MCMC methods for restoration of nonlinearly distorted autoregressive signals

Paul T. Troughton and Simon J. Godsill

Signal Processing Group, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK

Abstract

We approach the problem of restoring distorted autoregressive (AR) signals by using a cascade model, in which the observed signal is modelled as the output of a nonlinear AR (NAR) process excited by the linear AR signal we are attempting to recover.

The Volterra expansion of the NAR model has a very large number of possible terms even when truncated at fairly small maximum polynomial degrees and lags. We address the problem of subset selection and uncertainty in the nonlinear stage and model order uncertainty in the linear stage through a hierarchical Bayesian approach. A Markov chain Monte Carlo (MCMC) approach is used for implementation, with reversible-jump moves for the linear model order and a rapidly mixing Gibbs sampler for subset selection in the nonlinear model stage, exploiting the partially analytic properties of the model.

We demonstrate the method using both synthetic AR data and an audio extract, and extend the approach to process a long distorted audio time series, for which the source model cannot be considered to be time-invariant.

Key words: Stochastic algorithms; Reversible-jump Markov chain Monte Carlo (MCMC); Gibbs sampler; Bayesian model selection; Subset selection; Volterra expansion; Polynomial autoregressive models; Cascade modelling

1 Introduction

Autoregressive (AR) processes are widely used to model a variety of signals, including audio (see *e.g.* [17, 34, 48]), speech (see *e.g.* [9]) and statistical and econometric time series (see *e.g.* [5, 49]). Their source-filter structure is analogous to the physical processes involved in speech and many musical instruments [9]. We consider the problem of reconstructing such a signal from a

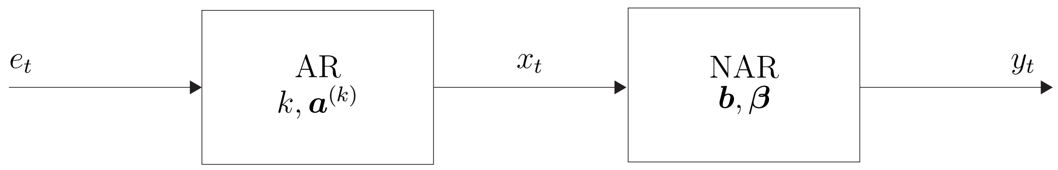

Fig. 1. Cascade of AR source model and NAR distortion model.

nonlinearly distorted version of it, when neither the precise form of the nonlinearity nor the order of the AR process is known.

The intended application is the restoration of nonlinearly distorted audio signal. Most previous work on this problem has only considered memoryless nonlinearities [32, 36, 50]. These methods have had limited success, as many practical sources of nonlinear distortion, such as transducer overloading and tape saturation, are not memoryless. We allow for memory effects by using a nonlinear autoregressive (NAR) model [8, 42] for the distortion process. The NAR model incorporates a nonlinear function which is approximated by a Volterra polynomial expansion [40] of past sample values.

A Volterra expansion to the degree and lag necessary to model adequately a given nonlinearity can have a very large number of terms. It has been found in various applications (see *e.g.* [4]) that of the order of 10 terms are sufficient to model even highly nonlinear systems. Incorporating all of the terms would lead to severe overfitting. Hence there is a need to select a subset.

Previous work on this problem in the audio area [32] has used suboptimal stepwise search methods and experienced problems with local minima. We present here a fully Bayesian approach, implemented using MCMC methods (see *e.g.* [15]), which has the advantage, in the case of model uncertainty, that model mixing can be used, in which the reconstruction is based on all the possible models, weighted according to their posterior probabilities, rather than just the single most probable one.

2 Model framework

As shown in Fig. 1, we use a cascade model consisting of an AR source model driving a NAR channel model [32].

2.1 Linear AR source model

A linear AR model of order k can be expressed as:

$$x_t = e_t + \sum_{i=1}^{k} a_i^{(k)} \, x_{t-i} \tag{1}$$

where

$\{e_t\}$ is a zero-mean i.i.d. Gaussian excitation sequence
$\{a_i^{(k)}\}$ are the parameters of the AR process.

