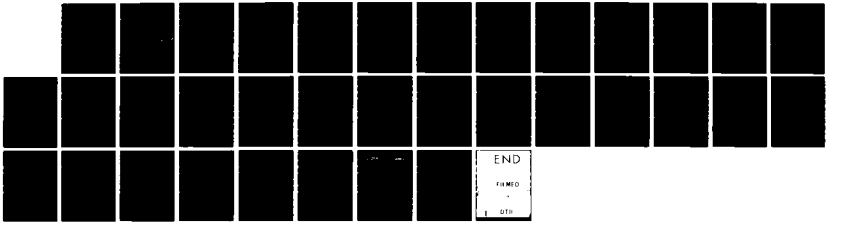
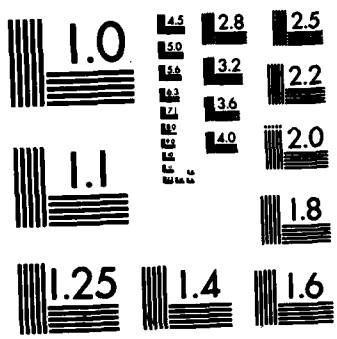
AD-A123 206 TIME-SERIES SEGMENTATION: A MODEL AND A METHOD(U) - 1/1
ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE
METHODS S L SCLOVE 22 DEC 82 TR-N82-7 ARO-19085.2-MA
UNCLASSIFIED DAAG29-82-K-0155 F/G 12/1 NL





TIME-SERIES SEGMENTATION: A MODEL AND A METHOD

by

STANLEY L. SCLOVE

A Presentation to the
Workshop on Applied Time Series Analysis,
sponsored by the Adaptive and Learning Systems Technical Committee of the
IEEE Systems, Man and Cybernetics Society,
held at the Technical University of Munich, Germany, October 22-23, 1982,
in conjunction with the 6th International Conference on Pattern Recognition

To appear in INFORMATION SCIENCES

TECHNICAL REPORT NO. 82-7
December 22, 1982

PREPARED FOR THE
OFFICE OF NAVAL RESEARCH
UNDER
CONTRACT N00014-80-C-0408, TASK NR042-443

Development of Procedures and Algorithms for
Pattern Recognition and Image Processing
based on Two-Dimensional Markov Models

Principal Investigator: Stanley L. Sclove

Also issued as Technical Report No. A82-3 under Army Research Office
Contract DAAG29-82-K-0155, Quantitative Methods Department,
University of Illinois at Chicago

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348, CHICAGO, IL 60680

TIME-SERIES SEGMENTATION: A MODEL AND A METHOD

STANLEY L. SCLOVE

Department of Quantitative Methods, College of Business Administration
University of Illinois at Chicago

CONTENTS

TIME-SERIES SEGMENTATION: A MODEL AND A METHOD

STANLEY L. SCLOVE

Department of Quantitative Methods, College of Business Administration
University of Illinois at Chicago
Box 4348, Chicago, IL 60680

ABSTRACT

The problem of partitioning time-series into segments is treated. The segments are considered as falling into classes. A different probability distribution is associated with each class of segment. Parametric families of distributions are considered, a set of parameter values being associated with each class. With each observation is associated an unobservable label, indicating from which class the observation arose. The label process is modeled as a Markov chain. Segmentation algorithms are obtained by applying a relaxation method to maximize the resulting likelihood function. In this paper special attention is given to the situation in which the observations are conditionally independent, given the labels. A numerical example, segmentation of U.S. Gross National Product, is given. Choice of the number of classes, using statistical model-selection criteria, is illustrated.

Key Words and Phrases: Markov chains; maximum likelihood; maximum a posteriori estimation; Viterbi algorithm; relaxation methods; isodata procedure; model-selection criteria; Akaike's information criterion (AIC).

12/31/82

TIME-SERIES SEGMENTATION: A MODEL AND A METHOD

STANLEY L. SCLOVE
University of Illinois at Chicago

1. Introduction

The problem of segmentation considered here is: Given a time series

$$\{x_t,\ t = 1, \ldots, n\},$$

partition the set of values of t into segments (sub-series, regimes) within which the behavior of x_t is homogeneous. The segments are considered as falling into several classes.

The observation X may be a scalar, vector, or matrix -- any element of a linear space, for which the operations of addition and scalar multiplication are defined. (If X is a scalar, operations such as $x_t - cx_{t-1}$, where c is a scalar, are required. If X is a vector or matrix, the operation $x_t - Cx_{t-1}$, where C is a matrix, is required.)

2. The Model

One can imagine a series which is usually relatively smooth but occasionally rather jumpy as being composed of sub-series which are first-order autoregressive, the autocorrelation coefficient being positive for the smooth segments and negative for the jumpy ones. One might try fitting such data with a segmentation of two classes, one corresponding to a positive autocorrelation, the other, to a negative autocorrelation.

The mechanism generating the process changes from time to time, and these changes manifest themselves at some unknown time points (epochs, change-points). The number, say m, of segments and the epochs are unknown. Generally there will be fewer than m generating mechanisms.

The number of mechanisms (classes) will be denoted by k; it will be assumed that k is at most m. In some situations, k is specified; in others, it is not. Estimation of k will be considered. Although the process changes from time to time, it should be stationary in the mean if it is to be segmented. Otherwise, one would merely be fitting the drifting level of the process, and the larger the series length, the larger the value of k that would be required. Thus, series must be differenced to achieve stationarity before applying segmentation techniques.

With the c-th class is associated a stochastic process, P_c, say. E.g., above we spoke of a situation with $k = 2$ classes, where, for $c = 1,2$, the process P_c is first-order autoregressive with coefficient ϕ_c, where ϕ_1 is positive and ϕ_2 is negative.

