)-T & otherwise \end{cases}$$

where $[z]^+=max(z, 0)$.

Proof. The overall likelihood is given by

$$L(\underline{\mu}, \delta) = (2\pi)^{-T}\delta^{-\frac{T}{2}}\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left\{(y_t-\mu_t)^2+\frac{1}{\delta}\mu_t^2\right\}\right].$$

The integrated likelihood is obtained as

$$IL(\delta) = (2\pi)^{-\frac{T}{2}}(1+\delta)^{-\frac{T}{2}}\exp\left\{-\frac{\|\underline{y}\|^2}{2(1+\delta)}\right\}.$$

Differentiating $\log IL(\delta)$ with respect to δ, we have

$$\frac{\partial}{\partial\delta}\log IL(\delta) = -\frac{T}{2(1+\delta)}+\frac{\|\underline{y}\|^2}{2(1+\delta)^2} = 0.$$

Then it follows that

$$\hat{\delta} = [\|\underline{y}\|^2-T]^+/T.$$

Maximizing $L(\underline{\mu}, \hat{\delta})$ with respect to $\underline{\mu}$, we obtain

$$\hat{\mu}_t = \frac{\hat{\delta}}{1+\hat{\delta}}y_t = [\|\underline{y}\|^2-T]^+ y_t/\|\underline{y}\|^2.$$

Since $2\cdot\log IL(\hat{\delta})=-T\cdot\log 2\pi-T\cdot\log(\|\underline{y}\|^2/T)-T$ for $\|\underline{y}\|^2>T$ and $2\cdot\log IL(0)=-T\cdot\log 2\pi-\|\underline{y}\|^2$, we have

$$S = \begin{cases} 0 & \|\underline{y}\|^2 \leq T \\ \|\underline{y}\|^2 - T\cdot\log(\|\underline{y}\|^2/T) - T & otherwise. \blacksquare \end{cases}$$

Corollary 1. 1. The rejection region of the test for $\delta=0$ with a standard level α, say .05, is given by using the χ^2-distribution as $\|\underline{y}\|^2 > \chi^2_{T;(1-\alpha)}$.

Proof. Differentiating S with respect to $\|\underline{y}\|^2$, we have $(\partial/\partial\|\underline{y}\|^2)S = 1 - T/\|\underline{y}\|^2 > 0$ for $\|\underline{y}\|^2 > T$. This implys that S is monotone increasing with respect to $\|\underline{y}\|^2$ when $\|\underline{y}\|^2 > T$. Therefore, for an appropriate α, the rejection region is given by $\|\underline{y}\|^2 > \chi^2_{T;(1-\alpha)}$. ■

1. 3. 2. One way design

A random effect model in one way design can be discussed within our framework. Let $\underline{Y}_t$ be a random sample of size n from the t-th normal population $N(\mu_t, \sigma^2)$. Suppose $\underline{\mu}$ is a random sample of size T from a normal hyperpopulation $N(\nu, \tau)$. The null hypothesis is expressed as $\tau=0$. When $\tau=0$, it follows that $\mu_1=\cdots=\mu_T=\nu$.

Theorem 1. 2.

$$\hat{\mu}_t = \bar{y}_t + (\bar{y} - \bar{y}_t)/\{1 + [(n-1)R - 1]^+\}$$

$$\hat{\sigma}^2 = S_w^2 + S_b^2/\{1 + [(n-1)R - 1]^+\}$$

$$\hat{\nu} = \bar{y}$$

$$\hat{\tau} = [R - 1/(n-1)]^+ S_w^2$$

$$S = \begin{cases} 0 & R \leq 1/(n-1) \\ Tn\cdot\log\{(n-1)(R+1)/n\} - T\cdot\log\{(n-1)R\} & otherwise \end{cases}$$

where $\bar{y}_t$ and $\bar{y}$ are the sample means of $\underline{y}_t$ and $\underline{y}$, respectively, $S_w^2 = \frac{1}{Tn}\sum_t\sum_i (y_{ti} - \bar{y}_t)^2$, $S_b^2 = \frac{1}{T}\sum_t (\bar{y}_t - \bar{y})^2$ and $R = S_b^2/S_w^2$.

Proof. The overall likelihood is given by

$$L(\underline{\mu}, \sigma^2, \nu, \lambda) = (2\pi\sigma^2)^{-\frac{T(n+1)}{2}} \lambda^{\frac{T}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left\{\sum_{i=1}^{n}(y_{it}-\mu_t)^2+\lambda(\mu_t-\nu)^2\right\}\right]$$

where $\lambda=\sigma^2/\tau$. The integrated likelihood is obtained as

$$IL(\sigma^2, \nu, \lambda) = (2\pi\sigma^2)^{-\frac{Tn}{2}} \left(\frac{\lambda}{n+\lambda}\right)^{\frac{T}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left\{\|\underline{y}_t\|^2+\lambda\nu^2-\frac{1}{n+\lambda}(n\bar{y}_t+\lambda\nu)^2\right\}\right].$$

Differentiating $\log IL(\sigma^2, \nu, \lambda)$ with respect to ν, we have

$$\frac{\partial}{\partial\nu}\log IL(\sigma^2, \nu, \lambda) = -\frac{Tn\lambda(\nu-\bar{y})}{\sigma^2(n+\lambda)} = 0.$$

Then it follows that $\hat{\nu}=\bar{y}$. Further, differentiating $\log IL(\sigma^2, \hat{\nu}, \lambda)$ with respect to σ and λ, we have

$$\frac{\partial}{\partial\sigma}\log IL(\sigma^2, \hat{\nu}, \lambda) = -\frac{Tn}{\sigma}+\frac{Tn\{nS_w^2+\lambda(S_w^2+S_b^2)\}}{\sigma^3(n+\lambda)} = 0.$$

$$\frac{\partial}{\partial\lambda}\log IL(\sigma^2, \hat{\nu}, \lambda) = \frac{Tn}{2\lambda(n+\lambda)} - \frac{Tn^2S_b^2}{2\sigma^2(n+\lambda)^2} = 0.$$

Then it follows that

$$\hat{\lambda} = n/[(n-1)R-1]^+$$

$$\hat{\sigma}^2 = S_w^2+S_b^2/\{1+[(n-1)R-1]^+\}$$

$$\hat{\tau} = [R-1/(n-1)]^+S_w^2.$$

Maximizing $L(\underline{\mu}, \hat{\sigma}^2, \hat{\nu}, \hat{\lambda})$ with respect to $\underline{\mu}$, we obtain

$$\hat{\mu}_t = \bar{y}_t+(\bar{y}-\bar{y}_t)/\{1+[(n-1)R-1]^+\}.$$

Since $2\cdot\log IL(\hat{\sigma}^2, \hat{\nu}, \hat{\lambda}) = -Tn\cdot\log 2\pi-Tn\cdot\log\{nS_w^2/(n-1)\}-T\cdot\log\{(n-1)R\}-Tn$ for $R>1/(n-1)$ and $2\cdot\log IL(\tilde{\sigma}^2, \tilde{\nu}, \infty) = -Tn\cdot\log 2\pi-Tn\cdot\log(S_w^2+S_b^2)-Tn$, we have

$$S = \begin{cases} 0 & R\leq 1/(n-1) \\ Tn\cdot\log\{(n-1)(R+1)/n\}-T\cdot\log\{(n-1)R\} & otherwise. \blacksquare \end{cases}$$

Corollary 1. 2. The rejection region of the test for $\tau=0$ with a standard level α is given by using the F-distribution as $R>F_{T-1,\,(n-1)T;\,(1-\alpha)}$.

Proof. Differentiating S with respect to R, we have $(\partial/\partial R)S=T\{(n-1)R-1\}/\{R(R+1)\}>0$ for $R>1/(n-1)$. This implys that S is monotone increasing with respect to R when $R>1/(n-1)$. Therefore, for an appropriate α, the rejection region is given by $R>F_{T-1,\,(n-1)T;\,(1-\alpha)}$. ■

The above two simple examples show that the obtained estimators and tests are appealing. The derivation of methods based on other models is easily done in a parallel way, especially when conjugate priors are assumed. However, more useful methods pertain to smoothing data. We will later focus on the smoothing problem.

1. 3. 3. The discrete spline

In smoothing, the strata are linearly ordered in t and the neighboring strata are assumed to have densities close to each other. Let Y_t be a random sample of size 1 from the t-th normal population $N(\mu_t, \sigma^2)$. We describe the relations between the densities of the neighboring strata by

$$(1.5) \qquad \mu_t - 2\mu_{t-1} + \mu_{t-2} \sim i.i.d.\ N(0, \tau) \qquad t=3, \cdots, T.$$

This model represents gradual change of μ_t with respect to t. The following lemma is immediately obtained.

Lemma 1. 1. Model (1. 5) can be written in the matrix form as

$$\underline{\mu}_P \sim i.i.d.\ N(-D_P^{-1}D_I\underline{\mu}_I,\ \tau(D_P'D_P)^{-1})$$

where $\underline{\mu}_I=(\mu_1, \mu_2)'$, $\underline{\mu}_P=(\mu_3, \cdots, \mu_T)'$ and

$$D_I = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \qquad D_P = \begin{bmatrix} 1 & & & & 0 \\ -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \vdots & \vdots & \vdots & \\ 0 & & 1 & -2 & 1 \end{bmatrix}.\blacksquare$$

We write $D=[D_I, D_P]$. In this problem, the prior is assumed only for $\underline{\mu}_P$, and consequently the integrated likelihood is obtained by integrating the overall likelihood with respect to $\underline{\mu}_P$. The null hypothesis is expressed as $\tau=0$. When $\tau=0$, we have $\mu_t-2\mu_{t-1}+\mu_{t-2}=0$, suggesting that all μ_t's are on a straight line.

Theorem 1.3.

$$\hat{\underline{\mu}} = (I_T+\lambda D'D)^{-1}\underline{y}$$

$$\hat{\sigma}^2 = (\underline{y}-A\hat{\underline{\mu}}_I)'V^{-1}(\underline{y}-A\hat{\underline{\mu}}_I)/T$$

$$2\cdot\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda) = -T\cdot\log 2\pi\hat{\sigma}^2-\log|V|-T$$

where $\lambda=\sigma^2/\tau$, I_T denotes the identity matrix of rank T, $V=I_T+\frac{1}{\lambda}BB'$ and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -2 \\ \vdots & \vdots \\ T-2 & 1-T \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 & & 0 \\ 0 & 0 & & \\ 1 & 0 & & \\ 2 & 1 & & \\ \vdots & \vdots & \vdots & \\ T-2 & T-3 & \cdots & 1 \end{bmatrix}.$$

Proof. The overall likelihood is given by

$$L(\underline{\mu}, \sigma^2, \lambda) = (2\pi\sigma^2)^{-(T-1)}\lambda^{\frac{T-2}{2}}\exp\left[-\frac{1}{2\sigma^2}\{(\underline{y}-\underline{\mu})'(\underline{y}-\underline{\mu})+\lambda(D\underline{\mu})'D\underline{\mu}\}\right].$$

Integrating $L(\underline{\mu}, \sigma^2, \lambda)$ with respect to $\underline{\mu}_P$, we have

$$(1.6) \qquad IL(\underline{\mu}_I, \sigma^2, \lambda) = (2\pi\sigma^2)^{-\frac{T}{2}}|V|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\underline{y}-A\underline{\mu}_I)'V^{-1}(\underline{y}-A\underline{\mu}_I)\right\}.$$

Maximizing $IL(\underline{\mu}_I, \sigma^2, \lambda)$ with respect to $\underline{\mu}_I$ and σ^2, we obtain

$$\hat{\underline{\mu}}_I = (A'V^{-1}A)^{-1}A'V^{-1}\underline{y}$$

$$\hat{\sigma}^2 = (\underline{y}-A\hat{\underline{\mu}}_I)'V^{-1}(\underline{y}-A\hat{\underline{\mu}}_I)/T.$$

Maximizing $L(\hat{\underline{\mu}}_I, \underline{\mu}_P, \hat{\sigma}^2, \lambda)$ with respect to $\underline{\mu}_P$, we obtain

$$\hat{\underline{\mu}} = (I_T+\lambda D'D)^{-1}\underline{y}.$$

Substituting $\hat{\underline{\mu}}_I$ and $\hat{\sigma}^2$ into (1.6), we have

$$2\cdot\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda) = -T\cdot\log 2\pi\hat{\sigma}^2 - \log|V| - T.\blacksquare$$

This result suggests that maximization of $IL(\underline{\mu}_I, \sigma^2, \lambda)$ with respect to the parameters reduces to the one-dimensional problem, namely, maximization of $\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda)$ with respect to λ. This maximization problem is hard to solve analytically, and therefore we estimate λ numerically by using, e.g., a line search method. Consequently, the test statistic cannot be written explicitly. We evaluate the critical values numerically by computer simulation. According to our study, the critical values with the level .05 for several T's are almost 0.

The model assumed in this section is called the discrete spline.

1. 3. 4. Seasonal adjustment

Let Y_t be a random sample of size 1 from the t-th normal population $N(\mu_t, \sigma^2)$. The purpose of seasonal adjustment is to decompose μ_t as

$$\mu_t = \nu_t + \xi_t$$

where ν_t and ξ_t denote the trend and seasonal components, respectively. To decompose it, we assume the following model.

$$\nu_t - 2\nu_{t-1} + \nu_{t-2} \sim i.i.d.\ N(0, \frac{\sigma^2}{\lambda}) \qquad t=3, \cdots, T$$

$$(1.7a) \qquad \xi_t - \xi_{t-s} \sim N(0, \frac{\sigma^2}{\tau_1^2}) \qquad t=s+1, \cdots, T$$

$$(1.7b) \qquad \xi_t + \cdots + \xi_{t-s+1} \sim N(0, \frac{\sigma^2}{\tau_2^2}) \qquad t=s, \cdots, T$$

where s denotes the cycle of the seasonal component. More precisely, our model is given in the matrix form as

$$\underset{\sim}{\nu}_P \sim i.i.d.\ N(-D_P^{-1}D_I\underset{\sim}{\nu}_I, \frac{\sigma^2}{\lambda}(D_P'D_P)^{-1})$$

$$(1.8) \qquad \underset{\sim}{\xi}_P \sim i.i.d.\ N(-(E_P'E_P)^{-1}E_P'E_I\underset{\sim}{\xi}_I, \sigma^2(E_P'E_P)^{-1})$$

where $\underset{\sim}{\nu}_I=(\nu_1, \nu_2)'$, $\underset{\sim}{\nu}_P=(\nu_3, \cdots, \nu_T)'$, $\underset{\sim}{\xi}_I=(\xi_1, \cdots, \xi_{s-1})'$, $\underset{\sim}{\xi}_P=(\xi_s, \cdots, \xi_T)'$ and

$$E_I \underset{(2T-2s+1)\times(s-1)}{=} \begin{bmatrix} -\tau_1 & 0 & \cdots & 0 \\ & -\tau_1 & \cdots & 0 \\ & & \vdots & \\ & & & -\tau_1 \\ & & & \\ & 0 & & \\ & & & \\ \tau_2 & \tau_2 & \cdots & \tau_2 \\ & \tau_2 & \cdots & \tau_2 \\ & & \vdots & \\ & & & \tau_2 \\ & & & \\ & 0 & & \end{bmatrix} \qquad E_P \underset{(2T-2s+1)\times(T-s+1)}{=} \begin{bmatrix} 0 & \tau_1 & & & & & \\ 0 & 0 & \tau_1 & & & 0 & \\ & & \vdots & & & & \\ 0 & 0 & 0 & \cdots & 0 & \tau_1 & \\ -\tau_1 & 0 & 0 & \cdots & 0 & 0 & \tau_1 \\ & & \vdots & & & & \\ & -\tau_1 & 0 & \cdots & 0 & 0 & 0 & \tau_1 \\ \tau_2 & & & & & & & \\ \tau_2 & \tau_2 & & & & 0 & & \\ & & \vdots & & & & & \\ \tau_2 & \tau_2 & \tau_2 & \cdots & \tau_2 & & & \\ \tau_2 & \tau_2 & \tau_2 & \cdots & \tau_2 & \tau_2 & & \\ & & \vdots & & & & & \\ & & \tau_2 & \cdots & \tau_2 & \tau_2 & \tau_2 & \tau_2 \end{bmatrix}.$$

We write $\underline{\eta}_I=(\underline{\nu}_I', \underline{\xi}_I')'$, $\underline{\eta}_P=(\underline{\nu}_P', \underline{\xi}_P')'$ and $\underline{\eta}=(\underline{\eta}_I', \underline{\eta}_P')'$. The null hypotheses can be constructed by setting $\lambda=\infty$ and/or $\tau_1=\tau_2=\infty$. When $\lambda=\infty$, we have $\nu_t-2\nu_{t-1}+\nu_{t-2}=0$. When $\tau_1=\tau_2=\infty$, we have $\xi_t-\xi_{t-s}=0$ and $\xi_t+\cdots+\xi_{t-s+1}=0$.

