Bayesian Mixtures of Autoregressive Models

By Sally Wood

Melbourne Business School, University of Melbourne, Melbourne, Victoria, 3053, Australia

Ori Rosen

Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, U.S.A.

and Robert Kohn

Australian School of Business, University of New South Wales, Sydney, New South Wales, 2052, Australia

Summary

In this paper we propose a class of time-domain models for analyzing possibly nonstationary time series. This class of models is formed as a mixture of time series models, whose mixing weights are a function of time. We consider specifically mixtures of autoregressive models with a common but unknown lag. To make the methodology work we show that it is necessary to first partition the data into small non-overlapping segments, so that all observations within one segment are always allocated to the same component. The model parameters, including the number of mixture components, are then estimated via Markov chain Monte Carlo methods. The methodology is illustrated with simulated and real data. Supplemental materials are available online.

Key Words: Forecasting; Mixtures-of-experts; Nonstationary time series; Reversible jump MCMC; Segmentation.

1 Introduction

Our article develops a model for analyzing possibly nonstationary time series. We allow for parameter evolution using a mixture model whose components are time series with constant but unknown parameters and mixture probabilities that depend on time. The number of mixture components is determined from the data. We take a Bayesian approach which is implemented using Markov chain Monte Carlo sampling.

Many current methods for fitting time series whose parameters change over time are based on some segmentation of the time series. Fitting piecewise autoregressive models was suggested by Kitagawa and Akaike (1978) who use AIC to determine the change points. Davis et al. (2006) propose fitting piecewise autoregressive models using minimum description length and use genetic algorithms for solving the resulting optimization problem. Ombao et al. (2001) segment a time series using orthogonal complex-valued transforms that are simultaneously localized in time and space. A cost is evaluated for each particular segmentation and the optimal segmentation is the one with minimal cost. Since it is infeasible to consider all possible segmentations, Ombao et al. (2001) assume dyadic segmentations. Punskaya et al. (2002) propose a Bayesian method for fitting piecewise linear regression models such as autoregressive models. They place prior distributions on the number of change points, their locations and the order of the linear regression in each segment, and use Markov chain Monte Carlo simulation to estimate the model.

A second approach which allows the parameters to change is to model their evolution. For example, West et al. (1999) allow the parameters of an autoregressive process to change over time by modeling them as a random walk. However, they assume that the maximum lag in the autoregressive process is fixed. The assumption of a fixed lag is relaxed by Prado and Huerta (2002). Gerlach et al. (2000) provide a sampling scheme that allows for smooth parameter evolution as well as structural breaks in the parameters. For an application of the methods in Gerlach et al. (2000) and some further extensions see Giordani and Kohn (2008).

Our method differs from the methods described above in that we do not seek to directly determine which parameters change and which do not, nor do we have to specify the model for parameter evolution to allow for structural changes. This is an important advantage of our approach because in models with more than a few parameters not all parameters will evolve in the same way nor would all parameters change abruptly at the same time. In addition, it is sometimes difficult to model the evolution of some parameters, e.g. covariance matrices. We note that our model formulation allows some of the parameters to be the same over time by making them common across all components.

There are several papers related to our methodology. A basic reference for regression mixture models having covariates in both the components and the component probabilities is Jacobs et al. (1991) who call them mixture-of-experts models. Rosen et al. (2009) estimate an evolving spectral density by partitioning the data into segments of contiguous observations, calculating the log periodogram of each segment and then fitting a smoothly varying mixture to these log periodograms. In this paper, we use a time-domain approach which does not require computation of local periodograms. For this reason, smaller segment lengths can be used, making it possible to detect changes over smaller time intervals. In addition, the frequency domain approach of Rosen et al. (2009) assumes local stationarity and is restricted to modeling the second moments of the process which is inherent to frequency-domain methods. The current approach does not have any of these restrictions and can model the entire distribution of the observations.

In other related work, Wong and Li (2001) use a two-component mixture model with logistic weights that may depend on time and exogenous variables. Parameter estimation is performed via an EM algorithm, and autoregressive lag selection is facilitated by BIC. Carvalho and Tanner (2005, 2006, 2007) use a mixture-of-experts approach to model nonlinearities in time series models. These authors use maximum likelihood estimation, investigate identifiability and asymptotic normality of the estimates and use AIC and BIC for selecting the number of components in the model. Similarly, Prado et al. (2006) use hierarchical mixtures-of-experts with vector autoregressive models, with the parameters estimated by the

EM algorithm, and model selection is performed by BIC. Lau and So (2008) use a Dirichlet process mixture of autoregressive processes to flexibly model the predictive density of a time series. Their approach does not handle structural breaks in the time series and their mixture weights are not functions of time.

The rest of the paper is organized as follows. Section 2 presents the proposed model and the priors. Section 3 describes the sampling scheme of the proposed Markov chain Monte Carlo procedure. Section 4 discusses forecasting. Section 5 describes the results of a simulation study and Section 6 considers applications.

2 The model and prior specification

2.1 The general model

We propose to model a possibly non-stationary time series $\{y_t, t = 1, \ldots, n\}$ as a mixture of autoregressive (AR) processes, where the lag of the AR processes and the number of components in the mixture are both unknown but finite. Let $\boldsymbol{\theta}_{pr}$ be the set of parameters needed to prescribe a mixture having r components each of lag p. The predictive density $p_{pr}(y_t|y_{t-1}, \ldots, y_1; \boldsymbol{\theta}_{pr})$ is the mixture model,

$$p_{pr}(y_t|y_{t-1}, \ldots, y_1; \boldsymbol{\theta}_{pr}) = \sum_{j=1}^{r} p_{jpr}(y_t|y_{t-1}, \ldots, y_1; \boldsymbol{\theta}_{jpr})\pi_{tjpr}, \tag{1}$$

where $\boldsymbol{\theta}_{pr} = (\boldsymbol{\theta}_{1pr}, \ldots, \boldsymbol{\theta}_{rpr})$. This model belongs to the class of mixtures-of-experts models (Jacobs et al. (1991)). An early reference for a mixtures-of-experts time series model is Zeevi et al. (1996). The mixing weights π_{tjpr}, described more fully in Section 2.3, are multinomial logit and depend on time. In (1), π_{tjpr} is the weight attached to the jth component at time t in a mixture containing r autoregressive processes each of lag p. We denote $\boldsymbol{\theta}_{pr} = \{(\boldsymbol{\omega}_{jpr}, \boldsymbol{\delta}_{jpr})\}_{j=1}^{r}$, where $\boldsymbol{\omega}_{jpr}$ are the parameters needed to specify the AR processes, and $\boldsymbol{\delta}_{jpr}$ are those needed to specify the weights attached to the components of the mixture.

The overall predictive density $p(y_t|y_{t-1}, \ldots, y_1; \boldsymbol{\theta})$ is obtained by averaging over the number of possible lags and components so that

$$p(y_t|y_{t-1},\ldots,y_1;\boldsymbol{\theta}) = \sum_{r=1}^{R}\sum_{p=1}^{P} p_{pr}(y_t|y_{t-1},\ldots,y_1;\boldsymbol{\theta}_{pr})\,\Pr(p,r), \tag{2}$$

where R and P are the maximum number of components and lags respectively. The density $p(y_t|y_{t-1},\ldots,y_1;\boldsymbol{\theta})$ in model (2) can be quite general with y_t discrete or continuous; for example, it can be a GARCH model with possible structural breaks in the parameters or slowly evolving parameters. One important contribution of the paper is that it can accommodate structural breaks or slowly evolving parameters without having to directly model the change points or the evolution of the parameters.

2.2 Model and priors for autoregressive components

Model

For simplicity, we assume that all components have the same unknown lag length p. A more general model would assume a different lag for each component, but for tractability, we do not pursue this here. We write the jth AR(p) process in a mixture containing r components as

$$y_t = \phi_{jpr,0} + \sum_{k=1}^{p} \phi_{jpr,k} y_{t-k} + \sigma_{jpr} e_{jt}, \quad e_{jt} \sim N(0,1)\ . \tag{3}$$

Thus, $\boldsymbol{\omega}_{jpr} = (\phi_{jpr,0}, \phi_{jpr,1}, \ldots, \phi_{jpr,p}, \sigma^2_{jpr})'$. Note that although the lag p does not affect the dimension of σ^2, we add the subscript p to indicate that different values of p will result in different values of σ^2.