Now with the t-th observation ($t = 1,\ldots, n$) associate the label γ_t, which is equal to c if and only if x_t arose from class c, $c = 1,\ldots, k$. Each time-point t gives rise to a pair

$$(x_t, \gamma_t),$$

where x_t is observable and γ_t is not. The process $\{x_t\}$ is the observed time series; the process $\{\gamma_t\}$ will be called the "label process."

Define a segmentation, then, as a partition of the time index set $\{t: t = 1,\ldots, n\}$ into subsets

$$S_1 = \{1,\ldots,t_1\},\ S_2 = \{t_1+1,\ldots,t_2\},\ \ldots,\ S_m = \{t_{m-1}+1,\ldots, n\},$$

where the t's are subscripted in ascending order. Each subset S_g,

$g = 1,\ldots, m$, is a segment. The integer m is not specified. In the context of this model, to segment the series is merely to estimate the γ's.

The focus in the present paper is not on the change-points t_i, $i = 1,\ldots,m$. Rather, the idea underlying the development here is that of transitions between classes. The labels γ_t will be treated as random variables Γ_t with transition probabilities

$$\Pr(\Gamma_t=d|\Gamma_{t-1}=c) = p_{cd},$$

taken as stationary, i.e., independent of t. The k-by-k matrix of transition probabilities will be denoted by $\underline{P}$, i.e.,

$$\underline{P} = [p_{cd}].$$

Restrictions on the process can be imposed by setting the appropriate transition probabilities equal to zero. E.g., some processes are strictly cyclic, such as the operation of an internal-combustion engine, with its cycle of intake to compression to combustion to exhaust to intake, etc. Similarly, one might wish to describe the economy in terms of transitions from recession to recovery to expansion, not allowing transition directly from recession to expansion.

Segmentation will involve the simultaneous estimation of several sets of parameters, the distributional parameters of the within-class stochastic processes, the transition probabilities, and the labels. In order to develop a procedure for maximum likelihood estimation, obviously the likelihood must first be obtained.

To do this, note that a joint probability density function (p.d.f.) for the whole process $\{(X_t,\Gamma_t),\ t = 1,\ldots, n\}$ can be obtained by

successively conditioning each variable on all the preceding ones. The label γ is considered as preceding the corresponding observation X. The variable

X_1 is conditioned on Γ_1;

Γ_2, on X_1 and Γ_1;

X_2 on Γ_2, X_1, and Γ_1;

Γ_3, on X_2, Γ_2, X_1, and Γ_1;

X_3, on Γ_3, X_2, Γ_2, X_1, and Γ_1;

etc. Using f as a generic symbol for any p.d.f., this leads to the joint p.d.f.

$$(2.1)\quad f(\gamma_1)\, f(x_1|\gamma_1) \prod_{t=2}^{n} f(\gamma_t|x_{t-1},\gamma_{t-1},\ldots,\gamma_1)\, f(x_t|\gamma_t,x_{t-1},\gamma_{t-1},\ldots,\gamma_1)$$

The working assumptions of this paper are the following.

(A.1) The label process $\{\gamma_t\}$ is a first-order Markov chain, homogeneous in the sense of having stationary transition probabilities, and conditionally independent of the observations; i.e.,

$$(2.2)\qquad f(\gamma_t|x_t,\gamma_{t-1},\ldots,x_1,\gamma_1) \;=\; f(\gamma_t|\gamma_{t-1}).$$

When $\gamma_{t-1} = c$ and $\gamma_t = d$, then

$$f(\gamma_t|\gamma_{t-1}) = p_{cd},$$

and these transition probabilities do not depend upon t. (The first-order assumption is not critical.)

(A.2) The distribution of the random variable X_t depends only upon its own label and previous X's, not previous labels:

$$(2.3)\qquad f(x_t|\gamma_t,x_{t-1},\gamma_{t-1},\ldots,x_1,\gamma_1) \;=\; f(x_t|\gamma_t,x_{t-1},\ldots,x_1).$$

With these assumptions (2.1) becomes

$$(2.4) \qquad f(\gamma_1) f(x_1|\gamma_1) \prod_{t=2}^{n} p_{\gamma_{t-1}\gamma_t} f(x_t|\gamma_t, x_{t-1}, \ldots, x_1) .$$

Note that this is

$$(2.5) \qquad \left(\prod_{c=1}^{k} \prod_{d=1}^{k} p_{cd}^{n_{cd}} \right) f(\gamma_1) f(x_1|\gamma_1) \prod_{t=2}^{n} f(x_t|\gamma_t, x_{t-1}, \ldots, x_1) ,$$

where the (unobservable) quantity n_{cd} is the number of transitions from class c to class d.

This model, with transition probabilities, has certain advantages over a model based on the change-points. The change-points are discrete parameters, and, even if the corresponding generalized likelihood ratio were asymptotically chi-square, the number of degrees of freedom would not be clear. On the other hand, the transition probabilities vary in an interval and it is clear that they constitute a set of $k(k-1)$ free parameters.

Examples. (i) If each class-conditional process P_c is a first-order Markov process, then

$$(2.6) \qquad f(x_t|\gamma_t, x_{t-1}, \ldots, x_1) = f(x_t|\gamma_t, x_{t-1}) .$$

(ii) If in addition the c-th class-conditional process is Gaussian first-order autoregressive with autoregression coefficient ϕ_c and constant δ_c, with common σ^2, then (2.6) holds with

$$f(x_t|\gamma_t=c, x_{t-1}) = (2\pi\sigma^2)^{-1/2} \exp[-u_{tc}^2/(2\sigma^2)] ,$$

where

$$u_{tc} = x_t - (\phi_c x_{t-1} + \delta_c) .$$

E.g., the value of the likelihood for

$$\gamma_1 = 1 = \gamma_2 = \ldots = \gamma_r \quad \text{and} \quad \gamma_{r+1} = 2 = \gamma_{r+2} = \ldots = \gamma_n$$

is, for given x_0,

$$p_{11}^{m-1} p_{12} p_{22}^{n-m-2} (2\pi\sigma^2)^{-(n-1)/2} \exp[-q/(2\sigma^2)],$$

where

$$q = \sum_{t=1}^{r} [x_t - (\phi_1 x_{t-1} + \delta_1)]^2 + \sum_{t=r+1}^{n} [x_t - (\phi_2 x_{t-1} + \delta_2)]^2 .$$

