The Annals of Statistics
1984, Vol. 12, No. 4, 1434–1447

ESTIMATION OF A NOISY DISCRETE-TIME STEP FUNCTION: BAYES AND EMPIRICAL BAYES APPROACHES[1]

BY YI-CHING YAO

Colorado State University

Consider the problem of estimating, in a Bayesian framework and in the presence of additive Gaussian noise, a signal which is a step function. Best linear estimates and Bayes estimates are derived, evaluated and compared. A characterization of the Bayes estimates is presented. This characterization has an intuitive interpretation and also provides a way to compute the Bayes estimates with a number of operations of the order of T^3 where T is the fixed time span. An approximation to the Bayes estimates is proposed which reduces the total number of operations to the order of T. The results are applied to the case where the Bayesian model fails to be satisfied using an empirical Bayes approach.

1. Introduction. We consider the problem of estimating, in a Bayesian framework, a signal which is a step function when one observes the signal plus Gaussian noise. Optimal linear and nonlinear estimates are derived and compared.

This problem is a simplified version of a more general one, applications of which appear in many fields such as seismology, tomography, image processing, econometric modeling, regression analysis and tracking problems. In these problems, the unknown underlying structure is a function, of one or more variables, which is discontinuous or has discontinuous derivatives. It is desired to estimate these nonsmooth functions (signal processes). They can be measured either directly with measurement error or indirectly through various transformations. There are two important and relevant problems: (1) Can one estimate such signals efficiently? (2) Can one detect whether or when a process changes its character?

We shall restrict ourselves to the simple case where the signal processes are flat except for jumps and can be measured directly. In other words, in discrete time denote the signal process by $\mu_1, \mu_2, \cdots, \mu_T$ and let $\mu_{n+1} = \mu_n$ except for occasional changes. Let the observations $X_n = \mu_n + \varepsilon_n$, $n = 1, 2, \cdots, T$ where the ε_n are measurement noise. We shall concentrate on estimating the signal process (i.e. the first problem) and pay little attention to detecting change points.

In this simple case, if the change points were known, we could estimate μ_n by the average of the data points between the two surrounding change points. If jumps are not large, it is hard to tell when jumps take place and take appropriate action. Moreover, if measurement noise has a heavy-tailed distribution, outliers may be disguised as jumps.

Received May 1983; revised May 1984.

[1] This research was supported in part by the Office of Naval Research under contract N00014-75-C-0555 (NR-609-001).

AMS 1980 *subject classifications.* Primary 62M20, 93E14; secondary 62G05, 62C12.

Key words and phrases. Change points, Bayes, empirical Bayes, filtering, smoothing.

In order to develop insight for estimating the signal from the observations, we take a Bayesian point of view and consider a simple model. To be specific, we will characterize the underlying problem through the following special assumptions, which form the discrete time version of a model of Duncan (see Barnard, 1959, page 255).

(1) The sequence of the change points forms a discrete renewal process with identically geometrically distributed interarrival times.

(2) The distinct levels of the signal are mutually independent from a common Gaussian distribution.

(3) The measurement noise is Gaussian white noise.

Barnard (1959) and Chernoff and Zacks (1964) studied a similar model where the number of operations required to compute the Bayes solution is of the order of 2^T. Here T is the fixed time span. In contrast, we will see that in our case the Bayes solution can be computed with a number of operations of the order of T^3.

This paper is organized as follows. In Section 2, the Bayesian model is formulated more precisely. In Section 3, the minimum variance linear estimates of the signal are derived and their average mean squared error is expressed in a closed form. In Section 4, a characterization of the Bayes solution is presented which has an intuitive interpretation. A good approximation to the Bayes solution is also proposed, which is computationally efficient. In Section 5, various estimates are compared under the Bayesian model. In Section 6, we discuss, by use of an empirical Bayes idea, a more general problem where either the Bayesian model has unknown parameters or it fails to be satisfied. Section 7 is a simulation study of the empirical Bayes estimate developed in Section 6.

2. The special Bayesian model (Model A). The three assumptions of the model (to be called Model A) are described more precisely below.

(1) Let $\mathbf{J} = (J_1, J_2, \cdots, J_{T-1})$ be a Bernoulli sequence indicating when *changes* take place; i.e.

$$\begin{aligned} J_n &= 1 \quad \text{if there is a change between } n \text{ and } n+1, \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{2.1}$$

where $\Pr(J_n = 1) = p$, for $1 \leq n \leq T-1$. For convenience, define $J_0 \equiv J_T \equiv 1$.

(2) Let $Y_1, Y_2, \cdots, Y_T$ be i.i.d. $\mathcal{N}(\theta, \sigma^2)$. Define the *signal process* $\{\mu_n\}$ recursively as follows.

$$\mu_1 = Y_1, \quad \mu_{n+1} = (1 - J_n)\mu_n + J_n Y_{n+1}, \quad n = 1, 2, \cdots, T-1. \tag{2.2}$$

This means that when a change takes place between n and $n+1$ (i.e. $J_n = 1$), the signal process shifts to a new level Y_{n+1} from $\mathcal{N}(\theta, \sigma^2)$ (independent of the previous levels) and keeps at the same level until the next change occurs.

(3) Let the *observation process* $\{X_n\}$ be given by

$$X_n = \mu_n + \varepsilon_n, \quad n = 1, 2, \cdots, T \tag{2.3}$$

where the *noise* $\{\varepsilon_n\}$ is i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$. The processes $\{J_n\}$, $\{Y_n\}$ and $\{\varepsilon_n\}$ are mutually independent.

In Sections 3 through 5, the parameters p, θ, σ^2, and σ_ε^2 are assumed known and without loss of generality, θ and σ_ε^2 are set equal to 0 and 1, respectively.