Remark 1. 2. Models (1. 7) and (1. 8) do not correspond with each other exactly, but their practical effects are almost equal.

Theorem 1. 4.

$$\hat{\underline{\eta}} = (F'F)^{-1}F'\underline{z}$$

$$\hat{\sigma}^2 = (\underline{z}-F\hat{\underline{\eta}})'(\underline{z}-F\hat{\underline{\eta}})/T$$

$$2\cdot\log IL(\hat{\underline{\eta}}_I, \hat{\sigma}^2, \lambda, \tau_1, \tau_2) = -T\cdot\log 2\pi\hat{\sigma}^2+(T-2)\cdot\log\lambda + \log|E_P'E_P|-\log|F_P'F_P|-T$$

where $\underline{z}=(\underline{y}', \underline{0}_{3T-2s-1}')'$ with $\underline{0}_i$ being the zero vector of size i, $F=[F_I, F_P]$ and

$$F_I = \begin{bmatrix} I_2 & I_{s-1} \\ O_{(T-2)\times 2} & O_{(T-s+1)\times(s-1)} \\ \lambda^{\frac{1}{2}} D_I & O_{(T-2)\times(s-1)} \\ O_{(2T-2s+1)\times 2} & E_P(E_P'E_P)^{-1}E_P'E_I \end{bmatrix} \quad F_P = \begin{bmatrix} O_{2\times(T-2)} & O_{(s-1)\times(T-s+1)} \\ I_{T-2} & I_{T-s+1} \\ \lambda^{\frac{1}{2}} D_P & O_{(T-2)\times(T-s+1)} \\ O_{(2T-2s+1)\times(T-2)} & E_P \end{bmatrix}$$

with $O_{i\times j}$ being the zero matrix of size $i\times j$.

Proof. The overall likelihood is given by

$$L(\underline{\eta}, \sigma^2, \lambda, \tau_1, \tau_2) = (2\pi\sigma^2)^{-\frac{3T-s-1}{2}} \lambda^{\frac{T-2}{2}} |E_P'E_P|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\underline{z}-F\underline{\eta})'(\underline{z}-F\underline{\eta})\right\}.$$

Maximizing $L(\underline{\eta}, \sigma^2, \lambda, \tau_1, \tau_2)$ with respect to $\underline{\eta}_P$, we obtain

$$\hat{\underline{\eta}}_P = (F_P'F_P)^{-1}F_P'(\underline{z}-F_I\underline{\eta}_I).$$

Then the integrated likelihood is obtained as

$$IL(\underline{\eta}_I, \sigma^2, \lambda, \tau_1, \tau_2) = (2\pi\sigma^2)^{-\frac{T}{2}} \lambda^{\frac{T-2}{2}} |E_P'E_P|^{\frac{1}{2}} |F_P'F_P|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\underline{z}-F_I\underline{\eta}_I-F_P\hat{\underline{\eta}}_P)'(\underline{z}-F_I\underline{\eta}_I-F_P\hat{\underline{\eta}}_P)\right\}.$$

Therefor, we have

$$\hat{\underline{\eta}} = (F'F)^{-1}F'\underline{z}$$

$$\hat{\sigma}^2 = (\underline{z}-F\hat{\underline{\eta}})'(\underline{z}-F\hat{\underline{\eta}})/T.$$

Substituting $\hat{\underline{\eta}}_I$ and $\hat{\sigma}^2$ into $IL(\underline{\eta}_I, \sigma^2, \lambda, \tau_1, \tau_2)$, we have

$$2\cdot\log IL(\hat{\underline{\eta}}_I, \hat{\sigma}^2, \lambda, \tau_1, \tau_2) = -T\cdot\log 2\pi\hat{\sigma}^2 + (T-2)\cdot\log\lambda + \log|E_P'E_P| - \log|F_P'F_P| - T. \blacksquare$$

We estimate the remaining parameters λ, τ_1 and τ_2 numerically by using a grid search method. Evaluation of the critical values may be

possible by computer simulation.

1. 3. 5. Smoothing of quantile data

Let $\underline{Y}_t$ be a random sample of size n_t from the t-th population with a continuous density $p_t(y)$. We consider smoothing of quantile data $y_{t(i_{tj})}$ $(t=1, \cdots, T,\ j=1, \cdots, m)$ with respect to t, where $i_{tj}=[n_t\alpha_j]+1$ and $0<\alpha_1<\cdots<\alpha_m<1$. The following asymptotic result is helpful to construct a model for smoothing quantile data.

Theorem 1. 5 (*Mosteller*, 1946). If $p_t(y)$ is differentiable in the neighborhoods of the population quantiles $\mu_{t\alpha_j}$ and $p_t(\mu_{t\alpha_j})\neq 0$ $(j=1, \cdots, m)$, then the joint distribution of the sample quantiles $Y_{t(i_{t1})}, \cdots, Y_{t(i_{tm})}$ tends to a m-dimensional normal distribution with means $\mu_{t\alpha_1}, \cdots, \mu_{t\alpha_m}$ and covariances

$$\mathrm{cov}(Y_{t(i_{tj})}, Y_{t(i_{tk})}) = \frac{\alpha_j(1-\alpha_k)}{n_t p_t(\mu_{t\alpha_j}) p_t(\mu_{t\alpha_k})} \qquad j\leq k. \blacksquare$$

Let $\underline{Y}_{\alpha t}=(Y_{t(i_{t1})}, \cdots, Y_{t(i_{tm})})'$, $\underline{\mu}_t=(\mu_{t\alpha_1}, \cdots, \mu_{t\alpha_m})'$, $\underline{\mu}_I=(\underline{\mu}_1', \underline{\mu}_2')'$ and $\underline{\mu}_P=(\underline{\mu}_3', \cdots, \underline{\mu}_T')'$. Our model is assumed as:

$$\underline{Y}_{\alpha t} \sim i.\,i.\,d.\ N(\underline{\mu}_t, \sigma^2 C_t) \qquad t=1, \cdots, T$$

$$\underline{\mu}_P \sim i.\,i.\,d.\ N(-D_P^{-1}D_I\underline{\mu}_I, \frac{\sigma^2}{\lambda}(D_P'D_P)^{-1})$$

where C_t is the $m\times m$ matrix and

$$D_I = \begin{bmatrix} I_m & -2I_m & & \\ & I_m & & \\ & & & \\ 0 & & & \end{bmatrix} \qquad D_P = \begin{bmatrix} I_m & & & 0 \\ -2I_m & I_m & & \\ I_m & -2I_m & I_m & \\ & & \vdots & \\ 0 & & I_m & -2I_m \quad I_m \end{bmatrix}.$$

The elements of C_t are given by

$$c_{tjk} = c_{tkj} = \frac{\alpha_j(1-\alpha_k)}{n_t \rho_{\alpha_j}\rho_{\alpha_k}} \qquad 1 \le j \le k \le m$$

where $\rho_{\alpha_j}=\varphi(\Phi^{-1}(\alpha_j))$ and $\Phi(y)$ is an assumed distribution with a density $\varphi(y)$. We usually assume several alternatives for $\Phi(y)$, then select the best fit one among them by using the integrated likelihood. Write $\underline{y}_\alpha=(\underline{y}_{\alpha 1}', \cdots, \underline{y}_{\alpha T}')'$, $\underline{\mu}=(\underline{\mu}_I', \underline{\mu}_P')'$, $D=[D_I, D_P]$ and

$$C = \begin{bmatrix} C_1 & & 0 \\ & C_2 & \\ & & \ddots \\ 0 & & C_T \end{bmatrix}.$$

Theorem 1. 6.

$$\hat{\underline{\mu}} = (C^{-1}+\lambda D'D)^{-1}C^{-1}\underline{y}_\alpha$$

$$\hat{\sigma}^2 = \{(\underline{y}_\alpha-\hat{\underline{\mu}})'C^{-1}(\underline{y}_\alpha-\hat{\underline{\mu}})+\lambda(D\hat{\underline{\mu}})'D\hat{\underline{\mu}}\}/Tm$$

$$2\cdot\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda) = -Tm\cdot\log 2\pi\hat{\sigma}^2+(T-2)m\cdot\log\lambda$$
$$-\log|C|-\log|E_P'C^{-1}E_P+\lambda D_P'D_P|-Tm$$

where

$$E_P = \begin{bmatrix} 0_{2m\times(T-2)m} \\ I_{(T-2)m} \end{bmatrix}.$$

Proof. The overall likelihood is given by

$$L(\underline{\mu}, \sigma^2, \lambda) = (2\pi\sigma^2)^{-(T-1)m}\lambda^{\frac{(T-2)m}{2}}|C|^{-\frac{1}{2}}\times$$
$$\exp\left[-\frac{1}{2\sigma^2}\{(\underline{y}_\alpha-\underline{\mu})C^{-1}(\underline{y}_\alpha-\underline{\mu})+(D\underline{\mu})'D\underline{\mu}\}\right].$$

Maximizing $L(\underline{\mu}, \sigma^2, \lambda)$ with respect to $\underline{\mu}_P$, we obtain

$$\hat{\underline{\mu}}_P = (E_P'C^{-1}E_P+\lambda D_P'D_P)^{-1}(E_P'C^{-1}\underline{y}_\alpha-\lambda D_P'D_I\underline{\mu}_I).$$

Then the integrated likelihood is obtained as

$$IL(\underline{\mu}_I, \sigma^2, \lambda) = (2\pi\sigma^2)^{-\frac{Tm}{2}}\lambda^{\frac{(T-2)m}{2}}|C|^{-\frac{1}{2}}|E_P'C^{-1}E_P+\lambda D_P'D_P|^{-\frac{1}{2}}\times$$

$$\exp\left[-\frac{1}{2\sigma^2}\{(\underline{y}_\alpha-E_I\underline{\mu}_I-E_P\hat{\underline{\mu}}_P)'C^{-1}(\underline{y}_\alpha-E_I\underline{\mu}_I-E_P\hat{\underline{\mu}}_P)+(D_I\underline{\mu}_I+D_P\hat{\underline{\mu}}_P)'(D_I\underline{\mu}_I+D_P\hat{\underline{\mu}}_P)\}\right]$$

where

$$E_I = \begin{bmatrix} I_{2m} \\ O_{(T-2)m\times 2m} \end{bmatrix}.$$

Therefore, we have

$$\hat{\underline{\mu}} = (C^{-1}+\lambda D'D)^{-1}C^{-1}\underline{y}_\alpha$$

$$\hat{\sigma}^2 = \{(\underline{y}_\alpha-\hat{\underline{\mu}})'C^{-1}(\underline{y}_\alpha-\hat{\underline{\mu}})+\lambda(D\hat{\underline{\mu}})'D\hat{\underline{\mu}}\}/Tm.$$

Substituting $\hat{\underline{\mu}}_I$ and $\hat{\sigma}^2$ into $IL(\underline{\mu}_I, \sigma^2, \lambda)$, we have

$$2\cdot\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda) = -Tm\cdot\log 2\pi\hat{\sigma}^2+(T-2)m\cdot\log\lambda$$

$$-\log|C|-\log|E_P'C^{-1}E_P+\lambda D_P'D_P|-Tm.\blacksquare$$

The log integrated likelihood $\log IL(\hat{\underline{\mu}}_I, \hat{\sigma}^2, \lambda)$ is used for selecting λ and $\Phi(y)$, and for testing the null hypothesis.

1. 3. 6. Smoothing of spatial data

Finally, we discuss smoothing of spatial data. In spatial smoothing, the strata are arranged in a rectangular lattice shape. Let Y_{ij} be a random sample of size 1 from the (i, j)-th normal population $N(\mu_{ij}, \sigma^2)$ on a two-dimensional rectangular lattice, where

$i=1, \cdots, T_r$, $j=1, \cdots, T_c$ and $T=T_r \cdot T_c$. Write $\underline{y}=(y_{11}, \cdots, y_{T_rT_c})'$ and $\underline{\mu}=(\mu_{11}, \cdots, \mu_{T_rT_c})'$. We describe the relations between the densities of the neighboring strata by

$$(1.9) \qquad 4\mu_{ij} - \mu_{i+1,j} - \mu_{i-1,j} - \mu_{i,j+1} - \mu_{i,j-1} \sim i.i.d.\ N(0, \frac{\sigma^2}{\lambda})$$

with $i=1, \cdots, T_r$, $j=1, \cdots, T_c$ and

$$(1.10) \qquad \mu_{0j} = \mu_{1j} \qquad \mu_{T_r+1,j} = \mu_{T_rj} \qquad \mu_{i0} = \mu_{i1} \qquad \mu_{i,T_c+1} = \mu_{i,T_c}.$$

Model (1.9) is the two-dimensional version of Model (1.5). Condition (1.10) is derived from the assumption that the normal difference of the first order is equal to zero on the boundary. Our model can be written in the matrix form as

$$D\underline{\mu} \sim i.i.d.\ N(\underline{0}_T, \frac{\sigma^2}{\lambda} I_T) \quad or \quad \underline{\mu} \sim i.i.d.\ N(\underline{0}_T, \frac{\sigma^2}{\lambda}(D'D)^{-1})$$

where D is the $T \times T$ matrix constructed so as to satisfy (1.9) and (1.10). The null hypothesis is expressed as $\lambda=\infty$.

Theorem 1.7.

$$\hat{\underline{\mu}} = (I_T+\lambda D'D)^{-1}\underline{y}$$

$$\hat{\sigma}^2 = \{(\underline{y}-\hat{\underline{\mu}})'(\underline{y}-\hat{\underline{\mu}})+\lambda(D\hat{\underline{\mu}})'D\hat{\underline{\mu}}\}/T$$

$$2 \cdot \log IL(\hat{\sigma}^2, \lambda) = -T \cdot \log 2\pi\hat{\sigma}^2 + T \cdot \log\lambda + \log|D'D| - \log|I_T+\lambda D'D| - T.$$

Proof. The overall likelihood is given by

$$L(\underline{\mu}, \sigma^2, \lambda) = (2\pi\sigma^2)^{-\frac{T}{2}}\lambda^{\frac{T}{2}}|D'D|^{\frac{1}{2}}\exp\left[-\frac{1}{2\sigma^2}\{(\underline{y}-\underline{\mu})'(\underline{y}-\underline{\mu})+\lambda(D\underline{\mu})'D\underline{\mu}\}\right].$$

Maximizing $L(\underline{\mu}, \sigma^2, \lambda)$ with respect to $\underline{\mu}$, we obtain

$$\hat{\underline{\mu}} = (I_T+\lambda D'D)^{-1}\underline{y}.$$

Then the integrated likelihood is obtained as

$$IL(\sigma^2, \lambda) = (2\pi\sigma^2)^{-\frac{T}{2}}\lambda^{\frac{T}{2}}|D'D|^{\frac{1}{2}}|I_T+\lambda D'D|^{-\frac{1}{2}}\times \exp\left[-\frac{1}{2\sigma^2}\{(\underline{y}-\hat{\underline{\mu}})'(\underline{y}-\hat{\underline{\mu}})+\lambda(D\hat{\underline{\mu}})'D\hat{\underline{\mu}}\}\right].$$

Therefore, we have

$$\hat{\sigma}^2 = \{(\underline{y}-\hat{\underline{\mu}})'(\underline{y}-\hat{\underline{\mu}})+\lambda(D\hat{\underline{\mu}})'D\hat{\underline{\mu}}\}/T.$$

Substituting $\hat{\sigma}^2$ into $IL(\sigma^2, \lambda)$, we have

$$2\cdot\log IL(\hat{\sigma}^2, \lambda) = -T\cdot\log 2\pi\hat{\sigma}^2+T\cdot\log\lambda+\log|D'D|-\log|I_T+\lambda D'D|-T.\blacksquare$$

1. 4. APPLICATIONS

Two examples of applying the smoothing method to actual data follow.