In regard to (A.2), in the simplest case the X's are (conditionally) independent, given the labels. That is, the distribution of X_t depends only upon its label, and not previous X's. Then

$$f(x_t|\gamma_t, x_{t-1}, \ldots, x_1, \gamma_1) = f(x_t|\gamma_t).$$

We shall pay special attention to this case in the present paper. In this case the p.d.f.'s $f(x|\gamma_t=c)$, $c = 1, \ldots, k$, are called class-conditional densities. In the parametric case the class-conditional density takes the form

$$f(x_t|\gamma_t=c) = g(x_t;\beta_c), \tag{2.7}$$

where β is a parameter indexing a family of p.d.f.'s of form given by the function g and β_c is its value for the c-th class. For example, in the case of Gaussian class-conditional distributions β_c consists of the mean and variance for the c-th class.

3. An Algorithm

3.1. Development of the algorithm

The likelihood L is (2.5), considered as a function of the parameters, for fixed $\{x_t\}$. From (2.5) and (2.7), the likelihood L

can be written in the form

(3.1) $$L = A(\{p_{cd}\}, \{\gamma_t\}) B(\{\gamma_t\}, \{\beta_c\}).$$

Hence, for fixed values of the γ's and β's, L is maximized with respect to the p's by maximizing the factor A. But

$$A = \prod_{c=1}^{k} \prod_{d=1}^{k} p_{cd}^{n_{cd}}.$$

The n_{cd} are determined by the γ's. So from the usual multinomial model, it follows that maximum likelihood estimation of the p's, for fixed values of the other parameters, is given by taking the estimate of p_{cd} to be

(3.2) $$n_{cd}/n_c ,$$

where

$$n_c = n_{c1} + n_{c2} + \ldots + n_{ck} .$$

Further, given the p's and γ's, the estimates of the distributional parameters -- the β's -- are easy to obtain because the observations have been sorted into k groups. This suggests the following algorithm.

Step 0. Set the β's at initial values, perhaps suggested by previous knowledge of the phenomenon under study. Set the p's at initial values, e.g., $1/k$. Set $f(\gamma_1)$ at initial values, e.g., $f(\gamma_1) = 1/k$, for $\gamma_1 = 1, \ldots, k$.

Step 1. Estimate γ_1 by maximizing $f(\gamma_1) f(x_1|\gamma_1)$.

Step 2. For $t = 2, \ldots, n$, estimate γ_t by maximizing the current estimate of

$$p_{\gamma_{t-1}\gamma_t} f(x_t|\gamma_t, x_{t-1}, \ldots, x_1),$$

as the likelihood can be expressed as a product of such factors.

Step 3. Now, having labeled the observations, estimate the distributional parameters, and estimate the transition probabilities according to (3.2).

Step 4. If no observation has changed labels from the previous iteration, stop. Otherwise, repeat the procedure from Step 1.

This method of maximizing with respect to one set of variables, while the others remain fixed, then maximizing with respect to the second set while the first remain fixed, etc., is a relaxation method.

Step 2 is Bayesian classification of x_t. Suppose the (t-1)-st observation had been tentatively classified into class c. Then the prior probability that the t-th observation belongs to class d is p_{cd}, d = 1,...,k. Hence all the techniques for classification of particular models are available (e.g., use of linear discriminant functions when the observations are multivariate normal with common covariance matrix).

Since the labels are treated as random and information equivalent to a prior distribution is put in, one might more properly term this a procedure of maximum a posteriori estimation, rather than maximum likelihood estimation.

Within each iteration Step 2 is the Viterbi algorithm (see [6]), a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite state Markov process. In the present context it obtains the most probable sequence of labels, conditionally upon the results of Steps 0 and 1.

3.2. The first iteration

When the k class-conditional processes consist of independent, identically distributed normally distributed random variables with common variance, one can start by choosing initial means and labelling the observations by a minimum-distance clustering procedure. (This is one iteration of ISODATA [2]; one could iterate further at this stage.) From this clustering initial estimates of transition probabilities and the variance are obtained. This starting procedure could also be used for fitting AR models by taking the initial values of the autoregression coefficients as zero.

3.3. Estimation at the boundary

In Step 1 the label γ_1 is estimated from x_1, without using even the neighboring x_2. Effects of possible error in estimating γ_1 will be mitigated as processing continues on toward $t = n$. In view of this, a way to mitigate further these effects is to "backcast," running every other iteration backwards. (This is possible since Markov chains are reversible.) Another approach would be to run the algorithm k times, once with each possible value of γ_1, and choose the best result. The results reported below, however, were obtained simply using Step 1, as is.

3.4. Restrictions on the transitions

As mentioned above, one might wish to place restrictions on the

transitions, e.g., to allow transitions only to adjacent states. (E.g., "recovery" is adjacent to "recession", "expansion" is adjacent to "recovery," but "expansion" is not adjacent to "recession.") The model does permit restrictions on the transitions. The maximization is conducted, subject to the condition that the corresponding transition probabilities are zero. This is easily implemented in the algorithm. If initially one sets a given transition probability at zero, the algorithm will fit no such transitions, and consequently the corresponding transition probability will remain zero at every iteration.

4. An Example

Here, in the context of a specific numerical example, the problems of (1) fitting the model for a fixed k and (2) choosing k will be discussed.