3. The minimum variance linear estimates (MVLE). The minimum variance linear estimates of the signal depend only on the first and second moments of the signal and observation processes. A standard argument shows that $\tilde{\mu}_n$, the MVLE of μ_n, satisfies $\tilde{\mu}_n = \mathbf{e}_n'(\mathbf{I} - \mathbf{M}^{-1})\mathbf{X}$ where $\mathbf{e}_n = (0, 0, \cdots, 1, 0, \cdots, 0)'$ is the nth natural coordinate vector, $\mathbf{I}$ = the $T \times T$ identity matrix, $\mathbf{X} = (X_1, X_2, \cdots, X_T)'$ and the (i, j)-component of $\mathbf{M} = \delta_{ij} + \sigma^2(1 - p)^{|i-j|}$.

Several explicit expressions have been derived in Snyders (1972) for the asymptotic behavior of the minimum mean squared errors as $T \to \infty$ in linear filtering, prediction and interpolation of weakly stationary discrete time processes corrupted by additive noise under very general conditions. In contrast, for finite T, explicit expressions have seldom been found. The following proposition presents a closed-form representation for AMSE$(\tilde{\mu}_n)$, *the average of the mean squared errors of* $\tilde{\mu}_n$. The proof can be found in Yao (1981).

PROPOSITION 3.1.

$$\begin{aligned}\text{AMSE}(\tilde{\mu}_n) &\equiv T^{-1} \textstyle\sum_{n=1}^{T} E(\tilde{\mu}_n - \mu_n)^2 = 1 + T^{-1}\sigma^{-2} f_T'(-\sigma^{-2})/f_T(-\sigma^{-2}) \\ &= 1 + \sigma^{-2} u_+'(-\sigma^{-2})/u_+(-\sigma^{-2}) + o(1), \quad (T \to \infty).\end{aligned}$$

where $\rho = 1 - p$ *and*

$$f_T(\lambda) = a(\lambda)(u_+(\lambda))^T + b(\lambda)(u_-(\lambda))^T,$$

$$u_\pm(\lambda) = [1 - \rho^2 - \lambda(1 + \rho^2) \pm \sqrt{(1 - \rho^2 - \lambda(1 + \rho^2))^2 - 4\rho^2\lambda^2}]/2,$$

$$a(\lambda) = [(1 - \lambda)^2 - \rho^2 - (1 - \lambda)u_-]/(u_+^2 - u_+u_-),$$

$$b(\lambda) = [(1 - \lambda)u_+ - (1 - \lambda)^2 + \rho^2]/(u_+u_- - u_-^2).$$

4. The Bayes solution—the minimum variance nonlinear estimates. The Bayes solution can be computed by brute force with a number of operations of the order of 2^T. In this section, we present a characterization of Bayes estimates which has an intuitive interpretation and also provides a way to compute the solution with $O(T^3)$ operations.

In the following, we consider the conditional distributions of μ_n based on (1) the past and present data, $\mathcal{L}(\mu_n | X_1, \cdots, X_n)$, (2) the future data, $\mathcal{L}(\mu_n | X_{n+1}, \cdots, X_T)$, and (3) all of the data, $\mathcal{L}(\mu_n | X_1, X_2, \cdots, X_T)$. We will

see that $\mathscr{L}(\mu_n \mid X_1, \cdots, X_n)$ and $\mathscr{L}(\mu_n \mid X_{n+1}, \cdots, X_T)$ can be computed recursively and $\mathscr{L}(\mu_n \mid X_1, X_2, \cdots, X_T)$ can be computed by use of $\mathscr{L}(\mu_n \mid X_1, \cdots, X_n)$ and $\mathscr{L}(\mu_n \mid X_{n+1}, \cdots, X_T)$.

Here are convenient notations:

(1) $X_i^j \equiv (X_i, X_{i+1}, \cdots, X_j)$ $(i \le j)$. In particular, $X_1^T = \mathbf{X}$.

(2) $S_0 \equiv 0$, $S_n \equiv \sum_{k=1}^n X_k$ (cumulative sums).

(3) $\mathscr{L}(Y) \equiv$ the distribution of random variable Y.

(4) $f_{\mu_n}(z \mid X_i^j) \equiv$ the conditional probability density of μ_n at z given X_i^j.

(5) $\phi(\cdot) =$ the standard normal density, the density of the noise distribution.

(6) $f_\mu(x) = (1/\sigma)\phi(x/\sigma)$, the density of the prior signal distribution.

(7) "$f(x, z) \propto g(x, z)$ in z" means that there exists $c(x)$ such that $f(x, z) = c(x)g(x, z)$ for all x, z.

4.1 *Expressions for* $\mathscr{L}(\mu_n \mid X_1^n)$ *and* $\mathscr{L}(\mu_n \mid X_{n+1}^T)$.

PROPOSITION 4.1.

$$\mathscr{L}(\mu_n \mid X_1^n) = \sum_{k=1}^n A_k^{(n)} \cdot \mathscr{N}\left[\frac{S_n - S_{n-k}}{k + \sigma^{-2}}, \frac{1}{k + \sigma^{-2}}\right] \tag{4.1}$$

where, for $n = 1, 2, \cdots, T$; $k = 1, \cdots, n$,

$$A_k^{(n)} = \frac{p(1-p)^{k-1}\alpha_{n-k+1}}{\sqrt{1 + k\sigma^2}\alpha_{n+1}} \exp\left[\frac{(S_n - S_{n-k})^2}{2(k + \sigma^{-2})}\right] \tag{4.2}$$

and α_n $(n = 1, 2, \cdots, T + 1)$ *are defined recursively by* $\alpha_1 = 1$, *and*

$$\alpha_{n+1} = \sum_{k=1}^n \alpha_{n-k+1} \frac{p(1-p)^{k-1}}{\sqrt{1 + k\sigma^2}} \exp\left[\frac{(S_n - S_{n-k})^2}{2(k + \sigma^{-2})}\right]. \tag{4.3}$$