2.2 Nonlinear AR channel model

The AR model can be generalised by replacing the weighted sum of past output values with an arbitrary function [8]:

$$y_t = x_t + f(y_{t-1}, y_{t-2}, \ldots, y_{t-\eta_b}). \tag{2}$$

The discrete-time Volterra polynomial expansion [40], a "with memory" extension of the Taylor series, allows the approximation of a broad range of nonlinearities as a simple sum of nonlinear terms. It is widely used in the control literature (see *e.g.* [4]), classification problems (see *e.g.* [37]) and artificial neural networks [6, 29]. Present applications in audio include input/output modelling of audio systems (see *e.g.* [38]) and nonlinear filtering to precompensate for known loudspeaker nonlinearities [26, 41].

The Volterra expansion can be used to approximate the nonlinear function in Eq. (2) by expanding the state space to include a Volterra expansion of past output values:

$$y_t = x_t + \sum_{i=1}^{\eta_b} \sum_{j=1}^{i} b_{(i,j)} \, y_{t-i} \, y_{t-j} + \sum_{i=1}^{\eta_b} \sum_{j=1}^{i} \sum_{k=1}^{j} b_{(i,j,k)} \, y_{t-i} \, y_{t-j} \, y_{t-k} + \text{higher degree terms} \tag{3}$$

where

$\{y_t\}$ is the distorted signal we observe
$\{x_t\}$ is the undistorted signal (Eq. 1)
$\{b_{(i,j)}, b_{(i,j,k)}, \ldots\}$ are the parameters of the NAR distortion process
η_b is the maximum lag of the NAR model.

To avoid duplication of equivalent polynomial terms, this expansion is in triangular form [33]. To avoid possible identifiability problems when used in cascade with the linear AR model, purely linear terms have been excluded from the expansion [32]. As discussed in Section 6.1, this is not always necessary.

A major advantage of this model formulation is that the inverse of the nonlinear stage is a straightforward nonlinear moving average (NMA) filter,

$$x_t = y_t - \sum_{i=1}^{\eta_b}\sum_{j=1}^{i} b_{(i,j)}\, y_{t-i}\, y_{t-j} - \sum_{i=1}^{\eta_b}\sum_{j=1}^{i}\sum_{k=1}^{j} b_{(i,j,k)}\, y_{t-i}\, y_{t-j}\, y_{t-k} - \text{higher degree terms}, \quad (4)$$

which is guaranteed to be stable. Hence it is simple to reconstruct the signal $\{x_t\}$ from $\{y_t\}$ for a given set of NAR parameters, $\{b_{(i,j)}, b_{(i,j,k)}, \ldots\}$.

2.3 Subset & matrix-vector representation

For simplicity, we concatenate the nonlinear parameters of all degrees into a single vector of length n_b:

$$\boldsymbol{b} = \begin{bmatrix} b_{(1,1)} & b_{(1,2)} & \cdots & b_{(i,j,k)} & \cdots & b_{(\eta_b,\eta_b,\eta_b,\cdots\eta_b)} \end{bmatrix}^T. \quad (5)$$

To represent a subset of the nonlinear terms, we introduce a vector of binary indicators, $\boldsymbol{\beta}$, such that if $\beta_i = 1$ then the term with parameter b_i, *i.e.* the ith element in the vector $\boldsymbol{b}$, is included in the model; otherwise the term is excluded.

We can extend this to express Eq. (1) and (3) in a fully matrix-vector form:

$$\boldsymbol{e} = \mathbf{A}^{(k)}\boldsymbol{x} = \boldsymbol{x}_1 - \mathbf{X}^{(k)}\boldsymbol{a}^{(k)} \quad (6)$$
$$\boldsymbol{x} = \boldsymbol{y}_1 - \mathbf{Y}(\boldsymbol{b} \circ \boldsymbol{\beta}) \quad (7)$$

where
$\circ$ denotes the Hadamard (elementwise) product
$\boldsymbol{x}_1$ omits the first k terms of $\boldsymbol{x}$
$\boldsymbol{y}_1$ omits the first η_b terms of $\boldsymbol{y}$
$\mathbf{A}^{(k)}$ and $\mathbf{X}^{(k)}$ are matrices containing elements from $\boldsymbol{a}^{(k)}$ and $\boldsymbol{x}$, respectively, as defined in Appendix A

and $\mathbf{Y}$ is a matrix in which row $t - \eta_b$ contains the components the predictors relating to y_t, *i.e.* the Volterra expansion of $\{y_{t-\eta_b} \ldots y_{t-1}\}$, with the terms in the order that they appear in $\boldsymbol{b}$.