1. 4. 1. Cancer mortality in Japan

We analyzed the yearly data cited from Japanese vital statistics for the crude number of cancer death in males between 1965 and 1986. Figure 1. 1 shows the results in the case of stomach cancer in males by the discrete spline. We observe that even in the crude number base, the annual mortality has been decreasing in recent years, though it is widely accepted that the adjusted mortality is decreasing. The goodness-of-fit of the simple linear regression, which is the null hypothesis in the discrete spline, is apparently bad. This is supported by the fact that the integrated likelihood ratio test statistic takes 7. 18. To compare with an existing method, we analyzed the same data by using *smooth* in the familiar statistical software, *S*. The results are given in Fig. 1. 2. The general trends

are similar, but the estimated line in Fig. 1. 2 seems to be overfitted. A clearer difference between the two analyses is in the fact that ours involves the null hypothesis test.

We also analyzed cancer mortality data of other sites. The annual mortality of lung and pancreatic cancers in males appears to be increasing exponentially rather than linearly. Therefore, we assumed $y_t \sim LN(\mu_t, \sigma^2)$, i. e., $\log y_t \sim N(\mu_t, \sigma^2)$. Our analysis shows that the estimated lines are close to the estimated exponential regression curves. The estimated trend in lung cancer is exponential at the earlier stage of the period in study, and is going down from the exponential curve. On the other hand, pancreatic cancer shows better agreement with the exponential curve. However, the tests for the null hypothesis are still highly significant. The case of lung cancer is given in Fig. 1. 3.

1. 4. 2. SMON patient incidence

According to leading Japanese epidemiologists, Subacute Myelo-Optico Neuropathy (SMON) is a tragic large-scale side effect of the drug, clioquinol. At the time when the etiology of SMON was in study, it was suspected that a relatively high incidence of SMON occurred in the summer. To illustrate the usefulness of the seasonal adjustment method, we analyzed the data for the monthly incidence of SMON cited from Table 7. 1 in the Research Report by the SMON Research Commision between November 1966 and August 1970. The estimated line with the estimated trend and seasonal effects is given in Fig. 1. 4. The discrete spline is also applied and is given in Fig. 1. 5. Both estimated lines appear to be acceptable. More precisely, very

short-term fluctuations are observed in the seasonal adjustment method. On the other hand, the upper and lower peaks cannot be interpreted well by the discrete spline. In this case, the integrated likelihood ratio test statistic takes 50.32. Since the difference of the numbers of parameters in the models is 13, the test for the existence of seasonal effect is obviously highly significant, though we do not have explicit results on the critical value. The estimated seasonal effects show the gradual increase from winter to summer and the highest peak seen in September, followed by a sharp decrease.

Figure 1. 1. Smoothing the data for the annual mortality of stomach cancer in males by the discrete spline (solid line) and by the simple linear regression (dotted line).

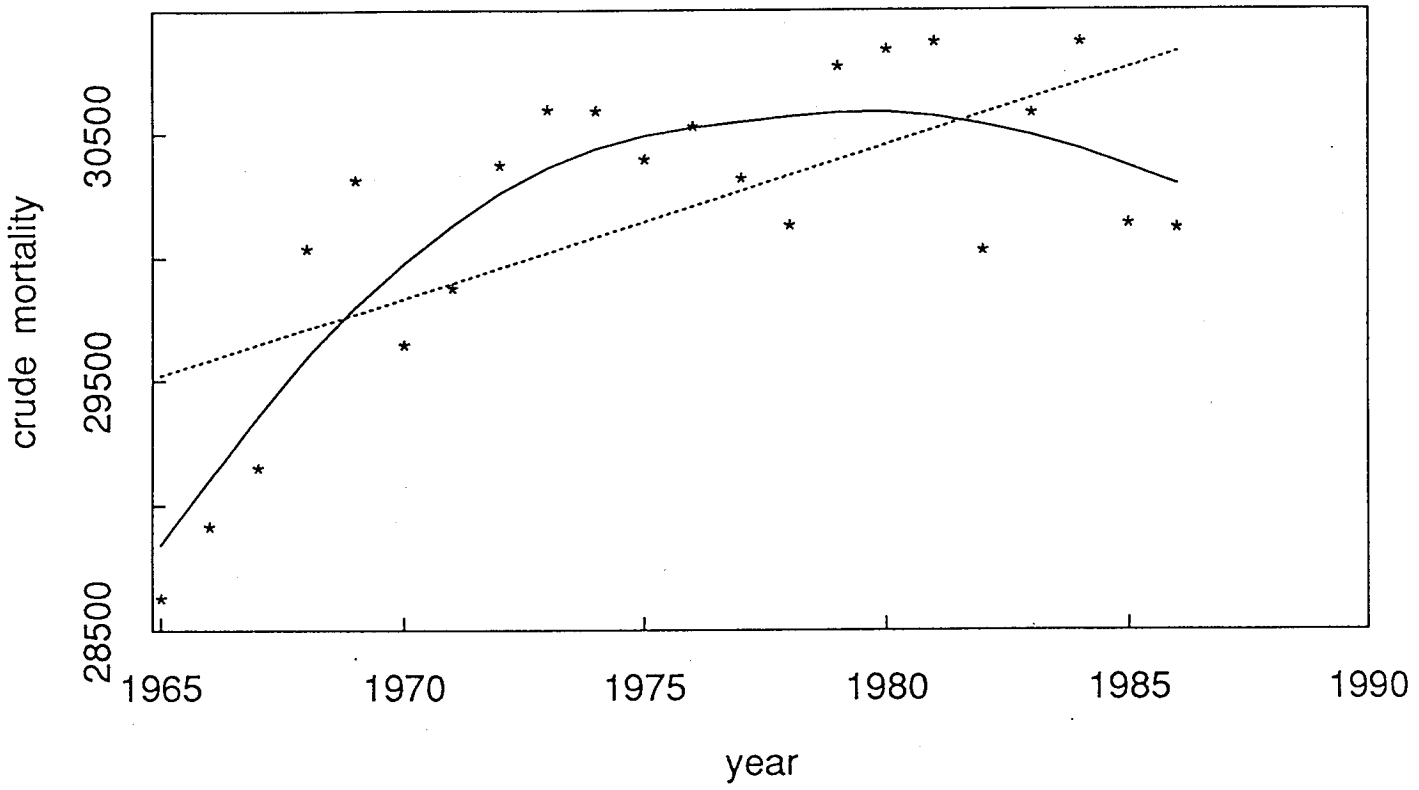


Figure 1.2. Smoothing the data of Figure 1.1 by *smooth* in the software *S*.

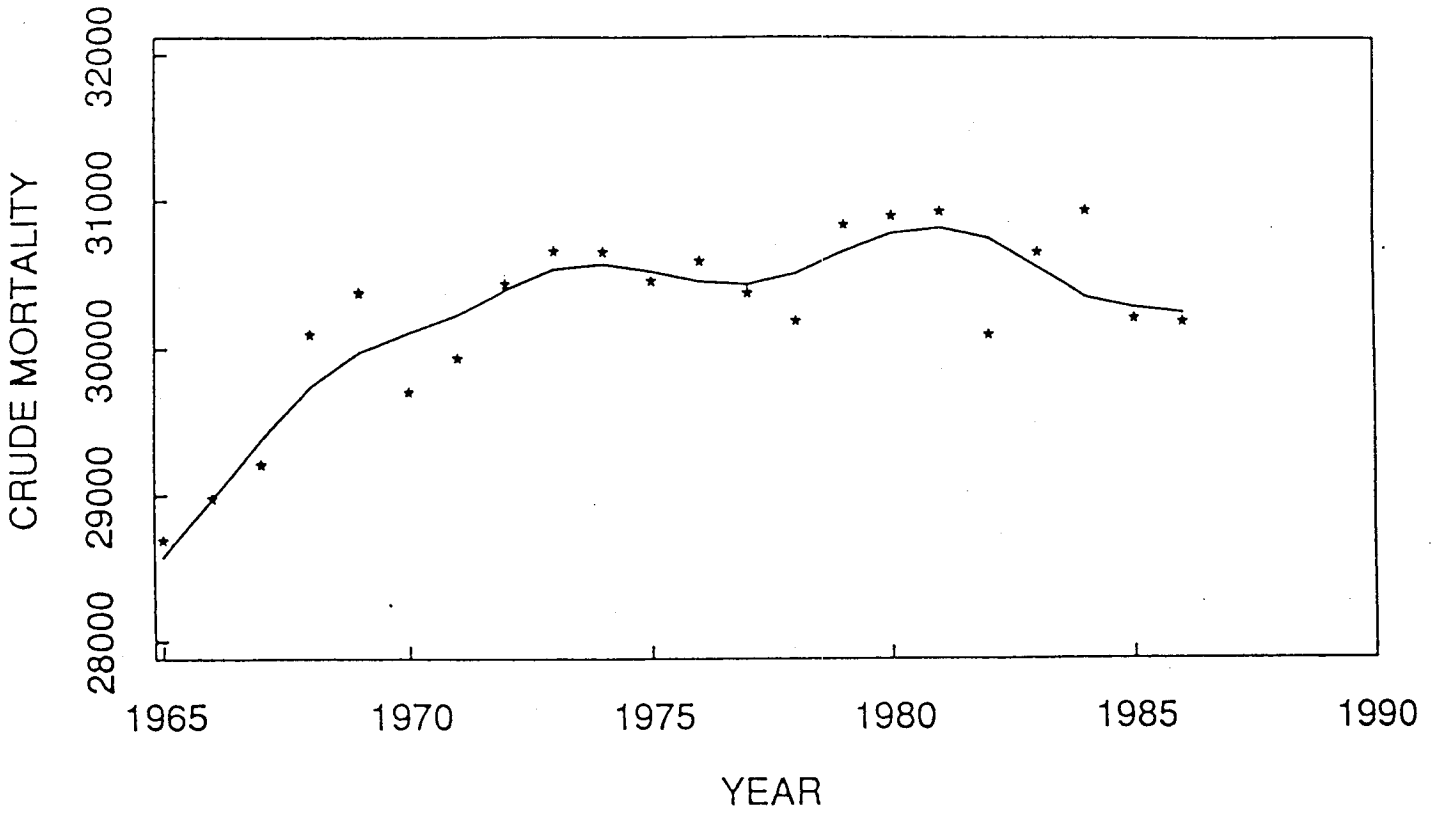


Figure 1.3. Smoothing the data for the annual mortality of lung cancer in males by the discrete spline (solid line) and by the simple linear regression (dotted line), both after logarithmic transformation.

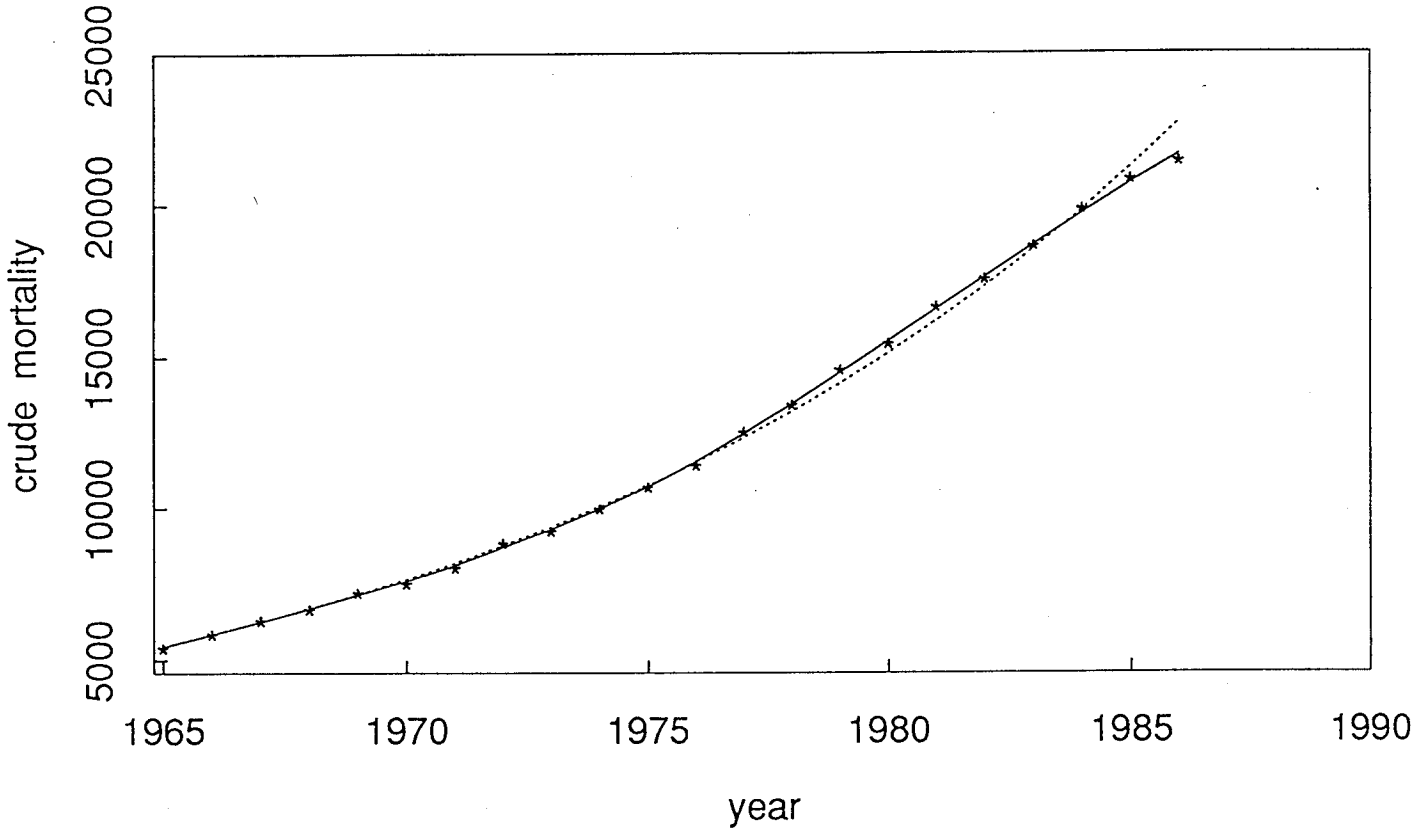


Figure 1.4. Fitting the seasonal adjustment model to the data for the monthly incidence of SMON (A) with the estimated trend (B) and the estimated seasonal effects (C).