PROOF. Applying Bayes' theorem and using the conditional independence of μ_{n+1} and $\{\mu_i; i \le n\}$ given $J_n = 1$, we can easily derive, for $1 \le n \le T - 1$,

$$f_{\mu_{n+1}}(z \mid X_1^{n+1}) \propto \phi(X_{n+1} - z)[(1-p)f_{\mu_n}(z \mid X_1^n) + pf_\mu(z)] \quad \text{in} \quad z. \tag{4.4}$$

This provides a recursive way to compute $\mathscr{L}(\mu_n \mid X_1^n)$ for all n. Since $\mathscr{L}(\mu_1 \mid X_1) = \mathscr{N}(S_1(1 + \sigma^{-2})^{-1}, (1 + \sigma^{-2})^{-1})$, the proposition follows by use of induction and (4.4). □

The number of operations to compute α_n, given α_k, $k < n$, is $O(n)$ according to (4.3). The total number of operations to compute

$$E(\mu_n \mid X_1^n) = \sum_{k=1}^n A_k^{(n)}(S_n - S_{n-k})/(k + \sigma^{-2})$$

for all n is $O(T^2)$. As for Proposition 4.1, we can derive, by use of backward

induction,

$$\mathscr{L}(\mu_n \mid X_{n+1}^T) = \textstyle\sum_{k=0}^{T-n} B_k^{(T-n)} \cdot \mathscr{N}\left[\frac{S_{n+k} - S_n}{k + \sigma^{-2}}, \frac{1}{k + \sigma^{-2}}\right] \tag{4.5}$$

where for $T - n = 0, 1, \cdots, T - 1$; $k = 0, \cdots, T - n$,

$$B_k^{(T-n)} = \frac{p(1-p)^k \beta_{T-n-k}}{\sqrt{1 + k\sigma^2}\beta_{T-n}} \exp\left[\frac{(S_{n+k} - S_n)^2}{2(k + \sigma^{-2})}\right] + (1-p)\delta_{nT} \tag{4.6}$$

and β_{T-n} $(n = T, \cdots, 1)$ are defined recursively by $\beta_0 = 1$, and

$$\beta_{T-n} = \textstyle\sum_{k=0}^{T-n-1} \beta_{T-n-1-k} \frac{p(1-p)^k}{\sqrt{1 + (k+1)\sigma^2}} \exp\left[\frac{(S_{n+k+1} - S_n)^2}{2(k + 1 + \sigma^{-2})}\right] \tag{4.7}$$

4.2 *Expressions for* $\mathscr{L}(\mu_n \mid X_1^T)$ *and* $E(\mu_n \mid X_1^T)$.

PROPOSITION 4.2. *The signal density satisfies*

$$f_{\mu_n}(z \mid X_1^T) \propto f_{\mu_n}(z \mid X_1^n) f_{\mu_n}(z \mid X_{n+1}^T)/f_\mu(z) \quad \textit{in} \quad z. \tag{4.8}$$

REMARK. This states that the "two-sided" conditional density of the signal is proportional to the product of the two "one-sided" conditional densities divided by its prior density. The idea of using forward and backward recursions has been introduced in the engineering literature. See Mayne (1966), Fraser (1967) and Forney (1973).

PROOF. By Bayes' theorem,

$$\begin{aligned} f_{\mu_n}(z \mid X_1^T = x_1^T) &= f_{X_1^T}(x_1^T \mid \mu_n = z) f_\mu(z)/f_{X_1^T}(x_1^T) \\ &\propto f_{X_1^T}(x_1^T \mid \mu_n = z) f_\mu(z) \quad \text{in} \quad z. \end{aligned}$$

From the Markov property of the process $\{\mu_n\}$,

$$\begin{aligned} f_{X_1^T}(x_1^T \mid \mu_n = z) &= f_{X_1^n}(x_1^n \mid \mu_n = z) \cdot f_{X_{n+1}^T}(x_{n+1}^T \mid \mu_n = z) \\ &= f_{\mu_n}(z \mid X_1^n = x_1^n) f_{X_1^n}(x_1^n) f_{\mu_n}(z \mid X_{n+1}^T = x_{n+1}^T) f_{X_{n+1}^T}(x_{n+1}^T)/[f_\mu(z)]^2 \\ &\propto f_{\mu_n}(z \mid X_1^n = x_1^n) f_{\mu_n}(z \mid X_{n+1}^T = x_{n+1}^T)/[f_\mu(z)]^2 \quad \text{in} \quad z. \end{aligned}$$

Applying Bayes' theorem again completes the proof. □

From (4.1), (4.5) and (4.8), we can derive

PROPOSITION 4.3.

$$\mathscr{L}(\mu_n \mid X_1^T) = \textstyle\sum_{1 \le i \le n \le j \le T} c_{ij} \cdot \mathscr{N}\left[\frac{S_j - S_{i-1}}{j - i + 1 + \sigma^{-2}}, \frac{1}{j - i + 1 + \sigma^{-2}}\right] \tag{4.9}$$

where

$$C_{ij} = C'_{ij}/D, \quad D = \sum_{1\le i\le n\le j\le T} C'_{ij},$$

$$C'_{ij} = \alpha_i \beta_{T-j} \frac{(1-p)^{j-i}}{\sqrt{1+(j-i+1)\sigma^2}} \exp\left[\frac{(S_j - S_{i-1})^2}{2(j-i+1+\sigma^{-2})}\right]. \tag{4.10}$$

NOTE. It can be shown that D is independent of n and therefore equal to α_{T+1}/p. Therefore, we have

$$E(\mu_n \mid X_1^T) = \sum_{1\le i\le n\le j\le T} C_{ij}(S_j - S_{i-1})/(j-i+1+\sigma^{-2}). \tag{4.11}$$

