RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

NOVEMBER 1995

ROBUST SEASONAL ADJUSTMENT BY BAYESIAN MODELLING

WORKING PAPER #9512-26

MARTIN R. YOUNG
UNIVERSITY OF MICHIGAN

Robust Seasonal Adjustment by Bayesian Modelling

Martin R. Young
University of Michigan
School of Business Administration
Department of Statistics and Management Science
Ann Arbor, MI, 48109
Phone: (313)-936-1332
Fax: (313)-763-5688
Internet: myoung@umich.edu

November, 1995

Robust Seasonal Adjustment by Bayesian Modelling

Abstract

Akaike's BAYSEA approach to seasonal decomposition is designed to capture the respective merits of several preexisting adjustment techniques. BAYSEA is computationally efficient, requires only weak assumptions about the data generating process, and is based on solid inferential (namely, Bayesian) foundations. We present a model similar to that used in BAYSEA, but based on a double exponential rather than a Gaussian error model. The resulting procedure has the advantages of Akaike's method, but in addition is resistant to outliers. The optimal decomposition is obtained rapidly using a sparse linear programming code. Confidence bands and predictive intervals can be obtained using Gibbs Sampling.

Keywords: Gibbs Sampling, L1–Regression, Linear Programming, Outlier

Introduction

In a 1980 paper, Akaike categorized then existing seasonal decomposition methods into three classes: methods based on moving averages (e.g., X-11, Shiskin, Young, and Musgrave, 1976, and, we might now add. SABL. Cleveland et al., 1982 and STL, Cleveland et al., 1990); methods based on multiple regression; and methods based on time series ARIMA modelling (e.g., Box, Hillmer, and Tiao, 1978). Akaike introduced a method, which he termed BAYSEA, for "Bayesian Seasonal Adjustment", which, like the moving average procedures, makes only very weak assumptions about the underlying probability model; which, like the regression techniques, admits computationally efficient estimation; and which, like the direct modelling techniques, allows for the incorporation of likelihood and Bayesian inferential tools as means of model selection and for providing predictive intervals.

While one feature of the modern moving average techniques is their flexibility, another important feature contributing to their popularity is their incorporation of sophisticated outlier detection methods, which produce programs that are resistant to defective observations. The BAYSEA program, on the other hand, is based on a Gaussian model; the associated constrained least squares estimation procedure has the same poor breakdown properties possessed by all least squares estimation procedures.

In this paper, we use a flexible modelling framework similar to that upon which BAYSEA is based, but in which the innovations for the various underlying random processes are assumed to have a double-exponential distribution. The trend/seasonal decomposition is achieved by maximizing the posterior distribution for the unknown parameters; this maximum is shown to be obtainable by solving a sparse linear program.

In Akaike's model, the posterior is Gaussian, so that inferential statements such as predictive intervals can easily be derived. In our model, the joint posterior is complex; however, samples from the joint posterior can be obtained by Gibbs sampling, thus permitting the construction of confidence bands and predictive intervals. Finally, we show that the flexibility of BAYSEA in treating such issues as trading day and leap year effects is retained in the double-exponential version of the model.

The Time Series Model: Likelihood and Prior

The usual additive model for seasonal decomposition states that an observed time series Y_t can be written as

$$Y_t = T_t + S_t + I_t, \quad t = 1, \ldots, n$$

where T_t is a smoothly varying underlying trend component, S_t is a seasonal component with some period, say p, and I_t is the irregular component. Akaike (1980) treated the irregular component I_t as a Gaussian white noise process. We will assume that I_t is a white sequence, with each I_t having the *double exponential* distribution:

$$f_I(i) = \lambda \exp(-\lambda\,|i|),$$

where λ is not necessarily known. The use of the broad-tailed double exponential distribution is a common procedure in robust statistical modelling. In the simple one-sample problem with

$$y_t = \theta + I_t,$$

the MLE for the location parameter θ under a Gaussian model for I_t is the sample mean of the y_t, whereas the MLE for θ under the double exponential model is the median of the y_t, an estimator with a breakdown fraction of 50%. Gonin and Money (1989) describe the use of the double-exponential, or "L_1", model for robust linear and nonlinear regression modelling.

As in Akaike (1980), we do not restrict the sequence T_t to be represented by a simple function, such as a linear or quadratic function. Neither is the sequence S_t required to be representable by a simple sum of harmonics. Thus, each T_t and each S_t, $t = 1, \ldots, n$, is an unknown parameter to be estimated. The log–likelihood for the parameters $\mathbf{T} = (T_1, \ldots, T_n)$, $\mathbf{S} = (S_1, \ldots, S_n)$, and λ is

$$\ell(\mathbf{T}, \mathbf{S}, \lambda | Y) = n \log(\lambda) - \lambda \sum_{t=1}^{n} |Y_t - T_t - S_t| . \tag{1}$$

Clearly, some additional prior knowledge must be added to this specification, as there are $2n + 1$ unknown parameters and only n observations. The additional knowledge will be contributed through *prior distributions* on the parameters T_t and S_t. In the case of T_t, our prior knowledge consists of the fact that the sequence, while perhaps nonlinear, is "smooth". The prior distribution $\pi(T_1, \ldots, T_n)$ is thus formulated to give high prior mass to those functions which are smooth. The log-prior we use, up to a constant, is

$$\log \pi(\mathbf{T}) = -d \sum_{t=2}^{n-1} |T_{t+1} - 2T_t + T_{t-1}| . \tag{2}$$

The log–prior is maximized for sequences T_t which are such that the second differences $T_{t+1} - 2T_t + T_{t-1}$ are zero; namely, linear sequences. The joint prior for the T_t does not belong to any well known functional form. However, by analogy with the quadratic smoothness log-prior used in Akaike (1980), the prior in (2) can be understood as a specification of the believed degree of smoothness of the time series components, with parameter d governing the extent of the smoothness. A value of 0 indicates no preference for smoothness, whereas a value of ∞ states that non-linear sequences have no prior mass (and, hence, no posterior mass given any observed data sequence). As is described in a later section of this paper, one can use Gibbs sampling (Gelfand and Smith, 1990) to simulate realizations of $\mathbf{T}$ from the posterior distribution of $\mathbf{T}$ given $\mathbf{Y}$; similar techniques can be used to simulate from the prior distribution in (2), thus enabling a user of the model to visualize the meaning of a certain choice of prior hyper-parameter d. Alternative priors can be chosen without requiring substantial changes in the estimation algorithm. For example, one might use

$$\log \pi(\mathbf{T}) = -d \sum_{t=2}^{n-2} |T_{t+2} - 3T_{t+1} + 3T_t - T_{t-1}| . \qquad (3)$$

Realizations from this prior will have a higher degree of smoothness than would those from the prior in (2).

