Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

MODELLING AND ROBUSTNESS ISSUES IN BAYESIAN TIME SERIES ANALYSIS[1]

BY MIKE WEST

Abstract. Some areas of recent development and current interest in time series are noted, with some discussion of Bayesian modelling efforts motivated by substantial practical problems. The areas include non-linear auto-regressive time series modelling, measurement error structures in state-space modelling of time series, and issues of timing uncertainties and time deformations. Some discussion of the needs and opportunities for work on non/semi-parametric models and robustness issues is given in each context.

1. Introduction. Three areas of recent development and current interest in Bayesian time series analysis are: non- or semi-parametric models for non-linear auto-regressions, and related time series structures, based on mixture models; the modelling and accommodation of measurement errors in state space models; and timing errors, uncertainties, and the use of time deformations to map linear time series models to practically interesting non-linear forms. Methodological developments in each area are made possible through the use of MCMC simulation methods, and we are likely to see growth in application of these, and related, kinds of models for this reason (if no other). Needs and opportunities for theoretical and empirical robustness and sensitivity studies are apparent and, in the light of the preceding comment, very practically desirable. It is hoped that this paper will stimulate some time series research interest among some members of the Bayesian robustness communities. In contrast to much of the growth in the "official" Bayesian robustness field, the majority of the practically interesting robustness issues raised here have to do with the forms of data models and likelihood functions, rather than priors (though the distinction is not always clear-cut).

The discussions below are all based in the context of a real-valued, scalar time series y_t, observed over a specified discrete time interval $t = 1, \ldots, n$. For any time point t, y^t denotes the first t observed values $y^t = \{y_1, \ldots, y_t\}$; for any fixed $p < t$, y_p^t denotes the most recent p values, $y_p^t = \{y_{t-p+1}, \ldots, y_t\}$.

2. Non-linear auto-regression. Non-linear time series has been a growth area in non-Bayesian statistics for over fifteen years, with interest generally focussed on non-linear auto-regressions (e.g. Tong 1990 and references therein). Though there is much Bayesian work in non-linear and non-stationary modelling, only recently have Bayesians really taken up the

[1] *Keywords:* Mixture models; Non-normal time series; Non-linear auto-regressions; Non-parametric Bayesian models; Time series outliers; Time deformations; Timing uncertainties.

challenge of empirical non-linear modelling in any generality; recent interest is, naturally, partly driven by computational feasibility and currently accessible simulation methods.

A general objective of AR modelling is to identify useful predictive models $p(y_t|y^t)$ via a fixed order model $p(y_t|y_p^t)$, an $AR(p)$ model for some order p, under the Markovian assumption $p(y_t|y^t) = p(y_t|y_p^t)$ for all t. Covariates may be involved, in which case the focus is on $p(y_t|y_p^t, x_t)$ where x_t represents known covariates at time t. The standard linear $AR(p)$ model, with covariates, has y_t conditionally normal, $N(y_t|\mu_t, v)$, where μ_t is linear in elements of y_p^t and x_t. Approaches to generalising this to non-linear time series would model the entire distributional form, rather than just the mean as in common non-Bayesian approaches.

The basic and simple idea of developing Bayesian mixture models defining interesting classes of conditional distributions $p(y_t|y_p^t, x_t)$ is introduced in Müller, West and MacEachern (1994). In essence, these authors build a modelling framework that embodies and formalises the notions underlying standard kernel auto-regression methods (e.g. Tong 1990). The mixture framework derives from Dirichlet mixture models in density estimation (e.g. as in West, Müller and Escobar 1994). Full background and technical details are given by these authors, and by references therein, for the interested reader.

In the archetype model, $p(y_t, y_p^t, x_t)$ is a mixture of multivariate normal distributions, whose number and parameters may be estimated to provide Bayesian, model based approaches to estimating the joint density. Such mixtures are well-known to provide a high degree of flexibility in modelling observed data configurations, so suggesting that the derived conditional distributions $p(y_t|y_p^t, x_t)$ will be of utility in representing non-linear features of auto-regressions. The conditional distribution $p(y_t|y_p^t, x_t)$ is then a mixture of univariate normals with mixing weights varying as functions of the conditioning past values y_p^t and covariates x_t. To be specific, write $z_t = (y_t, y_p^t, x_t)$ and suppose that, conditional on a parameter π,

$$p(z_t|\pi) = \sum_{j=1}^{k} w_j N(z_t|a_j, A_j);$$

here π represents the full set of parameters $\{k, (w_j, a_j, A_j)_{j=1}^k\}$. The conditional distributions of interest are then

$$p(y_t|y_p^t, x_t, \pi) = \sum_{j=1}^{k} w_j(y_p^t, x_t) N(y_t|b_j(y_p^t, x_t), B_j)$$

where: (A) the component regression functions $b_j(\cdot)$ are linear and vary with j, and the component variances B_j vary with j but do not depend on (y_p^t, x_t);

and (B) $w_j(\cdot) = cw_jk_j(y_p^t, x_t)$ where $k_j(\cdot)$ is a kernel factor proportional to the density ordinate at (y_p^t, x_t) under the marginal (multivariate normal) distribution of (y_p^t, x_t) from $N(z_t|a_j, A_j)$; this implies higher conditional weights $w_j(y_p^t, x_t)$ for components j best supporting the current state value (y_p^t, x_t).

The connections with kernel methods are transparent; $E(y_t|y_p^t, x_t, \pi)$ has the form of predictor usual in kernel (auto-)regression. In some applications, however, there may be substantial variation in the global shape of $p(y_t|y_p^t, x_t)$, as well as in just its mean, and then the full model-based framework provides additional, useful information. The conditional mean may poorly summarise location; for example, conditional distributions may be unimodal for (y_p^t, x_t) in some regions of the "design" space, bimodal or multi-modal elsewhere. In moving continuously through the design space, a trace of the conditional mode may be discontinuous at discrete points, so capturing "threshold *AR* structure" (Tong 1990) in useful ways. Such modal traces will bifurcate in cases where a single mode in one region develops into two modes in another region. The facility to capture such features, observed in practice, is a nice facet of conditional mixture models.