REMARK 1. Since for $i \le n \le j$

$$\mathcal{L}(\mu_n \mid X_1^T, J_{i-1} = 1, J_i = J_{i+1} = \cdots = J_{j-1} = 0, J_j = 1) = \mathcal{N}\left[\frac{S_j - S_{i-1}}{j-i+1+\sigma^{-2}}, \frac{1}{j-i+1+\sigma^{-2}}\right],$$

one can see from (4.9) that

$$C_{ij} = \Pr(J_{i-1} = 1, J_i = \cdots = J_{j-1} = 0, J_j = 1 \mid X_1^T).$$

Thus $\{C_{(i+1)j}: 0 \le i < n \le j \le T\}$ represents the conditional distribution of the two change points surrounding time n. So, $\Pr(J_n = 1 \mid X_1^T)$ can be computed by

$$\begin{aligned}\Pr(J_n = 1 \mid X_1^T) &= \sum_{k=0}^{n-1} \Pr(J_k = 1, J_{k+1} = \cdots = J_{n-1} = 0, J_n = 1 \mid X_1^T) \\ &= \sum_{k=0}^{n-1} C_{(k+1)n}.\end{aligned}$$

In particular, $\Pr(\text{No change in } [1, T] \mid X_1^T) = C_{1T}$ can be used to test whether changes have occurred.

REMARK 2.

$$E(\mu_n \mid X_1^T) = \sum_{1\le i\le n\le j\le T} C_{ij}(S_j - S_{i-1})/(j-i+1+\sigma^{-2}) = \sum_{k=1}^T d_k^{(n)} X_k$$

where

$$d_k^{(n)} = \sum_{1\le i\le \min(n,k), \max(n,k)\le j\le T} C_{ij}/(j-i+1+\sigma^{-2}), \quad 1 \le k \le T.$$

So, $0 < d_1^{(n)} < d_2^{(n)} < \cdots < d_{n-1}^{(n)} < d_n^{(n)} > d_{n+1}^{(n)} > \cdots > d_T^{(n)} > 0$, and

$$\sum_{k=1}^T d_k^{(n)} = \sum_{1\le i\le n\le j\le T} \frac{j-i+1}{j-i+1+\sigma^{-2}} C_{ij} < 1.$$

Thus, the Bayes estimate $E(\mu_n \mid X_1^T)$ is a sample dependent weighted average of the observations X_k and the prior mean 0, and the weights $d_k^{(n)}$ attain their maximum at $k = n$ and decrease strictly as k moves away from n on either side.

REMARK 3. The number of operations required to compute α_n, β_{T-n}, C_{ij} $(1 \le n \le T, 1 \le i \le j \le T)$ is $O(T^2)$. The number of operations to compute

$E(\mu_n \mid X_1^T)$ is $O(n(T - n))$. So the total number of operations to compute $E(\mu_n \mid X_1^T)$ for all n is $O(T^3)$.

4.3 *An approximation to the Bayes solution.* Harrison and Stevens (1976) proposed, for the filtering (i.e. one-sided) problem, an approximation technique for computing the posterior distributions of states in multi-process models. Their basic idea is to apply the following step recursively in time. First, the (estimated) posterior distribution of the state at time t is approximated by a normal distribution with the same first two moments. Next, this normal approximation is used together with the observation at $t + 1$ to estimate the posterior distribution of the state at $t + 1$. Applying this idea, we can approximate $\mathcal{L}(\mu_n \mid X_1^n)$ as follows. Suppose that $\mathcal{L}(\mu_n \mid X_1^n)$ approximately equals $\mathcal{N}(\theta_n, \tau_n^2)$. By use of (4.4), $\mathcal{L}(\mu_{n+1} \mid X_1^{n+1})$ approximately equals a mixture of two normal distributions. Then we approximate $\mathcal{L}(\mu_{n+1} \mid X_1^{n+1})$ by $\mathcal{N}(\theta_{n+1}, \tau_{n+1}^2)$ where θ_{n+1} and τ_{n+1}^2 are the mean and variance of this two-normal mixture derived from (4.4). Here θ_{n+1} and τ_{n+1}^2 are functions of θ_n, τ_n^2 and X_{n+1} only. See Yao (1982) for explicit expressions. Applying this step recursively with initial conditions $\theta_1 = X_1(1 + \sigma^{-2})^{-1}$ and $\tau_1^2 = (1 + \sigma^{-2})^{-1}$, we approximate $\mathcal{L}(\mu_n \mid X_1^n)$ by $\mathcal{N}(\theta_n, \tau_n^2)$ for each n. Similarly, we approximate $\mathcal{L}(\mu_n \mid X_n^T)$ by $\mathcal{N}(\omega_n, \delta_n^2)$ where ω_n and δ_n^2 are functions of ω_{n+1}, δ_{n+1}^2 and X_n only. Notice that these normal approximations are exact for $p = 0$ or 1, for Model A is a Gaussian system when $p = 0$ or 1.

Now, we extend Harrison–Stevens approximation to the smoothing (i.e. two-sided) case. We need the following variation of (4.8), the proof of which is similar to that of (4.8). For $2 \le n \le T - 1$,

$$f_{\mu_n}(z \mid X_1^T) \propto \phi(X_n - z)[(1 - p)f_{\mu_{n-1}}(z \mid X_1^{n-1}) + pf_\mu(z)] \times [(1 - p)f_{\mu_{n+1}}(z \mid X_{n+1}^T) + pf_\mu(z)]/f_\mu(z) \quad \text{in} \quad z. \tag{4.12}$$

Since $\mathcal{L}(\mu_{n-1}, X_1^{n-1})$ and $\mathcal{L}(\mu_{n+1} \mid X_{n+1}^T)$ are approximated by $\mathcal{N}(\theta_{n-1}, \tau_{n-1}^2)$ and $\mathcal{N}(\omega_{n+1}, \delta_{n+1}^2)$, $\mathcal{L}(\mu_n \mid X_1^T)$ is naturally approximated by the normalization of the right side of (4.12) with $f_{\mu_{n-1}}(z \mid X_1^{n-1})$ and $f_{\mu_{n+1}}(z \mid X_{n+1}^T)$ being replaced by their normal approximations. This approximation is a mixture of four normal distributions. Denote its mean by $\hat{\mu}_n$, an explicit expression of which in terms of θ_{n-1}, τ_{n-1}^2, ω_{n+1}, δ_{n+1}^2 and X_n is given in Yao (1982). The estimate $\hat{\mu}_n$ may be regarded as an approximation to the Bayes estimate $E(\mu_n \mid X_1^T)$. In fact we will see in the next section that the $\hat{\mu}_n$ are close to the Bayes estimates in the sense of mean squared error. Thus $\hat{\mu}_n$ are nearly optimal. More importantly, the number of operations required to compute $\hat{\mu}_n$ for all n is $O(T)$. Hence, the computational requirements of these approximate Bayes estimates are much less than those of the exact Bayes ones.