Priors on $\boldsymbol{\phi}_{pr} = (\boldsymbol{\phi}'_{1pr}, \ldots, \boldsymbol{\phi}'_{rpr})'$

Zellner's G-prior distributions (see Marin and Robert (2007)) $N(\mathbf{0}, c\,\sigma^2_{jpr}(X'_pX_p)^{-1})$ are placed on the $\boldsymbol{\phi}_{jpr}$'s where $c = n$ and

$$X_p = \begin{pmatrix} 1 & y_P & y_{P-1} & y_{P-2} & \cdots & y_{P-p+1} \\ 1 & y_{P+1} & y_P & y_{P-1} & \cdots & y_{P-p+2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n-1} & y_{n-2} & y_{n-3} & \cdots & y_{n-p} \end{pmatrix}.$$

Note that X_p is an $(n-P) \times (p+1)$ matrix with a fixed number of rows and a variable number of columns, for $p = 1, \ldots, P$.

Priors on $\boldsymbol{\sigma}^2_{pr} = (\sigma^2_{1pr}, \ldots, \sigma^2_{rpr})'$

The priors on the σ^2_{jpr}'s are independent inverse gamma distributions with densities $p(\sigma^2_{jpr}) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2_{jpr})^{-(\alpha+1)} \exp(-\beta/\sigma^2_{jpr})$, where $\alpha = \beta = 0.01$. This choice of α and β reflects vague knowledge of σ^2_{jpr}. For identifiability, the σ^2_{jpr}s are ordered. As described in Section 3, the first stage of the sampling scheme consists of obtaining posterior means and variances of the $\boldsymbol{\omega}$'s and $\boldsymbol{\delta}$'s, which are then used in a second stage to form proposal distributions. Maintaining identifiability in the first stage is essential to forming good proposals in the second stage.

2.3 Model and priors for the mixture weights

Model

The mixing weights depend on time and on an unknown parameter vector $\boldsymbol{\delta}_{pr} = (\boldsymbol{\delta}'_{1pr}, \ldots, \boldsymbol{\delta}'_{rpr})'$, and have the multinomial logit form

$$\pi_{tjpr} = \frac{\exp(\boldsymbol{\delta}'_{jpr}\boldsymbol{s}_t)}{\sum_{h=1}^{r} \exp(\boldsymbol{\delta}'_{hpr}\boldsymbol{s}_t)} \ , \tag{4}$$

where $\boldsymbol{s}_t = (1 \quad t)'$. For identifiability we take $\boldsymbol{\delta}_{1pr} = \mathbf{0}$. Again, we note that although the value of p does not affect the dimension of $\boldsymbol{\delta}$, we add the subscript p to indicate that different values of p will result in different values of $\boldsymbol{\delta}$.

Prior on $\boldsymbol{\delta}_{pr}$

The priors on $\boldsymbol{\delta}_{jpr}$, $j = 2, \ldots, r$, are independent bivariate normal $N(\mathbf{0}, \sigma^2_\delta I_2)$, where $\sigma^2_\delta = n$. The variance of the prior on δ_{pr} is chosen to be proportional to the sample size so that the prior remains diffuse with respect to the likelihood. Such a prior is similar to Zellner's G prior, which is discussed in detail in Marin and Robert (2007). This prior should not be improper for the following two reasons.

1. Not all the mixing parameters, $\boldsymbol{\delta}_{pr}$, are common to all models. The single-component

mixture has no mixing parameters, a two-component mixture has two, a three component mixture has four and so on. Placing improper priors on these parameters will result in all of the probability in the posterior being assigned to the simplest model (the one-component mixture). This will occur even if the posterior distribution of the mixing parameters, for a given model, is proper.

2. In logistic regression, if there is complete separation of the data, then the likelihood is not bounded. Placing a proper prior on the mixing parameters means that, for a given model, the posterior will be proper.

2.4 Priors on r and p

The maximum number of components is R and $\Pr(j = r) = 1/R$ for $r = 1, \ldots, R$. Similarly, the maximum value of the lag is P and $\Pr(k = p) = 1/P$ for $p = 1, \ldots, P$.

2.5 Segmentation

In mixture models for independent observations, the allocation of an observation to a component is usually done by computing the probability that the observation arose from that component. However, with time series data, computing the probability that a single observation belongs to a component does not take into account the dependence of the data over time. This time dependence must be accounted for, and for this reason, we propose to divide the time series into S small segments, each of length L. In particular, all observations y_t for $t \in \{1 + (s-1)L, \ldots, sL\}$ are included in segment s, $s = 1, \ldots, S$, and allocation to a component is done by computing the probability that the segment was generated from a particular AR process. The predictive density in (1) becomes

$$p_{pr}(y_t | y_{t-1}, \ldots, y_1; \boldsymbol{\theta}_{pr}) = \sum_{j=1}^{r} p_{jpr}(y_t | y_{t-1}, \ldots, y_1; \theta_{jpr}) \pi_{sjpr} \tag{5}$$

for all y_t for $t \in \{1+(s-1)L, \dots, sL\}$. The corresponding likelihood function is

$$\prod_{s=1}^{S} \prod_{t=1+(s-1)L}^{sL} \sum_{j=1}^{r} p_{jpr}(y_t|y_{t-1}, \dots, y_1; \theta_{jpr})\pi_{sjpr}.$$

The mixing weights are a function of s, $s = 1, \dots, S$ rather than of time. Thus,

$$\pi_{sjpr} = \frac{\exp(\delta'_{jpr}\boldsymbol{u}_s)}{\sum_{h=1}^{r}\exp(\delta'_{hpr}\boldsymbol{u}_s)} \ ,$$

where $\boldsymbol{u}_s = (1 \quad s)'$. Note that different segments may, and often do, belong to the same component, in the same way that individual observations in independent data may belong to the same component. Thus, segments and structural breaks are not equivalent, but structural breaks in the data are accommodated automatically through the mixing weights. Unlike Rosen et al. (2009), our method does not require segmentation, but we segment the data because it improves the performance of our method.

In selecting the segment length, L, it is necessary that L satisfy the following two criteria, (i) L contains enough observations to estimate the dependence in the time series and (ii) L is as small as possible to accurately detect changes in the time series. We found that selecting $L = P + 2$, where P is the maximum allowable number of lags, met these two criteria. We also found the results were insensitive to the choice of L for $P + 2 < L < 3 \times P$.

We now give an example to demonstrate the need for segmentation, while Section 5 which describes simulation results, provides more details about the benefits of segmentation.

Example:

We generated a realization of 500 observations from the following process

$$y_t = \begin{cases} 0.9y_{t-1} + \epsilon_t^{(1)} & \text{for} \quad 1 \leq t \leq 470 \\ -0.9y_{t-1} + \epsilon_t^{(2)} & \text{for} \quad 471 \leq t \leq 500 \ , \end{cases} \tag{6}$$

where $\epsilon_t^{(1)} \overset{\text{iid}}{\sim} N(0, 0.8^2)$ and $\epsilon_t^{(2)} \overset{\text{iid}}{\sim} N(0,1)$. Stage I of the sampling scheme (see details in Section 3) was then performed, once without segmenting the time series, and a second time, after dividing it into 50 segments of length 10 each. The number of mixture components

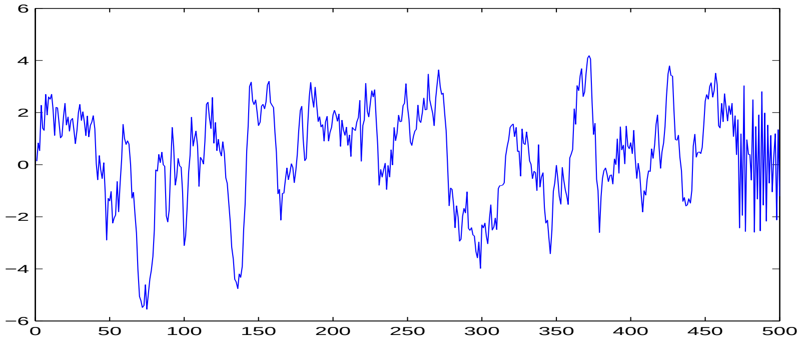

Figure 1: A realization from model (6)