In the case of the parameters S_t, a reasonable and non-restrictive specification of typical prior knowledge consists of the facts that (a) the value of a seasonal index S_t should be similar to S_{t-p}, the value of the index at the same point of the previous cycle; and (b) the sum of the seasonal values over a single period should be close to zero. These facts constitute a minimal definition of "additive seasonality". This prior knowledge about the sequence S_t can be encoded in a prior probability distribution as follows:

$$\log \pi(\mathbf{S}) = -r \sum_{t=p+1}^{n} |S_t - S_{t-p}| - z \sum_{j=0}^{P-1} |S_{jp+1} + \ldots + S_{jp+p}| , \tag{4}$$

where $P = [n/p]$ is the number of fully observed periods in the dataset. According to this specification, the most likely seasonal series, a-priori, are those for which the seasonal component is equal from period to period at all times, and in which the sum over each set of p adjacent observations is exactly zero. The hyper-parameters r and z regulate the extent to which deviations from these conditions are permitted, or believed possible.

The optimal point estimate of the components $(\mathbf{T}, \mathbf{S})$ does not depend on the particular prior chosen for the scale parameter λ. However, the system for computing the posterior variances for the parameters $(\mathbf{T}, \mathbf{S})$ described below does require specifying a complete prior distribution, so for that purpose we will use the usual non-informative prior $\pi(\lambda) \propto 1/\lambda$. Finally, to complete the specification of the prior, we assume that $\mathbf{T}$, $\mathbf{S}$ and λ are independent a-priori; thus the joint log-prior is just the sum of the respective marginal log-priors.

The prior distribution (2) is improper; a sequence $(T_1, \ldots, T_n)$ has the same probability as the sequence $(T_1 + c, \ldots, T_n + c)$ for any $c \in (-\infty, +\infty)$. In effect, the prior identifies the shape, but not the level, of the function T_t. However, the data $(Y_1, \ldots, Y_n)$ are informative about the level of the sum of the trend and seasonal functions, and the prior (4) is informative about the level of $(S_1, \ldots, S_n)$; as a result, the joint posterior for $(T_1, \ldots, T_n, S_1, \ldots, S_n)$ is proper. The prior in (2) can be modified so that it is a proper joint density, for example by adding to it terms of the form $d_0 \left|T_1 - \tilde{T}_1\right| + d_0 \left|T_n - \tilde{T}_n\right|$, where d_0 is a small but finite prior precision parameter, and $\tilde{T}_1$ and $\tilde{T}_n$ are prior mean parameters. A data-driven choice for $\tilde{T}_1$ and $\tilde{T}_n$ might be $\tilde{T}_1 = Y_1$, $\tilde{T}_n = Y_n$.

Point Estimation by Linear Programming

The log-posterior distribution for the parameters $(\mathbf{T}, \mathbf{S})$ is given by the sum of the log-likelihood and the log-prior:

$$\begin{aligned}\log \pi(\mathbf{T}, \mathbf{S}|\mathbf{Y}) = C - \lambda \sum_{t=1}^{n} |Y_t - T_t - S_t| - d \sum_{t=2}^{n-1} |T_{t+1} - 2T_t + T_{t-1}| \\ - r \sum_{t=p+1}^{n} |S_t - S_{t-p}| - z \sum_{j=0}^{P-1} |S_{jp+1} + \ldots + S_{jp+p}|,\end{aligned} \tag{5}$$

where C is a constant independent of $\mathbf{T}$ and $\mathbf{S}$.

One reasonable point estimate for the unknown components is the maximizer of the joint posterior – the so-called MAP (maximum a posteriori) estimate. We now show that, as in the case of least absolute deviations (L_1) regression, this maximizer can be obtained using linear programming techniques.

Define δ_t^+ and δ_t^- as the positive and negative parts of the irregular sequence I_t, so that

$$Y_t - S_t - T_t = \delta_t^+ - \delta_t^-, \tag{6}$$

with both δ_t^+ and δ_t^- non-negative. Similarly define quantities $(\epsilon_t^+, \epsilon_t^-)$, (ω_t^+, ω_t^-), and (ν_j^+, ν_j^-) as follows:

$$T_{t+1} - 2T_t + T_{t-1} = \epsilon_t^+ - \epsilon_t^- \tag{7}$$

$$S_t - S_{t-p} = \omega_t^+ - \omega_t^- \tag{8}$$

$$S_{jp+1} + \ldots + S_{jp+p} = \nu_j^+ - \nu_j^-, \tag{9}$$

with

$$\delta_t^+, \delta_t^-, \epsilon_t^+, \epsilon_t^-, \omega_t^+, \omega_t^-, \nu_j^+, \nu_j^- \geq 0, \quad \forall t, \forall j. \tag{10}$$

Then the maximum of the log-posterior can be obtained by solving the minimization problem

$$\min_{T_t, S_t, \delta_t^+, \delta_t^-, \epsilon_t^+, \epsilon_t^-, \omega_t^+, \omega_t^-, \nu_j^+, \nu_j^-}$$

$$\lambda \sum_{t=1}^{n} (\delta_t^+ + \delta_t^-) + d \sum_{t=2}^{n-1} (\epsilon_t^+ + \epsilon_t^-) + r \sum_{t=p+1}^{n} (\omega_t^+ + \omega_t^-) + z \sum_{j=0}^{P-1} (\nu_j^+ + \nu_j^-) \tag{11}$$