5. Comparison among various estimates under Model A. In this section, the performances of the MVLE $\tilde{\mu}_n$, the Bayes estimate $E(\mu_n \mid \mathbf{X})$, and the approximate Bayes estimate $\hat{\mu}_n$, are compared in terms of their average mean squared errors for $T = 20$ for different pairs of p and σ^2.

Since $E(\hat{\mu}_n - \mu_n)^2 = E[E(\mu_n \mid \mathbf{X}) - \mu_n]^2 + E[\hat{\mu}_n - E(\mu_n \mid \mathbf{X})]^2$, we have

$$\text{AMSE}(\hat{\mu}_n) = \text{AMSE}(E(\mu_n \mid \mathbf{X})) + \text{AMSE}(\hat{\mu}_n - E(\mu_n \mid \mathbf{X})) \tag{5.1}$$

where $AMSE(\hat{\mu}_n - E(\mu_n \mid \mathbf{X})) \equiv T^{-1} \sum_{n=1}^{T} E[\hat{\mu}_n - E(\mu_n \mid \mathbf{X})]^2$. The AMSE of $E(\mu_n \mid \mathbf{X})$ and $\hat{\mu}_n - E(\mu_n \mid \mathbf{X})$ are estimated by simulation with 400 replications for each pair of p and σ^2. The AMSE of $\hat{\mu}_n$ is estimated by use of (5.1). The AMSE of $\tilde{\mu}_n$ is calculated from Proposition 3.1. Partial simulation results are presented in Figures 1 and 2 where either p or σ^2 is fixed. In these two figures, the AMSE of $E(\mu_n \mid \mathbf{X}, \mathbf{J})$ is also presented in order to see how much additional information for estimating μ_n is obtained from the knowledge of the change points.

REMARK 1. All these four estimates (including $E(\mu_n \mid \mathbf{X}, \mathbf{J})$) are identical for $p = 0$ or 1, for the model is a Gaussian system in these two cases.

REMARK 2. It can be shown that the AMSE of $\tilde{\mu}_n$ is increasing as p or σ^2 increases. So is the AMSE of $E(\mu_n \mid \mathbf{X})$ as p increases. However, from the simulation results, it appears that as σ^2 increases, $\text{AMSE}(E(\mu_n \mid \mathbf{X}))$ first increases and then decreases and eventually approaches $\text{AMSE}(E(\mu_n \mid \mathbf{X}, \mathbf{J}))$. One explanation is that when σ^2 is large enough, $\mathbf{J}$ can be well estimated from $\mathbf{X}$, and this information can offset the loss of the relatively small amount of prior information about μ_n.

REMARK 3. The linear estimate $\tilde{\mu}_n$ performs poorly when either σ^2 is moderately large or p is away from 0 and 1. It seems that in non-Gaussian systems, linear estimates are rather inflexible and therefore cannot perform well.

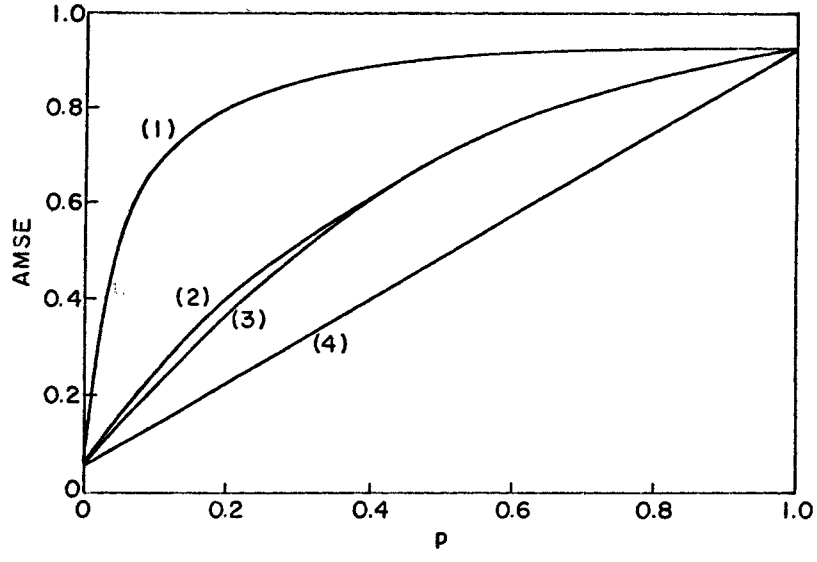

FIG. 1. *AMSE as a function of p.* ($\sigma = 4$). (1) *Best linear estimate*; (2) *Approximate Bayes estimate*; (3) *Bayes estmate*; (4) *Estimate given change points.*