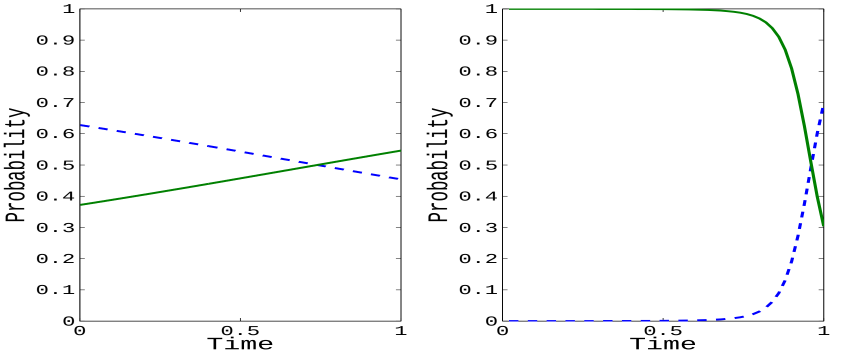

Figure 2: The mixing probabilities vs. time (rescaled to the unit interval). Left panel: no segmentation; right panel: 10 observations per segment.

and the lag were fixed at 2 and 1, respectively. Figure 1 displays one realization from model (6). Figure 2 presents the mixing probabilities for each component when the time series is not segmented (left panel) and when it is segmented (right panel). The plots show that without segmentation, each component is weighted either too heavily or too lightly, whereas segmentation results in weights that closely reflect the change in the time series. The improved behavior of the mixing weights also results in better forecasts. The top row of Figure 3 displays in solid lines the true k-step-ahead predictive densities for $k = 1, \ldots, 5$, as well as their estimates in dotted lines, based on the unsegmented data. The value of k increases from left to right. The bottom row shows the analogous densities based on the segmented times series. The figure shows that the estimates based on the segmented time series are closer to the true densities than their counterparts based on the unsegmented series.

Note that the bottom left estimate is bimodal which may happen because the estimate is a mixture of two predictive densities. More details on forecasting are given in Section 4.

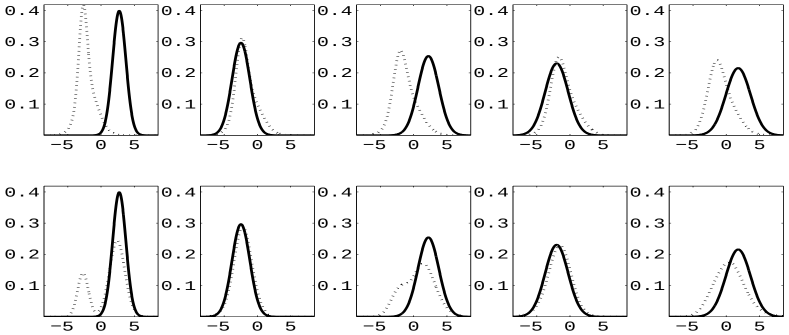

Figure 3: k-step-ahead predictive densities, $k = 1, \ldots, 5$. The true densities are plotted in solid lines, while their estimates are in dotted lines. The top row corresponds to the unsegmented time series, whereas the bottom row is based on the segmented data. The value of k increases from left to right.

Segmentation also results in improved spectrum estimation. For the realization from model (6), we estimated the local log spectral densities at $t =$ 45, 95, 155, 215, 275, 335, 395, 475, 485, 495 based on the unsegmented time series, as well as at segments $s =$ 5, 10, 16, 22, 28, 34, 40, 48, 49, 50 based on the segmented time series. The change in the time series occurs at $t = 471$, and the values of t and s were chosen so that the first 7 spectral densities correspond to the first AR(1) process, and the last 3 spectral densities correspond to the second AR(1) process. The estimates of the log spectral density for the no-segmentation case were obtained as the mixture $\sum_{j=1}^{2} \hat{\pi}_{tj12} \log \hat{f}_{j12}(\nu)$, where $\hat{f}_{j12}(\nu)$ is the estimate of the jth spectral density at frequency ν ($0 \leq \nu \leq 0.5$) in a mixture of two AR(1) components. In the case of segmentation, the weight π_{tj12} is replaced by π_{sj12}. The estimate of $f_{j12}(\nu)$ is given by

$$\hat{f}_{j12}(\nu) = \hat{\sigma}_{j12}^2 |\hat{\phi}_{j12}(e^{-2\pi i\nu})|^{-2} ,$$

where $i = \sqrt{-1}$ and $\phi_{j12}(x) = 1 - \phi_{1j12}x$ is the AR(1) characteristic polynomial. For a discussion of time-varying local spectra, see Dahlhaus (1997). Figure 4 presents plots of the true and estimated log spectral densities for both the unsegmented and segmented time

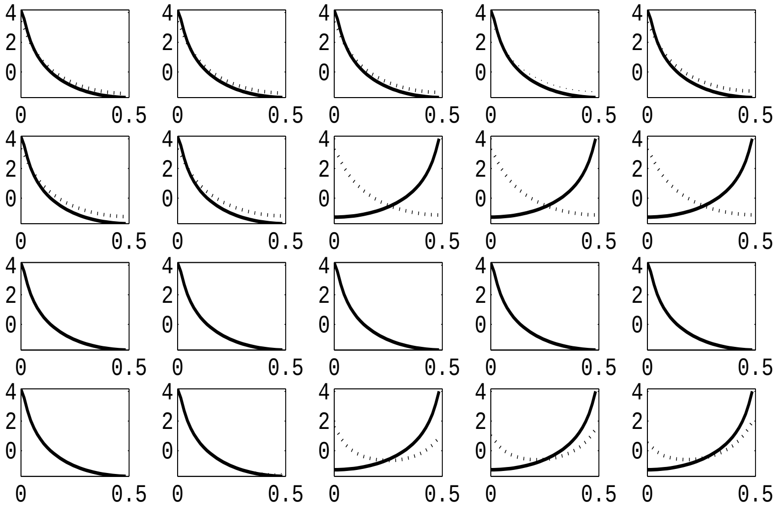

Figure 4: True (solid) and estimated (dotted) log spectral densities for the realization from model (6). The top two rows are based on the unsegmented series. The bottom two rows are based on the segmented time series.

series. The last three panels of the second row in Figure 4 show that the change in the spectral density is not detected for the unsegmented time series. The bottom two rows of Figure 4 show that the true and estimated log spectral densities are almost indistinguishable for the first 7 panels, and in the last three panels the estimated spectral densities are much closer to the true spectral densities than in the unsegmented case. If r and p are not fixed, the estimates obtained without segmentation are even poorer.

3 The sampling scheme

All the parameters, including r and p, are sampled from their posterior distribution, in two stages. In stage I, R separate chains are run for $r = 1, \ldots, R$. In each of these chains, r is fixed while all other parameters, including p, are sampled. The results from this preliminary analysis are utilized in stage II to perform a reversible jump step corresponding to varying values of r. The reversible jump step is needed to perform model averaging (as opposed to model selection), which requires the estimation of $\Pr(r|Y)$ and $\Pr(p|Y)$. Model selection techniques such as AIC or BIC do not yield estimates of $\Pr(r|Y)$ and $\Pr(p|Y)$. The advantages of model averaging over model selection have been researched by a number of authors, see Hoeting et al. (1999) and Kadane and Lazar (2004), who conclude that model averaging has better predictive ability than model selection. In the real examples of Section 6, most of the probability mass of the posterior distributions of r and p is concentrated on a single value and therefore model selection and model averaging will give very similar results. However, this can only be known after performing reversible jump MCMC, not before, and it cannot be assumed that all datasets will have this property.

To simplify the sampling scheme of Stage I, we introduce latent indicator variables to indicate the component to which a segment belongs. Let $z_{sjpr} = 1$ if y_t, $t = 1 + (s-1)L, \ldots, sL$, $s = 1, \ldots, S$, is generated by the jth component, and $z_{sjpr} = 0$ otherwise. Note that $z_{sjpr} = 1$ means that $z_{tjpr} = 1$ for all $t \in \{1+(s-1)L, \ldots, sL\}$. The augmented conditional likelihood is

$$L(\boldsymbol{y}^*, \boldsymbol{z}|y_1, \ldots, y_P, \boldsymbol{\theta}) = \prod_{s=1}^{S}\prod_{j=1}^{r}\{\pi_{sjpr} \prod_{t=1+(s-1)L}^{sL} p(y_t|y_{t-1}, \ldots, y_{t-p}; \boldsymbol{\phi}_{jpr}, \sigma^2_{jpr})\}^{z_{sjpr}}\ , \qquad (7)$$

where $\boldsymbol{y}^* = (y_{P+1}, \ldots, y_n)'$ and $\boldsymbol{z}$ contains all the z_{sjpr}'s for $j = 1, \ldots, r$ and $s = 1, \ldots, S$. Note that for $s = 1$, i.e., for the first segment, $t = P + 1, \ldots, L$.