Adapting Dirichlet mixture modelling for random sampling contexts, Müller, West and MacEachern (1994) show how this framework can be implemented. Inference is largely predictive, focussed on evaluating features of the predictive distribution for the next value in the series, i.e. integrating the mixture distribution with respect to the posterior $p(\pi|y^t, x^t)$. Implementation is feasible, though complicated, via MCMC, combining the configuration based Gibbs sampling methods for mixtures (MacEachern and Müller 1994; West, Müller and Escobar 1993) with various Metropolis steps. Extensions to incorporate inference about uncertain hyper-parameters, including variance parameters that play the role of local smoothing parameters, are practically essential, and correspond to automatic smoothing parameter estimation.

Some early applied development of these models appears in Müller, West and MacEachern (1994). One example involves an auto-regression plus covariates model for the waiting time between consecutive eruptions of the Old Faithful geyser. This data series is evidently not well described with linear models (restricting only to the available covariates and past values of the series) and the mixture analysis highlights that fact; predictive distributions for waiting times become bimodal in regions of the design space covered by this data set, and so quite marked variation in predictive structure is exhibited. A second example in that paper concerns the oft-analysed lynx trapping series. This series measures annual estimates of lynx trappings in a Canadian region for a period of over a hundred years, and exhibits cyclical behaviour with a period of around 7–11 years. The cycles are time-varying and, though various linear models provide reasonable fits, are generally viewed as suffering mild non-linearities. One analysis reported mixes basic cyclical $AR(2)$ models, viz $y_t \sim \sum_{j=1}^{k} w_j(y_{t-1}, y_{t-2})N(y_t|\beta_j y_{t-1} - y_{t-2}, B_j)$. As in West (1995a), $AR(2)$ models $E(y_t) = \beta_j y_{t-1} - y_{t-2}$ represent sinusoidal patterns with time-varying amplitudes and phases, but constant periods $\lambda_j = 2\pi/\cos^{-1}(\beta_j/2)$. Hence the state-dependent mixture of such models provides opportunity for identifying variation in the period parameter λ across the design space. Some summary inferences from the analyses indicate such variation; they highlight the suggestion that, in periods of rising lynx trappings (i.e. when $y_{t-1} > y_{t-2}$), the model favours mixture components j with larger values of the period λ_j (through larger values of the regression parameter β_j) than when trappings have been locally falling. This is suggestive of an asymmetry in the form of the cycles – a slower rate of increase, with period closer to 11 years, than the rate of decrease, with period closer to 7 years – consistent with early studies.

In addition to further studies of this framework for non-linear time series, there are various related areas that are of potential for future development, now mentioned.

First, though some interesting application have been studied, these models have significant overheads in terms of specifying hyperpriors on parameters of mixture models, and resulting inferences can be very sensitive to these priors. On the positive side, a focus on prediction makes such sensitivity much less of an issue than it is for parameter estimation. However, this is an area of significant need for sensitivity and robustness investigations, especially in higher dimensional models.

Second, note that other kinds of time series problems are being approached using semi/non-parametric models based on Dirichlet mixtures. In particular, the modelling of jumps, structural changes, outliers, etc., in dynamic linear models is explored in Corradi and Mealli (1995).

Third, the above framework says nothing about stationarity, and there is a need to develop alternative approaches for assumedly stationary processes. Take the case of an $AR(1)$ series, so that the process is characterised by a bivariate distribution $F(y_t, y_{t-1})$. Stationarity implies a common univariate margin, $G(y_t)$ for all t. Interest lies in inference about F quite generally, subject to the constraint to a common margin. Assuming densities, exist, for example, the full joint density of the series y^n given G is simply

$$p(y^n) = p(y_1)\prod_{t=2}^{n} p(y_t|y_{t-1}) \equiv g(y_1)\prod_{t=2}^{n} f(y_t, y_{t-1})/g(y_{t-1})$$

in an obvious notation. So the kinds of questions of interest might include: identifying non-parametric models for F constrained to common margins; identifying flexible classes of parametrised bivariate distribution F with common margins, and consideration of non-parametric priors for the margin G; robustness approaches where the uncertain distributions lie in specified classes, subject to the constraints. The case an $AR(p)$ series is a direct extension, with apparent additional complications.

Finally, data in some applied contexts depart from linearity in only small and often subtle ways. In the non-parametric modelling approach, this suggests starting out with priors, and hyper-priors, that are weighted towards linearity; in the above mixture approach, for example, this would be implied by priors on the Dirichlet hyper-parameters that induce priors with high probability on the series being generated by a mixture of just one or two components. On the robustness side, studies of analyses under classes of models whose elements lie "close" to linear auto-regression are suggested; some developments in this direction, such as local sensitivity analyses exploring "interesting" directions of departure from linearity, would be of interest to the time series community quite widely.

3. Measurment Error Models. The recent introduction of serious simulation methodology into time series analysis, in particular, in state space modelling, is providing opportunity for wider use of both standard (i.e. normal, linear) and somewhat non-standard (i.e. non-normal, non-linear) error models for observational/measurement errors (e.g. Carlin *et al* 1992; Carter and Kohn 1994; Frühwirth-Schnatter 1994; West 1995a,b). For example, the class of auto-regressive dynamic linear models, of interest in modelling (linear) auto-regressions corrupted by purely additive noise (measurement and sampling errors, truncation and rounding errors, gross outliers), is being more widely explored (e.g. West 1995a,b, 1996). MCMC methods are developing for inference about model variance components and other parameters defining state evolution equations, in addition to sequences of state vectors. Extensions of the normal error models to heavier-tailed normal mixtures for accommodating time series outliers are quite straightforward. Though these kinds of models have been around for some time, together with various approximate and non-Bayesian approaches to their analyses, it is only through the recent MCMC developments that we can begin to explore them fully in practical contexts. And there are quite significant computational, convergence and implementation issues raised in this area, largely due to the dimension of resulting parameters spaces. There are also interesting and important modelling and robustness issues raised, some specific examples mentioned below.