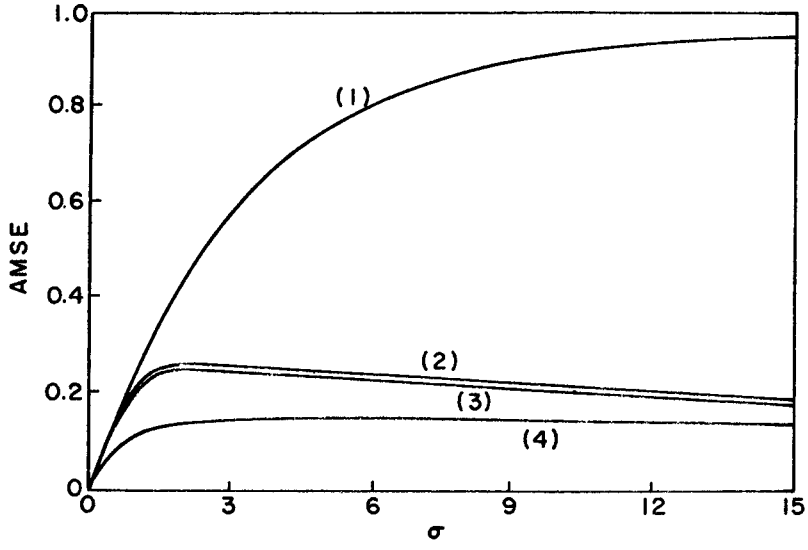

FIG. 2. *AMSE as a function of* σ. ($p = 0.1$).

REMARK 4. In our simulation study, $\mathrm{AMSE}(\hat{\mu}_n - E(\mu_n \mid \mathbf{X})) = \mathrm{AMSE}(\hat{\mu}_n) - \mathrm{AMSE}(E(\mu_n \mid \mathbf{X}))$ is at most about 10% of $\mathrm{AMSE}(E(\mu_n \mid \mathbf{X}))$. This indicates that $\hat{\mu}_n$ is a good approximation to $E(\mu_n \mid \mathbf{X})$. Since the cost of computing $\hat{\mu}_n$ is much less than that of $E(\mu_n \mid \mathbf{X})$, it may be desirable to substitute $\hat{\mu}_n$ for $E(\mu_n \mid \mathbf{X})$.

6. An empirical Bayes estimate using Model A with unknown parameters. In general, a step-function signal can be either deterministic or stochastic and therefore Model A can fail to be satisfied. Why then should we consider this model? The basic idea is that it is hoped the unknown signal would resemble a "typical" realization of the model with properly assigned parameters. Indeed, this is an interpretation of the empirical Bayes approach. The most famous example is the James–Stein estimate which shows uniform superiority to the classical estimate of the mean of a multivariate normal distribution.

It is almost impossible to produce a sensible estimate of the signal without any information about the structure of the signal and/or the noise. Hence, our first assumption is that the noise is Gaussian white noise. One main reason to have the Gaussian assumption is that it is hard to distinguish outliers from jumps if the noise has a heavy tailed distribution. Furthermore, if the step-function signal has many jumps, the noise variance cannot be well estimated. Indeed, the noise variance in Model A is not identifiable without further information. For instance, the observation process $\{X_n\}$ is i.i.d. $\mathcal{N}(0, 1)$ when $(p, \theta, \sigma, \sigma_\epsilon) = (1, 0, 1, 0)$ or $(p, 0, 0, 1)$. So, we make the second assumption that the rate of jump in the signal is at most p_0 where p_0 is a prespecified number between 0 and 1.

As the next step in generalizing our estimation procedure, let us assume that Model A applies with unknown parameters p, θ, σ, σ_ϵ and apply maximum

likelihood to estimate these parameters. Notice that in this section we do *not* assume $\theta = 0$ and $\sigma_\epsilon^2 = 1$ unless otherwise specified. To be more precise, we estimate the signal μ_n as follows. First, fit Model A to the observations X_i $(1 \le i \le T)$ by finding the maximum likelihood estimates (MLE) $\hat{p}$, $\hat{\theta}$, $\hat{\sigma}$ and $\hat{\sigma}_\epsilon$ with the constraint that $p \le p_0$. Next, estimate μ_n by

$$\hat{\mu}_n^{\mathrm{EB}} \equiv E(\mu_n \mid \mathbf{X}) \quad \text{at} \quad (\hat{p}, \hat{\theta}, \hat{\sigma}, \hat{\sigma}_\epsilon) \tag{6.1}$$

where EB stands for empirical Bayes and $E(\cdot)$ at $(\hat{p}, \hat{\theta}, \hat{\sigma}, \hat{\sigma}_\epsilon)$ means expectation according to the probability structure determined by parameter values $p = \hat{p}$, $\theta = \hat{\theta}$, $\sigma = \hat{\sigma}$ and $\sigma_\epsilon = \hat{\sigma}_\epsilon$. Since the MLE satisfy, (for constants $a \neq 0$ and any c) $\hat{p}(a\mathbf{X} + c) = \hat{p}(\mathbf{X})$, $\hat{\theta}(a\mathbf{X} + c) = a\hat{\theta}(\mathbf{X}) + c$, $\hat{\sigma}(a\mathbf{X} + c) = |a| \hat{\sigma}(\mathbf{X})$ and $\hat{\sigma}_\epsilon(a\mathbf{X} + c) = |a| \hat{\sigma}_\epsilon(\mathbf{X})$, and since Model A is time reversible, the empirical Bayes estimate of μ_n, $\hat{\mu}_n^{\mathrm{EB}}$, is translation and scale invariant, and time reversible. That is,

$$\hat{\mu}_n^{\mathrm{EB}}(a\mathbf{X} + c) = a\hat{\mu}_n^{\mathrm{EB}}(\mathbf{X}) + c, \quad \hat{\mu}_n^{\mathrm{EB}}(X_1, \cdots, X_T) = \hat{\mu}_{T-n+1}^{\mathrm{EB}}(X_T, \cdots, X_1). \tag{6.2}$$

The computation of the MLE can be very time-consuming. A naive method may require $O(2^T)$ operations to compute the likelihood for each quadruple $(p, \theta, \sigma, \sigma_\epsilon)$. We present in Proposition 6.1 a representation of the likelihood function which reduces the number of operations to the order of T^2. Since the log likelihood $L(p, \theta, \sigma, \sigma_\epsilon; \mathbf{X})$ satisfies

$$L(p, \theta, \sigma, \sigma_\epsilon; \mathbf{X}) = L(p, 0, \sigma/\sigma_\epsilon, 1; \mathbf{X}') - T \log \sigma_\epsilon \tag{6.3}$$

where $X_n' = (X_n - \theta)/\sigma_\epsilon$, we need only consider $L(p, 0, \sigma, 1; \mathbf{X})$. The following proposition is a simple consequence of Proposition 4.1.