Subject to

$$Y_t - S_t - T_t = \delta_t^+ - \delta_t^-, \quad t = 1, \ldots, n$$

$$T_{t+1} - 2T_t + T_{t-1} = \epsilon_t^+ - \epsilon_t^-, \quad t = 2, \ldots, n-1$$

$$S_t - S_{t-p} = \omega_t^+ - \omega_t^-, \quad t = p+1, \ldots, n$$

$$S_{jp+1} + \ldots + S_{jp+p} = \nu_j^+ - \nu_j^-, \quad j = 0, \ldots, P-1,$$

$$\delta_t^+, \delta_t^-, \epsilon_t^+, \epsilon_t^-, \omega_t^+, \omega_t^-, \nu_j^+, \nu_j^- \geq 0, \quad t = 1, \ldots, n, \quad j = 0, \ldots, P.$$

This minimization problem is a linear program. It is fairly easy to see that, at the optimum, at most one of the terms δ_t^+ and δ_t^- will be non-zero, and that the absolute value of this term will necessarily equal that of the corresponding residual $Y_t - T_t - S_t$. Similar results apply to the other absolute deviation terms. This establishes the equivalence between the LP problem and the maximization of the log-posterior, which is a sum of absolute deviations. Furthermore, if one divides the objective function through by the scalar λ, it becomes apparent that one needs to specify only the three ratios $(d/\lambda, r/\lambda, z/\lambda)$, and not the four separate values (λ, d, r, z), in order to find the maximizer of the posterior. Given the MAP decomposition $(\hat{\mathbf{T}}, \hat{\mathbf{S}})$, an estimate of the scale parameter λ can be obtained using the MAP estimate:

$$\hat{\lambda} = \frac{n}{\sum_{t=1}^{n} \left| Y_t - \hat{T}_t - \hat{S}_t \right|}. \tag{12}$$

The linear program has $8n + 2P - 2p - 4$ variables, and $3n + P - p - 2$ equality constraints. The constraint matrix is sparse however, with the number of non-zero entries depending only linearly on n. The maximum can thus be found efficiently using a sparse LP solver. In the examples given below, the optimal decomposition is obtained using Vanderbei's "LOQO" optimizer (Vanderbei and Carpenter, 1990), a code based on an interior point linear programming algorithm. This procedure does not make use of the special structure in this problem, which implies, for example, that at most one of the terms δ_t^+ and δ_t^- are positive. Nevertheless, convergence to the optimum is rapid. Gonin and Money (1989) describe efficient linear programming algorithms for L_1 regression which might be modified to further improve the efficiency of the seasonal decomposition algorithm presented here.

Missing Data and Future Forecasts

Missing data is easily treated with this estimation procedure. For any t for which Y_t is unobserved, simply exclude the terms δ_t^+ and δ_t^- from the objective function, and exclude the constraint

$$Y_t - T_t - S_t = \delta_t^+ - \delta_t^- \tag{13}$$

from the set of constraints. Alternatively, one can include the value Y_t as an unknown to be chosen by the linear programming routine; thus the decomposition routine can serve as a missing data imputation system. This latter approach is the means by which a future value may be forecast: include the values $Y_{n+1}, T_{n+1}, \delta_{n+1}^+, \delta_{n+1}^-, \epsilon_{n+1}^+, \epsilon_{n+1}^-, \omega_{n+1}^+, \omega_{n+1}^-$, as unknowns, add the appropriate terms to the objective function, and add constraints

$$Y_{n+1} - T_{n+1} - S_{t+1} = \delta_{n+1}^+ - \delta_{n+1}^- \tag{14}$$

$$T_{n+1} - 2T_n + T_{n-1} = \epsilon_{n+1}^+ - \epsilon_{n+1}^-, \tag{15}$$

$$S_{n+1} - S_{n+1-p} = \omega_{n+1}^+ - \omega_{n+1}^-. \tag{16}$$

The optimal value of Y_{n+1} will exactly equal $T_{n+1} + S_{n+1}$. T_{n+1} will equal $2T_n - T_{n-1}$; i.e., will represent a linear extrapolation of the last two estimated values of the trend component.

Trading Day, Holiday, and other Time Series Phenomena

The discussion so far presumes the existence of only two time series components, a trend and a seasonal component. The modelling framework, though, easily admits the inclusion of additional components. For example, there can be two seasonal components, daily and monthly: $Y_t = T_t + S_{1t} + S_{2t} + I_t$. Each seasonal component would have a prior distribution as in (4), guaranteeing that the component be appropriately "periodic". One can also add indicator variables to the linear model corresponding to such effects as trading day or holiday. If an intervention to a process occurred at some time j, then one can add an indicator variable

$$E_t = \begin{cases} 1 & t \geq j \\ 0 & t < j \end{cases} \tag{17}$$

to the model, obtaining

$$Y_t = T_t + S_t + \beta E_t + I_t; \tag{18}$$

the parameter β, which quantifies the impact of the intervention, can be estimated using the linear programming algorithm. In general, any time series effect that can be modelled in a linear regression framework can be incorporated in this L_1 version of

BAYSEA.

The linear programming algorithm also makes it easy to include structural constraints on model parameters. For example, if one believes that the coefficient β in (18) must be positive, one can simply add the constraint $\beta > 0$ to the set of linear programming constraints. If one has reason to believe that the trend must be monotone increasing, or concave, one can add constraints of the form $T_{t+1} - T_t > 0$ or $T_{t+1} - 2T_t + T_{t-1} < 0$ to the constraint set. The monotonicity constraint, when appropriate, can provide a useful guarantee that the estimate of trend is smooth.