A particular class of auto-regressive models has the following, basic DLM, or state-space, representation. An underlying, latent $AR(p)$ process $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \epsilon_t$, with zero-mean, independent innovations $\epsilon_t \sim N(\epsilon_t|0, v)$, is assumed to be observed with additive, zero-mean and independent noise ν_t, giving the data series $y_t = x_t + \nu_t$. In DLM form, $y_t = F'z_t + \nu_t$ and $z_t = Gz_{t-1} + \omega_t$, where $z_t = (x_t, x_{t-1}, \ldots, x_{t-p+1})'$, the state vector at time t,

$\omega_t = (\epsilon_t, 0, \ldots, 0)'$, and with

$$F = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}.$$

MCMC analysis involves running Markov chain simulations to produce sequences of values from the posterior of all model parameters and state vectors. Under the usual assumptions of normality of the observation and innovation error sequences, this is now straightforward to implement. Full technical details appears in West (1995b), especially the appendix on MCMC in state space models, and are quite similar to the developments in related models in West (1995a). This is a new area for MCMC and there are many issues associated with the simulation analyses that need study. A typical problem with even a moderate AR dimension p has many uncertain quantities; the entire set of state vectors over the observed data series, plus the AR parameters and variance components. MCMC convergence issues with the associated high dimensional posterior distributions need study.

Note that the MCMC analyses of state space models are computationally very demanding, in even moderate dimensional models, relative to a standard AR model. In the latter, we observe x_t exactly, and have a basic linear regression model framework, though with some complications due to the starting value problem. The posterior of interest has dimension $2p+1$ (p elements ϕ_j, p starting values, plus the innovations variance). As soon as we admit non-zero observational errors, the dimension is sample size dependent; with n observed values y^n, the number of uncertain quantities is increased by n, the number of latent values of the x_t. Current versions of MCMC analyses operate in the strict state-space format above; here the state dimension is p and so calculations involve sequences of n or more, highly related $p-$variate normal distributions, and their iterative simulation. This involves repeat inversions of the associated covariance matrices, raising questions of numerical stability as well as dramatically increasing the computational burden. More efficient algorithms are needed.

The basic normal model is trivially extended to accommodate the usual kinds of contamination error distributions, such as heavy-tailed normal mixtures. The contamination models of Kleiner, Martin and Thompson (1979), for example, have become popularised through S-Plus implementation (e.g. S-Plus 1993, section 16.7). Here $\nu_t \sim (1-\pi)\delta_0(\nu_t) + \pi N(\nu_t|0, w)$, a mixture of a point mass at zero with the contaminating normal of variance w; thus either the AR process is observed exactly, or it is observed with a (usually rather diffuse) normal error. This is a special case of the well-known scale-inflation model $\nu_t \sim (1-\pi)N(\nu_t|0,\sigma^2) + \pi N(\nu_t|0,k^2\sigma^2)$, admitting a background level of "routine" measurement error together with (occasional) "outliers;" here σ^2 is the variance of the routine error, and $k > 0$ a scale inflation factor. Some exploratory analyses of a range of data series from (largely) the physical sciences indicate cases when this latter model, with non-negligible σ, is to be preferred over that with σ very close to zero, but also cases more in conformity with the latter. The following is an example of the latter case.

Figure 1 displays a series of measurements on the strontium isotope ratio $^{87}Sr/^{86}Sr$ derived from seawater measurements from the Indian ocean (Clemens *et al* 1993, and communicated to the author by Steve Brooks of Cambridge University). These data have been constructed as interpolates of original raw data, and represent approximate levels of the strontium ratio at equal spacings of 3kyr (3000 years); thus the time interval represents approximately the last 450,000 years. They are of interest as climatic indicators, driven by changes in seawater chemistry due, largely, to input from the weathering of continental crust. The series has been preprocessed to remove an underlying trend. A state-space $AR(5)$ has been fitted in the above framework with the scale-inflation measurement error model $\nu_t \sim (1-\pi)N(\nu_t|0,\sigma^2) + \pi N(\nu_t|0,k^2\sigma^2)$; reported analysis assumed $\pi = 0.05$, $k = 10$, and uniform priors (over finite ranges, bounded below above zero but at a very small value) for each of the standard deviations, σ and $\sqrt{v}$ (n.b. the model actually included an additional term for a non-zero, locally constant trend, though the effect is negligible as the series had been pre-processed and reduced to one with essentially no trend).

Here we use this series to illustrate an analysis under a contamination model, and one with rather diffuse priors on the key variance factors. The figure displays the data and some summary estimates from analysis. All estimates and components quoted are approximate posterior values based on the 5,000 posterior draws. The figure shows, on the same vertical scale, the series, the estimates of the measurement errors ν_t, and the estimated decomposition of the latent $AR(5)$ component x_t into its two major components. This decomposition is based on the developments in West (1995b); the $AR(5)$ can model two distinct, damped sub-cycles of time-varying amplitudes and phases, and, as it turns out here, the posterior heavily supports two such cycles. As described in West (1995b), we can easily extract posterior estimates of the amplitude of the corresponding sub-components, and these appear in the figure. There is a third, negligible (and acyclic) component that is not graphed; as a result, the strontium series is approximately the sum of the two sub-components and the observation errors displayed. Also shown is the estimated innovation series ω_t that "drives" the $AR(5)$ process. The model is a reasonable data description (a small degree of residual correlation in