3.1 Stage I: fixed r

Given r, drawing p, $\boldsymbol{\phi}_{pr}$, $\boldsymbol{\sigma}^2_{pr}$, $\boldsymbol{\delta}_{pr}$ and $\boldsymbol{z}$ is based on (7) in combination with the prior distributions. In particular, the sampling scheme consists of the following steps.

1. Fix r and initialize $\boldsymbol{z}$.
2. Draw the lag p from the multinomial distribution $p(p|\boldsymbol{y}^*, r, \boldsymbol{z})$.
3. For $j = 1, \ldots, r$, draw σ^2_{jpr} from the inverse gamma distribution $p(\sigma^2_{jpr}|\boldsymbol{y}^*, \boldsymbol{z}_{jpr}, r, p)$.
4. For $j = 1, \ldots, r$, draw $\boldsymbol{\phi}_{jpr}$ from the multivariate normal distribution $p(\boldsymbol{\phi}_{jpr}|\sigma^2_{jpr}, \boldsymbol{z}_{jpr}, \boldsymbol{y}^*, p)$.

5. For $j = 2, \ldots, r$, draw $\boldsymbol{\delta}_{jpr}$ from the multivariate normal distribution $p(\boldsymbol{\delta}_{jpr}|\boldsymbol{z})$.

6. Draw z_s from the multinomial distribution $p(z_s|\boldsymbol{\phi}_{pr}, \boldsymbol{\sigma}^2_{pr}, \boldsymbol{\delta}_{pr}, r, p)$, for $s = 1, \ldots, S$.

R chains are run for $r = 1, \ldots, R$. From the iterates of each of these chains, we obtain the posterior distribution of the lag, as well as the posterior mean vectors $\hat{\boldsymbol{\phi}}_{pr}$, $\hat{\boldsymbol{\delta}}_{pr}$ and $\hat{\boldsymbol{\nu}}_{pr} = \log \hat{\boldsymbol{\sigma}}^2_{pr}$. The corresponding variance-covariance matrices $\hat{\Sigma}_{\phi_{pr}}$, $\hat{\Sigma}_{\delta_{pr}}$ and $\hat{\Sigma}_{\nu_{pr}}$ are obtained as sample variance-covariance matrices based on the iterates of $\boldsymbol{\phi}_{pr}$, $\boldsymbol{\delta}_{pr}$ and $\boldsymbol{\nu}_{pr} = \log \boldsymbol{\sigma}^2_{pr}$, respectively. Computing the sample mean and covariance of $\boldsymbol{\nu}_{pr}$ rather than of $\boldsymbol{\sigma}^2_{pr}$ allows us to use a multivariate normal distribution as a proposal distribution for $\boldsymbol{\nu}$ in stage II.

3.2 Stage II: variable r

Stage II consists of a reversible jump step, corresponding to the values r and p of the unknown number of components and autoregressive lag. Specifically, the Metropolis-Hastings step is performed as follows.

1. Draw $r^{(n)}$ from a discrete uniform distribution over $\{1, 2, \ldots, R\}$. Here and in the following steps, the superscript (n) on a parameter denotes a newly proposed value for that parameter.

2. Draw $p^{(n)}$ from the posterior distribution of the lag over $\{1, 2, \ldots, P\}$ obtained from Stage I.

3. Draw a vector $\boldsymbol{\phi}^{(n)}$ from the multivariate normal distribution $N(\hat{\boldsymbol{\phi}}_{p^{(n)}r^{(n)}}, \hat{\Sigma}_{\phi_{p^{(n)}r^{(n)}}})$, where $\hat{\boldsymbol{\phi}}_{p^{(n)}r^{(n)}}$ and $\hat{\Sigma}_{\phi_{p^{(n)}r^{(n)}}}$ are from stage I.

4. Draw a vector $\boldsymbol{\delta}^{(n)}$ from the multivariate normal distribution $N(\hat{\boldsymbol{\delta}}_{p^{(n)}r^{(n)}}, \hat{\Sigma}_{\delta_{p^{(n)}r^{(n)}}})$, where $\hat{\boldsymbol{\delta}}_{p^{(n)}r^{(n)}}$ and $\hat{\Sigma}_{\delta_{p^{(n)}r^{(n)}}}$ are from stage I.

5. Draw a vector $\boldsymbol{\nu}^{(n)}$ from the multivariate normal distribution $N(\hat{\boldsymbol{\nu}}_{p^{(n)}r^{(n)}}, \hat{\Sigma}_{\nu_{p^{(n)}r^{(n)}}})$, where $\hat{\boldsymbol{\nu}}_{p^{(n)}r^{(n)}}$ and $\hat{\Sigma}_{\nu_{p^{(n)}r^{(n)}}}$ are from stage I.

Note that the maximum number of components, R, and the maximum allowable lag, P, are chosen so that the model is flexible enough to capture features which may be present in complex data, while remaining computationally tractable.

4 Forecasting

One of our goals is to improve prediction of future observations based on the mixture model (2), compared to prediction based on a model with a single component ($r = 1$). For a time series $\boldsymbol{y} = (y_1, \ldots, y_n)'$ with y_t modeled by a density function $p(y_t|y_{t-1}, \ldots, y_1; \boldsymbol{\omega})$ indexed by a parameter vector $\boldsymbol{\omega}$, the k-step-ahead predictive distribution is

$$p(y_{n+k}|\boldsymbol{y}) = \int p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{y}) d\boldsymbol{\omega} \ ,$$

where $p(\boldsymbol{\omega}|\boldsymbol{y})$ is the posterior distribution of $\boldsymbol{\omega}$. For an autoregressive model of order p, the density $p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})$ is normal with mean and variance μ_k and σ_k^2, respectively, which are computed using the Kalman filter. Using the MCMC iterates $(p^{(l)}, r^{(l)}, \boldsymbol{\delta}_{pr}^{(l)}, \boldsymbol{\omega}_{pr}^{(l)})$, $l = 1, \ldots, M$, a k-step-ahead prediction based on the mixture model is

$$\widehat{p}(y_{n+k}|\boldsymbol{y}) = \frac{1}{M} \sum_{l=1}^{M} \sum_{j=1}^{r^{(l)}} \pi_{jpr}(t_{n+k}|\boldsymbol{\delta}_{pr}^{(l)}) p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega}_{jpr}^{(l)}) \ , \tag{8}$$

where M is the number of iterates used, and t_{n+k} is the time corresponding to y_{n+k}. To quantify the distance between a known normal predictive density $p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})$ and its estimate $\widehat{p}(y_{n+k}|\boldsymbol{y})$ based on (8), we use the Kullback-Leibler (KL) divergence, given by

$$KL(\widehat{p}(y_{n+k}|\boldsymbol{y}), p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})) = \int p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega}) \log \frac{\widehat{p}(y_{n+k}|\boldsymbol{y})}{p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})} dy_{n+k} \ . \tag{9}$$

Note that $p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})$ is the predictive density evaluated at the known true parameter ω. This divergence satisfies $KL(\widehat{p}, p) \leq 0$ with equality if and only if the two densities are equal. For a normal density $p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})$, the integral in (9) can be approximated by

$$KL_{GH}(\widehat{p}(y_{n+k}|\boldsymbol{y}), p(y_{n+k}|\boldsymbol{y}, \boldsymbol{\omega})) = \frac{1}{\sqrt{\pi}} \sum_{m=1}^{N} w_m g(\mu_k + \sqrt{2}\sigma_k u_m) \ , \tag{10}$$

which is an N-point Gauss-Hermite quadrature. In (10), $g(\cdot) = \log \frac{\widehat{p}(\cdot|\boldsymbol{y})}{p(\cdot|\boldsymbol{y},\boldsymbol{\omega})}$, and w_m and u_m are constants depending on N, which can be found for example in Abramowitz and Stegun (1965). The KL divergence can also be approximated via Monte Carlo integration as

$$KL_{MC}(\widehat{p}(y_{n+k}|\boldsymbol{y}), p(y_{n+k}|\boldsymbol{y},\boldsymbol{\omega})) = \frac{1}{I}\sum_{i=1}^{I} g(y_{n+k}^{(i)}) \ , \tag{11}$$

where $y_{n+k}^{(i)}$ is drawn from $p(y_{n+k}|\boldsymbol{y},\boldsymbol{\omega})$, $i = 1, \ldots, I$.