The data. Quarterly gross national product (GNP) in current (i.e., non-constant) dollars for the years 1946 to 1982 was considered. The data are given in Table 1. They are quarterly, but scaled up to an annual basis. The notation 1946-1 denotes the first quarter of 1946; 1946-2, the second quarter of 1946; etc. The time series will be denoted by

$$y_t, \quad t = 1,2,\ldots,146.$$

Thus, y_1 is GNP for 1946-1, y_2 is GNP for 1946-2, etc. The datum for 1970-3 is 1003.6, or just over 1000. This means that in the third quarter of 1970 the economy was producing goods and services at a

rate of just over one trillion (1000 billion) dollars per year. (Since the readership of this journal is international, it is worth mentioning that in this paper one million means a thousand thousands; a billion is a thousand millions; a trillion is a thousand billions.)

Choice of a transformation. In the context of the linear statistical model

$$y_t = \alpha + \beta_1 x_{1t} + \ldots + \beta_p x_{pt} + u_t,$$

where y is a dependent variable, the x's are explanatory variables, and u is noise, Box and Cox [3] developed a method for choosing a transformation from among the power transformations

$$y^{(\lambda)} = (y^\lambda - 1)/(\lambda y_{g.m.}^{\lambda-1}), \quad \lambda \neq 0,$$
$$= y_{g.m.} \ln(y), \quad \lambda = 0.$$

Here $y_{g.m.}$ denotes the geometric mean of y_t, $t=1,2,\ldots,n$. The value $\lambda = 2$ corresponds to the square, 1 to no transformation, 0.5 to the square root, 0 to the log, and -1 to the reciprocal. One proceeds by fitting linear models

$$y_t^{(\lambda)} = \alpha^{(\lambda)} + \beta_1^{(\lambda)} x_{1t} + \ldots + \beta_p^{(\lambda)} x_{pt} + u_t^{(\lambda)}$$

for various values of λ, say, for example, $\lambda = -1$ to 2 in steps of 0.5. For any fixed value of λ, this is just an ordinary least squares analysis for the data $y_t^{(\lambda)}$, $t = 1,\ldots,n$. An assumption is that, for the true value of λ, the linear model holds with the $u_t^{(\lambda)}$ at least approximately normally distributed with constant standard

deviation $\sigma_u(\lambda)$. Maximum likelihood estimation of λ reduces to comparison of the residual sums of squares $RSS(\lambda)$ for various λ:

$$RSS(\lambda) = \sum_{t=1}^{n} [y_t(\lambda) - \text{pred.val. of } y_t(\lambda)]^2,$$

where

$$\text{pred.val. of } y_t(\lambda) = a(\lambda) + b_1(\lambda)x_{1t} + \ldots + b_p(\lambda)x_{pt},$$

$a(\lambda)$ and $b_j(\lambda)$, $j = 1,\ldots,p$, being the maximum likelihood (least squares) estimates of $\alpha(\lambda)$ and $b_j(\lambda)$, $j = 1,\ldots,p$. A 95% confidence interval for λ [4, pp. 239-240] is

$$\{\lambda: RSS(\lambda) < \min_\lambda RSS(\lambda)[1 + t^2(\nu;.025)/\nu]\}$$

where $t(\nu;.025)$ denotes the upper 97.5 percentage point of Student's t-distribution with ν degrees of freedom, and $\nu = n - (p+1)$, the number of degrees of freedom for error. When ν is large, as is the case with applications to time series, $t(\nu;.025)$ is close to its asymptotic value of 1.96. The choice of 95% is conventional but somewhat arbitrary. For a 90% interval when ν is large one would use $t(\nu;.05)$, which for large ν is approximately 1.645.

This method was applied to time series by means of autoregression, taking the x's in the linear model to be lagged versions of y. Eight lags of y were used. The value 8 was chosen to incorporate the direct effects of lagged variables involved in anticipated regular or seasonal differencing of order one or two and regular or seasonal autoregression of order one or two. For example, a second-order autoregression of

the first differences

$$z_t = y_t - y_{t-1}$$

takes the form

$$z_t = \alpha + \beta_1 z_{t-1} + \beta_2 z_{t-2} + u_t,$$

which is

$$y_t - y_{t-1} = \alpha + \beta_1 (y_{t-1} - y_{t-2}) + \beta_2 (y_{t-2} - y_{t-3}) + u_t,$$

or

$$y_t = \alpha + (\beta_1 + 1) y_{t-1} + (\beta_2 - \beta_1) y_{t-2} - \beta_2 y_{t-3} + u_t,$$

which is a special case of

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + u_t,$$

with

$$\phi_1 = \beta_1 + 1, \quad \phi_2 = \beta_2 - \beta_1, \quad \text{and} \quad \phi_3 = -\beta_2.$$

Due to the use of 8 lags, the value of n for this regression analysis was $146 - 8 = 138$, and $\nu = 138 - (8+1) = 129$. The RSS for y itself was 29,273. This is equal to RSS(1). The RSS for log(y) was 0.00316, so, letting log denote common logs and ln denote natural logs, one has

$$\begin{aligned} RSS(0) &= \text{RSS for } y_{g.m.} \ln(y) \\ &= \text{RSS for } y_{g.m.} \ln(10) \log(y) \\ &= (y_{g.m.} \ln(10))^2 \text{RSS for } \log(y). \\ &= [(711.450)(2.3026)]^2 0.00316 \\ &= 8,480. \end{aligned}$$

A limit for the 95% confidence interval is given by

$$\min_\lambda RSS(\lambda)\,[1 + 1.96^2/129] = 1.0298 \min_\lambda RSS(\lambda).$$

Computations have been done only for $\lambda = 0$ and 1, so the λ yielding $\min_\lambda RSS(\lambda)$ has not been located. However, since the focus here is merely on choosing between $\lambda = 0$ and $\lambda = 1$, one can proceed as follows. One notes that the confidence interval is given by a maximum acceptable value of $RSS(\lambda)$. This maximum acceptable value is less than $1.0298 RSS(0)$, which equals $(1.0298)(8{,}480)$, or 8,732. The confidence interval

$$\{\lambda: \; RSS(\lambda) < 8{,}732.\}$$

based on this limit is conservative, in the sense that it includes more λ-values than may be necessary. Values that are excluded by this interval would also be excluded by the one based on $\min_\lambda RSS(\lambda)$. Note in particular that $\lambda = 0$, corresponding to no transformation, is excluded. The log-transformed data will be used in what follows.