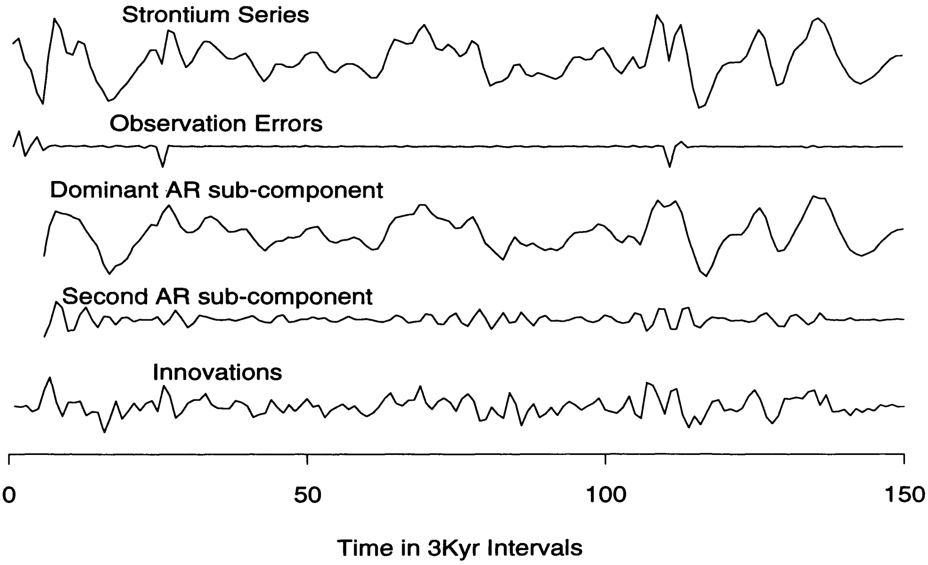

Figure 1: *Strontium series and some features of its analysis. The graph displays the posterior estimates of additive observational errors ν_t, the two main, quasi-cyclical sub-components of latent $AR(5)$ process x_t, and the corresponding estimated innovation series ν_t. All estimates are approximate posterior means from the MCMC analysis.*

the estimated innovations, at lag one, notwithstanding). Of note here is the fact that, though a general contamination model is assumed with a diffuse prior for σ, the posterior heavily supports the degenerate version, indicating that σ is negligible and that only a small fraction of the data values are corrupted with observational noise. (The first three or four errors are interesting; the data series, it turns out, was constructed by interpolating basic raw data, and the suspicion is that the initial three or four observations are corrupted due to the spline interpolation method used). The impact of the few observations on the analysis can be gauged by comparing some inferences of interest with an analysis assuming no measurement errors, and this was done using the standard reference analysis (e.g. West 1995b). For instance, the standard reference posterior means (s.d.s) of the five *AR* parameters ϕ_j are approximately 1.43(0.08), −1.13(0.14), 0.62(0.16), −0.14(0.14), and −0.09(0.08), whereas the state-space analysis leads roughly 2.08(0.15), −2.36(0.33), 1.80(0.38), −0.85(0.26), and 0.16(0.09). Perhaps the main point is to note the "damping" towards zero of the reference estimates, i.e. the suppression of the "signal" that is theoretically predicted if measurement errors are ignored. This effect, and the impact on posterior uncertainties, is quite marked, especially in view of the very small number of really discrepant measurement errors.

Of interest in the application context of this series are inferences about cyclical structure in the series. As noted, the analysis indicates that the *AR*(5) process has two main quasi-cyclical sub-components, damped cycles of time-varying amplitudes and phases; based on posterior estimates and samples of the *AR* coefficients, we can easily extract posterior estimates of the periods of these sub-cycles. Consider the dominant of the two sub-cycles. In the state-space analysis, the approximate posterior distribution for its period has median of 40.9/3, quartiles at 37.2/3 and 44.7/3, and mean (s.d.) of 40.8/3 (1.8/3); the posterior mean of the ϕ_j translates into an estimate of 42.0/3 Kyrs. These values are quoted in terms of Kyrs/3, and correspond closely to the accepted of close to 41Kyrs value that is the established earth-orbital (Milankovitch) period known to drive climatic variation. By comparison, the corresponding figures in the reference analysis are biased upwards; for example, the direct estimate based on the estimated *AR* coefficients is 44.5/3Kyrs, some way above the accepted period of the known "driving" influence.

Some practical issues arising with these kinds of developments, and some needs for further development and robustness studies, include the following; much of this is based on the state-space *AR* context above, though naturally the comments have more general purview.

The extension of the basic *AR* model to the state space version to accommodate observational errors is made at the expense of a very large increase in computational cost, and in user time in monitoring and analysing the simulation outputs. Hence, in parallel to working on more efficient MCMC algorithms for state space models, it might be useful to explore these models from a robustness viewpoint to ask when and whether or not this cost is worthwhile or can be avoided. In some applications, there may be strong evidence to support non-negligible additive observational errors, and they may impact severely on some inferences, thus should not be ignored; then the state space analysis is needed. The potential for additive errors to distort subsequent inferences in the spectral domain, for example, are well known (e.g. Kleiner, Martin and Thompson, 1979). For other inferences, however, and in other applications, their impact may be minor. It would be of some interest, therefore, to develop robustness investigations focussed on local sensitivity; beginning with the strict AR model, i.e. the state space model with a measurement error distribution degenerate at zero, focus on how specific posterior characteristics vary with measurement error models "close" to degeneracy. Studies like this would impact on the question of whether or not the rather dramatic increase in computational expense (in moving from the simple AR model to the state space form) is justified. Presumably some theoretical study in the $AR(1)$ case might yield insights.

Specification issues concerning the form of non-normal elaborations of the observational error distribution arise. Normal mixtures are traditional forms for accommodating occasional outliers, but the constraint to symmetry is sometimes questionable based on a posterior residual analysis under a normal model in a given study. Even retaining symmetry, there are issues of sensitivity to assumed form. Some possible directions for future study include normal mixtures with rather non-parametric approaches to modelling the mixing distributions (e.g. Corradi and Mealli 1995) and possible exploration of other mixture classes (e.g. Brunner 1994).

Related to the above comments are the specific features of error distributions arising in problems where the observations are censored, truncated or rounded. For example, a recent discussion with research radiologists concerned simulations of ultrasound signals reflected from human body tissues of various compositions, with a focus on characterising the returned ultrasound signals in order to distinguish/discriminate tissue types. For various reasons, the return signals are significantly truncated, so suggesting a finite range uniform observational error model (assuming other sources of measurement error are negligible). In principle, MCMC analyses under such models might be developed. Can robustness approaches using mixture classes be developed? The general isses of dealing with censored and truncated measurements are not, of course, time series specific; our interest in robustness issues in this context relates closely to the work of Dempster and Rubin (1983) on rounding errors in a regression context.