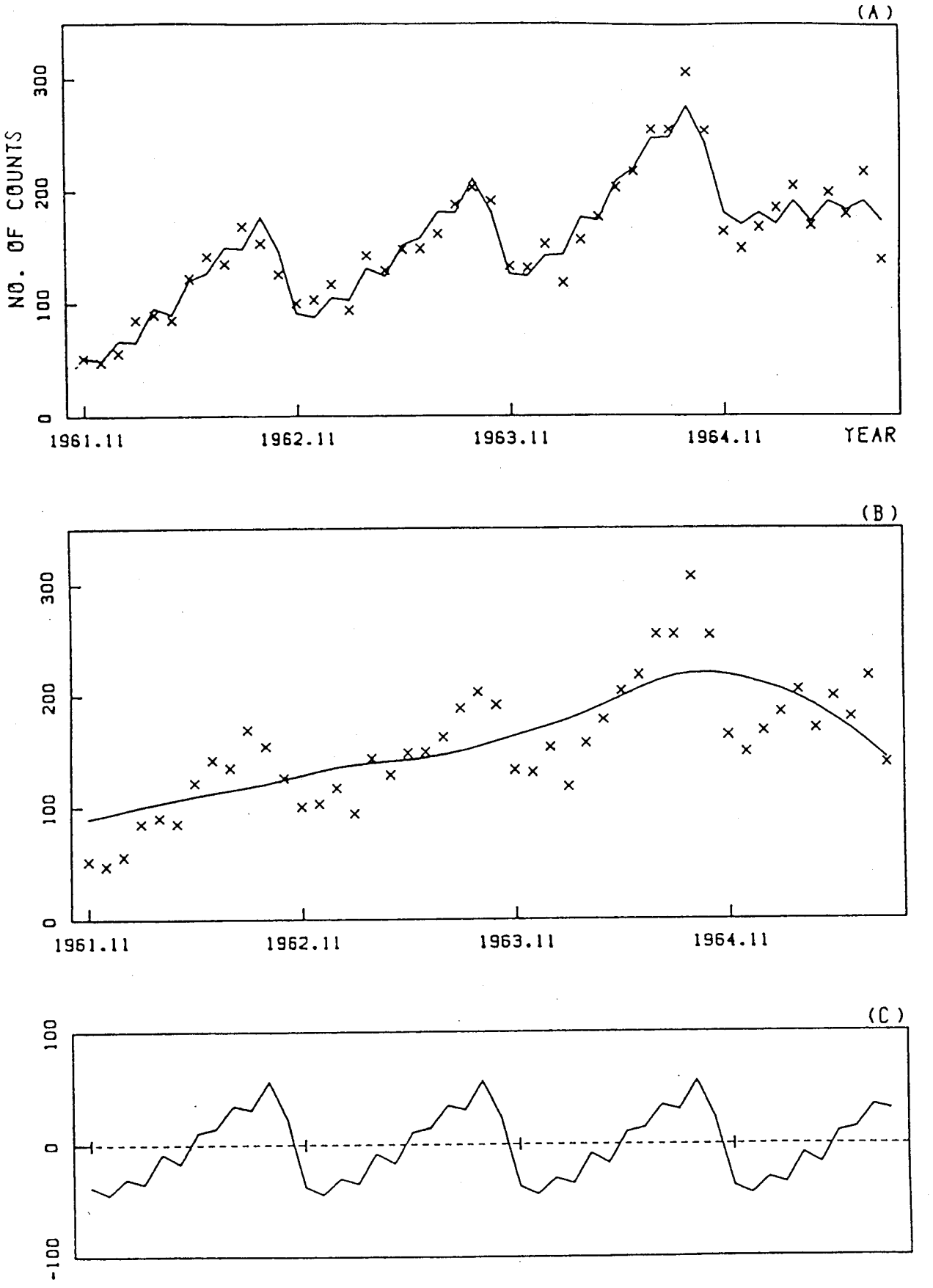
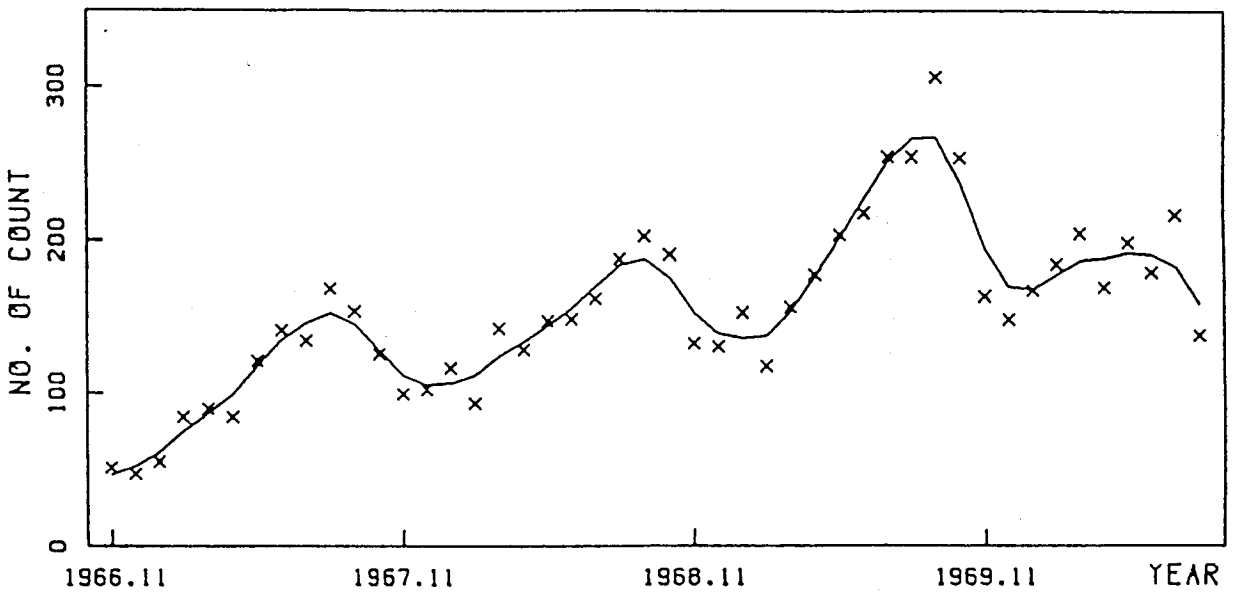


Figure 1.5. Fitting the discrete spline to the data of Figure 1.4.



CHAPTER 2

SMOOTHING SERIAL COUNT DATA THROUGH A STATE-SPACE MODEL

2. 1. INTRODUCTION

In this chapter, we discuss smoothing of serial Poisson count data with the non-Gaussian state-space approach and propose a likelihood ratio test for homogeneity of the Poisson means.

The state-space approach has been discussed with the emphasis on applications to engineering and econometrics. Consequently, primary attention has been paid to filtering and prediction rather than smoothing. In addition, the estimation of the initial state has received little attention, presumably because sample sizes are fairly large in those applications. Actually, the initial state has been assumed to be an outcome from a known prior distribution. However, this assumption causes an arbitrariness in making statistical inferences. To avoid such an arbitrariness, we assume that the initial state is an unknown parameter and give recursive formulas for obtaining the maximum likelihood estimate of the initial state. Then we can construct a likelihood ratio test which does not depend on the unidentifiable assumption.

Section 2. 2 reviews a state-space model for smoothing serial count data. In Section 2. 3, an estimation procedure and a test for homogeneity of the means are proposed. Computational details for implementing the proposed procedure are given in Section 2. 4. Section 2. 5 presents examples of application. Section 2. 6 is devoted to further extension of our approach. A brief comparison with the

method under the normality assumption is given in Section 2. 7.

2. 2. ASSUMED MODEL

Suppose serial data $y_1, \cdots, y_T$ are outcomes from the Poisson distributions $Po(\mu_1), \cdots, Po(\mu_T)$, respectively. We assume that μ_t changes gradually. To model the gradual change, we impose the difference constraint on the canonical link $\log\mu_t$:

$$\Delta^d \log\mu_t \sim i.\ i.\ d.\ N(0, \sigma^2)$$

where Δ denotes the difference operator, i. e., $\Delta\log\mu_t = \log\mu_t - \log\mu_{t-1}$ and d is the difference order. The smoothness of the resulting estimate is defined by d and σ^2. These parameters can be estimated by using the likelihood. However, in consideration of the testing problem, we focus on the case $d=1$. When $d=1$ and $\sigma^2=0$, we have

$$\mu_1 = \mu_2 = \cdots = \mu_T.$$

Using this fact, we construct the test for homogeneity of the means, i. e., the test for $H_0: \sigma^2=0$ against $H_1: \sigma^2>0$. The alternative hypothesis means that μ_t is not stable but changes gradually. The case $d\geq 2$ will be discussed in §2. 6.

The difference constraint is regarded as a prior distribution for μ_t in the Bayesian context. As a prior distribution, a conjugate prior may be more familiar. However, in smoothing, there is no merit in assuming a conjugate prior. Actually, the calculation of the likelihood cannot be simplified, even if a gamma distribution is assumed for μ_t.

The state-space form of our model is given as follows.

Lemma 2. 1. The model for d=1 is written in the state-space form as

$$(2.1a)\quad g(y_t|\mu_t) = e^{-\mu_t}\mu_t^{y_t}/y_t! \qquad t=1, \cdots, T$$

$$(2.1b)\quad h(\mu_t|\mu_{t-1}, \sigma^2) = (2\pi)^{-\frac{1}{2}}(\sigma\mu_t)^{-1}\exp\left\{-\frac{1}{2\sigma^2}(\log\mu_t-\log\mu_{t-1})^2\right\} \qquad t=2, \cdots, T. \blacksquare$$

Recall that μ_1 is an unknown parameter to be estimated.

2. 3. PROPOSED PROCEDURE

The procedure proposed in §1. 2 is applicable to the current problem, but a minor modification is necessary. As in the procedure in §1. 2, we estimate μ_1 and σ^2 by maximizing the integrated likelihood

$$(2.2)\qquad IL(\mu_1, \sigma^2) = \int_0^\infty\cdots\int_0^\infty \prod_{t=1}^{T} g(y_t|\mu_t)\cdot\prod_{t=2}^{T} h(\mu_t|\mu_{t-1}, \sigma^2)\,d\mu_2\cdots d\mu_T.$$

However, since it is difficult to directly maximize the overall likelihood in non-Gaussian smoothing, we introduce the smoothing density to estimate $\mu_2, \cdots, \mu_T$.

Definition 2. 1. The smoothing density of μ_t is defined by

$$(2.3)\quad s(\mu_t|\underline{y}, \mu_1, \sigma^2) \propto \int_0^\infty\cdots\int_0^\infty \prod_{t=1}^{T} g(y_t|\mu_t)\cdot\prod_{t=2}^{T} h(\mu_t|\mu_{t-1}, \sigma^2)\,d\mu_2\cdots d\mu_{t-1}d\mu_{t+1}\cdots d\mu_T$$

where $\underline{y}=(y_1, \cdots, y_T)'$. ■

We estimate $\mu_2, \cdots, \mu_T$ by taking the expectations with respect to the empirical smoothing densities $s(\mu_t|\underline{y}, \hat{\mu}_1, \hat{\sigma}^2)$. The smoothing density is a marginal posterior density. Therefore, our estimates of $\mu_2, \cdots, \mu_T$ are the empirical posterior means.

The test statistic for homogeneity of the means is defined by

$$S = 2 \cdot \log\left\{ IL(\hat{\mu}_1, \hat{\sigma}^2) / IL(\tilde{\mu}_1, 0) \right\}$$

where $\tilde{\mu}_1$ attains the maximum of $IL(\mu_1, 0)$. The critical value c_α is hard to obtain analytically, and consequently we obtain empirical critical values by computer simulation. Table 2.1 presents these empirical values for the level $\alpha=.05$ with several sizes T and means μ_1, each of which is obtained from 10000 trials. We observe that they slightly increase with the mean μ_1 but are stable as a whole.

2.4. COMPUTATIONAL DEVELOPMENTS

To implement our procedure, it is necessary to calculate the multiple integrals in Eqs. (2.2) and (2.3). The integrations are facilitated by applying recursive formulas.

Theorem 2.1. The integrated likelihood $IL(\mu_1, \sigma^2)$ can be calculated recursively by using the formulas

$$(2.4a) \qquad q(\underline{y}^t | \mu_{t-1}, \sigma^2) = \int_0^\infty r(\underline{y}^t | \mu_t, \sigma^2) h(\mu_t | \mu_{t-1}, \sigma^2) d\mu_t$$

$$(2.4b) \qquad r(\underline{y}^{t-1} | \mu_{t-1}, \sigma^2) = g(y_{t-1} | \mu_{t-1}) q(\underline{y}^t | \mu_{t-1}, \sigma^2)$$

where $q(\underline{y}^t | \mu_{t-1}, \sigma^2)$ and $r(\underline{y}^t | \mu_t, \sigma^2)$ are the conditional densities of $\underline{y}^t = (y_t, \cdots, y_T)'$.

Proof. Repeating (2.4a) and (2.4b) alternately for $t=T, \cdots, 2$ with the initial condition $r(\underline{y}^T | \mu_T, \sigma^2) \equiv g(y_T | \mu_T)$, we get $r(\underline{y}^1 | \mu_1, \sigma^2)$, which is just the integrated likelihood. ■

In numerical computation, by executing the above procedure for once,

we can obtain $r(\underline{y}^1|\mu_1, \sigma^2)$ as a function with respect to μ_1 given $\underline{y}^1$ and σ^2. Therefore, we can estimate $\hat{\mu}_1$ given σ^2 with a small amount of computation. However, since $q(\cdot)$ and $r(\cdot)$ cannot be normalized at any step in the above procedure, it is difficult to evaluate $r(\underline{y}^1|\mu_1, \sigma^2)$ precisely in its absolute value. Consequently, we use $r(\underline{y}^1|\mu_1, \sigma^2)$ only for estimating $\hat{\mu}_1$. To estimate $\hat{\sigma}^2$ and to evaluate the log integrated likelihood, we use the following procedure.

Theorem 2.2 (*Kitagawa*, 1987). Let $\underline{y}_t=(y_1, \cdots, y_t)'$. The log integrated likelihood $\log IL(\mu_1, \sigma^2)$ can be calculated by

$$\log IL(\mu_1, \sigma^2) = \log g(y_1|\mu_1) + \sum_{t=2}^{T} \log l(y_t|\underline{y}_{t-1}, \mu_1, \sigma^2)$$

where

$$l(y_t|\underline{y}_{t-1}, \mu_1, \sigma^2) = \int_0^\infty g(y_t|\mu_t) p(\mu_t|\underline{y}_{t-1}, \mu_1, \sigma^2) d\mu_t$$

and each $p(\mu_t|\underline{y}_{t-1}, \mu_1, \sigma^2)$ is provided recursively by using the formulas

$$(2.5a) \qquad p(\mu_t|\underline{y}_{t-1}, \mu_1, \sigma^2) = \int_0^\infty h(\mu_t|\mu_{t-1}, \sigma^2) f(\mu_{t-1}|\underline{y}_{t-1}, \mu_1, \sigma^2) d\mu_{t-1}$$

$$(2.5b) \qquad f(\mu_t|\underline{y}_t, \mu_1, \sigma^2) = g(y_t|\mu_t) p(\mu_t|\underline{y}_{t-1}, \mu_1, \sigma^2) / l(y_t|\underline{y}_{t-1}, \mu_1, \sigma^2).$$

Here, $l(y_t|\underline{y}_{t-1}, \mu_1, \sigma^2)$, $p(\mu_t|\underline{y}_{t-1}, \mu_1, \sigma^2)$ and $f(\mu_t|\underline{y}_t, \mu_1, \sigma^2)$ denote conditional densities. ■

In the procedure (2.5), the initial condition is $f(\mu_1|\underline{y}_1, \hat{\mu}_1, \sigma^2) \equiv \delta(\mu_1-\hat{\mu}_1)$ with $\delta(\cdot)$ being the Dirac distribution, and the time t runs from 2 to T. We estimate $\hat{\sigma}^2$ by maximizing $\log IL(\hat{\mu}_1, \sigma^2)$ using a line search method.

After estimating $\hat{\mu}_1$ and $\hat{\sigma}^2$, we may calculate the empirical

smoothing densities.

Theorem 2.3. The smoothing densities can be calculated by

$$s(\mu_t|\underline{y}, \mu_1, \sigma^2) \propto q(\underline{y}^{t+1}|\mu_t, \sigma^2) f(\mu_t|\underline{y}_t, \mu_1, \sigma^2) \qquad t=2, \cdots, T-1.$$

Proof. This result follows immediately from the Bayes theorem. ■

When $t=T$, the smoothing density is identical with the filtering density $f(\mu_T|\underline{y}_T, \mu_1, \sigma^2)$.

We implement the above formulas by using standard numerical methods (cf., e.g., Dahlquist and Bjorck, 1974). Each function in the formulas is approximated by a piecewise linear function with m equally spaced knots defined on the interval $[\mu_{min}, \mu_{max}]$, and the trapezoidal rule is used for integration. The constants we actually used are $m=257$, $\mu_{min}=c$, $\mu_{max}=m\cdot c$ and $c=2\cdot \max_t y_t/(m-1)$.

2.5. APPLICATIONS

We apply the proposed method to three sets of weekly disease incidence data as illustrations.

The first data set consists of the weekly incidence of acute hemorrhagic conjunctivitis in Chiba-prefecture in Japan during 1987 collected by the National Infectious Disease Surveillance Program. A rise in incidence of this disease was reported in 1985 and 1986. However, in 1987, the number of cases was relatively small, and there was no clear incidence trend. It is of interest to examine the possible existence of some systematic pattern in these data. Figure 2.1 shows the data and the estimated trend plotted against time. As seen in this figure, the estimated trend is at an increased

level between about the 28th and 38th weeks. It would be difficult to detect such an increase, which is not readily apparent in the data, without a suitable smoothing method. The test statistic S takes the value 4.8, which is greater than the critical value. This indicates that the mean weekly incidence was not stable during the period, but changed gradually.