Choice of Smoothing Parameters

The estimates of the time series components depend on the choice of the parameters (d, r, z). Akaike (1980) suggests applying the BAYSEA procedure for a finite set of values for the smoothing parameters, and then choosing the model which minimizes the "ABIC" criterion, which is defined as twice the negative log-posterior. A similar approach can be applied in the present context, with the log-posterior given by equation (5). An alternative approach to automatically selecting the smoothing parameters is the cross-validation technique (e.g., Stone, 1974, or Geisser and Eddy, 1979). In such a scheme, a subset of the data $\{Y_t | t \in A\}$ is withheld from the estimation procedure, and the model is fit based on the remaining data $\{Y_t | t \notin A\}$, using a number of different choices for the smoothing parameters. The predicted values $\{\hat{Y}_t | t \in A\}$ are then compared to the actual values $\{Y_t | t \in A\}$, where $\hat{Y}_t = T_t + S_t$, and the smoothing parameters which lead to the closest fit between Y_t and $\hat{Y}_t$ are chosen as the final values. The earlier discussion on missing values provides the means by which such withholding of data can be easily implemented. Ansley and Kohn (1987) and Kohn, Ansley, and Tharm (1991) discuss the use of cross–validation in time–series smoothing. The discussion below on the Gibbs sampling implementation of this model offers a further alternative approach to choosing the smoothing parameters.

It may be noted that the sensitivity of the adjustment procedure to the choice of smoothing parameters can be reduced by the addition of constraints on the components, such as a monotonicity restriction on $\mathbf{T}$.

Examples

Exhibit 1 displays the logarithm of quarterly sales figures for General Motors Corporation, over a period from 1962 to 1991. There is clearly a seasonal variation; there is also at least one observation that seems to outly the typical pattern. The trend component, obtained using the parameter values $d = 10.0, r = 10.0, z = 10.0$, is superimposed. The estimate seems reasonable, capturing such details as the hump occurring near 1978. Exhibit 2 shows the estimated seasonal component, the amplitude of which is seen to vary over time. An examination of the original data in Exhibit 1 appears to support this inference. This observation would seem to indicate that the logarithm is not exactly the correct transformation for the raw data. The example shows that the L_1 BAYSEA model can capture phenomena such as an interaction between the trend and seasonal components.

[Insert Exhibits 1- 2 Here]

The MAP estimation procedure was applied to a simulated data set featuring two gross outliers, one each at the beginning and end of the series. Exhibit 3 shows the raw data, together with the estimate of the trend component, obtained using smoothing parameters $d = r = z = 1$. The estimate appears to be unaffected by the outliers. Note, though, the unexpected dip in the trend at the end of the series; this would be undesirable in a forecasting context. Exhibit 4 was obtained using smoothing parameters $d = 10$, $r = 1$, $z = 1$; this appears quite satisfactory. Alternatively, the downturn in the trend estimate could be prevented by constraining the trend component to be monotone.

[Insert Exhibits 3- 4 Here]

Estimation by Gibbs Sampling

The linear programming procedure of the previous section locates the *mode* of the joint posterior $\pi(\mathbf{T},\mathbf{S}|\mathbf{Y})$. One might also wish to identify other features of the posterior, in particular marginal posterior distributions for the parameters of interest. The joint posterior is not of a standard form, and it is not feasible to directly compute marginal quantities by integration, nor is it convenient to sample from the complex joint posterior. However, it is the case that the *conditional* distributions of the joint posterior are fairly simple, and it is feasible to sample from these conditionals. One can use these simple conditionals to generate samples from the joint distribution, using the Gibbs sampling procedure (Gelfand and Smith, 1990). The Gibbs sampling procedure is an iterative technique for sampling from some complicated multivariate distribution. Suppose $f(x_1, \ldots, x_n)$ is a joint distribution, with associated conditionals $f(x_1|x_2, x_3, \ldots, x_n)$, $f(x_2|x_1, x_3, \ldots, x_n), \ldots,$ $f(x_n|x_1, x_2, \ldots, x_{n-1})$. The

Gibbs sampling procedure starts with an arbitrary initial vector $(x_1,\dots,x_n)$. A new realization of x_1 is then generated from the conditional $f(x_1|x_2,x_3,\dots,x_n)$, a realization of x_2 from the conditional $f(x_2|x_1,x_3,\dots,x_n)$, and so on, with the cycle repeated several – typically hundreds or thousands – of times. The stationary distribution of the vectors $(x_1,\dots,x_n)$ generated by this Markovian procedure is equal to the joint distribution $f(x_1,\dots,x_n)$. Because the process is ergodic (Gelfand and Smith, 1990), the marginal variance of any variate x_k can be estimated by the sample variance of x_k over the realized stream of numbers generated by the Gibbs Sampler.

A convenient notation introduced by Gelfand and Smith (1990) is to refer to an unconditional distribution of a variate X as $[X]$, and to the conditional distribution of X given Y as $[X \mid Y]$. Also, we will refer to the set $\{T_t, \quad t \neq j\}$ as $\mathbf{T}^{(j)}$, and the set $\{S_t, \quad t \neq j\}$ as $\mathbf{S}^{(j)}$. In the following, the scale parameter λ will be treated as an unknown parameter to be estimated; the smoothing parameters (d,r,z) will be assumed to be fixed.

By examining the terms in equation (5), it can be seen that the conditional posterior distribution of $\left[T_t \mid \mathbf{Y}, \mathbf{T}^{(t)}, \mathbf{S}, \lambda\right]$, for $t = 3,\dots,n-2$, depends only on the quantities $Y_t, S_t, T_{t-2}, T_{t-1}, T_{t+1}, T_{t+2}, \lambda$. In particular the conditional distribution is given by

$$\left[T_t \mid \mathbf{Y}, \mathbf{T}^{(t)}, \mathbf{S}, \lambda\right] \propto \exp(-\lambda\,|Y_t - T_t - S_t| - d\,|T_{t+2} - 2T_{t+1} + T_t| \\ - d\,|T_{t+1} - 2T_t + T_{t-1}| - d\,|T_t - 2T_{t-1} + T_{t-2}|) \qquad (19)$$

This conditional distribution is piecewise exponential; i.e., the log-conditional is piecewise linear. The "knots" in the piecewise linear function occur at the abscissae