4. Timing Uncertainty Rather more specific to time series problems are issues arising through errors or uncertainties about the timing of observations. Two examples provide context.

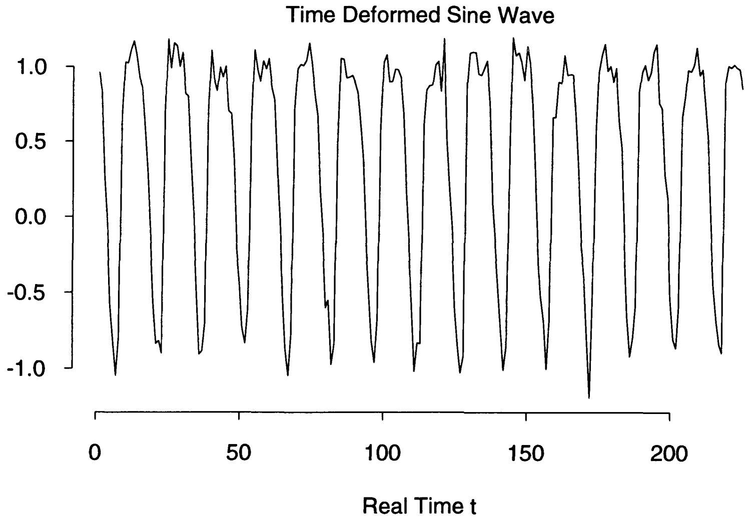

Figure 2: *Time deformed sine wave of period $\lambda = 15$, exhibiting flatter, noisy peaks and sharper troughs.*

First, errors and uncertainties in timing, as exemplified in West (1996), can be a serious practical complication. In that study, the time series data y^n represent chemical constituents of lake sediment related to patterns of climatic change over time. The true calendar time of observation y_i is unknown, measured indirectly and with substantial uncertainty based on its associated sedimentary core depth and processes of carbon-14 calibration. This leads to a rather elaborate model for the true times $t^n = \{t_1, \ldots, t_n\}$ of all observation y^n, defining a (class of) prior distributions $p(t^n)$; West (1996) describes in detail the development of prior distrbutions within this class. Analysis is then developed to incorporate t^n along with the parameters and state variables of the time series model for y^n conditional on t^n. MCMC analysis, as illustrated in this application, is feasible.

The second example is a rather simple illustration of how a time transformation maps a simple linear model to a practically interesting non-linear (but close to linear model). In a beginning collaboration with psychiatrists at Duke University, we are studying waveform characteristics of lengthy EEG

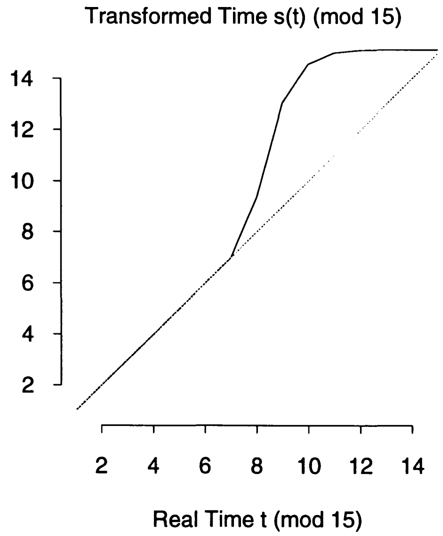

Figure 3: *Time deformation inducing distortion of sine wave.*

(electro-encephalogram) signals in order to characterise such signals for use in discrimination between EEG outputs from electro-convulsive therapy applied to patients on different seizure control treatments. Various models, such as AR, time-varying AR, and others, are being considered, as are perhaps more basic harmonic regressions. At sampling rates much cruder than the raw EEG recordings, some of the data have the appearance of noisy cycles that would be quite well approximated by single harmonic regressions but for one apparent feature; the cycles have very sharp "troughs" and much flatter and noiser "peaks". This could be modelled, perhaps, by superposition of several harmonics of a base frequency. Another idea is to "deform" the time axis in order to map a sine wave to this flat peak/sharp trough appearance; this can be done by "stretching" time around the peak and "compressing" time around the trough. For example, Figure 2 displays the function $\cos(2\pi s(t)/\lambda) + \nu_t$ over time $t = 1, \ldots, 225$, where the ν_t are independent $N(\nu_t|0, 0.1)$ errors, the wavelength $\lambda = 15$, and the time scale is deformed periodically as follows: for $t > \lambda$, $s(t) = s(t \bmod \lambda)$; for $t \leq \lambda/2$, $s(t) = t$; and for $\lambda/2 < t \leq t$, $s(t)$ is given by $\log(s(t)/\lambda - s(t)) = \sigma \log(t/\lambda - t)$ where, in this example, $\sigma = 5$. Figure 3 graphs $s(t)$ versus t over the range $(0, \lambda)$.

Time varying amplitude and phase characteristics might be incorporated by using auto-regressive models with cyclical or quasi-cyclical components.

Otherwise, the kinds of features exhibited in Figure 2 are in close resemblance to those mentioned in the EEG context, so suggesting that a basic harmonic model with a modified time scale could be relevant.

More widely, this relates to the idea that non-linearities in observed series may be modelled indirectly via (possibly stochastic) deformations of the time scale in some generality, assuming that appropriate models for deformations can be identified and estimated. This builds on basic ideas in Stock (1988), who demonstrates that, under certain deterministic deformations, traditional linear models can be mapped into models with characteristics similar to some common non-linear models, such as $ARCH$ models and threshold AR models.