The second data set consists of the weekly incidence of acute febrile Muco-Cutaneous Lymphnode Syndrome (MCLS) in Tottori-prefecture during 1982 collected by the Study Committee on Cause of MCLS. Figure 2.2 shows the data and the estimated trend. In this year, a nation-wide outbreak was reported. The data shows a clear rise in incidence. The estimated trend has a peak around the 17th week, followed by a sharp decrease, and after that it maintains a fairly constant value until the end of the year. In this example, inspection by eye may yield a trend similar to the above. The test statistic S takes the value 20.4.

The last data set contains the same MCLS data as in the previous data set, for 1983. Figure 2.3 shows the data and the estimated trend. While the total number of cases in this year was not much smaller than in 1982, the incidence pattern was not as clear as in the earlier example. In the present case, $\hat{\sigma}^2=0$ and the estimated trend is a horizontal straight line. This result agrees with the report by the Study Committee on Cause of MCLS. By definition, S takes the value zero.

2.6. EXTENSIONS

An advantage of the proposed method is that it can be extended

to a wide class of models. Some of these extensions are now discussed.

First, we consider the second difference constraint case, i.e., $d=2$, which is often employed in the smoothing problem.

Lemma 2.2. The model for $d=2$ is written in the state-space form as

$$g(y_t|\mu_t) = e^{-\mu_t}\mu_t^{y_t}/y_t! \qquad t=1, \cdots, T$$

$$h(\mu_t|\mu_{t-1}, \mu_{t-2}, \sigma^2) = (2\pi)^{-\frac{1}{2}}(\sigma\mu_t)^{-1}\exp\left\{-\frac{1}{2\sigma^2}(\log\mu_t - 2\log\mu_{t-1} + \log\mu_{t-2})^2\right\} \qquad t=3, \cdots, T$$

where μ_1 and μ_2 are unknown parameters to be estimated. ■

The recursive formulas for this model can be derived by modifying the discussions in §2.4. Figure 2.4 presents the results of application of the procedure with $d=2$ to the data given in Fig. 2.1. The estimated trend is smoother than that in Fig. 2.1. It may be more appealing, but the trends are very close to each other. The extension to other higher order cases is straightforward, though the required computer memory size increases exponentially with d. The required memory size is roughly estimated to be proportional to $T \cdot m^d$.

Next, we consider the extension to the binomial case.

Lemma 2.3. The model for smoothing serial binomial data corresponding to Model (2.1) is written in the state-space form as

$$g(y_t|\mu_t) = {}_{m_t}C_{y_t}\mu_t^{y_t}(1-\mu_t)^{m_t-y_t} \qquad t=1, \cdots, T$$

$$h(\mu_t|\mu_{t-1}, \sigma^2) = (2\pi)^{-\frac{1}{2}}\{\sigma\mu_t(1-\mu_t)\}^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\log\frac{\mu_t}{1-\mu_t} - \log\frac{\mu_{t-1}}{1-\mu_{t-1}}\right)^2\right\} \qquad t=2, \cdots, T. \blacksquare$$

The recursive formulas mentioned in §2.4 are applicable to this model

with the minor modification of replacing the domain of integration $(0, \infty)$ by $(0, 1)$.

Finally, we add that our approach can be extended to include explanatory variables. Several authors have discussed the regression problem with time-dependent coefficients for serial non-Gaussian data. West, Harrison and Migon (1985) discussed dynamic generalized linear models. Zeger and Qaqish (1988) discussed quasi-likelihood Markov models. However, they did not attempt to evaluate exact likelihoods. Our approach enables us to evaluate exact likelihoods even in regression models with time-dependent coefficients, though it requires a computer with a large memory capacity. Therefore, it seems necessary to develop a numerical method with reduced memory size requirements.

2.7. THE METHOD UNDER THE NORMALITY ASSUMPTION

It may be still appealing to assume normality in (2.1a) and (2.1b) even in the analysis of serial count data, because of its simplicity and familiarity. However, the characteristics of the Poisson and normal distributions are quite different, especially when the means are small. Therefore, it is desirable to assume the Poisson distribution for count data, because it is more realistic.

The comparison of smoothing methods is unfortunately difficult since preferences between fitted trends are largely subjective. To illustrate this, we present Fig. 2.5, which gives the estimated trend with the first difference constraint under the normality assumption by using the data in Fig. 2.2. In this figure, a relatively large wave is observed towards the end of the period. On the other hand,

the general shape is close to that in Fig. 2. 2.

The comparative study of tests is easier since we have objective criteria such as power. Freedman (1981) assumed the alternative $(p_1, \cdots, p_{12})$=(. 101, . 103, . 101, . 085, . 076, . 073, . 076, . 076, . 073, . 075, . 073, . 088) to compare the power of several tests. This alternative gives a gradually changing trend. We assume $(\mu_1, \cdots, \mu_{12})=(p_1, \cdots, p_{12})N$ and N=100. The empirical powers at the level . 05 with 10000 trials are . 202 and . 167 for the proposed test and the test obtained under the normality assumption, respectively. Another simulation study for the case $(\mu_1, \cdots, \mu_{12})=(1, \cdots, 1)$ shows that the test obtained under normality is slightly liberal. It is reasonable to believe that the reduction in power is due to the inappropriate normality assumption.

Table 2. 1. Empirical critical values c_α for the level α=. 05 with 10000 trials.

	T				
μ_1	10	20	30	40	50
1	. 41	. 33	. 34	. 36	. 35
3	. 43	. 32	. 32	. 31	. 33
5	. 41	. 37	. 34	. 37	. 34
7	. 47	. 38	. 33	. 33	. 37
10	. 43	. 42	. 35	. 39	. 37

Figure 2. 1. Observations (×) and estimated trend (solid line) with the 1st order difference constraint for the weekly incidence of acute hemorrhagic conjunctivitis in Chiba-prefecture in Japan during 1987.

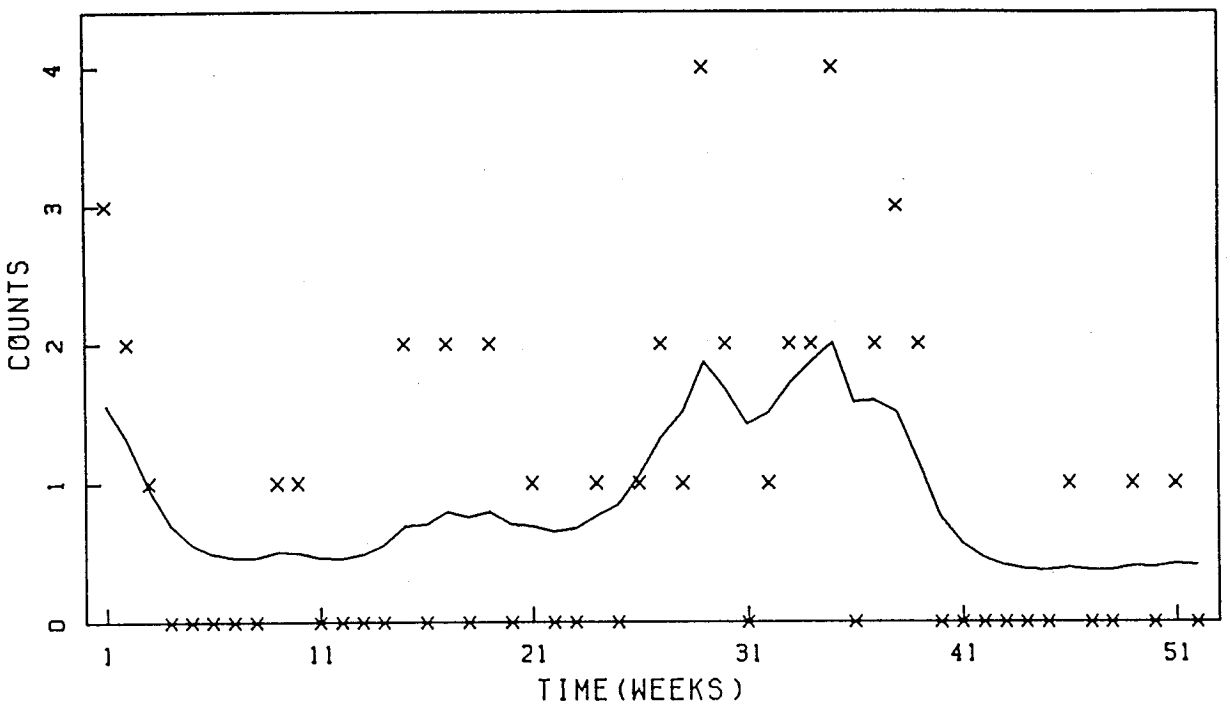


Figure 2. 2. Observations (×) and estimated trend (solid line) with the 1st order difference constraint for the weekly incidence of MCLS in Tottori-prefecture in Japan during 1982.

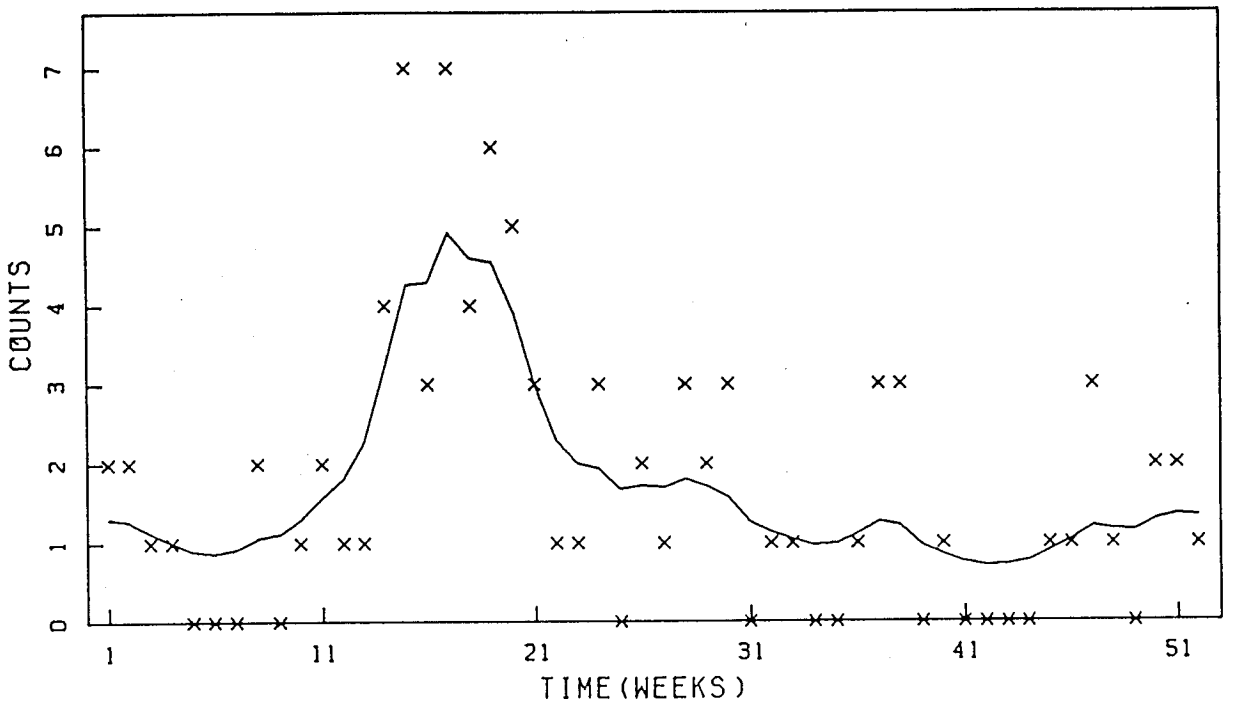


Figure 2.3. Observations (×) and estimated trend (solid line) with the 1st order difference constraint for the weekly incidence of MCLS in Tottori-prefecture in Japan during 1983.

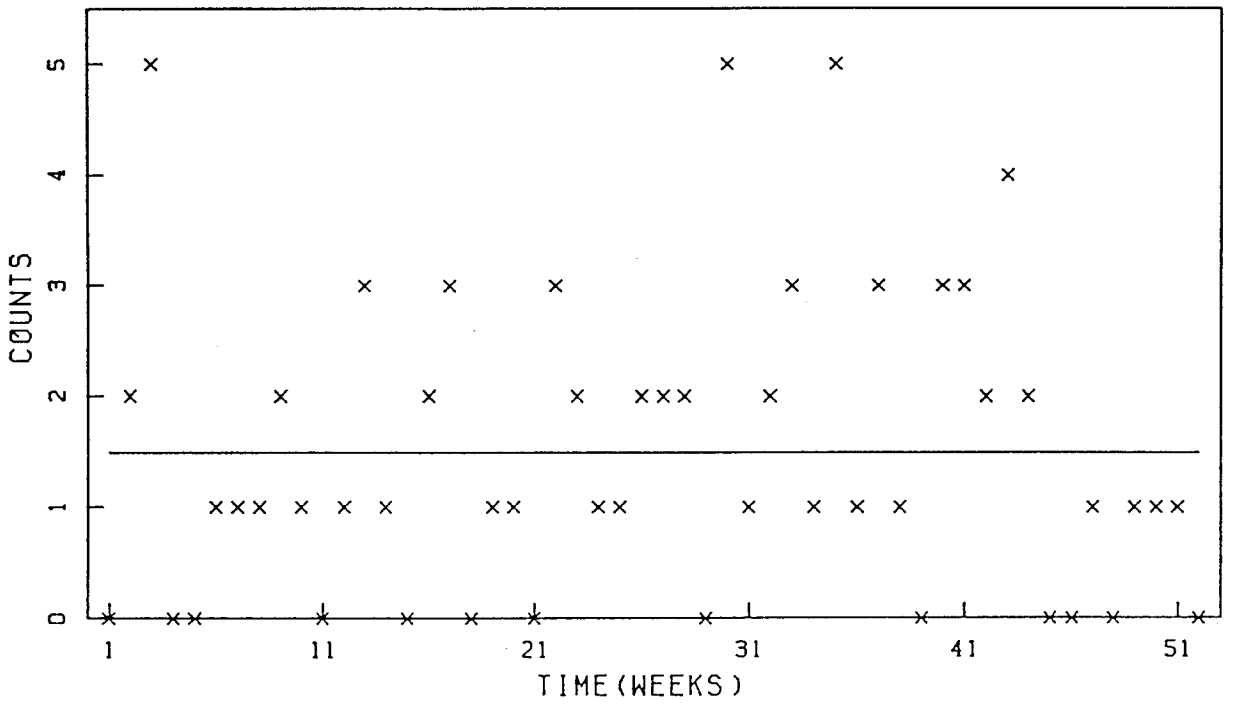


Figure 2.4. Observations (×) and estimated trend (solid line) with the 2nd order difference constraint for the data of Figure 2.1.