Box-Jenkins analysis. The main focus of this paper is on the segmentation of the time series, but as a preliminary a Box-Jenkins analysis will be presented. Such an analysis aids with the choice of variable (difference, second difference, etc.) for segmentation. "Box-Jenkins analysis" refers to the fitting of data with one or another model chosen from the Box-Jenkins ARIMA models. "ARIMA" means "integrated autoregressive moving average". A fuller notation is ARIMA(p,d,q), where p is the order of the autoregression, d is the order of differencing, and q is the order of the moving average part of the

model. Systematic treatments of Box-Jenkins analysis include (in order of decreasing mathematical level) [5], [11], [10], and [8].

Nelson [11] analyzed quarterly GNP for the twenty years 1947 to 1966. He used an ARI(1,1) model, that is, he fit a first-order autoregression to the first differences. (The notation AR means "autoregressive." The notation ARI means "integrated autoregressive;" i.e., ARI(p,d) means that the d-th differences are AR(p).)

Here the mixed second differences of the logarithms were analyzed. A plot of the first seasonal differences $y_t - y_{t-4}$, corresponding to the annual velocity of the economy, still seemed to trend upward. So did a plot of the regular differences $y_t - y_{t-1}$, which correspond to the quarterly velocity of the economy. Hence second differences were considered. The regular-seasonal mixed second differences

$$(y_t - y_{t-1}) - (y_{t-4} - y_{t-5})$$

appeared stationary. Second differences, corresponding to acceleration, provide a not unnatural way of looking at the data.

The Minitab computing system (see [13]) was used for the analysis. In fact, the transformation, differencing and plotting already referred to were done using Minitab. The arithmetic average of the common logarithms of the data is 2.8522. Their geometric mean is 711.450. A model allowing for first-order regular autogression and first-order seasonal autoregression was fit. In the Box-Jenkins notation for seasonal models,

$$\text{ARIMA}(p,d,q,)\,(P,D,Q)_S,$$

this is

$$\text{ARIMA}(1,1,0)\,(1,1,0)_4.$$

In general, S denotes the seasonality, P, D, and Q the orders of seasonal autoregression, differencing, and moving average.

The value of the estimate of the regular autoregression coefficient was 0.4276; that of the seasonal autoregression coefficient was -0.5332. The value of the estimate of the constant in the model was -0.0001405. The residual sum of squares was 0.00602478.

Check on the constant term. The model was also fit without a constant term in the model. The value of the estimate of the regular autoregression coefficient was 0.4275; that of the seasonal autoregression coefficient was -0.5535. The residual sum of squares was 0.00602508.

Model selection criteria will be used in several ways in this paper. At this point their use will be illustrated with the decision of whether to retain the constant in the model. First some general remarks on model-selection criteria will be made.

Model selection criteria are figures of merit for alternative models. That model which optimizes the criterion is chosen. One such criterion is Akaike's information criterion (AIC). (See, e.g., [1].) Suppose there are K alternative models M_k, $k = 1,\ldots,K$. The model chosen is the one which minimizes $AIC(k)$, where

$$AIC(k) = -2 \ln[\max L(k)] + 2c(k).$$

Here $L(k)$ is the likelihood when M_k is the model, max denotes

its maximum over the parameters, and $c(k)$ is the number of independent parameters when M_k is the model.

The statistic $AIC(k)$ is a natural estimate of the "cross-entropy" (see [12]) between f and $g(k)$, where f is the (unknown) true density and $g(k)$ is the density corresponding to the model M_k. According to AIC, inclusion of an additional parameter is appropriate if $\ln[\max L]$ increases by one unit or more, i.e., if $\max L$ increases by a factor of e or more. Schwarz' model-selection criterion ([14], [7]),

$$-2 \ln[\max L(k)] + \ln(n)c(k),$$

being derived from a first-order approximation to the posterior probability of M_k, enjoys certain advantages. Note that both AIC and Schwarz' criterion are of the form

$$-2 \ln[\max L(k)] + a(n)c(k),$$

where $a(n) = \ln(n)$ for Schwarz' criterion and $a(n) = 2$ for AIC. According to Schwarz' criterion, an additional parameter will be included if it increases $\ln(\max L)$ by an amount greater than $\ln(n)/2$, that is, if $\max L$ increases by a factor of square root of n or more. In particular, for n at least 8, Schwarz' criterion favors models with fewer parameters, relative to AIC.