A framework for Bayesian inference currently being explored uses non-parametric models for deformations functions; this is work with F Li and Y Chen at Duke University. Assume an underlying continuous time scale. This is the case in harmonic regression models, for example, but not in usual AR or state space models. In the latter case, and in other cases of discrete time models, we may be able to access an underlying continuum to support arbitrary timings of observations by embedding the discrete time model in an underlying continuous stochastic process model (e.g. Stock 1988, for example). The resulting analysis is complicated in major ways, but some initial work (with F Li) is encouraging. Alternatively, rather standard state-space modelling techniques for missing data may apply; this is true in cases when an underlying fine time scale, discrete and equally spaced, is available, and when the variation in uncertain timings is restricted to this underlying discrete scale. This is the case in the study in West (1996) for example.

Assuming a continuous time scale, one class of non-parametric models under study assumes Gaussian processes priors for deformations. In the EEG case above, for example, a plausible model is as follows. First, take $y_t = \alpha\cos(2\pi s(t)/\lambda) + \beta\sin(2\pi s(t)/\lambda) + \nu_t$ where the errors are independent $N(\nu_t|0,v)$; set $\theta = (\alpha,\beta,\lambda,v)$, the data-model parameter, and write $s^n = \{s(1),\ldots,s(n)\}$ for the uncertain true timings. The above defines $p(y^n|\theta,s^n) = \prod_{t=1}^n p(y_t|\theta,s(t))$, hence provides a likelihood function for (θ,s^n) given the observations. A non-parametric prior for the global time deformation function will now imply a prior $p(s^n|\theta)$ for the discrete set of values for any n. In many contexts, this may not depend on the time series model parameters θ, though there are contexts in which it will; this includes the current model in which the deformation is assumed periodic with (unknown) period λ, an element of θ. One prior with the right kinds of features here is a Gaussian process for the function $\log(s(t)/(\lambda-s(t)))$ over $0 < t < \lambda$, with periodic behaviour outside the interval. The mean function might be a prior guess at the form suggested by Figure 3, or simply $\log(t/(\lambda - t))$, so not anticipating the form of deformation.

Given specified time series and deformation models, Bayesian analysis may be developed through extensions of the MCMC methods introduced in West (1996) in the discrete time context. These kinds of developments are under investigation in this, and other, applied contexts. Of the many modelling and robustness arising, some rather immediate directions include the following.

First, in contexts of expected uncertainties in timing, such as rounding and truncation errors on the time scale, various candidate forms of timing error distribution, and possibly non-parametric approaches, are worth investigating. Second, local robustness studies may be able to contribute in cases where small degrees of timing error are anticipated or suspected – how much timing truncation/rounding can be "tolerated"? Third, in developing time deformation models, the idea of exploring a neighbourhood of a specified (linear) model through use of a stochastic deformation function that is, under the prior deformation model, expected to be "very close" to linear, i.e. heavily biased towards essentially no deformation, bears close resemblance to the earlier ideas of local exploration of models close to linearity in the mixture context. Both non-parametric modelling of the deformation and local robustness studies might be worthwhile here.

Acknowledgments. Research supported in part under NSF grants DMS-9024793 and DMS-9311071. The comments of two referees led to revisions represented in the final version of this paper.

REFERENCES

BRUNNER, L.J. (1994) Bayesian linear regression with error terms that have symmetric unimodal densities. *Journal of Nonparametric Statistics* 4, 335-348.

CARLIN, B.P, POLSON, N.G., and STOFFER, D.S. (1992) A Monte Carlo approach to non-normal and non-linear state-space modelling. *Journal of the American Statistical Association* 87, 493-500.

CARTER, C.K. and KOHN, R. (1994) On Gibbs sampling for state space models. *Biometrika* 81, 541-553.

CLEMENS, S.C., FARRELL, J.W. and GROMET, L.P. (1993) Synchronous changes in seawater strontium isotope composition and global climate change. *Nature* 363, 607-610.

CORRADI, F. and MEALLI, F. (1995) Non-parametric specification of error terms in dynamic models. *Unpublished discussion paper,* Universitá di Firenze.

DEMPSTER, A.P. and RUBIN, D.B. (1983) Rounding errors in regression: the appropriateness of Sheppard's corrections. *Journal of the Royal Statistical Society* (Ser. B), 45, 51-59.

FRÜHWIRTH-SCHNATTER, S. (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis* 15, 183-102.

KLEINER, B., MARTIN, R.D., and THOMPSON, D.J. (1979) Robust estimation of power spectra (with discussion). *Journal of the Royal Statistical Society* (Ser. B), 41, 313-351.

MACEACHERN, S.N. and MÜLLER, P. (1994) Estimating mixture of Dirichlet process models. *Discussion Paper #94-A11*, ISDS, Duke University.

MÜLLER, P., WEST, M. and MACEACHERN, S.N. (1994) Bayesian models for non-linear auto-regressions. *Discussion Paper #94-30*, ISDS, Duke University.

STATISTICAL SCIENCES (1993) *S-PLUS Guide to Statistical and Mathematical Analysis (Version 3.2).* Seattle: StatSci, a division of MathSoft, Inc.

STOCK, J.H. (1988) Estimating continuous-time processes subject to time deformation: An application to postwar U.S. GNP. *Journal of the American Statistical Association* 83, 77-85.

TONG, H. (1990) *Non-Linear Time Series*, Oxford University Press, Oxford, England.

WEST, M., MÜLLER, P., and ESCOBAR, M.D. (1994) Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, eds. A.F.M. Smith and P. Freeman, Wiley, New-York.

WEST, M. (1995A) Bayesian inference in cyclical component dynamic linear models. *Journal of the American Statistical Association* 90, 1301-1312.

WEST, M. (1995B) Time series decomposition. *ISDS Discussion Paper #95-18* Duke University.

WEST, M. (1996) Some statistical issues in Palæoclimatology (with discussion). In *Bayesian Statistics 5*, (eds: J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith), Oxford University Press.