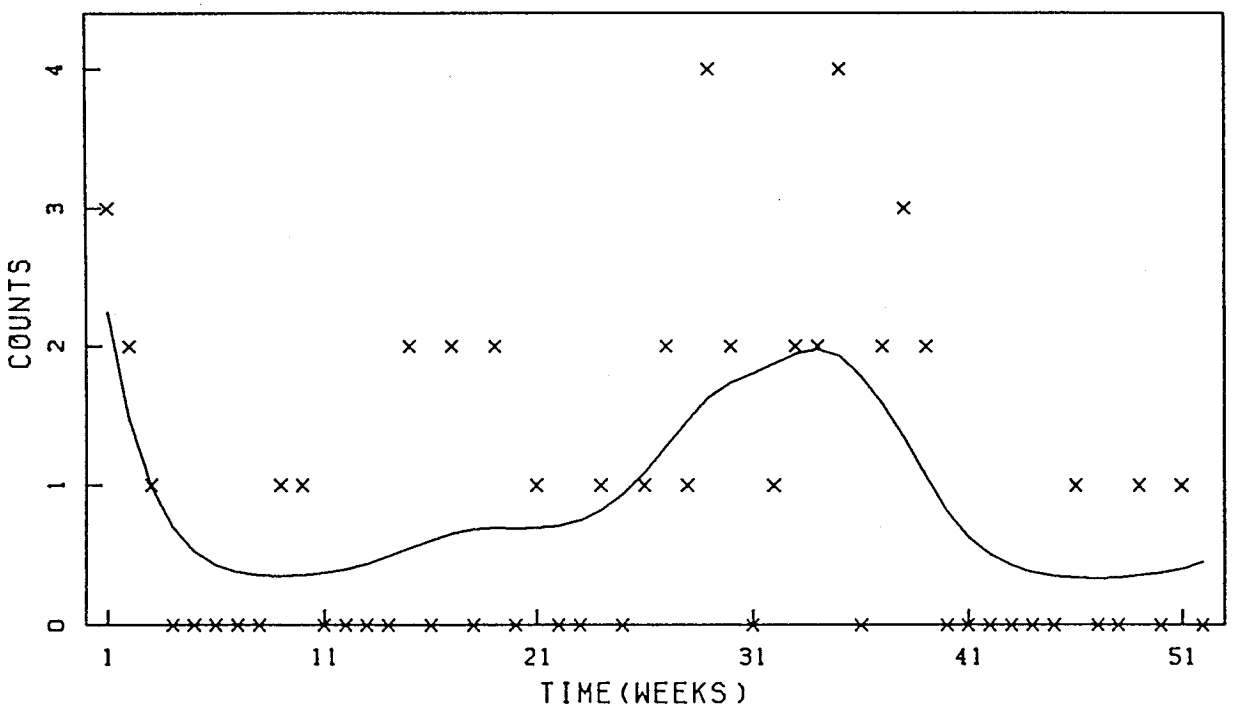
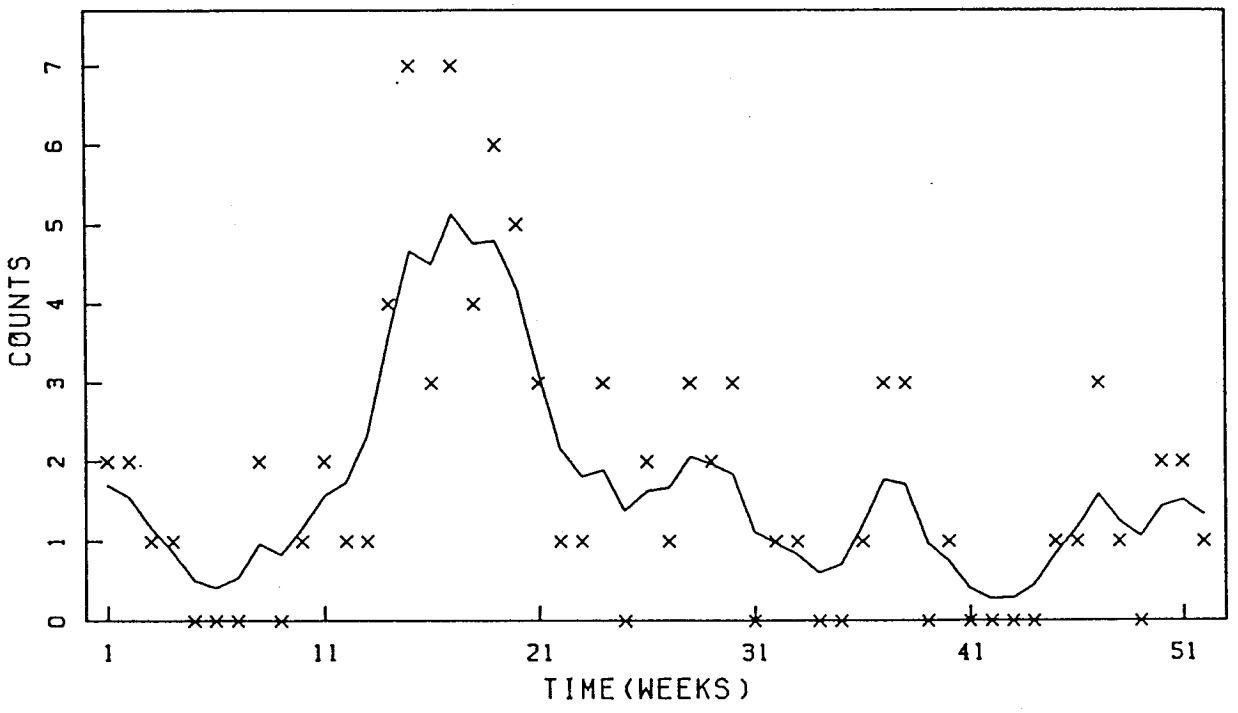


Figure 2. 5. Observations (x) and estimated trend (solid line) with the 1st order difference constraint under the normality assumption for the data of Figure 2. 2.



CHAPTER 3

BAYESIAN DETECTION OF STRUCTURAL CHANGES

3. 1. INTRODUCTION

Let $y_1, \cdots, y_T$ be a sequence of observations taken at equally spaced intervals.

Definition 3. 1. A sequence of random variables $Y_1, \cdots, Y_T$ is said to have n change points at $j(1), \cdots, j(n)$ $(1 \leq j(1) < \cdots < j(n) < T)$ if the density of $\underline{y}=(y_1, \cdots, y_T)'$ has the form

$$(3.1) \qquad p(\underline{y} | J_{j(1)} \cap \cdots \cap J_{j(n)}, N=n, \underline{\theta}_0, \cdots, \underline{\theta}_n) = \prod_{i=0}^{n} p_i(\underline{y}_i | \underline{\theta}_i)$$

where $J_{j(i)}$ is the event that the sequence has a change point at $j(i)$; N is the number of change points; $\underline{y}_i=(y_{j(i)+1}, \cdots, y_{j(i+1)})'$ with $j(0)=0$ and $j(n+1)=T$; $p_i(\underline{y}_i | \underline{\theta}_i)$ is the density of $\underline{y}_i$ with the parameter $\underline{\theta}_i$ and $\underline{\theta}_i \neq \underline{\theta}_{i'}$ $(i \neq i')$. ■

In this chapter, we consider the problem of making inferences about change points under the conditions that the places of change points, the number of change points and the values of $\underline{\theta}_i$'s are unknown.

Since Page (1954), the change point problem has been considered by many authors from various viewpoints. [For references, see Poirier (1976), Zacks (1983) and Broemeling and Tsurumi (1987)]. However, most studies have been concerned with the detection of a single change or the detection of multiple changes by using a stepwise procedure; as a result, few studies are available on the problem of detecting multiple changes without using a stepwise

procedure. Smith (1980) is one of those few works, in which he suggested the usefulness of the Bayesian approach for the multiple change case.

In this chapter, we are concerned with the problem of detecting multiple changes without using a stepwise procedure. To solve this problem, we propose a method for evaluating the posterior distribution of N and the posterior probability of each J_t unconditionally. We also present an approximation procedure to decrease the amount of computation.

In Section 3.2, a Bayesian formulation of the problem is presented. In Section 3.3, the detailed analysis is given for binomial data and the Lindisfarne scribes problem is analyzed. In Section 3.4, an approximation procedure is presented. In Section 3.5, changes in the regression case are studied for two specific models; the simple regression model and the discrete spline, and numerical illustrations are also provided.

3.2. A BAYESIAN FORMULATION

In this section, we derive the posterior distribution of N and the posterior probability of J_t.

When the sequence is assumed to have n change points at $j(1), \cdots, j(n)$, the density of $\underline{y}$ is given by (3.1). Assuming a prior density $\omega(\underline{\theta})$ for $\underline{\theta}$; where $\underline{\theta}=(\underline{\theta}_0, \cdots, \underline{\theta}_n)$, the integrated likelihood of $\{J_{j(1)} \cap \cdots \cap J_{j(n)}, N=n\}$ is obtained as

$$p(\underline{y} | J_{j(1)} \cap \cdots \cap J_{j(n)}, N=n) = \int \cdots \int p(\underline{y} | J_{j(1)} \cap \cdots \cap J_{j(n)}, N=n, \underline{\theta}) \omega(\underline{\theta}) d\underline{\theta}.$$

By Bayes' theorem, the posterior probability of $J_{j(1)} \cap \cdots \cap J_{j(n)}$ given $\underline{y}$

and n is provided as

$$p(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,\underline{y},\ N=n) = \frac{p(\underline{y}\,|\,J_{j(1)}\cap\cdots\cap J_{j(n)},\ N=n)\,\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,N=n)}{p(\underline{y}\,|\,N=n)}$$

where $\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,N=n)$ is a prior probability of $J_{j(1)}\cap\cdots\cap J_{j(n)}$ given n and

$$p(\underline{y}\,|\,N=n) = \sum_{\Omega_n} p(\underline{y}\,|\,J_{j(1)}\cap\cdots\cap J_{j(n)},\ N=n)\,\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,N=n)$$

$$\Omega_n = \{(j(1),\cdots,j(n))\,|\,1\leq j(1)<\cdots<j(n)<T\}.$$

Taking the sum of $p(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,\underline{y},\ N=n)$'s which involve J_t, the posterior probability of J_t given $\underline{y}$ and n is obtained as

$$p(J_t\,|\,\underline{y},\ N=n) = \sum_{\Omega_{n,\,t}} p(J_{j(1)}\cap\cdots\cap J_{j(n)}\,|\,\underline{y},\ N=n)$$

$$\Omega_{n,\,t} = \{(j(1),\cdots,j(n))\,|\, \exists k \text{ such that } j(k)=t\ \ 1\leq k\leq n,\ \ 1\leq j(1)<\cdots<j(n)<T\}.$$

On the other hand, the posterior probability of $N=n$ given $\underline{y}$ is provided by Bayes' theorem as

$$p(N=n\,|\,\underline{y}) = \frac{p(\underline{y}\,|\,N=n)\,\omega(N=n)}{p(\underline{y})}$$

where $\omega(N=n)$ is a prior probability of $N=n$ and $p(\underline{y})=\sum_{n=0}^{T-1} p(\underline{y}\,|\,N=n)\,\omega(N=n)$. The posterior probability of J_t given $\underline{y}$ is obtained as

$$p(J_t\,|\,\underline{y}) = \sum_{n=1}^{T-1} p(J_t\,|\,\underline{y},\ N=n)\,p(N=n\,|\,\underline{y}).$$

The necessary ingredients to evaluate the posterior

probabilities in the above formulation are $p(\underline{y}|J_{j(1)}\cap\cdots\cap J_{j(n)}, N=n)$, $\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}|N=n)$ and $\omega(N=n)$. Here, we assume the following prior probabilities used in Smith (1980) for $\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}|N=n)$ and $\omega(N=n)$.

$$\omega(J_{j(1)}\cap\cdots\cap J_{j(n)}|N=n) = \frac{1}{{}_{T-1}C_n} \qquad 1\leq n<T$$

$$\omega(N=n) = \frac{1}{T} \qquad 0\leq n<T.$$

The remaining ingredient, the likelihood of $J=\{J_{j(1)}\cap\cdots\cap J_{j(n)}, N=n\}$ is provided concretely for some models in Sections 3. 3, 3. 5. 1 and 3. 5. 2.

3. 3. LINDISFARNE SCRIBES PROBLEM

The Lindisfarne scribes problem is one of the well-known examples of the change point problem. The aim in this problem is to make inferences about changes of scribe by using the data on the number of occurrences of present indicative 3*rd* singular endings *s* and δ in each section of Lindisfarne. Table 3. 1 shows the data taken from Smith (1980). These data have been analyzed by Smith (1980), Silvey (1958), Pettitt (1979) and Carlstein (1988). The latter three authors drew the conclusion by using some test statistics that the change occurred after the 5*th* section. Smith (1980) evaluated the posterior probabilities of up to two changes and concluded that the change occurred after the 4*th* section and again after the 5*th* section. In this section, we apply our method to the data of Lindisfarne and compare our results with theirs.

Let m_t and y_t be the numbers of occurrences of present indicative 3*rd* singular endings and δ-forms at the t-*th* section

$(t=1, \cdots, T)$, respectively. Similarly to Smith (1980), we assume the binomial distribution with parameters m_t and θ_t for y_t $(t=j(i)+1, \cdots, j(i+1),\ i=0, \cdots, n)$. Then the likelihood of J is written as

$$p(\underline{y}|J) = \int\cdots\int \prod_{i=0}^{n} \prod_{t=j(i)+1}^{j(i+1)} {}_{m_t}C_{y_t}\theta_t^{y_t}(1-\theta_t)^{m_t-y_t}\omega(\underline{\theta})d\underline{\theta}$$

where $\underline{\theta}=(\theta_0, \cdots, \theta_n)$.

As $\omega(\underline{\theta})$, Smith (1980) assumed a conjugate prior. This is one of several possible selections. On the other hand, we specify the model by the maximum likelihood estimate of $\underline{\theta}$, i.e., we place our confidence on the maximum likelihood estimate. Then we define the likelihood of J by using the maximum likelihood. The following lemma is immediately obtained.

Lemma 3. 1. The maximum likelihood estimate of $\underline{\theta}$ is given as $\hat{\theta}_t=\sum_t y_t/\sum_t m_t$ $(i=0, \cdots, n)$ and the maximum log likelihood is given by

$$\log p(\underline{y}|J, \hat{\underline{\theta}}) = \sum_{t=1}^{T} \log {}_{m_t}C_{y_t} + \sum_{i=0}^{n} \sum_{t=j(i)+1}^{j(i+1)} \left\{ y_t \cdot \log\hat{\theta}_t + (m_t-y_t)\cdot\log(1-\hat{\theta}_t) \right\}. \blacksquare$$

It may be a possible selection to use the maximum likelihood $p(\underline{y}|J, \hat{\underline{\theta}})$ as an estimate of $p(\underline{y}|J)$. However, the maximum log likelihood has a bias in the following sense.

Lemma 3. 2.

$$\log p(\underline{y}|J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z}|J, \hat{\underline{\theta}}) = \sum_{t=1}^{T} (\log {}_{m_t}C_{y_t} - E_Z \log {}_{m_t}C_{Z_t}) + \sum_{i=0}^{n} f_i(\hat{\theta}_i-\theta_i)\cdot\log\frac{\hat{\theta}_i}{(1-\hat{\theta}_i)}$$

where E_Z denotes the expectation under the assumed distribution of $\underline{Z}$, $p(\underline{z}|J, \underline{\theta})$ and $f_i=\sum_{t=j(i)+1}^{j(i+1)} m_t$ $(i=0, \cdots, n)$.

Proof. This is a direct consequence of the definition. ■

This bias increases in average as $dim(\underline{\theta})$ becomes large, suggesting that the use of the maximum likelihood causes an overestimation of the number of change points. To prevent such an overestimation, it is necessary to correct the bias. However, since the true parameter $\underline{\theta}$ is unknown, the present form of the bias is useless. Consequently, we employ the predictive log likelihood to correct the bias.

Definition 3. 2 (*Kitagawa and Akaike*, 1982). The predictive log likelihood is defined by

$$\log p^{pred}(\underline{y}|J) = \log p(\underline{y}|J, \hat{\underline{\theta}}) - E_Y\left[\log p(\underline{Y}|J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z}|J, \hat{\underline{\theta}})\right]$$

where E_Y denotes the expectation under the assumed distribution of data. ■

Theorem 3. 1.

$$E_Y\left[\log p(\underline{Y}|J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z}|J, \hat{\underline{\theta}})\right]$$
$$= \sum_{i=0}^{n}\left\{1+\frac{\theta_i^2-\theta_i+\frac{1}{2}}{f_i\theta_i(1-\theta_i)}+\frac{\theta_i^4-2\theta_i^3+4\theta_i^2-3\theta_i+\frac{5}{6}}{f_i^2\theta_i^2(1-\theta_i)^2}+O(f_i^{-3})\right\}.$$

Proof. The expectation of the bias can be written as

$$E_Y\left[\log p(\underline{Y}|J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z}|J, \hat{\underline{\theta}})\right] = \sum_{i=0}^{n} f_i E_Y\left[(\hat{\theta}_i-\theta_i)\cdot\log\frac{\hat{\theta}_i}{(1-\hat{\theta}_i)}\right].$$

Using the power series, we have

$$E_Y\left[(\hat{\theta}_i-\theta_i)\cdot\log\frac{\hat{\theta}_i}{(1-\hat{\theta}_i)}\right] = \sum_{j=1}^{\infty}\frac{1}{j}\cdot\left\{\frac{(-1)^{j-1}}{\theta_i^j}+\frac{1}{(1-\theta_i)^j}\right\}E_Y(\hat{\theta}_i-\theta_i)^{j+1}.$$

Therefore, substituting the central moments, we obtain

$$E_Y\left[\log p(\underline{Y}|J,\hat{\underline{\theta}})-E_Z\log p(\underline{Z}|J,\hat{\underline{\theta}})\right]$$

$$=\sum_{i=0}^{n}\left\{1+\frac{\theta_i^2-\theta_i+\frac{1}{2}}{f_i\theta_i(1-\theta_i)}+\frac{\theta_i^4-2\theta_i^3+4\theta_i^2-3\theta_i+\frac{5}{6}}{f_i^2\theta_i^2(1-\theta_i)^2}+O(f_i^{-3})\right\}.\blacksquare$$

Using this result, we define the predictive log likelihood in the current problem as

$$\log p^{pred}(\underline{y}|J)=\log p(\underline{y}|J,\hat{\underline{\theta}})-\sum_{i=0}^{n}\left\{1+\frac{\hat{\theta}_i^2-\hat{\theta}_i+\frac{1}{2}}{f_i\hat{\theta}_i(1-\hat{\theta}_i)}+\frac{\hat{\theta}_i^4-2\hat{\theta}_i^3+4\hat{\theta}_i^2-3\hat{\theta}_i+\frac{5}{6}}{f_i^2\hat{\theta}_i^2(1-\hat{\theta}_i)^2}\right\}.$$

As the estimate of $p(\underline{y}|J)$, we use $\exp\{\log p^{pred}(\underline{y}|J)\}$.