Note that for Gaussian models

$$-2 \ln(\max L(k)) = n \ln(2\pi) + n \ln(v(k)) + n,$$

where $v(k)$ is the maximum likelihood estimate of the error variance in the model M_k: $v(k) = RSS(k)/n$, where $RSS(k)$ is the

residual sum of squares in fitting the model M_k. In terms of RSS(k), this is

$$-2 \ln(\max L(k)) = n \ln(2\pi) + n \ln(RSS(k)) - n \ln(n) + n.$$

This gives

$$-2 \ln(\max L(k)) + a(n)c(k) = n \ln(2\pi) + n \ln(RSS(k)) - n \ln(n) + n + a(n)c(k).$$

To compare models, it suffices to compute only the portion depending upon k, namely, the statistic

$$n \ln(RSS(k)) + a(n)c(k).$$

To apply model-selection criteria to decide whether to include a constant term in the model, one takes K=2, corresponding to two models, one with the constant (say k = 1) and the other without the constant (k = 2). One has n = 146 - 5 = 141, due to the regular and seasonal differencing. This gives

$$n \ln(RSS(k)) + a(n)c(k) = 141 \ln(RSS(k)) + a(141)c(k).$$

Here AIC will be used; it is favorable to inclusion of more parameters so if AIC rejects the constant, then Schwarz' criterion would also. For AIC, a(n) = 2, so the statistic becomes 141 ln(RSS(k)) + 2c(k). For k = 1 (model with constant term) this is 141 ln(0.00602478) + 2(4), counting the number of parameters as four (regular and seasonal autoregression coefficients, constant, and error variance). This is equal to 141(-5.1118876) + 8 = -712.776. For k = 2 (model without constant term) this is 141 ln(0.00602508) + 2(3), the number of parameters being 3 instead of 4 due to the omission of the constant.

This is 141(-5.1118378) + 6 = -714.769, which is less than the value of -712.776 obtained for the model with the constant. (Note that the difference 714.769 - 712.776 = 1.993 is essentially all due to the difference of 2 associated with the difference in number of parameters. The very slight improvement in residual sum of squares associated with the fitting of the additional parameter is more than offset by the use of an additional parameter.) Hence one concludes that the constant may be excluded. Note that this decision is made without any choice of arbitrary level of significance, such as 5%. (Rational choice of level of significance involves simultaneous consideration of the power of the test, and power computations can be rather involved. In any case, most practitioners seem either unwilling or unable to do them.)

Segmentation analysis. The values of the differences and second differences for 1950 are strikingly higher than those for earlier and later years. On plots these observations appear to be "outliers." They locate very well the mobilization of the economy at the onset of the Korean conflict. The need for segmentation of the time series is apparent. The segmentation analysis will be performed on the mixed regular-seasonal second differences, as these appear to be stationary.

4.1. Fitting the model

In this section the fitting of a model with $k = 3$ classes is

treated, discussion of the choice of k being deferred to the next section. The three classes may be considered as corresponding to Recession, Recovery, and Expansion, although some may prefer to think of the segments labeled as Recovery as level periods corresponding to peaks and troughs. The approximate maximum likelihood solution found by the iterative procedure was (units are billions of current (non-constant) dollars) -0.01125, 0.00184, and 0.01780 for the means, 4.202×10^{-3} for the standard deviation, and

$$\begin{matrix} .4167 & .5556 & .0278 \\ .2151 & .7312 & .0538 \\ .0000 & .5455 & .4545 \end{matrix}$$

for the transition probability matrix.

Remember that the input to the segmentation procedure was the mixed regular-seasonal second difference of the common logs. If the value of this variable equals x, then

$$y_t = 10^x (y_{t-4}/y_{t-5}) y_{t-1}.$$

For example, if $y_{t-4}/y_{t-5} = 1$, this gives $y_t = 1.047 y_{t-1}$ if $x = 0.02$, $y_t = 1.0046 y_{t-1}$ if $x = 0.002$, and $y_t = 0.977 y_{t-1}$ if $x = -0.01$.

The estimated labels are given in Table 2; labels (1, 2, 3, 4, 5) resulting from fitting $k = 5$ classes (discussed below) are also given. The process was in state 1 for 26% of the time, state 2 for 66% of the time, and state 3 for 8% of the time.

The conventional wisdom regarding recessions during the period of time covered by these data is as follows. (See, e.g., [9], pp.

209-211.) In 1948-1949 there was a reduction of inventory investment. In 1953-1954 there was a reduction in government expenditures when the Korean conflict came to a close. In mid-1957 to late 1958 an ongoing recession was aggravated by a drop in defense expenditures in late 1957. In 1960 monetary and fiscal authorities had put on the brakes; interest rates had risen substantially during 1958 and 1959. Readers can probably remember some more recent recessions.

An interesting feature of the model and the algorithm is that, as the iterations proceed, some isolated labels change to conform to their neighbors. This should be the case when p_{cc} is large relative to p_{cd}, $d \neq c$.

4.2. Choice of number of classes

Various values of k were tried, the results being scored by means of Akaike's and Schwarz' model-selection criteria.

The results are given in Table 3. The best segmentation model, as indicated by minimum value of Schwarz' criterion, is that with five classes. (It may be possible to associate these in some way with Recession, Trough, Recovery, Expansion, and Peak.) AIC would choose 7 classes.

5. Extensions

The segmentation procedure has been illustrated here for the univariate case, and with an assumption of common variance. Class-specific variances can be allowed. One can use model-selection

criteria to decide whether or not to use separate class variances. Multiple time series can be treated. Again, one can use model selection criteria to decide whether or not to use separate class covariance matrices. Computer programs to perform these analyses have already been written by the author.

Here we fit only the independent, identically distributed model within segments. An extension will be the fitting of Box-Jenkins models within segments.

Though the segmentation method presented is general, the focus here has been on Gaussian data. There are other important particular cases. In epidemiology, one might wish to segment series for which the observed variable X is a discrete count. In sampling by attribute in industrial quality control X is binary. One might wish to segment the output stream according to classes, "in control," "close to control," "out of control," and estimate the proportion of defectives in these classes.

Acknowledgements. This research was supported by Office of Naval Research Contract N00014-80-C-0408, Task NR042-443, and Army Research Office Contract DAAG29-82-K-0155, at the University of Illinois at Chicago.

References

[1] Akaike, H. (1981). "Likelihood of a Model and Information Criteria." Journal of Econometrics, vol. 16, pp. 1-14.

[2] Ball, G. H., and Hall, D. J. (1967). "A Clustering Technique for Summarizing Multivariate Data." Behavioral Science, vol. 12, pp. 153-155.