$\{Y_t - S_t,\ 2T_{t+1} - T_{t+2},\ (T_{t+1} + T_{t-1})/2,\ 2T_{t-1} - T_{t-2}\}$. The cumulative distribution, say $F(T)$, associated with this conditional density is just that associated with a piecewise exponential; namely $F(T) = a_0 - a_1 \exp(-a_2 T)$, where the constants (a_0, a_1, a_2) depend on where T lies relative to the set of knots. This CDF is easily inverted; thus one can generate samples from the conditional density in (19) using the inverse transform method (Law and Kelton, 1982). Simple modifications are required to treat the endpoints T_1, T_2, T_{n-1}, and T_n. For example, the conditional distribution of $\left[T_1 \mid \mathbf{Y}, \mathbf{T}^{(1)}, \mathbf{S}, \lambda\right]$ is just

$$\left[T_1 \mid \mathbf{Y}, \mathbf{T}^{(1)}, \mathbf{S}, \lambda\right] \propto \exp\left(-\lambda \left|Y_1 - T_1 - S_1\right| - d\left|T_3 - 2T_2 + T_1\right|\right).$$

Also, the conditional distribution of the future value T_{n+1} is given by

$$\left[T_{n+1} \mid \mathbf{Y}, \mathbf{T}^{(n+1)}, \mathbf{S}, \lambda\right] \propto \exp\left(-d\left|T_{n+1} - 2T_n + T_{n-1}\right|\right), \tag{20}$$

which is the usual double-exponential distribution with location parameter $2T_n - T_{n-1}$ and scale parameter d.

If one's prior knowledge suggests that the underlying trend component should be monotone, then one can restrict the distribution $\left[T_t \mid \mathbf{Y}, \mathbf{T}^{(t)}, \mathbf{S}, \lambda\right]$ to the domain (T_{t-1}, T_{t+1}). The restricted density is also piecewise exponential, and so one can use the inverse transform method to generate a T_t within the interval (T_{t-1}, T_{t+1}); each sample of $\mathbf{T}$ generated in this fashion will necessarily be monotone. A similar approach can be used to guarantee that the realizations of $\mathbf{T}$ be concave or convex.

The conditional distribution of $\left[S_t \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}^{(t)}, \lambda\right]$, for $t = p+1, \ldots, n-p$, is given by

$$\begin{aligned}\left[S_t \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}^{(t)}, \lambda\right] \propto \exp(& -\lambda\,|Y_t - T_t - S_t| - r\,|S_t - S_{t-p}| \\ & - r\,|S_{t+p} - S_t| - z\,|S_{jp+1} + \ldots + S_{jp+p}|),\end{aligned} \tag{21}$$

where $j = [t/p]$. This is again a piecewise exponential density, with four knots at the abscissae $\{(Y_t - T_t),\ S_{t-p},\ S_{t+p},\ (S_{jp+1} + \ldots + S_{t-1} + S_{t+1} + \ldots + S_{jp+p})\}$. Thus samples from the conditional distribution of S_t can be generated using the inverse-transform technique. Obvious modifications are made for the endpoints S_t, $t = 1, \ldots, p$ and $t = n-p+1, \ldots, n$. For example, the conditional distribution of $\left[S_1 \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}^{(1)}, \lambda\right]$ is given by

$$\left[S_1 \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}^{(1)}, \lambda\right] \propto \exp\left(-\lambda\,|Y_1 - T_1 - S_1| - r\,|S_{1+p} - S_1| - z\,|S_1 + \ldots + S_p|\right). \tag{22}$$

The conditional distribution of the future value S_{n+1} is given by

$$\left[S_{n+1} \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}^{(n+1)}, \lambda\right] \propto \exp\left(-r\,|S_{n+1} - S_{n+1-p}|\right), \tag{23}$$

which is the double–exponential distribution with location parameter S_{n+1-p} and scale parameter r.

The parameter λ plays the role of a scale parameter, and can be estimated using the usual Bayesian techniques for scale parameters. Given the prior $[\lambda] \propto 1/\lambda$, the conditional posterior is

$$[\lambda \mid \mathbf{Y}, \mathbf{T}, \mathbf{S}] \propto \lambda^{n-1} \exp\left(-\lambda \sum_{t=1}^{n} |Y_t - T_t - S_t|\right). \tag{24}$$

This is just the Gamma distribution, with parameters $(n, \sum_{t=1}^{n} |Y_t - T_t - S_t|)$. Law and Kelton (1982) describe schemes for generating deviates from the Gamma distribution.

If one is interested in making inference on a future observable value Y_{n+1}, the conditional distribution is given by

$$[Y_{n+1} \mid \mathbf{Y}, \mathbf{T}, T_{n+1}, \mathbf{S}, S_{n+1}, \lambda] \propto \exp\left(-\lambda\, |Y_{t+1} - T_{t+1} - S_{t+1}|\right), \tag{25}$$

which is the double exponential distribution with location parameter $T_{t+1} + S_{t+1}$, and scale parameter λ. T_{t+1} and S_{t+1} are themselves generated using equations (20) and (23).

Estimation of Smoothing Parameters

Lenk (1993) has noted, in the context of nonparametric density estimation, that the Gibbs sampling approach can be used to automatically generate appropriate smoothing parameters, based on observed smoothness attributes of the data. The smoothing parameters (d, r, z) appear in the likelihood in the form of scale parameters, and so can also be estimated in the same way as is λ; namely, by generating from their conditional posteriors, which will be of the Gamma form. The use of a Gamma prior $[d] \propto d^{\alpha_0 - 1} \exp(-\alpha_1 d)$ leads to a Gamma posterior for d; in particular, the conditional posterior distribution for d will be Gamma with parameters $(\alpha_0+n-2, \alpha_1+\sum_{t=2}^{n-1}|T_{t+1}-2T_t+T_{t-1}|)$. Similarly, the conditional posterior for r will be Gamma with parameters $(\alpha_0+n-p, \alpha_1+\sum_{t=p+1}^{n}|S_t-S_{t-p}|)$, and the conditional posterior for z will be Gamma with parameters $(\alpha_0+P, \alpha_1+\sum_{j=0}^{P-1}|S_{jp+1}+\ldots+S_{jp+p}|)$. The hyper-parameters α_0 and α_1 will generally be chosen to be small positive numbers, to reflect prior ignorance about these parameters. If $\alpha_0 \ll n$ and $\alpha_1 \ll \sum_{t=2}^{n-1}|T_{t+1}-2T_t+T_{t-1}|$, then the prior will have a small effect on the posterior, and the choice of the smoothing parameter will be essentially data-driven.