5 Simulations

5.1 The effect of segmentation

In Section 2, we saw, based on a single realization, that segmentation leads to improved estimates of the mixing weights in the model which in turn results in better forecasts and spectral estimates. This section describes results based on 50 realizations from model (6). Two different segment lengths are used, $L = 10$ and $L = 5$. Figure 5 displays the mixing weights against (rescaled) time (or segments) for all 50 realizations without (left) and with (right) segmentation, when $L = 10$. The figure shows that segmentation leads to better-behaved mixing weights for all the 50 realizations.

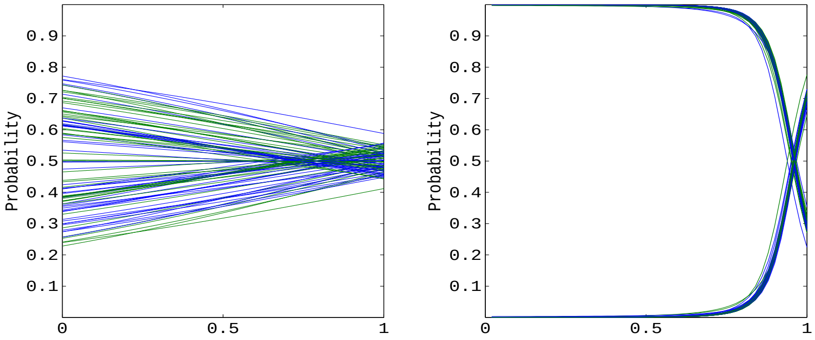

Figure 5: The mixing weights for all 50 samples without segmentation (left) and with segmentation (right), where $L = 10$.

For each realization and segment length, we compute the predictive densities for k-step-ahead forecasts, where $k = 1, \ldots, 5$. Thus, for each value of k, $k = 1, \ldots, 5$, we compare the predictive density for $L = 1$ with the ones corresponding to $L = 5$ and $L = 10$. For a given realization and a given value of k, the Kullback-Liebler divergence is computed for the pair of densities corresponding to $L = 1$ and $L = 5$, as well as for the pair corresponding to $L = 1$ and $L = 10$. Figures 6 and 7 display boxplots of

$$100\Delta\log(|KL|) = 100(\log(|KL_1|) - \log(|KL_L|)), \tag{12}$$

where the subscripts 1 and L indicate no segmentation and segmentation with segments of length L, respectively. A positive value of expression (12) indicates a reduction in the absolute value of the KL divergence as a result of segmentation, compared to no segmentation. When $L = 10$, the median and third quartile of $100\Delta\log(|KL|)$ are positive for all values of k. The third quartile of $100\Delta\log(|KL|)$ is greater than 100 for all values of k, which means that segmentation leads to reduction in the absolute value of the KL divergence by a factor of almost 3. When $L = 5$, the third quartile of $100\Delta\log(|KL|)$ exhibits a similar behavior to the third quartile when $L = 10$, but the first quartile and the median do not always indicate improvement. More details are given in Table B.1 of Appendix B. For each k and L, $k = 1, \ldots, 5$, $L = 5, 10$, the middle line of this table presents the 25th, 50th and 75th percentiles (across the 50 realizations) of criterion (12). For each k and L, the first and third lines are, respectively, the lower and upper confidence limits of 95% confidence intervals for the respective percentiles. These confidence limits were obtained by the bootstrap percentile method using 2000 bootstrap samples.

Section 2.5 shows that segmentation also improves spectral density estimation. To compare the estimates of the local spectral densities with and without segmentation, we use the $L2$ distance between the true log spectral density and its estimate. By analogy to equation (12), we then compute $100\Delta\log(L2)$. Figure 8 presents boxplots of $100\Delta\log(L2)$ for ten different segments, as described in Section 2.5. It is evident that segmentation leads to significant improvement in estimating the local spectral densities. More details are presented in Table B.2.

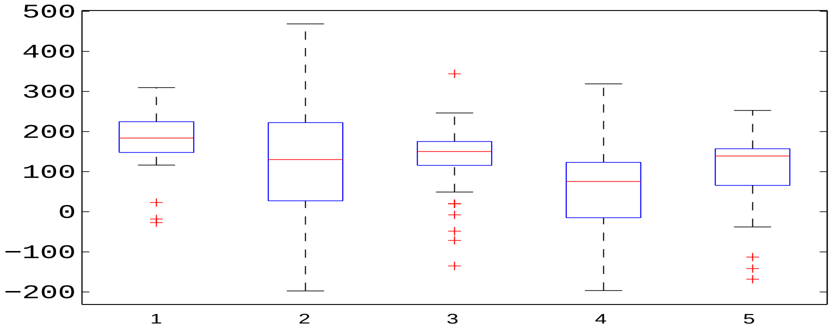

Figure 6: Boxplots of $100\Delta \log(|KL|)$ for the k-step-ahead forecasts, $k = 1, \ldots, 5$, based on the simulated data from model (6). The segment length used was $L = 10$.

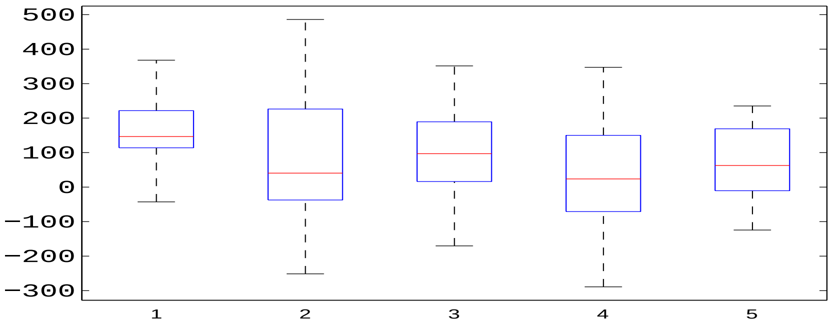

Figure 7: Boxplots of $100\Delta \log(|KL|)$ for the k-step-ahead forecasts, $k = 1, \ldots, 5$, based on the simulated data from model (6). The segment length used was $L = 5$.

5.2 Simulation from a model with a single component

Fifty time series, each of length 1008, are generated from an AR(3) model with parameter values $\sigma^2 = 56.73$ and $\boldsymbol{\phi} = (-0.0178, 0.4408, 0.1657, 0.1603)'$. These values are obtained by fitting a single AR(3) component to the Southern Oscillation Index data (http://www.bom.gov.au/climate/current/soihtm1.shtml). The segment length, maximum number of components and maximum lag length were 12, 2 and 10, respectively. Figure 9 displays the posterior probability of the number of components based on all 50 simulated times series. The posterior probability of a single component is 0.96. For each simulated time series, the spectral density was estimated twice; one estimate was based on a single-

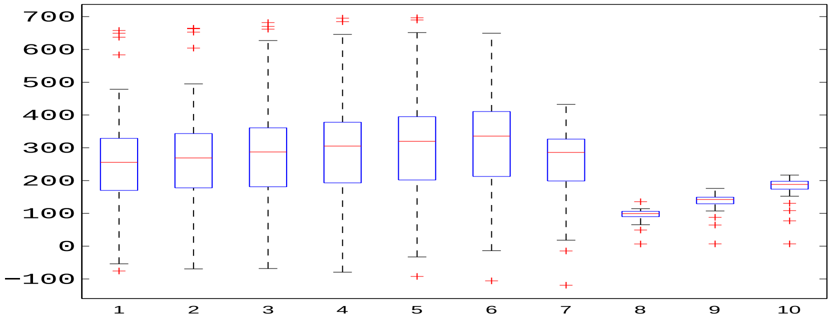

Figure 8: Boxplots of $100\Delta\log(L2)$ for the log spectral densities in ten different segments for the simulated data from model (6).