[3] Box, G. E. P., and Cox, D. R. (1964). "An Analysis of Transformations." Journal of the Royal Statistical Society, Series B, vol. 26, pp. 211-243.

[4] Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley and Sons, New York.

[5] Box, G. E. P., and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control, rev. ed. Holden-Day, San Francisco.

[6] Forney, G. D., Jr. (1973). "The Viterbi Algorithm." Proceedings of the Institute of Electrical and Electronics Engineers, vol. 61, pp. 268-278.

[7] Kashyap, R. L. (1982). "Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models." Institute of Electrical and Electronics Engineers Transactions on Pattern Analysis and Machine Intelligence, vol. 4, pp. 99-104.

[8] Makridakis, M., and Wheelwright, S. C. (1978). Interactive Forecasting: Univariate and Multivariate Methods, 2nd ed. Holden-Day, San Francisco.

[9] Mansfield, E. (1974). Economics: Principles, Problems, Decisions. W. W. Norton and Company, Inc., New York.

[10] McCleary, R., and Hay, R., Jr. (1980). Applied Time Series Analysis for the Social Sciences. Sage Publications, Beverly Hills, CA.

[11] Nelson, C. R. (1973). Applied Time Series Analysis for Managerial Forecasting. Holden-Day, Inc., San Francisco.

[12] Parzen, E. (1982). "Maximum Entropy Interpretation of Autoregressive Spectral Densities," Statistics and Probability Letters, vol. 1, pp. 7-11.

[13] Ryan, T. A., Jr., Joiner, B. L., and Ryan, B. F. (1980). Minitab Reference Manual. Minitab Project, Statistics Department, Pennsylvania State University, University Park, Pa.

[14] Schwarz, G. (1978). "Estimating the Dimension of a Model." Annals of Statistics, vol. 6, pp. 461-464.

TABLE 1. Quarterly GNP, 1946-1 through 1982-2.
Units: billions of current (non-constant) dollars
(Time Series #534, National Bureau of Economic Research, from BCD: Business Cycle Developments, U.S. Department of Commerce)

1946-1	197.7	1946-2	205.3	1946-3	215.6	1946-4	220.7
1947-1	225.1	1947-2	229.3	1947-3	233.6	1947-4	244.0
1948-1	250.0	1948-2	257.5	1948-3	264.5	1948-4	265.9
1949-1	260.5	1949-2	257.0	1949-3	258.9	1949-4	256.8
1950-1	267.6	1950-2	277.1	1950-3	294.8	1950-4	306.3
1951-1	320.4	1951-2	328.3	1951-3	335.0	1951-4	339.2
1952-1	341.9	1952-2	342.1	1952-3	347.8	1952-4	360.0
1953-1	366.1	1953-2	369.4	1953-3	368.4	1953-4	363.1
1954-1	362.5	1954-2	362.3	1954-3	366.7	1954-4	375.6
1955-1	388.2	1955-2	396.2	1955-3	404.8	1955-4	411.0
1956-1	412.8	1956-2	418.4	1956-3	423.5	1956-4	432.1
1956-1	440.2	1956-2	442.3	1956-3	449.4	1956-4	444.0
1958-1	436.8	1958-2	440.7	1958-3	453.9	1958-4	467.0
1959-1	477.0	1959-2	490.6	1959-3	489.0	1959-4	495.0
1960-1	506.9	1960-2	506.3	1960-3	508.0	1960-4	504.8
1961-1	508.2	1961-2	519.2	1961-3	528.2	1961-4	542.6
1962-1	554.2	1962-2	562.7	1962-3	568.9	1962-4	574.3
1963-1	582.0	1963-2	590.7	1963-3	601.8	1963-4	612.4
1964-1	625.3	1964-2	634.0	1964-3	642.8	1964-4	648.8
1965-1	668.8	1965-2	681.7	1965-3	696.4	1965-4	717.2
1966-1	738.5	1966-2	750.0	1966-3	760.6	1966-4	774.9
1967-1	780.7	1967-2	788.6	1967-3	805.7	1967-4	823.3
1968-1	841.2	1968-2	867.2	1968-3	884.9	1968-4	900.3
1969-1	921.2	1969-2	937.4	1969-3	955.3	1969-4	962.0
1970-1	972.0	1970-2	986.3	1970-3	1003.6	1970-4	1009.0
1971-1	1049.3	1971-2	1068.9	1971-3	1086.6	1971-4	1105.8
1972-1	1142.4	1972-2	1171.7	1972-3	1196.1	1972-4	1233.5
1973-1	1283.5	1973-2	1307.6	1973-3	1337.7	1973-4	1376.7
1974-1	1387.7	1974-2	1423.8	1974-3	1451.6	1974-4	1473.8
1975-1	1479.8	1975-2	1516.7	1975-3	1578.5	1975-4	1621.8
1976-1	1672.0	1976-2	1698.6	1976-3	1729.0	1976-4	1772.5
1977-1	1834.8	1977-2	1895.1	1977-3	1954.4	1977-4	1988.9
1978-1	2031.7	1978-2	2139.5	1978-3	2202.5	1978-4	2281.6
1979-1	2335.5	1979-2	2337.9	1979-3	2454.8	1979-4	2502.9
1980-1	2575.9	1980-2	2573.4	1980-3	2643.7	1980-4	2739.4
1981-1	2864.9	1981-2	2901.8	1981-3	2980.9	1981-4	3003.2
1982-1	2995.5	1982-2	3041.2				

TABLE 2. Estimated labels
(47-2 denotes the second quarter of 1947; 47-3, the third quarter of 1947, etc.)