Now we apply our method to the data of Lindisfarne. Table 3.2 presents the estimate of each $p(N=n|\underline{y})$ as well as the posterior mean, mode and median of N and Smith's results. Table 3.3 presents the estimate of each $p(J_i|\underline{y})$. It is difficult to precisely compare our results with those of Smith because he has not presented the posterior probabilities of more than two changes; nevertheless, there seem to be some differences between them. While the posterior probability of two changes is quite dominant in Smith's results, it is not so dominant in our results. This difference may be caused by the difference between the assumed distributions for $\underline{\theta}$ and by the different policy for the bias correction. However, in spite of this difference between both results, we can agree with Smith's conclusion. Actually, if we take the posterior mode of N, the conclusion that there are two changes is obtained. From Table 3.3, it is seen that the top two $p(J_i|\underline{y})$'s are obtained at the $4th$ and $5th$ sections.

3. 4. AN APPROXIMATION PROCEDURE

We call the evaluation of the posterior probabilities by the method mentioned in Section 3. 2 *the full computation*. In the Lindisfarne scribes problem, *the full computation* was feasible. However, the number of estimations of $p(\underline{y}|J)$'s in *the full computation*, which is given by $\sum_{n=0}^{T-1} {}_{T-1}C_n$, increases exponentially with the size of the sequence and quickly *the full computation* becomes infeasible. In this section, we present an approximation procedure which enables us to evaluate $p(J_t|\underline{y})$'s even when *the full computation* is infeasible.

The flow of the approximation is as follows:

0. Calculate $p(\underline{y}|N=0)$, and set $n \leftarrow 1$.
1. Calculate $p(J_t|\underline{y}, N=n)$ $(1 \leq t < T)$ by the method in §3. 2.
2. Let m be the number of repetitions of Step 1. If $n<m$ then set $n \leftarrow n+1$ and return to Step 1.
3. Determine whether n is sufficiently large to terminate. If so, then go to Step 7. If not, then set $n \leftarrow n+1$.
4. Let α be a small value. Make the index set $I_{n,\alpha}=\{i \mid p(J_i|\underline{y}, N=n-1) \leq \alpha \quad 1 \leq i < T\}$ and calculate $g(n, t) \equiv \sum_{\Omega_{n,t}} p(\underline{y}|J)$ $(1 \leq t < T)$ under the following assumption:

$$g(n, t) = \begin{cases} \dfrac{{}_{T-2}C_{n-1} \cdot g(n-1, t)}{{}_{T-2}C_{n-2}} & t \in I_{n,\alpha} \\ \displaystyle\sum_{\Omega^{\alpha}_{n,t}} p(\underline{y}|J) + \sum_{k \in I_{n,\alpha}} \frac{p(J_t|\underline{y}, N=n-1)\, g(n, k)}{1-\dfrac{p(J_k|\underline{y}, N=n-1)}{n-1}} & t \notin I_{n,\alpha} \end{cases}$$

$$\Omega^{\alpha}_{n,t} = \{(j(1), \cdots, j(n)) \mid (j(1), \cdots, j(n)) \in \Omega_{n,t},\ j(i) \notin I_{n,\alpha}\ 1 \leq i \leq n\}.$$

5. Using $g(n, t)$' s and the relations

$$p(J_t|\underline{y}, N=n) = \frac{g(n, t)}{{}_{T-1}C_n \cdot p(\underline{y}|N=n)} \qquad p(\underline{y}|N=n) = \frac{1}{{}_{T-1}C_n \cdot n}\sum_{t=1}^{T-1} g(n, t),$$

calculate $p(J_t|\underline{y}, N=n)$ $(1 \leq t < T)$.

6. Return to Step 3.

7. Calculate $p(J_t|\underline{y})$ $(1 \leq t < T)$ assuming the prior

$$\omega(N=k) = \begin{cases} \frac{1}{n+1} & k \leq n \\ 0 & k > n \end{cases}.$$

The number of estimations of $p(\underline{y}|J)$' s is decreased in Step 4 by approximating $g(n, t)$' s. The approximation of $g(n, t)$ is introduced as follows. Consider the case where J_k is assumed to be an unimportant event, i. e., $I_{n,\alpha}=\{k\}$. In this case, it may be reasonable to consider assigning approximate values to the predictive likelihoods of J' s which involve J_k in order to decrease the amount of computation. To obtain such approximate values, we set the following two assumptions. The first assumption is that the mean of the predictive likelihoods of J' s which involve J_k when $N=n$ is equivalent to the mean of those when $N=n-1$. By this assumption, we have

$$g(n, k) = \frac{{}_{T-2}C_{n-1} \cdot g(n-1, k)}{{}_{T-2}C_{n-2}}.$$

On the other hand, $g(n, t)$ $(t \neq k)$ can be written as

$$g(n, t) = g_1(n, t) + g_2(n, t)$$

$$g_1(n, t) = \sum_{\Omega^{\alpha}_{n, t}} p(\underline{y}|J) \qquad g_2(n, t) = \sum_{\Omega_{n, t} - \Omega^{\alpha}_{n, t}} p(\underline{y}|J).$$

We evaluate $g_1(n, t)$ by the method mentioned in Section 3. 2. However,

since $g_2(n, t)$ is the sum of the predictive likelihoods of J's which involve J_k, we assign an approximate value to it. An approximate value can be obtained by using the relation

$$\sum_{t \neq k} g_2(n, t) = (n-1)g(n, k).$$

This relation suggests that we may distribute $(n-1)g(n, k)$ into $g_2(n, t)$'s $(t \neq k)$. Using the posterior probabilities when $N=n-1$, we set the second assumption

$$g_2(n, t) = \frac{\frac{p(J_t | \underline{y}, N=n-1)}{n-1}}{1-\frac{p(J_k | \underline{y}, N=n-1)}{n-1}}(n-1)g(n, k) \qquad t \neq k.$$

The approximation of $g(n, t)$ has been obtained.

The above two assumptions may be ad hoc. However, a close approximation increases the amount of computation. We consider that they are acceptable ones in practical application.

In the approximation procedure, there are some arbitrary constants, m and α. A basic strategy as to their choice is to chose the largest m and smallest α as large and small, respectively, as the computer may permit. By some experiments, we have found that: 1) When m is greater than the mode of N, the possibility to miss change points is very small. 2) When α is less than a certain value, as the number of elements of $I_{n,\alpha}$ is less than about $T-n-6$, relatively good approximate values are obtained.

3. 5. DETECTION OF CHANGES BY REGRESSION MODELS

In this section, we give the estimate of $p(\underline{y}|J)$ for two

regression models, the simple regression model and the discrete spline. In addition, we show an example of the application of our method by using the discrete spline.

3. 5. 1. The simple regression model

Assume the simple regression model

$$y_t = \alpha_i+\beta_i t+\varepsilon_t \quad \varepsilon_t \sim i.i.d.N(0, \sigma^2) \quad j(i)+1\leq t\leq j(i+1)$$

for $\underline{y}_i$. Then the density of $\underline{y}_i$ can be written as

$$(3.2) \qquad p_S(\underline{y}_i|\underline{\nu}_i, \sigma^2) = (2\pi)^{-\frac{\kappa_i}{2}}\sigma^{-\kappa_i}\exp\left\{-\frac{1}{2\sigma^2}(\underline{y}_i-A_i\underline{\nu}_i)'(\underline{y}_i-A_i\underline{\nu}_i)\right\}$$

where $\underline{\nu}_i=(\alpha_i, \beta_i)'$, $\kappa_i=dim(\underline{y}_i)$ and

$$A_i = \begin{bmatrix} 1 & j(i)+1 \\ 1 & j(i)+2 \\ \vdots & \vdots \\ 1 & j(i+1) \end{bmatrix}.$$

Further, since the simple regression model is inapplicable to $\underline{y}_i$ when $\kappa_i=1$, we assume the following outlier model for such $\underline{y}_i$.

$$(3.3) \qquad p_N(\underline{y}_i|\underline{\nu}_i, \sigma^2) = \frac{1}{\sigma}\cdot\varphi\left(\frac{y_{j(i)+1}-\alpha_i}{\sigma}\right)$$

where $\underline{\nu}_i=(\alpha_i)$ and φ denotes the standard normal probability density function. Using (3. 2) and (3. 3), the density of $\underline{y}$ can be written as

$$p(\underline{y}|J, \underline{\theta}) = \prod_{i\in I_1} p_N(\underline{y}_i|\underline{\nu}_i, \sigma^2)\cdot\prod_{i\in I_G} p_S(\underline{y}_i|\underline{\nu}_i, \sigma^2)$$

where $\underline{\theta}=(\underline{\nu}_0', \cdots, \underline{\nu}_n', \sigma^2)$, $I_1=\{i|\kappa_i=1\ 0\leq i\leq n\}$ and $I_G=\{i|\kappa_i\geq 2\ 0\leq i\leq n\}$.

Lemma 3. 3. The maximum likelihood estimate of $\underline{\theta}$ is given as

$$\hat{\alpha}_i = y_{j(i)+1} \qquad \text{for } i \in I_1$$

$$\hat{\underline{\nu}}_i = (A_i'A_i)^{-1}A_i'\underline{y}_i \qquad \text{for } i \in I_G$$

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i \in I_G} (\underline{y}_i - A_i\hat{\underline{\nu}}_i)'(\underline{y}_i - A_i\hat{\underline{\nu}}_i)$$

and the maximum log likelihood is given by

$$\log p(\underline{y} \mid J, \hat{\underline{\theta}}) = -\frac{T}{2}\cdot\log 2\pi\hat{\sigma}^2 - \frac{T}{2}.$$

Proof. This result is obtained immediately by maximizing the log likelihood

$$\log p(\underline{y} \mid J, \underline{\theta}) = -\frac{T}{2}\cdot\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\{\sum_{i \in I_1} (y_{j(i)+1} - \alpha_i)^2 + \sum_{i \in I_G} (\underline{y}_i - A_i\underline{\nu}_i)'(\underline{y}_i - A_i\underline{\nu}_i)\right\}. \blacksquare$$

Lemma 3. 4.

$$\log p(\underline{y} \mid J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z} \mid J, \hat{\underline{\theta}}) = \frac{1}{2\hat{\sigma}^2}\left\{T\sigma^2 + \sum_{i \in I_1} (\alpha_i - \hat{\alpha}_i)^2 + \sum_{i \in I_G} (\underline{\nu}_i - \hat{\underline{\nu}}_i)'A_i'A_i(\underline{\nu}_i - \hat{\underline{\nu}}_i)\right\} - \frac{T}{2}.$$

Proof. This is a direct consequence of the definition. ■

Theorem 3. 2.

$$E_Y\left[\log p(\underline{Y} \mid J, \hat{\underline{\theta}}) - E_Z \log p(\underline{Z} \mid J, \hat{\underline{\theta}})\right] = \frac{T(T + \#I_1 + 2\#I_G)}{2\left(\sum_{i \in I_G} \kappa_i - 2\#I_G - 2\right)} - \frac{T}{2}$$

where $\#I_*$ denotes the number of elements included in the set I_*.

Proof. This result follows from Lemma 3. 4 and the following properties.

$$\frac{(a_i-\hat{a}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad for\ i\in I_1$$

$$\frac{(\nu_i-\hat{\nu}_i)'A_i'A_i(\nu_i-\hat{\nu}_i)}{\sigma^2} \sim \chi^2_{(2)} \qquad for\ i\in I_G$$

$$\frac{T\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\left(\sum_{i\in I_G} k_i-2\#I_G\right)}$$

where $\chi^2_{(k)}$ denotes the χ^2-distribution of order k.■

From Lemma 3.3 and Theorem 3.2, the predictive log likelihood is obtained as

$$\log p^{pred}(\underline{y}|J) = -\frac{T}{2}\log 2\pi\hat{\sigma}^2 - \frac{T(T+\#I_1+2\#I_G)}{2\left(\sum_{i\in I_G} \kappa_i-2\#I_G-2\right)}.$$

We use $\exp\{\log p^{pred}(\underline{y}|J)\}$ as the estimate of the likelihood $p(\underline{y}|J)$.

3.5.2. The discrete spline

Harrison and Stevens (1976) presented three examples of sequences including a single change, which are shown in Fig. 3.1. Although they generated these sequences by the linear growth model, it is possible to represent them by the model mentioned in the previous section. For example, the outlier case can be represented by applying model (3.2) to the data at $1\leq t\leq 4$ and $6\leq t\leq 10$ and applying model (3.3) to the data at $t=5$. However, if the data at $1\leq t\leq 4$ and $6\leq t\leq 10$ are on a curve instead of a straight line, the model mentioned in the previous section becomes inappropriate. For such a case, the discrete spline discussed in §1.3.3 is useful.

Assume the discrete spline

$$y_t \sim i.i.d.\ N(\mu_t, \sigma^2) \qquad j(i)+1 \le t \le j(i+1)$$

$$\mu_t - 2\mu_{t-1} + \mu_{t-2} \sim i.i.d.\ N(0, \frac{\sigma^2}{\lambda}) \qquad j(i)+3 \le t \le j(i+1)$$

for $\underline{y}_i$. Then the density of $\underline{y}_i$ can be written by using the form (1.6) as

$$(3.4)\quad p_D(\underline{y}_i | \underline{\nu}_i, \sigma^2, \lambda) = (2\pi\sigma^2)^{-\frac{\kappa_i}{2}} |V_i|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\underline{y}_i - A_i\underline{\nu}_i)' V_i^{-1} (\underline{y}_i - A_i\underline{\nu}_i) \right\}$$

where $\underline{\nu}_i = (\mu_{j(i)+1}, \mu_{j(i)+2})'$, $V_i = I_{\kappa_i} + \frac{1}{\lambda} B_i B_i'$ and

$$A_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -2 \\ \vdots & \vdots \\ \kappa_i - 2 & 1-\kappa_i \end{bmatrix} \qquad B_i = \begin{bmatrix} 0 & 0 & & 0 \\ 0 & 0 & & \\ 1 & 0 & & \\ 2 & 1 & & \\ \vdots & \vdots & \vdots & \\ \kappa_i - 2 & \kappa_i - 3 & \cdots & 1 \end{bmatrix}.$$

On the other hand, since the discrete spline is inapplicable to $\underline{y}_i$ when $\kappa_i \le 2$, we assume the following model for such $\underline{y}_i$.

$$(3.5)\quad p_N(\underline{y}_i | \underline{\nu}_i, \sigma^2) = \prod_{t=j(i)+1}^{j(i+1)} \frac{1}{\sigma} \cdot \varphi\left(\frac{y_t - \alpha_i}{\sigma} \right)$$

where $\underline{\nu}_i = (\alpha_i)$. Using (3.4) and (3.5), the density of $\underline{y}$ can be written as

$$p(\underline{y} | J, \underline{\theta}, \lambda) = \prod_{i \in I_1 \cup I_2} p_N(\underline{y}_i | \underline{\nu}_i, \sigma^2) \cdot \prod_{i \in I_G} p_D(\underline{y}_i | \underline{\nu}_i, \sigma^2, \lambda)$$

where $\underline{\theta} = (\underline{\nu}_0', \cdots, \underline{\nu}_n', \sigma^2)$, $I_1 = \{i \mid \kappa_i = 1\ 0 \le i \le n\}$, $I_2 = \{i \mid \kappa_i = 2\ 0 \le i \le n\}$ and $I_G = \{i \mid \kappa_i \ge 3\ 0 \le i \le n\}$. In this model, the maximum likelihood estimate of λ is hard to obtain analytically. Consequently, we first assume that λ is fixed.