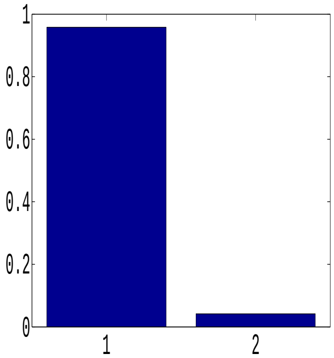

Figure 9: Posterior probability of the number of components for the simulated AR(3) time series

component, i.e., a standard autoregressive model, and the other estimate was based on the mixture model with unknown number of components and lag length. For each of these two estimates, the $L2$ distance between the true spectral density and the estimate was computed. Let $L2_1$ and $L2_r$ denote the $L2$ distances corresponding to a single-component model, and an r-component mixture model, respectively, where r is a random variable taking values in $\{1, 2\}$. Figure 10 displays a boxplot of the differences $L2_1 - L2_r$. It is seen that although in this case the underlying process is a single autoregression, fitting the mixture model ($r > 1$) still yields estimates which are not significantly different from estimates obtained using a single autoregressive process. This is not surprising given that the average posterior probability $\Pr(r = 1|\boldsymbol{y})$ has an average of 0.96 across the 50 replications. We tried other more

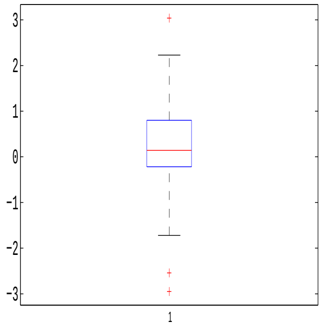

Figure 10: Boxplot of the difference $L2_1 - L2_r$.

complex single-component generating processes and found that our methodology produces spectral density estimates close to those obtained by fitting a single-component model to the data. For more details, see Appendix C.

5.3 Simulation from a model with multiple components

This section analyzes 50 time series, each containing 2000 observations, generated from the piecewise autoregressive model described by

$$y_t = \begin{cases} \sum_{k=1}^{6} \phi_{k1} y_{t-k} + \sigma_1 \epsilon_t^{(1)} & \text{for} \quad 1 \leq t \leq 200 \\ \sum_{k=1}^{6} \phi_{k2} y_{t-k} + \sigma_2 \epsilon_t^{(2)} & \text{for} \quad 201 \leq t \leq 1000 \\ \sum_{k=1}^{6} \phi_{k3} y_{t-k} + \sigma_3 \epsilon_t^{(3)} & \text{for} \quad 1001 \leq t \leq 1300 \\ \sum_{k=1}^{6} \phi_{k4} y_{t-k} + \sigma_4 \epsilon_t^{(4)} & \text{for} \quad 1301 \leq t \leq 1600 \\ \sum_{k=1}^{6} \phi_{k5} y_{t-k} + \sigma_5 \epsilon_t^{(5)} & \text{for} \quad 1601 \leq t \leq 2000, \end{cases}$$

with parameter values

j	ϕ_{1j}	ϕ_{2j}	ϕ_{3j}	ϕ_{4j}	ϕ_{5j}	ϕ_{6j}	σ_j
1	0.8874	-0.8523	0.2484	-0.6520	0.3224	-0.3287	0.0429
2	0.6955	-0.5518	0.3117	-0.6293	0.1137	-0.1003	0.0169
3	1.3415	-1.3702	0.8900	-0.9627	0.5807	-0.4173	0.0686
4	0.9776	-0.8560	0.4272	-0.6103	0.2016	-0.1631	0.0326
5	0.7995	-0.6821	0.2463	-0.5712	0.1656	-0.2169	0.0188

This piecewise autoregressive model is based on the model estimated for the explosion data and reported in Section 6. The segment length, maximum number of components and maximum lag length are 20, 6 and 10, respectively. Figure 11 presents boxplots of criterion (12)

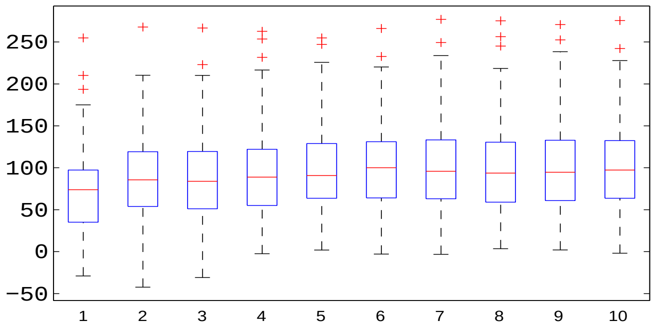

Figure 11: Boxplots of $100\Delta \log(|KL|)$ for the k-step-ahead forecasts, $k = 1, \ldots, 10$, for simulated data similar to the explosion data.

for k-step-ahead forecasts, $k = 1, \ldots, 10$. This figure shows that the mixture model leads to significantly better k-step-ahead forecasts compared to a model with a single autoregressive component. More details are given in Table B.3.

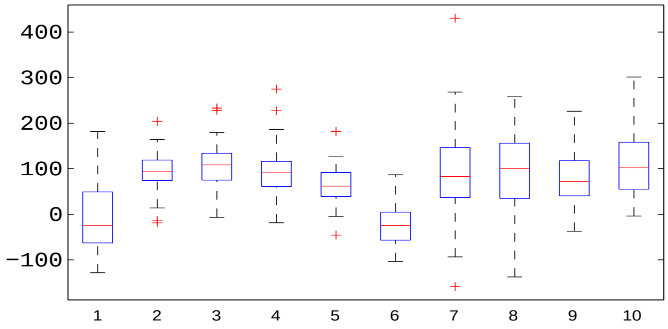

Figure 12: Boxplots of $100\Delta \log(L2)$ for the log spectral densities in ten different segments for the simulated data similar to the explosion data.

Based on the parameter estimates from each simulated data set, we also estimate the log spectral density in ten representative segments for (i) a model with a single component and for (ii) a model where the number of components is allowed to vary. For each of the ten segments and models (i) and (ii) we then compute the $L2$ distance between the estimate and the true log spectral density. Figure 12 presents boxplots of $100\Delta \log(L2)$ for each of the 10 segments and shows that in 8 of the 10 segments, model (ii) resulted in improved estimates of the log spectral densities. Table B.4 gives more details.

6 Application

This section analyzes the seismic traces of a mining explosion and an earthquake. A seismic trace is a plot of the earth's motion over time. The data presented here are measurements of the earth's vertical displacement where the recording frequency is 40 per second. The datasets are from a recording station in Scandinavia, and are reported by Shumway and Stoffer (2006). Plots of the seismic traces of the explosion and earthquake appear in figures 13 and 14, respectively. Both earthquake and explosion seismic traces consist of two waves, the compression wave, also known as primary or P wave, which occurs at the start of the series shown in Figure 13, and the shear, or S wave, which arrives at the midpoint of the series. Our analysis has two goals. The first is to obtain an estimate of the time-varying log spectrum of the process, and the second is to distinguish between a nuclear explosion and an earthquake.

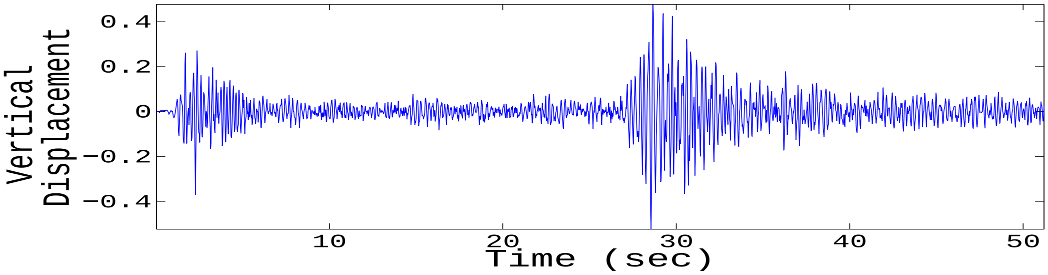

Figure 13: Seismic trace of an explosion.

Our analysis sets the maximum number of components to 6 and the maximum number of lags to 10. Two segmentation schemes are used; the first consists of 12 observations per segment, and the second consists of 20 observations per segment. We present the results from the first scheme because the results for the two schemes are almost identical. Table 1 shows the posterior probability of the number of components for both datasets. Interestingly,

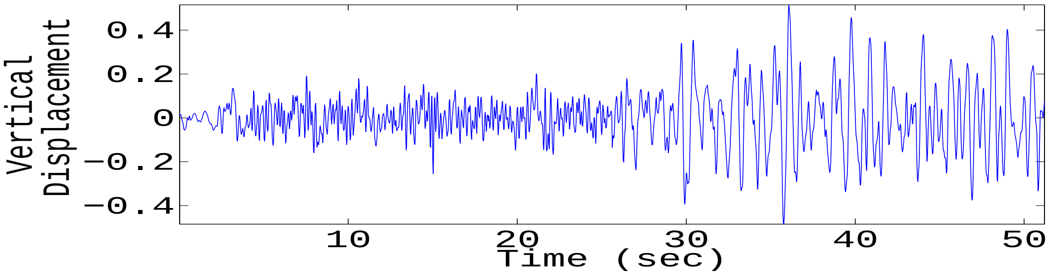

Figure 14: Seismic trace of an earthquake.