ISDS
DUKE UNIVERSITY
DURHAM NC 27708-0251
EMAIL mw@stat.duke.edu
INTERNET http://www.stat.duke.edu

Modelling and Robustness Issues in Bayesian Time Series Analysis

discussion by
JAMES O. BERGER
Purdue University

This paper is a nice introduction to several exciting current developments in Bayesian time series analysis, many of which relate to robustness issues. The paper also proposes a variety of challenging problems in Bayesian robustness; alas, I cannot suggest any solutions to these problems. My discussion will instead consist of two specific questions, and two general cautions about Bayesian analysis of highly complex models.

The first question concerns the idea of using mixture models, such as

$$p(y) = \sum_{j=1}^{k} w_j N(y|a_j, A_j), \tag{1}$$

to model non-linear auto regressions. Auto regressive models typically arise through the modelling of the conditional distributions of, say, y_t given previous observations. Nonnormal and nonlinear auto regressive models are also typically modelled through focusing on these conditional distributions. The question is whether the typical deviations from conditional linearity, that are encountered in practice, can be captured by low-dimensional mixtures of the form (1)? The concern is that (1) defines a joint distribution, and the particular conditionals of this joint distribution may not be the type of conditionals that easily model the desired nonlinear autoregressions. For large enough k, the model in (1) can, of course, capture any type of behavior, but unless it is effective for small or moderate k its usefulness in practice will be limited.

My second question concerns use of mixture models for outliers. If a measurement error, ν_t, is modelled as

$$\nu_t \sim (1-\pi)N(\nu_t|0, v) + \pi N(\nu_t|0, k^2 v),$$

the analysis can often be quite sensitive to the choice of the variance multiplier, k^2, in the "outlier" distribution. The alternative of using a flat-tailed distribution, such as a t-distribution with 4 degrees of freedom (my favorite, since it looks roughly normal), would seem to have the advantage of not requiring specification of k (or π). Is there any disadvantage to using t-distributions here? In particular, in time series contexts and using MCMC computation, is the use of t-distributions as computationally feasible as the use of mixture distributions?

The first general concern I have is that of propriety of posteriors when dealing with complicated models. As a simple example, consider the trivial version of the measurement error model discussed in section 2,

$$\begin{aligned} y_t &= x_t + \nu_t \qquad && (\nu_t \sim N(\nu_t|0, v)) \\ x_t &= \mu + \epsilon_t \qquad && (\epsilon_t \sim N(\epsilon_t|0, \sigma^2)). \end{aligned}$$

Clearly v and σ^2 are not identifiable (since only the y_t are observed), so that use of standard improper priors will result in an improper posterior. A second example is in the use of mixture models, as in (1). Here, use of almost any common noninformative priors for the a_j and A_j will yield improper posteriors.

A commonly used non-solution to this difficulty is to use vague proper priors. This is a non-solution, because the answer will then typically depend strongly on the degree of "vagueness" chosen. Indeed, I never use vague proper priors, feeling that they can only hide–and never solve–the problem. At the very least, when one uses vague proper priors, study of the sensitivity of the answers to the "scale" of the priors is virtually mandatory (and the scales should be varied separately for nonexchangeable parameters).

The real solution to this difficulty is to learn for which parameters (in a complex model) one can use improper priors and which require proper priors (or some default analogue). It is thus important for those who obtain experience in using a complex model to communicate this type of information, especially since we know that subjective specification of prior destributions for all parameters in a complex model is typically not feasible.

The second general concern in Bayesian use of compex models is that they no longer carry usual Bayesian guarantees of good performance, in part because of the necessity to choose many prior distributions, and small errors in such choices can accumulate across parameters to yield a bad answer. A simple example, developed by J.K. Ghosh, consists of independent observations

$$x_{ij} \sim N(x_{ij}|\mu_i, \sigma^2), \qquad i = 1, 2, ..., n; \qquad j = 1, 2.$$

The goal is inference concerning σ^2, for which trivial consistent estimators exist, such as $\hat{\sigma}^2 = (1/4n)\sum_{i=1}^{n}(x_{i1} - x_{i2})^2$. Interestingly, if one were to choose independent proper priors for the μ_i, then the resulting Bayes estimator for σ^2 would be inconsistent, unless the priors were perfectly calibrated in a certain sense. Hence, for large n, it would not be wise to blindly use even a proper subjective Bayesian analysis, without making some effort to check the performance of the procedure.

In conclusion, I am very excited by the potential of the new complex Bayesian models, such as those being developed by Mike West, but note that they carry with them certain difficulties of which we must continually remain aware.

REJOINDER

Mike West

Thanks to Jim Berger for his insightful comments and observations. They raise several issues worthy of further study in various modelling contexts, not restricted to the time series specifics of the paper. I take the specific points raised in turn.

Approximating auto-regressions using conditional mixtures
Elaborating the question a little, we really have two related issues:

(a) that the derived conditional distributions of a mixture may not adequately represent a specified conditional structure, at the specific values of conditioning quantities, even though the overall joint mixture may be a good approximation to the corresponding joint structure; and

(b) that, in any case, a conditional mixture may require a large number of components to model key observed features of conditional structure, so limiting its practical utility and interpretation.

With respect to (a), it is certainly true at a theoretical level that a joint mixture model may arbitrarily well approximate a specified multivariate distribution, whereas a derived conditional mixture, at a specific set of conditioning values, may be arbitrarily poor as an approximation to the relevant conditional distribution. At a general level, I suspect that meaningful discrepancies will arise only in models approaching the pathological, from an applied perspective, though theoretical studies might prove otherwise; I know of no specifically relevant published works.