Lemma 3. 5. The maximum likelihood estimate of $\underline{\theta}$ is given as

$$\hat{\alpha}_i = y_{j(i)+1} \qquad \text{for } i\in I_1$$

$$\hat{\alpha}_i = \frac{y_{j(i)+1}+y_{j(i)+2}}{2} \qquad \text{for } i\in I_2$$

$$\underline{\hat{\nu}}_i = (A_i'V_i^{-1}A_i)^{-1}A_i'V_i^{-1}\underline{y}_i \qquad \text{for } i\in I_G$$

$$\hat{\sigma}^2 = \frac{1}{T}\left\{\sum_{i\in I_2}\sum_{t=j(i)+1}^{j(i+1)}(y_t-\hat{\alpha}_i)^2+\sum_{i\in I_G}(\underline{y}_i-A_i\underline{\hat{\nu}}_i)'V_i^{-1}(\underline{y}_i-A_i\underline{\hat{\nu}}_i)\right\}$$

and the maximum log likelihood is given by

$$\log p(\underline{y}\,|\,J,\hat{\underline{\theta}},\lambda) = -\frac{T}{2}\cdot\log 2\pi\hat{\sigma}^2-\frac{1}{2}\sum_{i\in I_G}\log|V_i|-\frac{T}{2}.$$

Proof. This result is obtained immediately by maximizing the log likelihood

$$\log p(\underline{y}\,|\,J,\underline{\theta},\lambda) = -\frac{T}{2}\cdot\log 2\pi\sigma^2-\frac{1}{2}\sum_{i\in I_G}\log|V_i|$$
$$-\frac{1}{2\sigma^2}\left\{\sum_{i\in I_1\cup I_2}\sum_{t=j(i)+1}^{j(i+1)}(y_t-\alpha_i)^2+\sum_{i\in I_G}(\underline{y}_i-A_i\underline{\nu}_i)'V_i^{-1}(\underline{y}_i-A_i\underline{\nu}_i)\right\}.\blacksquare$$

Lemma 3. 6.

$$\log p(\underline{y}\,|\,J,\hat{\underline{\theta}},\lambda)-E_Z\log p(\underline{Z}\,|\,J,\hat{\underline{\theta}},\lambda)$$
$$=\frac{1}{2\hat{\sigma}^2}\left\{T\sigma^2+\sum_{i\in I_1}(\alpha_i-\hat{\alpha}_i)^2+\sum_{i\in I_2}2(\alpha_i-\hat{\alpha}_i)^2+\sum_{i\in I_G}(\nu_i-\hat{\nu}_i)'A_i'V_i^{-1}A_i(\nu_i-\hat{\nu}_i)\right\}-\frac{T}{2}.$$

Proof. This is a direct consequence of the definition. ■

Theorem 3. 3.

$$E_Y\left[\log p(\underline{Y}\,|\,J,\hat{\underline{\theta}},\lambda)-E_Z\log p(\underline{Z}\,|\,J,\hat{\underline{\theta}},\lambda)\right] = \frac{T(T+\#I_1+\#I_2+2\#I_G)}{2(\#I_2+\sum_{i\in I_G}\kappa_i-2\#I_G-2)}-\frac{T}{2}.$$

Proof. This result follows from Lemma 3.6 and the following properties.

$$\frac{(a_i-\hat{a}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad for\ i\in I_1$$

$$\frac{2(a_i-\hat{a}_i)^2}{\sigma^2} \sim \chi^2_{(1)} \qquad for\ i\in I_2$$

$$\frac{(\nu_i-\hat{\nu}_i)'A_i'V_i^{-1}A_i(\nu_i-\hat{\nu}_i)}{\sigma^2} \sim \chi^2_{(2)} \qquad for\ i\in I_G$$

$$\frac{T\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{\left(\#I_2+\sum_{i\in I_G}\kappa_i-2\#I_G\right)}. \blacksquare$$

From Lemma 3.5 and Theorem 3.3, the conditional predictive log likelihood is obtained as

$$\log p^{pred}(\underline{y}|J,\lambda) = -\frac{T}{2}\cdot\log 2\pi\hat{\sigma}^2-\frac{1}{2}\sum_{i\in I_G}\log|V_i|-\frac{T(T+\#I_1+\#I_2+2\#I_G)}{2(\#I_2+\sum_{i\in I_G}\kappa_i-2\#I_G-2)}.$$

We estimate $p(\underline{y}|J)$ by $\int\exp\{\log p^{pred}(\underline{y}|J,\lambda)\}\omega(\lambda)d\lambda$. The prior $\omega(\lambda)$ we actually assumed is $\omega(\lambda)=1/8$ ($\lambda^{1/2}=1, 2, 4, 8, 16, 32, 64, 128$).

3.5.3. An example of application

In this section, we apply our method by using the discrete spline to the data of opinion polls on the proportion of voters who support the Japan Liberal Democratic Party collected by Chuochosa-sha, a Japanese institute conducting sample surveys every month from December 1978 to November 1982. Figure 3.2 shows the data plotted against time. In this example, since *the full computation* is infeasible, we use the approximation procedure under the following conditions: 1) m is set as $m=4$. 2) α is set as $\alpha=.004n$. 3) The procedure is terminated at $n=9$. The results are given in Tables 3.4 and 3.5.

Table 3.4 presents the estimates of the posterior probabilities of up to nine changes as well as the posterior mean, mode and median of N. These results suggest that plural structural changes underlie the given series.

Table 3.5 presents the estimate of each $p(J_t|\underline{y})$. The largest posterior probability is obtained at the $19th$ observation. Its value is almost equal to 1. This suggests that the $19th$ observation is a change point. Actually, it is widely recognized that the change in the opinion poll between June and July in 1980 was a remarkable one ever in the last two decades. This change is believed to have been caused by the sudden death of Prime Minister Oohira at the beginning of the election campaign that started in June 1980.

The second largest posterior probability is obtained at the $24th$ observation. Its value is not so large as the one at the $19th$ observation, but the $24th$ observation is also likely to be a change point since there are at least three change points according to the values of the mean, mode and median shown in Table 3.4. From Fig. 3.2, it is seen that the observed value largely shifts at the $24th$ observation.

Five other candidates for change points following the above two are at observation points 36, 3, 27, 4 and 28. Figure 3.2 shows that the changes at these points are prominent. The slope of the trend obviously changes at the $36th$ observation and the observed values largely shift at the other four points.

The Bayesian procedure identifies plural change points in the opinion poll data. These change points seem to agree with those views on the shifts of support for the LDP which were expressed by

political observers and shown by analysis. The plot of the observation points also shows that these change points indicate the beginning of a shift in trend in the data.

Additionally, we note that the results obtained when repeating Step 1 seven times are very similar to the results given in Tables 3.4 and 3.5.

Table 3. 1. Number of occurrences of present indicative 3*rd* singular endings *s* and *δ* for different sections of Lindisfarne

section	*s*	*δ*	Total
1	12	9	21
2	26	10	36
3	31	13	44
4	24	6	30
5	28	24	52
6	34	11	45
7	39	9	48
8	46	11	57
9	41	7	48
10	19	3	22
11	17	3	20
12	17	4	21
13	16	4	20

Table 3. 2. Posterior distribution of N and Smith's results.

n	$p(N=n \mid \underline{y})$
0	. 003
1	. 185
2	. 210
3	. 194
4	. 155
5	. 109
6	. 068
7	. 038
8	. 020
9	. 010
10	. 004
11	. 002
12	. 001

Mean	3. 4
Mode	2
Median	3

	Smith's
n	$p(N=n \mid \underline{y})$
0	. 000
1	. 069
2	. 931

Table 3. 3. Posterior probabilities of J_t' s.

t	$p(J_t \mid \underline{y})$
1	. 265
2	. 176
3	. 215
4	. 544
5	. 744
6	. 382
7	. 205
8	. 210
9	. 158
10	. 151
11	. 158
12	. 146

Table 3.4. Posterior probabilities of up to nine changes.

n	$p(N=n\|\underline{y})$
0	.000
1	.001
2	.059
3	.237
4	.192
5	.154
6	.113
7	.098
8	.082
9	.064

Mean	5.0
Mode	3
Median	5

Table 3. 5. Posterior probabilities of J_t' s.

t	$p(J_t \mid \underline{y})$	t	$p(J_t \mid \underline{y})$
1	. 074	25	. 034
2	. 089	26	. 023
3	. 301	27	. 275
4	. 209	28	. 173
5	. 104	29	. 035
6	. 067	30	. 134
7	. 150	31	. 150
8	. 118	32	. 114
9	. 022	33	. 127
10	. 020	34	. 149
11	. 016	35	. 088
12	. 020	36	. 306
13	. 015	37	. 058
14	. 013	38	. 028
15	. 014	39	. 015
16	. 015	40	. 013
17	. 030	41	. 012
18	. 025	42	. 012
19	. 998	43	. 011
20	. 079	44	. 012
21	. 138	45	. 011
22	. 102	46	. 024
23	. 044	47	. 027
24	. 460		

Figure 3. 1. Examples of sequences including a single change.

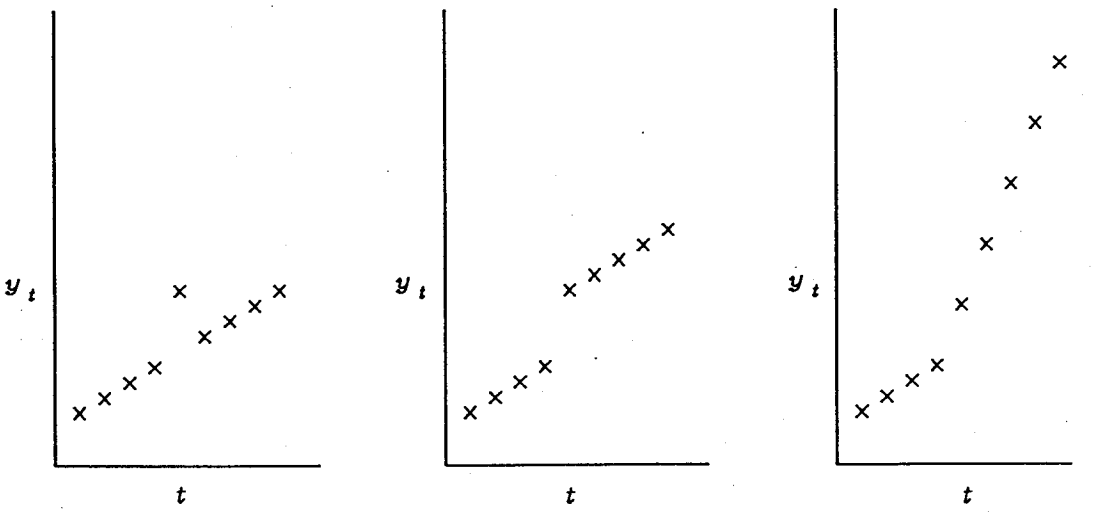
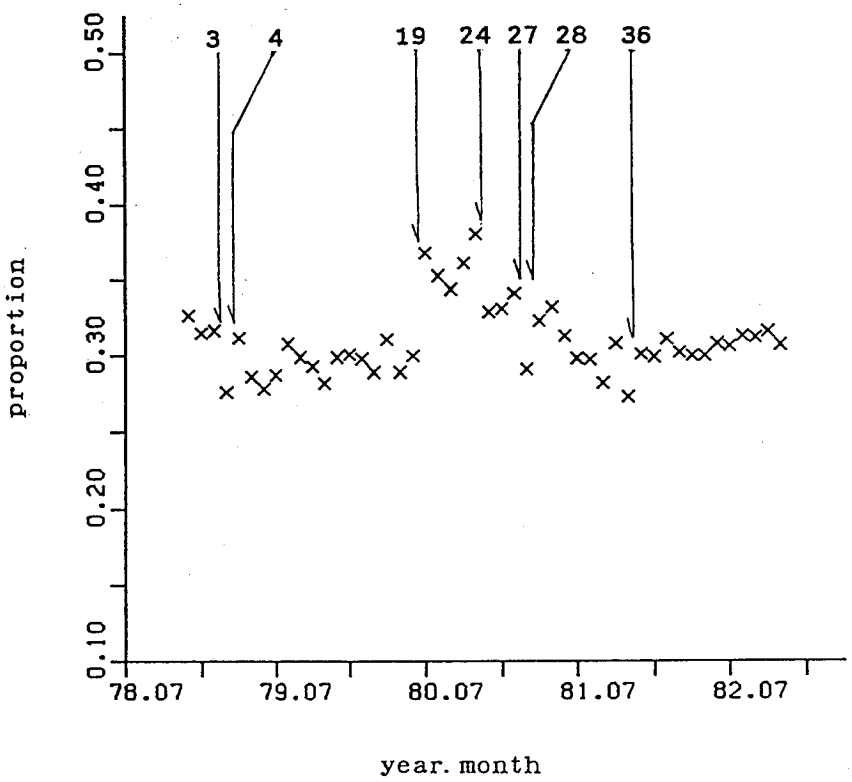


Figure 3. 2. A time-series plot of a public support rate for the Japan Liberal Democratic Party.



ACKNOWLEDGEMENTS

I would like to thank Professor N. Inagaki of Osaka University for his valuable suggestions and warm guidance. I am also indebted to Professors K. Ishii, S. Shirahata and M. Taniguchi of Osaka University, Professor K. Matusita of Teikyo University, Professors T. Yanagimoto and K. Hirano of the Institute of Statistical Mathematics, and Professor S. Aki of Chiba University for their hospitality and encouragement.

REFERENCES

Akaike, H. (1980). Seasonal adjustment by a Bayesian modeling. *Journal of Time Series Analysis* **1**, 1-13.

Broemeling, L. D. and Tsurumi, H. (1987). *Econometrics and structural change*. New York: Marcel Dekker.

Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics* **16**, 188-197.

Freedman, L. S. (1981). Watson's U_N^2 statistic for a discrete distribution. *Biometrika* **68**, 708-711.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society* **B38**, 205-247.

Kashiwagi, N. (1982). Estimation of fertilities in field experiments (in Japanese). *Proceedings of the Institute of Statistical Mathematics* **30**, 73-78.

Kashiwagi, N. and Itani, T. (1986). A Bayesian method for estimating percentiles of a conditional distribution using order statistics and a smoothness prior (in Japanese). *Proceedings of the Institute of Statistical Mathematics* **34**, 213-220.

Kashiwagi, N. (1990). Bayesian detection of structural changes. *To appear in Annals of the Institute of Statistical Mathematics.*

Kashiwagi, N. and Yanagimoto, T. (1990). Smoothing serial counts data through a state-space model. *Submitted to Biometrics.*

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* **82**, 1032-1041.

Kitagawa, G. and Akaike, H. (1982). A quasi Bayesian to outlier

detection. *Annals of the Institute of Statistical Mathematics* **34**, 389-398.

Mosteller, F. (1946). On some useful inefficient statistics. *Annals of Mathematical Statistics* **17**, 377-408.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100-114.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Applied Statistics* **28**, 126-135.

Poirier, D. J. (1976). *The econometrics of structural changes.* North-Holland.

Shiller, R. J. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica* **41**, 775-788.

Silvey, S. D. (1958). The Lindisfarne scribes' problem. *Journal of the Royal Statistical Society* **B20**, 93-101.

Smith, A. F. M. (1980). Change-point problems: approaches and applications. *Trabajos de Estadistica* **31**, 83-98.

West, M., Harrison, P. J. and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association* **80**, 73-83.

Whittaker, E. T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society* **41**, 81-89.

Yanagimoto, T. and Kashiwagi, N. (1990). Empirical Bayes methods for smoothing data and for simultaneous estimation of many parameters. *Environmental Health Perspective* **87**, 109-114.

Zacks, S. (1983). Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation. *Recent Advances in Statistics*

(M. H. Rizvi, J. S. Rustagi and D. O. Siegmund eds.), New York: Academic Press, pp. 245-269.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019-1031.