Quarter:	47-2	47-3	47-4	48-1	48-2	48-3	48-4	49-1	49-2	49-3	49-4	50-1
label,k=3:	1	1	2	2	2	2	1	1	1	1	1	3
label,k=5:	2	1	4	3	3	3	1	1	1	2	2	5

50-2	50-3	50-4	51-1	51-2	51-3	51-4	52-1	52-2	52-3	52-4	53-1	53-2	53-3	53-4	54-1
3	3	3	2	2	1	1	1	1	2	2	2	2	1	1	1
5	5	5	3	2	1	1	1	1	2	4	3	3	2	1	2

54-2	54-3	54-4	55-1	55-2	55-3	55-4	56-1	56-2	56-3	56-4	57-1	57-2	57-3	57-4	58-1
1	2	3	3	2	2	2	1	2	2	2	2	2	2	1	1
2	3	5	5	4	3	2	1	2	2	3	3	2	3	1	1

58-2	58-3	58-4	59-1	59-2	59-3	59-4	60-1	60-2	60-3	60-4	61-1	61-2	61-3	61-4	62-1
2	2	3	3	2	1	1	2	1	2	1	1	2	2	3	2
3	3	5	5	4	1	2	3	1	3	2	2	4	3	4	3

62-2	62-3	62-4	63-1	63-2	63-3	63-4	64-1	64-2	64-3	64-4	65-1	65-2	65-3	65-4	66-1
2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	2	2	3	3	3	4	3

66-2	66-3	66-4	67-1	67-2	67-3	67-4	68-1	68-2	68-3	68-4	69-1	69-2	69-3	69-4	70-1
2	2	2	1	2	2	2	2	2	2	2	2	1	2	2	1
2	2	2	2	2	3	3	3	4	3	2	3	2	3	2	2

70-2	70-3	70-4	71-1	71-2	71-3	71-4	72-1	72-2	72-3	72-4	73-1	73-2	73-3	73-4	74-1
2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	1
2	3	3	4	3	3	3	2	3	3	3	3	2	3	3	1

74-2	74-3	74-4	75-1	75-2	75-3	75-4	76-1	76-2	76-3	76-4	77-1	77-2	77-3	77-4	78-1
2	2	1	2	2	2	2	3	2	1	2	2	2	2	2	1
3	2	2	2	3	4	3	4	2	2	3	3	4	3	2	2

78-2	78-3	78-4	79-1	79-2	79-3	79-4	80-1	80-2	80-3	80-4	81-1	81-2	81-2	81-4	82-1
2	2	2	2	1	2	1	2	1	2	2	2	2	2	1	1
4	3	4	3	1	3	2	3	2	2	4	3	3	3	1	1

82-2
2
3

TABLE 3. Fitting models

Number of classes, k	Akaike's criterion	Schwarz' criterion
2	912.3	927.1
3	825.0	854.5
4	749.8	800.0
5	715.8	792.5*
6	696.4	805.5
7	664.8*	812.3
8	670.9	862.5
9	671.0	912.8

* denotes minimum.

ONR Technical Report List
****

TECHNICAL REPORTS

OFFICE OF NAVAL RESEARCH CONTRACT N00014-80-C-0408, TASK NR042-443
with the University of Illinois at Chicago

Development of Procedures and Algorithms for
Pattern Recognition and Image Processing
based on Two-Dimensional Markov Models

Principal Investigator: Stanley L. Sclove

No. 80-1. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/80

No. 80-2. Stanley L. Sclove. "Modeling the Distribution of Fingerprint Characteristics." 9/19/80

No. 81-1. Stanley L. Sclove. "On Segmentation of Time Series." 11/30/81

No. 82-1. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-Sample Cluster Analysis using Akaike's Information Criterion." 1/30/82

No. 82-2. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-Sample Cluster Analysis with Varying Parameters using Akaike's Information Criterion." 3/8/82

No. 82-3. Stanley L. Sclove. "Some Aspects of Inference for Multivariate Infinitely Divisible Distributions." 6/15/82

No. 82-4. Stanley L. Sclove. "On Segmentation of Time Series and Images in the Signal Detection and Remote Sensing Contexts." 8/1/82

No. 82-5. Stanley L. Sclove. "Application of the Conditional Population-Mixture Model to Image Segmentation." 8/15/82
Revision of Technical Report No. 80-1.

No. 82-6. Hamparsum Bozdogan and Stanley L. Sclove. "Multi-sample Cluster Analysis using Akaike's Information Criterion." 12/20/82

No. 82-7. Stanley L. Sclove. "Time-Series Segmentation: a Model and a Method." 12/22/82

12/29/82

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ DISTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report N82-7	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Time-Series Segmentation: a Model and a Metho		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stanley L. Sclove		8. CONTRACT OR GRANT NUMBER(s) Contract N00014-80-C-0408 (NR042-443)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Quantitative Methods Department University of Illinois at Chicago Box 4348, Chicago, IL 60680		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE December 22, 1982 13. NUMBER OF PAGES iii + 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Statistics & Probability Branch Office of Naval Research Department of the Navy Arlington, VA 22217		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also issued as Technical Report No. A82-3 under Army Research Office Contract DAAG29-82-K-0155, Quantitative Methods Department, University of Illinois at Chicago.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Markov chains; maximum likelihood; maximum a posteriori estimation; Viterbi algorithm; relaxation methods; isodata procedure; model-selection criteria; Akaike's information criterion (AIC)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(PLEASE REFER TO NEXT SHEET.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The problem of partitioning time-series into segments is treated. The segments are considered as falling into classes. A different probability distribution is associated with each class of segment. Parametric families of distributions are considered, a set of parameter values being associated with each class. With each observation is associated an unobservable label, indicating from which class the observation arose. The label process is modeled as a Markov chain. Segmentation algorithms are obtained by applying a relaxation method to maximize the resulting likelihood function. In this paper special attention is given to the situation in which the observations are conditionally independent, given the labels. A numerical example, segmentation of U.S. Gross National Product, is given. Choice of the number of classes, using statistical model-selection criteria, is illustrated.

END

FILMED

2-83

DTIC