Number of components	Posterior Probability	
	Explosion	Earthquake
1	0.00	0.00
2	0.00	0.05
3	0.00	0.95
4	0.02	0.00
5	0.98	0.00
6	0.00	0.00

Table 1: Posterior probability of the number of components

the number of components with the highest posterior probability for the nuclear explosion was 5 and not 4 as might be expected from Figure 13. This is primarily because the 5-component mixture selected 6 as the maximum number of lags, whereas the 4-component mixture selected 10. The number of parameters needed to prescribe a 5-component mixture with 6 lags is 48, while the number of parameters needed to prescribe a 4-component mixture with 10 lags is 54. Thus, the 5-component mixture is more parsimonious.

Figure 15 shows the mixing functions for the explosion data for the 5-component mixture. Although it is difficult to attribute features of the time series to individual components, it appears that the component indicated by the solid thin line in Figure 15 corresponds to the high-noise component of the P wave, while the component indicated by the dashed line corresponds to the low-noise component of the P wave. The components indicated by the dotted and solid thick lines appear to be capturing the high- and low-noise components of the S wave, respectively. The dotted-dashed line appears to be capturing the transition from the high-noise component to the low-noise component of the S wave. Note that the transition from one component to another is gradual, and therefore a technique based on piecewise segments cannot capture these gradual transitions.

Given that the components identified in this example may correspond to differing levels of σ^2, we reran the analysis for the five-component mixture, varying the prior for σ^2. In particular, we increased/decreased the values of the hyper-parameters, α and β, by an order of magnitude (0.1/0.001) and found that the results were insensitive to these changes.

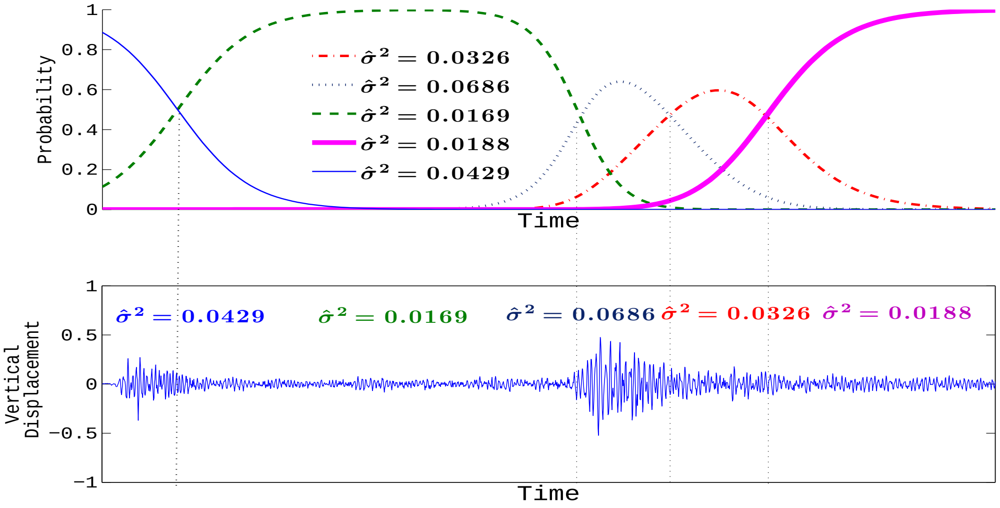

Figure 15: Mixing functions of the 5-component mixture for the nuclear explosion data. The $\hat{\sigma}^2$'s in the legend are estimates of the σ^2's for the components as given by equation (3) in Section 2.2.

Figure 16 shows the estimated time-varying log spectrum of the explosion time series. The top panel corresponds to the P wave, while the bottom panel corresponds to the S wave. The top panel of Figure 16 shows that the power of the spectrum of the P wave decreases over time; the power of the spectrum in the first plot of the top panel is twice that of the next four plots. The estimated log spectrum of the S wave is similar to that of the P wave, with two exceptions. First, the second peak of the estimated log spectrum of the S wave gradually diminishes, whereas the second peak of the P wave remains, and second, the power of the S wave is higher and lasts longer than that of the P wave.

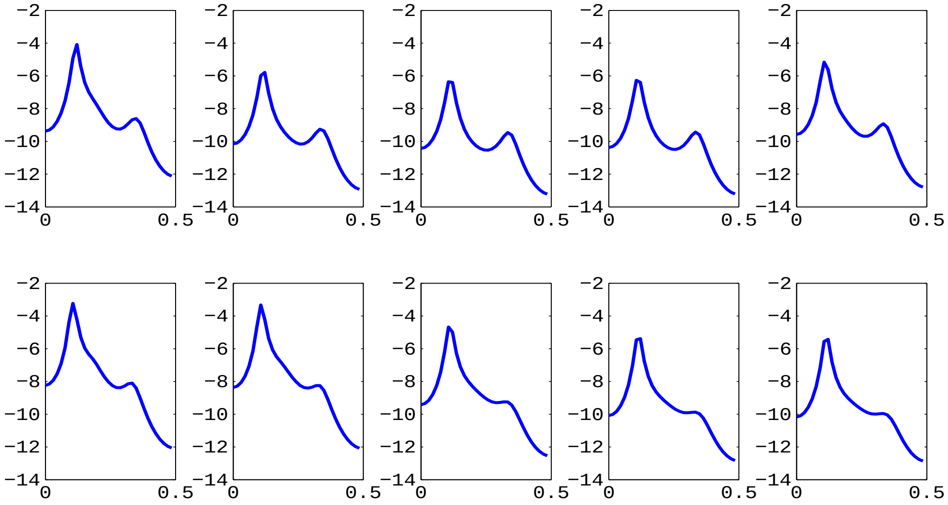

Figure 16: Estimated time-varying log spectrum for the explosion data.

Figure 17 depicts the mixing functions for the earthquake data for the 3-component mixture. This figure shows clearly that there is a strong correspondence between attributes of the time series and the individual components in the mixture. The components indicated by the solid and dashed lines appear to be capturing the P and S waves respectively, while the third component, indicated by the dashed-dotted line captures the transition from the P to the S wave. The correspondence between attributes of the time series and the components is more transparent because in contrast to the explosion data, the transition from one component to another is sudden.

Figure 18 shows the estimated time-varying log spectrum for the earthquake time series. The most notable feature of this figure is the marked difference in the estimated log spectra between the P wave (top panel) and the S wave (bottom panel). The first peak of the P wave covers a broader range of frequencies than the first peak of the S wave. In addition, while there is a second peak in power at higher frequencies for the P wave, there is no second peak for the S wave. However, the estimated log spectra for the P and S waves seem to suggest that within each wave, the time series is stationary. This observation is supported by the fact that the mixing functions for the earthquake data are constant within each wave.

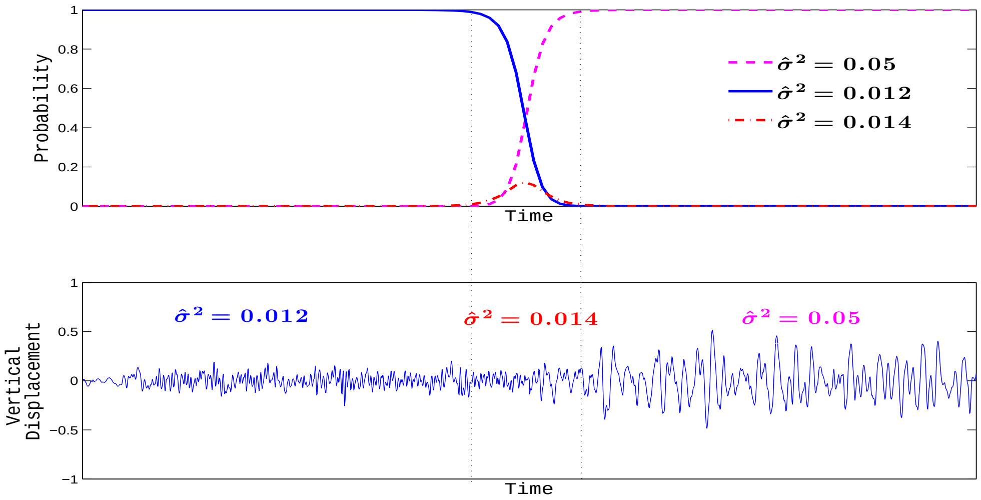

Figure 17: Mixing functions of a 3-components mixture for the earthquake data. The $\hat{\sigma}^2$'s in the legend are estimates of the σ^2's for the components as given by equation (3) in Section 2.2.