Implementation Issues in Gibbs Sampling

The work of Carlin et al. (1992) suggests a quite different approach to implementing the double exponential model via the Gibbs Sampler. Namely, one can model the error term I_t as coming from a Gaussian distribution, with 0 mean and variance $\lambda\omega_t$, where ω_t is an exponential variate with fixed mean. This scale mixture of normals implies a double exponential model (Andrews and Mallows, 1974). A Gibbs sampling procedure can alternate between generating the desired parameters $\mathbf{T}, \mathbf{S}$ for given values of $\mathbf{\Omega} = (\omega_t, \quad t=1,\ldots,n)$, and generating values of $\mathbf{\Omega}$ for fixed $\mathbf{T}, \mathbf{S}$. A possible advantage of this approach is that in this case the posterior distribution of $[\mathbf{T}, \mathbf{S} \mid \mathbf{\Omega}, \lambda]$ will be multivariate normal; thus one could use a multivariate generator to obtain all the variates $(\mathbf{T}, \mathbf{S})$ at once. However, one would probably wish to adapt the multivariate normal generator to make use of the sparseness of the conditional posterior covariance matrix.

In addition to the particular issues that arise in Bayesian modelling of seasonal time series, there are general implementation issues involved in any application of Gibbs sampling: diagnosis of convergence, choice of the order of sampling, and use of sequential versus parallel streams of random numbers. Gelfand et al. (1990) and Tanner (1991) provide guidance with respect to many of these questions. Roberts (1992) and Zellner and Min (1995) describe diagnostic measures for assessing convergence of the Markov chain.

Example

The Gibbs sampling procedure with direct modelling of the double exponential errors was applied to the simulated data set featuring two outliers. The smoothing parameters were themselves estimated, using the Gibbs sampling scheme described above. 20,000 Gibbs iterations were performed, and the first 5000 samples were discarded to avoid startup transient effects. The procedure was initialized by setting the values for (T_t, S_t), $t = 1, \dots, n$ at the respective MAP estimates obtained via linear programming, and λ at the MAP estimate given in equation (12). The smoothing parameters d, r and z were initially set at 10, and non-informative priors on d, r, and z were obtained by setting the values of the hyper-parameters (α_0, α_1) at $\alpha_0 = 1$, $\alpha_1 = .01$.

Exhibit 5 displays the data, the posterior mean estimate of the trend component, and the posterior mean plus and minus one posterior standard deviation for the trend component. The pointwise standard deviation for each T_t was obtained by using the sample path standard deviation of the T_t's generated by the Gibbs Sampler. Posterior standard errors can also be obtained from the Gibbs Sampler for the seasonal factors

S_t, and for predictions of Y_{n+1}. The data were informative about the smoothing parameters; for example, the 95% highest posterior density interval for $\log(d)$ was 3.1 ± 0.36.

[Insert Exhibit 5 Here]

Conclusion

The Bayesian paradigm allows for flexible modelling of seasonal time series phenomena, in a fashion which is insensitive to outlying data values. Important contributions to the Bayesian analysis of time series with outliers include West (1981), West and Harrison (1986), Tsay (1986), Kitagawa (1987), Meinhold and Singpurwalla (1989), Carlin et al. (1992), and McCulloch and Tsay (1994). In this paper, a seasonal decomposition method is presented which requires only weak assumptions about the nature of the underlying trend and seasonal components; robust point estimates of the time series decomposition are obtained rapidly via linear programming, with marginal and predictive inferences obtainable by Monte Carlo sampling from the joint posterior. The method is shown to effectively reject outliers in time series data. Extensions to multiple time series models are possible.

References

AKAIKE, H. (1980) Seasonal Adjustment by a Bayesian Modelling, Journal of Time Series Analysis, 1, 1, 1–13.

ANDREWS, D.F. and MALLOWS, C.L. (1974) Scale Mixtures of Normality, Journal of the Royal Statistical Society, Series B, 36, 99–102.

ANSLEY, C.F. and KOHN, R. (1987) Efficient Generalized Cross-Validation for State Space Models, Biometrika, 74, 139–148.

BOX, G.E.P., HILLMER, S.C., and TIAO, G.C. (1978) Analysis and Modelling of Seasonal Time Series, in Seasonal Analysis of Time Series, A. Zellner, ed.; U.S. Bureau of the Census, Economic Research Report ER–1, 309–334.

CARLIN, B.P., POLSON, N.G., and STOFFER, D.S. (1992) A Monte Carlo Approach to Nonnormal and Nonlinear State Space Modelling, Journal of the American Statistical Association, 87, 493–500.

CLEVELAND, R.B., CLEVELAND, W.S., MCRAE, J.E., and TERPENNING, I. (1982) STL: A Seasonal Trend Decomposition Procedure Based on Loess, Journal of Official Statistics, 1, 539–564.

CLEVELAND, W.S., DEVLIN, S.J., and TERPENNING, I. (1982) The SABL Seasonal Adjustment and Calendar Adjustment Procedures, Time Series Analysis: Theory and Practice, 1, 539–564.

GEISSER, S. and EDDY, W.F. (1979) A Predictive Approach to Model Selection, Journal of the American Statistical Association, 74, 153–160.

GELFAND, A.E. and SMITH, A.F.M. (1990) Sampling Based Approaches to Calculating Marginal Densities, Journal of the American Statistical Association, 85, 398–409.

GELFAND, A.E., HILLS, S.E., RACINE–POON, A., and SMITH, A.F.M. (1990) Illustration of Bayesian Inference in Normal Data Models using Gibbs Sampling, Journal of the American Statistical Association, 85, 972–985.