More directly in connection with our own work, I note that the practical impetus *starts* with the conditional mixture framework, and *embeds* in a joint structure for computational/algorithmic reasons. Thus our focus on approximation accuracy is initially and immediately in the conditional world, and there we have much to go on in terms of assurances that simple, low-dimensional mixtures exhibiting "piece-wise linear" auto-regressive structure are often adequate in representing observed phenomena (e.g. Tong 1990). A good part of the original motivation for the reported developments lay in the interest in a Bayesian, model-based framework encompassing the widely accepted "standard" non-linear models, namely threshold and smooth-threshold auto-regressions. Without a doubt threshold models occupy a central position in non-linear time series, and as mechanistic models of bifurcations that provide plausible explanations of non-linear, dynamic phenomena more widely (Tong 1990). The relevance of low-dimensional (e.g. two or three component) threshold models as useful approximating forms for a variety of observed non-linear phenomena is undoubted. The Bayesian mixture framework starts with a set of conditional distributions in which such threshold forms are embedded, with a view to overlaying the relevant formal machinery of Bayesian learning, "automatic" threshold detection and estimation, "smooth-threshold" modifications, hyperparameter estimation, and so forth. Appeal to an underlying joint mixture structure is made, of necessity here, in order to utilise modifications of existing simulation methods in the conditional model; in no other sense is the joint mixture structure needed nor used, though it becomes a more integrated part of the formal structure in contexts where stationarity of the time series is to be imposed.

With respect to point (b), the issue of whether or not the model's "*usefulness in practice will be limited*" unless the number of mixture components k is small or moderate turns on the nature of the practical uses to which the model is to be put. Our perspective has been largely predictive. From that viewpoint, the issues of parameter interpretation are somewhat moot. There are real issues of over-fitting, however; seriously increasing numbers of components, hence expanding numbers of parameters, can degrade the reliability of predictions as the model tailors itself too closely to past data. One "strategy" we have adopted is to explore sequences of analyses, successfully changing the prior for k from one very concentrated near $k = 1$ to successively less concentrated forms, aiming to identify only truly meaningful departures from the baseline linear, one component models. Of course, in some cases the data configuration really screams non-linearity, and then even very precise priors on small k can be drastically over-ridden, as in other areas of application (e.g. examples in West and Turner 1992).

Alternative outlier models

It is certainly the case that alternative outlier models, such as T distributions, are as easily implemented as normal mixtures, and various authors have experimented with such models in time series. Technically, depending on the specific mixture parameters and degrees of freedom parameters chosen, there may be little difference between the two models until we move way out into the tails. As a result, in some applications there will be little data based evidence to distinguish between error models, and then the mixture appears relatively disadvantaged due to the additional parameter specifications required (or the requirements to specify priors for parameters such as π). One point in favour of the mixture model is the question of very low levels of routine observation noise, and the opportunity to identify this in the presence of occasional outliers. Another is the resulting explicit computation of posterior probabilities of "outlying" for individual observations.

Perhaps more practically important than concerns about specific distributional forms (which can always be compared via repeat analyses in any case) are the needs for closer investigations and more incisive modelling of outlier generating mechanisms. Some application areas naturally involve considerations of dependencies between the outlier characteristics of consecutive or near-by observations – the occurrence of outliers in small "batches" is a case in point. Here the issue is not so much one of distributional form for the ν_t, but of time-varying and state-dependent outlier probabilities, i.e. π_t rather than π for ν_t.

Propriety of posteriors

Jim Berger is right to note that the general issues of posterior propriety and related problems of parameter identification are becoming more evident as Bayesian models become larger and more complex. The problems arise in models/likelihood functions, and evaporate if modellers toe-the-line and adhere to proper priors, though then the questions Jim raises about where the "action" is in the priors are foremost.

I have little to add to Jim's concluding comments about communication of insights and experiences in specific kinds of models. In practical Bayesian forecasting models we have long followed this rule in connection with the specification of variance components (related to Jim's example) in dynamic models. Our use of discount factors (related to ratios of component variances), and the focus on such factors as "key" determinants of model sensitivity, is evident in ranges of published work. We have stressed the need to explore data analyses and inferences with such factors constrained to vary in "relevant" ranges, and have often been involved in communicating the "dangers" of relaxing such constraints through the use of vaguer priors. Jim's comment is a plea for more of this, at a more formal and constructive level and addressed to prior sensitivity questions in broader classes of models.

More specifically, I am unclear as to the relevance of the particular model Jim displays to illustrate the identification point. That model is one of independent observations, and not one we use, nor discuss in the paper. The models discussed do not suffer mathematical identification problems. For example, a similar model form that does get used a great deal is the first-order polynomial, in which $y_t = x_t + \nu_t$ and $x_t = x_{t-1} + \epsilon_t$, with the same error assumptions. Here there is no mathematical identification issue; the joint density of a series of observations depends on (functions of) the variances of both error sequences that (for more than two observations) leads to identification and proper posteriors under the usual improper priors (though I don't recommend improper priors). The potential for significant sensitivity to the priors for variance components remains, of course.

As a final comment I note that the reported data analysis in the paper, and other similar analyses (e.g. West 1995), utilise priors on variance components that are developed from uniform priors, on finite ranges, for the corresponding standard deviations. These might be viewed as "vague proper priors" in these analyses, as the chosen ranges are essentially irrelevant. We learn this by running the simulation analyses using rejection methods to sample variance components from the relevant conditional posterior distributions, and this generates checks on the prior dependence through the rejection rates. Very low rejection rates are indicative of the insensitivity to the chosen prior ranges.

Consistency Issues

I have little to add to Jim's commentary and warning on the issue of posterior inconsistency in models whose parameter dimension grows with the sample size. The particular example is disturbing in view of its simplicity and for the inference that apparently subtle changes in proper priors may so completely and seriously impact on the asymptotic behaviour of the posterior. I'd like to hear more of the example, and, in particular, to ask about just how subtle the variations in the prior are in this case. The example model does bear analogy with even rather standard dynamic models used in times series, so suggesting the need to explore the issue in the time series domain. This reinforces the need for the Bayesian robustness community to begin to explore time series and dynamic modelling areas, in line with my "invitations" in the paper. Whether or not these issues are relevant in any of the specific contexts of the paper remains an open question.

I thank Jim for his insightful and incisive commentary.

REFERENCES

West, M. and Turner, D.A. (1992) Deconvolution of mixtures in analysis of neural synaptic transmission, *The Statistician*, **43**, 31-43.

West, M. (1995) Time series decomposition. *ISDS Discussion Paper #95-18* Duke University.