To ensure that the choice of $R = 6$ and $P = 10$ provides sufficient flexibility, the sensitivity of the priors on the number of lags p and the number of components r on the posterior distribution of these parameters is explored for the earthquake data. These data were chosen because the posterior mode of the number of lags is on the boundary of the support for the prior, that is $\Pr(p = 10|Y) = 0.98$, while the prior for p is a discrete uniform with a maximum allowable value of 10. Two sets of priors for the number of lags are considered; a discrete uniform with the maximum number of components varying from 10 to 15, and a Poisson prior with three values of λ. Table 2 shows that the posterior distribution of the number of lags is not sensitive to the prior.

The sensitivity of the posterior of the number of components to the prior was also examined. In addition to the discrete uniform described above, a Poisson prior with $\lambda = 2$ and $\lambda = 3$ was also used. For this analysis, we used a discrete uniform prior for p with $P = 12$. The results were identical for all three priors and so are not reported.

Both the explosion and the earthquake data are nonstationary time series. The major difference between the two is the source of the nonstationarity. The earthquake data suggest

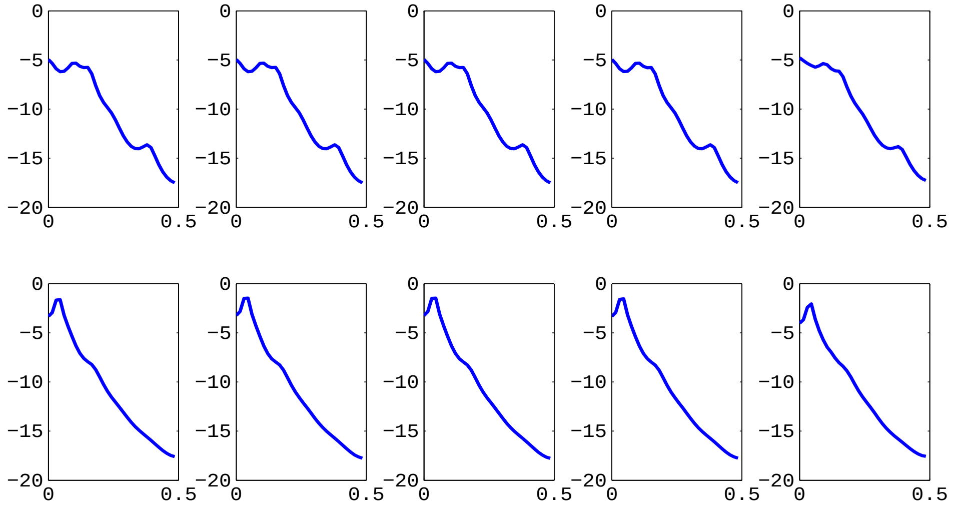

Figure 18: Estimated time-varying log-spectrum for the earthquake data.

Prior on p	Posterior Probability of p				
	9	10	11	12	13
Discrete Uniform					
$P = 10$	0.020	0.980	0.000	0.000	0.000
$P = 12$	0.000	0.982	0.010	0.008	0.000
$P = 15$	0.001	0.979	0.012	0.008	0.000
Poisson					
$\lambda = 3$	0.013	0.995	0.003	0.001	0.000
$\lambda = 4$	0.002	0.993	0.003	0.001	0.000
$\lambda = 5$	0.002	0.992	0.004	0.002	0.000

Table 2: Posterior probability of the number of lags, p, for different priors on p, for a mixture of 3 components using the earthquake data.

that the source of the nonstationarity is attributable to the arrival of the different waves, namely the P and S waves. However, within the P or S waves, the series appears stationary; the mixing functions are constant over time and the estimated time-varying log spectrum does not appear to change. In contrast, the nonstationarity of the explosion data is apparent both between and within the P and the S waves. Figure 15 shows that the mixing functions vary across time within each of the P and S waves, as well as across time between the two waves. These observations are consistent with those of Korrat et al. (2008), who state that one of the defining features of nuclear explosions is that the energy from the P and S waves decays more rapidly than it does for earthquakes.

Supplemental Materials

Data and Computer Code: The data and Matlab code for implementing the methods described in the article are available in a single archive. Please read the README file contained in the zip file for more details (code_data_wood_etal.zip).

Appendix The appendix contains in a single file (appendices.pdf) details of the sampling scheme (Appendix A), simulation results (Appendix B) and an addition to Section 5.2 (Appendix C).

Acknowledgements

We thank the editor and the referee for their helpful comments and suggestions. Ori Rosen was partially supported by NSF grants DMS-0706752 and DMS-0804140. The research of Robert Kohn was partially supported by ARC grant DP0667069.

References

Abramowitz, M. and Stegun, I. (1965), *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Dover publications.

Carvalho, A. and Tanner, M. (2005), “Modeling nonlinear time series with local mixtures of generalized linear models,” *The Canadian Journal of Statistics*, 33, 1–17.

— (2006), “Modeling nonlinearities with mixtures-of-experts of time series models,” *International Journal of Mathematics and Mathematical Sciences*, 2006, 1–22.

— (2007), "Modelling nonlinear count time series with local mixtures of Poisson autoregressions," *Computational Statistics and Data Analysis*, 51, 5266–5294.

Dahlhaus, R. (1997), "Fitting time series models to nonstationary processes," *Annals of Statistics*, 25, 1–37.

Davis, R., Lee, T., and Rodriguez-Yam, G. (2006), "Structural breaks estimation for nonstationary time series models," *Journal of the American Statistical Association*, 101, 223–239.

Gerlach, R., Carter, C., and Kohn, R. (2000), "Efficient Bayesian inference for dynamic mixture models," *Journal of the American Statistical Association*, 95, 819–828.

Giordani, P. and Kohn, R. (2008), "Efficient Bayesian inference for multiple change-point and mixture innovation models," *Journal of Business and Economic Statistics*, 26, 66–77.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), "Bayesian model averaging: A tutorial (with discussion)," *Statistical Science*, 14, 382–417.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991), "Adaptive mixtures of local experts," *Neural Computation*, 3, 79–87.

Kadane, J. B. and Lazar, N. A. (2004), "Methods and criteria for model selection," *Journal of the American Statistical Association*, 99, 279–290.

Kitagawa, G. and Akaike, H. (1978), "A procedure for the modeling of nonstationary time series," *Annals of the Institute of Statistical Mathematics*, 30, 351–363.

Korrat, I., Gharib, A., Abou Elenean, A., Hussein, H., and Gabry, M. (2008), "Spectral characteristics of natural and artificial seismic events in the Lop Nor test site China," *Acta Geophysica*, 56, 344–356.

Lau, J. and So, M. (2008), "Bayesian mixture of autoregressive models," *Computational Statistics and Data Analysis*, 53, 38–60.

Marin, J. and Robert, C. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer.

Ombao, H., Raz, J., Von Sachs, R., and Malow, B. (2001), “Automatic statistical analysis of bivariate nonstationary time series,” *Journal of the American Statistical Association*, 96, 543–560.

Prado, R. and Huerta, G. (2002), “Time-varying autoregressions with model order uncertainty,” *Journal of Time Series Analysis*, 23, 599–618.

Prado, R., Molina, F., and Huerta, G. (2006), “Multivariate time series modeling and classification via hierarchical VAR mixtures,” *Computational Statistics and Data Analysis*, 51, 1445–1462.

Punskaya, E., Andrieu, C., Doucet, A., and Fitzgerald, W. (2002), “Bayesian curve fitting using MCMC with applications to signal segmentation,” *IEEE Transactions on Signal Processing*, 50, 747–758.

Rosen, O., Stoffer, D., and Wood, S. (2009), “Local spectral analysis via a Bayesian mixture of smoothing splines,” *Journal of the American Statistical Association*, 104, 249–262.

Shumway, R. and Stoffer, D. S. (2006), *Time Series Analysis and Its Applications: With R Examples*, Springer, 2nd ed.

West, M., Prado, R., and Krystal, A. (1999), “Evaluation and comparison of EEG traces: latent structure in non-stationary time series,” *Journal of the American Statistical Association*, 94, 1083–1095.

Wong, C. and Li, W. (2001), “On logistic mixture autoregressive model,” *Biometrika*, 88, 833–846.

Zeevi, A., Meir, R., and Adler, R. (1996), “Time series prediction using mixtures of experts,” in *Proceedings of Advances in Neural Information Processing Systems*, MIT Press.