GONIN, R. and MONEY, A.H. (1989) Nonlinear Lp Norm Estimation, Marcel Dekker.

KITAGAWA, G. (1987) Non–Gaussian State–Space Modelling of Nonstationary Time Series, Journal of the American Statistical Association, 82, 1032–1041.

KOHN, R., ANSLEY, C.F., AND THARM, D. (1991) The Performance of Cross-Validation and Maximum Likelihood Estimators of Spline Smoothing Parameters, Journal of the American Statistical Association, 86, 1042–1050.

LAW, A.M. and KELTON, W.D. (1982) Simulation Modeling and Analysis, McGraw Hill, New York.

LENK, P.J., (1993) A Bayesian Nonparametric Density Estimator, Nonparametric Statistics, 3, 53–69.

MCCULLOCH, R. E. AND TSAY, R.S. (1994) Bayesian Analysis of Autoregressive Time Series via the Gibbs Sampler, Journal of Time Series Analysis, 15, 235–250.

MEINHOLD, R.J. and SINGPURWALLA, N.D. (1989) Robustification of Kalman Filter Models, Journal of the American Statistical Association, 84, 479–486.

NAYLOR, J.C. and SMITH, A.F.M. (1982) Applications of a Method for the Efficient Computation of Posterior Distributions, Applied Statistics, 31, 214–225.

ROBERTS, G.O. (1992) Convergence Diagnostics of the Gibbs Sampler, in Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, ed.; Oxford University Press, 775–782.

SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1976) The X-11 Variant of the Census Method II Seasonal Adjustment Program, Technical Paper No. 15, Bureau of the Census, U.S. Department of Commerce

STONE, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion), Journal of the Royal Statistical Society, Series B, 36, 111–147.

TANNER, M.A. (1991) Tools for Statistical Inference: Observed Data and Data Augmentation Methods, Springer Verlag.

TSAY, R.S. (1986) Time Series Model Sepcification in the Presence of Outliers, Journal of the American Statistical Association, 81, 132–141.

VANDERBEI, R.J., and CARPENTER, T.J. (1991) Symmetric Indefinite Systems for Interior Point Methods, Technical Report SOR-91-7. Princeton University Dept. of Statisics and Operations Research

WEST, M. (1981) Robust Sequential Approximate Bayesian Estimation Journal of the Royal Statistical Society, Series B, 43, 157–166 .

WEST, M. and HARRISON, P.J. (1986) Monitoring and Adaptation in Bayesian Forecasting Models , Journal of the American Statistical Association, 81, 741–750.

ZELLNER, A. and MIN, C.-K. (1995) Gibbs Sampler Convergence Criteria, Journal of the American Statistical Association, 90, 921–927.

Figure 1: MAP Estimate of Trend Component, $Y = \log(\text{GM Sales})$

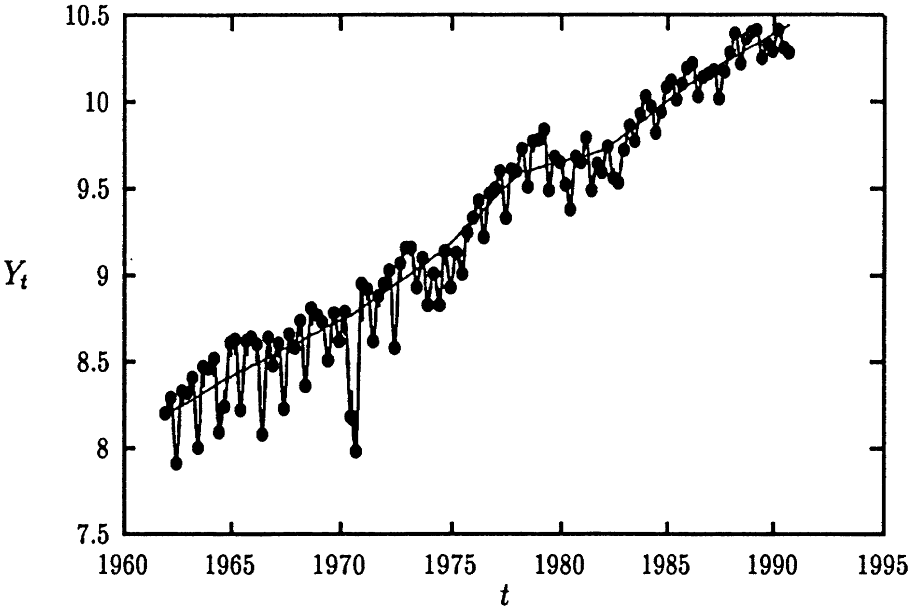


Figure 2: Estimate of Seasonal Component, Y = log(GM Sales)