PROPOSITION 6.1.

$$L(p, 0, \sigma, 1; X_1^T = x_1^T) = \log f_{X_1}(x_1) + \textstyle\sum_{n=1}^{T-1} \log f_{X_{n+1}}(x_{n+1} \mid X_1^n = x_1^n)$$

where

$$\mathscr{L}(X_1) = \mathscr{N}(0, \sigma^2 + 1),$$

$$\mathscr{L}(X_{n+1} \mid X_1^n) = (1 - p) \textstyle\sum_{k=1}^{n} A_k^{(n)} \cdot \mathscr{N}\left(\frac{S_n - S_{n-k}}{k + \sigma^{-2}}, \frac{1}{k + \sigma^{-2}} + 1\right) + p\mathscr{N}(0, \sigma^2 + 1) \tag{6.4}$$

and $A_k^{(n)}$ *are defined in Proposition* 4.1.

Even though this proposition suggests a way to compute the exact likelihood with $O(T^2)$ operations, it is still time-consuming to compute the MLE without further reduction in computation. Therefore it is desired to find a more efficient way to approximate the likelihood. We will again make use of the idea of Harrison and Stevens in Section 4.3 to develop an approximation procedure which reduces the number of operations to the order of T.

Assume $\theta = 0$ and $\sigma_\epsilon = 1$ now. First, we approximate $\mathscr{L}(X_{n+1} \mid X_1^n)$ as follows. In Section 4.3, $\mathscr{L}(\mu_n \mid X_1^n)$ is approximated by $\mathscr{N}(\theta_n, \tau_n^2)$ where θ_n and τ_n^2 are

defined recursively. Since

$$\mathscr{L}(X_{n+1} \mid X_1^n) = (1-p)\mathscr{N}(\mu_n \mid X_1^n) * \mathscr{N}(0, 1) + p\mathscr{N}(0, \sigma^2 + 1)$$

where $*$ means convolution of laws, we are naturally led to approximate $\mathscr{L}(X_{n+1} \mid X_1^n)$ by $(1-p)\mathscr{N}(\theta_n, \tau_n^2 + 1) + p\mathscr{N}(0, \sigma^2 + 1)$. Next, we approximate the log likelihood $L(p, 0, \sigma, 1; \mathbf{X})$ by use of Proposition 6.1 and the above approximation and denote this approximate log likelihood by $\tilde{L}(p, 0, \sigma, 1; \mathbf{X})$. By (6.3), define $\hat{L}(p, \theta, \sigma, \sigma_\epsilon; \mathbf{X}) = \tilde{L}(p, 0, \sigma/\sigma_\epsilon, 1; \mathbf{X}') - T \log \sigma_\epsilon$ where $X_n' = (X_n - \theta)/\sigma_\epsilon$. This approximation $\tilde{L}$ is close to L in the sense that the Kullback–Leibler information number between $\exp(L)$ and $\exp(\tilde{L})$,

$$E[L(p, \theta, \sigma, \sigma_\epsilon; \mathbf{X}) - \hat{L}(p, \theta, \sigma, \sigma_\epsilon; \mathbf{X})] \quad \text{at} \quad (p, \theta, \sigma, \sigma_\epsilon)$$

is small. Detailed numerical results on this approximation can be found in Yao (1982).

We shall define the pseudo MLE $\hat{p}'$, $\hat{\theta}'$, $\hat{\sigma}'$, $\hat{\sigma}_\epsilon'$ as the values of the parameters which maximize $\hat{L}$ subject to $p \le p_0$. Then we estimate μ_n by

$$\hat{\mu}_n' \equiv E(\mu_n \mid \mathbf{X}) \quad \text{at} \quad (\hat{p}', \hat{\theta}', \hat{\sigma}', \hat{\sigma}_\epsilon'). \tag{6.5}$$

7. Simulation on empirical Bayes estimators. In this section, we compare, through computer simulation, the performance of $\hat{\mu}_n'$ (an approximation to $\hat{\mu}_n^{\mathrm{EB}}$), and several other estimates. We considered deterministic signal sequences of length $T = 20$. For each signal sequence, we generated 100 samples of Gaussian white noise of variance 1.

In defining $\hat{\mu}_n'$, we estimated the parameters of Model A by use of pseudo maximum likelihood. It is interesting to see how well the method of moments can do compared to the pseudo maximum likelihood method. It is also interesting to see how much the additional information $\sigma_\epsilon = 1$ can contribute to estimating μ_n. Hence, we considered the following four estimators of μ_n.

(i) Estimator 1: $\hat{\mu}_n', p_0 = 0.2$.

(ii) Estimator 2: This is defined in the same way as Estimator 1 except for the additional constraint $\sigma_\epsilon = 1$ in the pseudo maximum likelihood estimation of the parameters.

(iii) Estimator 3: $E(\mu_n \mid \mathbf{X})$ at $(p_1, \theta_1, \sigma_1, \sigma_{\epsilon 1})$ where these four parameter values are estimates of the true ones by the method of moments based on the first two sample moments and sample autocovariances at lag 1 and lag 2. The estimated p_1 is truncated at 0.2.

(iv) Estimator 4: This is defined in the same way as Estimator 3 except for the additional constraint $\sigma_\epsilon = 1$.

We use the average of mean squared errors (AMSE) as the criterion. Partial simulation results are presented in Table 1 where we also present the mean and standard deviation of $\hat{\sigma}_\epsilon'$, the pseudo MLE of σ_ϵ.