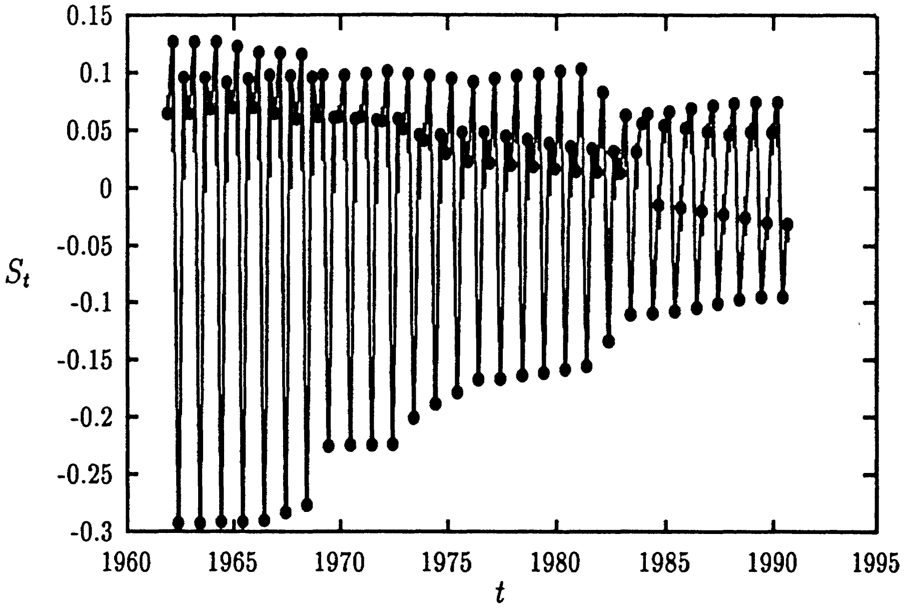


Figure 3: MAP Estimate of Trend Component, $d = r = z = 1$

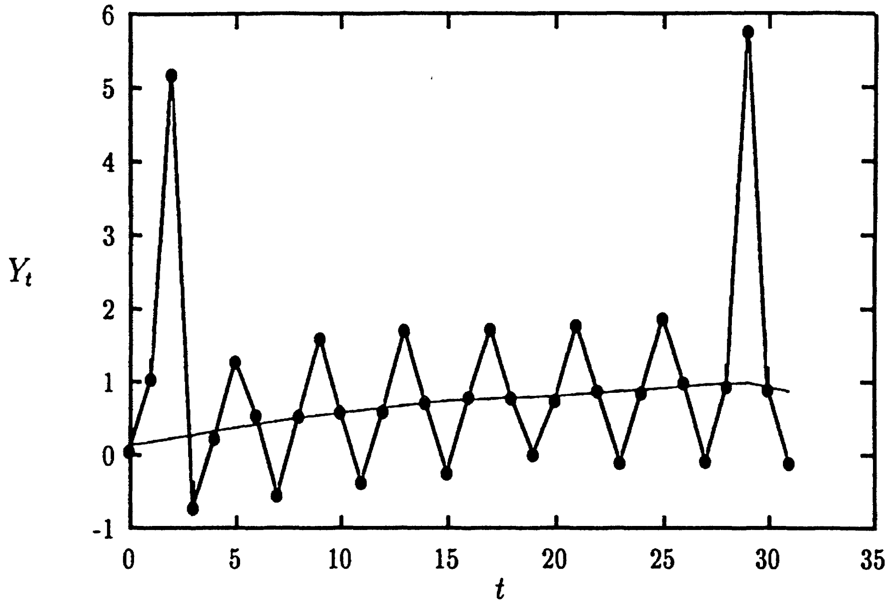


Figure 4: MAP Estimate of Trend Component, $d = 10$, $r = z = 1$

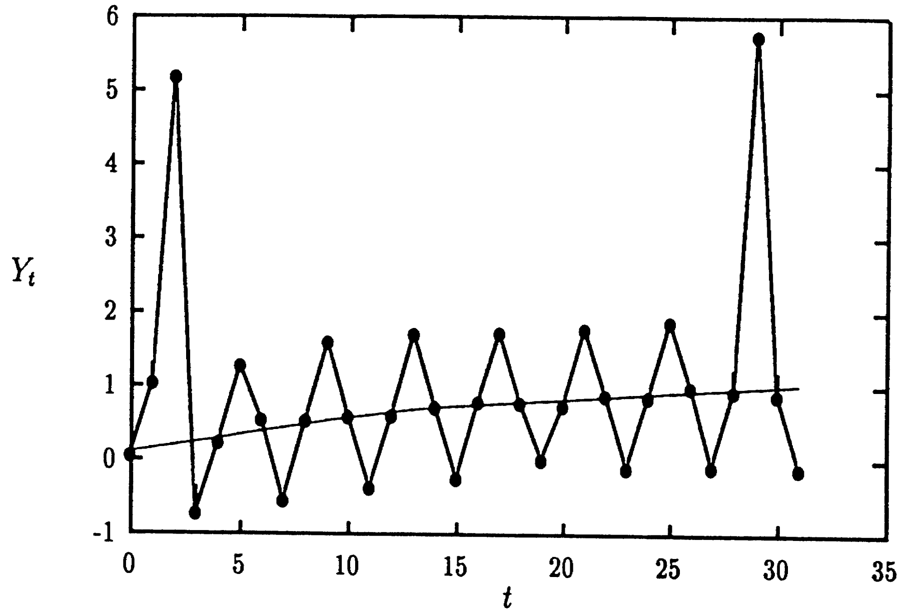
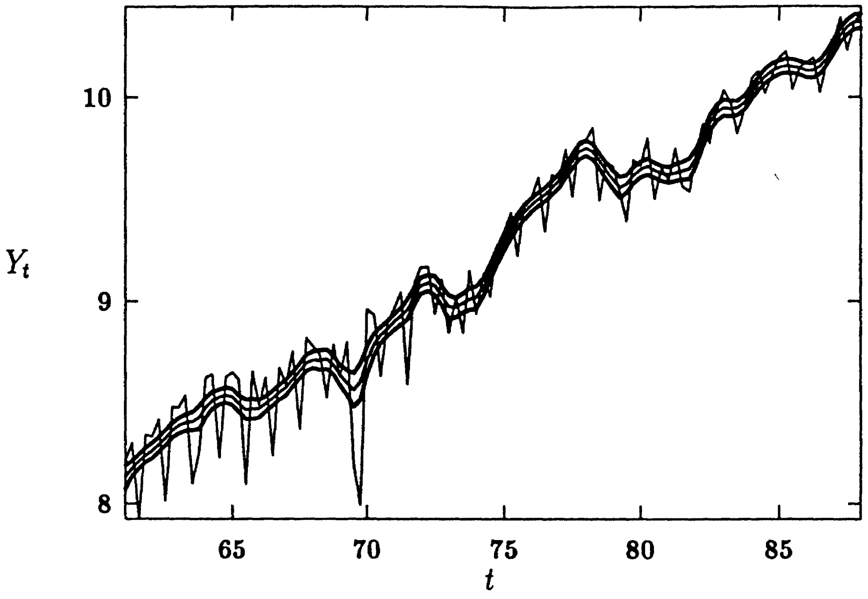


Figure 5: Estimate and Confidence Intervals for Trend Component, $Y = \log(\text{GM Sales})$, Obtained from Gibbs Sampling


10
9
8
Y_t
65
70
75
80
85
t