TABLE 1
The AMSE of the estimators over 100 Samples[a]

Signal	Successive Levels	Points of Change	ML method		Moment method		Given C.P.[b]	$E(\hat{\sigma}_\epsilon')$[c]	$SD(\hat{\sigma}_\epsilon')$
			Est. 1	Est. 2 (σ_ϵ known)	Est. 3	Est. 4 (σ_ϵ known)			
1	0	none	.071 (.012)	.059 (.008)	.124 (.025)	.067 (.008)	.05	.943	.166
2	0, 1	10	.228 (.012)	.219 (.011)	.352 (.024)	.251 (.009)	.1	.949	.196
3	0, 3	10	.254 (.018)	.241 (.017)	.670 (.058)	.385 (.033)	.1	.930	.180
4	0, 5	10	.187 (.019)	.185 (.019)	.486 (.050)	.189 (.018)	.1	.970	.165
5	0, 2, 4	4, 10	.370 (.017)	.361 (.016)	.802 (.054)	.610 (.040)	.15	.936	.220
6	0, 3, 0	5, 15	.395 (.031)	.380 (.027)	.952 (.078)	.780 (.067)	.15	.913	.222
7	0, 1, 2, 3	4, 10, 16	.348 (.023)	.308 (.017)	.729 (.041)	.493 (.032)	.2	.997	.204
8	0, 1, 0, 1	4, 10, 16	.280 (.007)	.265 (.008)	.365 (.026)	.259 (.008)	.2	1.014	.180
9	0, 1, 3, 4, 6	3, 7, 12, 16	.426 (.018)	.415 (.017)	.907 (.081)	.578 (.039)	.25	1.014	.228
10	0, 3, −3, 6, 0	3, 7, 12, 16	.378 (.022)	.366 (.023)	.984 (.034)	.364 (.022)	.25	.961	.200
11[d]			.935 (.024)	.893 (.022)	1.311 (.065)	2.124 (.033)	1	1.339	.268
12[e]			.854 (.024)	.836 (.023)	1.037 (.045)	1.920 (.035)	1	1.238	.280

[a] The number in parentheses next to an entry is the estimated standard error for that entry.
[b] This column is the AMSE of the estimator using the averages of the data points between successive time points of change.
[c] The estimated standard error of the estimated $E(\hat{\sigma}_\epsilon')$ is $SD(\hat{\sigma}_\epsilon')/10$.
[d] Signal 11 is the following. $\mu_n = n - 1$, $1 \le n \le 11$; $\mu_n = 21 - n$, $12 \le n \le 20$.
[e] Signal 12 is the following. $\mu_n = 10 - 0.1\,(n - 11)^2$, $1 \le n \le 20$.

NOTE. All these estimators have one property in common, namely, they first estimate p, θ, σ, σ_ϵ and then estimate μ_n by the Bayes estimate $E(\mu_n \mid \mathbf{X})$ at the estimated parameter values. In the simulation above, we actually computed the approximate Bayes estimate (see Section 4.3) instead of the exact one.

REMARKS ON TABLE 1:

(1) Roughly speaking, when the number of jumps increases, the AMSE of $\hat{\mu}_n'$ increases. When the size of jumps increases, the AMSE of $\hat{\mu}_n'$ first increases and then decreases. For when the size of jumps is moderate (i.e. compatible with the noise) it is hard to tell where jumps take place and take appropriate action. This property is similar to that of the Bayes estimator. (See Remark 2 of Section 5.)

(2) Estimator 1 ($\hat{\mu}_n'$) is better than Estimator 3. This suggests that the method of pseudo maximum likelihood is better than the method of moments in finding suitable parameter values.

(3) Estimator 1 is just slightly worse than Estimator 2. So the exact information about the noise variance is not very important for estimating the signal unless the rate of jump in the signal is high. In that case, it is hard to estimate σ_ϵ well.

(4) The empirical Bayes estimator, $\hat{\mu}_n'$, is robust against the signals' behavior. However, it is not known how to deal with cases involving non-Gaussian noise where outliers may be easily confused with jumps.

(5) If the rate of jump in the signal exceeds the prespecified number p_0, $\hat{\mu}_n'$ may be misleading, although our limited simulations do not indicate so.

(6) It is interesting that $\hat{\sigma}_\epsilon'$, the pseudo MLE of σ_ϵ, estimates σ_ϵ well with small bias. This is essentially due to the information $p \leq p_0$.

Acknowledgment. This paper was part of the author's dissertation written at MIT under the direction of Professor Herman Chernoff. The author is deeply grateful to him for his guidance and encouragement. He would also like to thank Professor Bill DuMouchel for his suggestion to consider Harrison-Stevens' approximation. Thanks also to the associate editor for his comments.

REFERENCES

BARNARD, G. A. (1959). Control charts and stochastic processes. *J. Roy. Statist. Soc. B* **21** 239–271 (with discussion).

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* **35** 999–1018.

FORNEY, G. D. (1973). The Viterbi algorithm. *IEEE Proc.* **61** 268–278.

FRASER, D. C. (1967). A new technique for the optimal smoothing of data. Sc.D. Dissertation, MIT, Cambridge, MA.

HARRISON, P. J. and STEVENS, C F. (1976). Bayesian Forecasting. *J. Roy. Statist. Soc. B* **38** 205–247 (with discussion).

MAYNE, D. Q. (1966). A solution of the smoothing problem for linear dynamic systems. *Automatica* **4** 73–92.

SNYDERS, J. (1972). Error formulae for optimal linear filtering prediction and interpolation of stationary time series. *Ann. Math. Statist.* **43** 1935–1943.

YAO, Y. C. (1981). The linear theory of estimating means in time series subjected to changes in geometrically distributed time intervals. Tech. Rept. ONR 21, Statist. Center, MIT, Cambridge, MA.

YAO, Y. C. (1982). Estimation in the presence of noise of a signal which is flat except for jumps—Part I, a Bayesian study; Part II, the empirical Bayes approach. Tech. Rept. ONR 25, ONR 27, Statist. Center, MIT, Cambridge, MA.

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523