2.4 Likelihoods

The approximate likelihood for the linear model can be derived as:

$$p(\boldsymbol{x} \mid k, \boldsymbol{a}^{(k)}, \sigma_e^2) \approx p(\boldsymbol{x} \mid \boldsymbol{x}_0, k, \boldsymbol{a}^{(k)}, \sigma_e^2) \tag{8}$$

$$= \mathbf{N}\Big(\boldsymbol{x}_1 - \mathbf{X}^{(k)}\boldsymbol{a}^{(k)} \mid \mathbf{0}, \sigma_e^2 \mathbf{I}_{n_e}\Big) \tag{9}$$

where $\boldsymbol{x}_0$ contains the initial k elements of $\boldsymbol{x}$ and $\mathbf{I}_{n_e}$ is the identity matrix with the same number of rows as $\boldsymbol{e}$. Eq. (9) can be rearranged to show that $\boldsymbol{x}$ is coloured zero-mean Gaussian noise which is approximately described by

$$\boldsymbol{x} \sim \mathbf{N}(\boldsymbol{x} \mid \mathbf{0}, \mathbf{C}_{\boldsymbol{x}}) \qquad \text{where} \qquad \mathbf{C}_{\boldsymbol{x}}^{-1} = \frac{\mathbf{A}^{(k)T}\mathbf{A}^{(k)}}{\sigma_e^2} \tag{10}$$

and σ_e^2 is the variance of $\boldsymbol{e}$. The approximation is due to conditioning on $\boldsymbol{x}_0$, which is well known to be an insignificant end-effect unless n_e is small [5]. This could be avoided by using exact likelihood techniques (see *e.g.* [5, 21, Appendix C]), resulting in an additional term in $\mathbf{C}_{\boldsymbol{x}}$. For clarity, however, we neglect this term in this paper.

The approximate likelihood for $\boldsymbol{y}$ can hence also be expressed as a multivariate Gaussian:

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2) \approx p(\boldsymbol{y} \mid \boldsymbol{y}_0, \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2) \tag{11}$$

$$= \mathbf{N}\Big(\mathbf{A}^{(k)}(\boldsymbol{y}_1 - \mathbf{Y}_{\boldsymbol{\beta}}\boldsymbol{b}_{\boldsymbol{\beta}}) \mid \mathbf{0}, \sigma_e^2 \mathbf{I}_{n_e}\Big) \tag{12}$$

$$= \mathbf{N}(\boldsymbol{y}_1 - \mathbf{Y}_{\boldsymbol{\beta}}\boldsymbol{b}_{\boldsymbol{\beta}} \mid \mathbf{0}, \mathbf{C}_{\boldsymbol{x}}) \tag{13}$$

where $\boldsymbol{y}_0$ is the first η_b elements of $\boldsymbol{y}$ and the notation $(\cdot)_{\boldsymbol{\beta}}$ denotes a partition containing only elements corresponding to ones in $\boldsymbol{\beta}$, such that $\mathbf{Y}_{\boldsymbol{\beta}}\boldsymbol{b}_{\boldsymbol{\beta}} = \mathbf{Y}(\boldsymbol{b} \circ \boldsymbol{\beta})$.

2.5 Priors

For the order of the linear AR model, we choose a simple bounded uniform distribution:

$$p(k) = \begin{cases} \frac{1}{k_{\max}+1} & k \in \{0, \dots k_{\max}\} \\ 0 & \text{elsewhere} \end{cases} \tag{14}$$

where $k_{\max}$ is set high enough to have no effect on model selection. If more specific prior model probabilities are available, they can be used instead.

In order to perform Bayesian model selection, we must use proper priors for the model parameters. For the linear AR parameters, we use the conjugate prior, which is a multivariate Gaussian distribution:

$$p(\boldsymbol{a}^{(k)} \mid \sigma_a^2) = \mathbf{N}\Big(\boldsymbol{a}^{(k)} \mid \boldsymbol{\mu}_{\boldsymbol{pa}^{(k)}}, \mathbf{C}_{\boldsymbol{pa}^{(k)}}\Big). \tag{15}$$

In the absence of any genuine prior knowledge, we set $\boldsymbol{\mu}_{\boldsymbol{pa}^{(k)}} = \mathbf{0}$ and $\mathbf{C}_{\boldsymbol{pa}^{(k)}} = \sigma_a^2 \mathbf{I}_k$, but any multivariate Gaussian could be used without significant changes to the MCMC framework. The hyperparameter σ_a^2 is assigned the conjugate inverse Gamma prior distribution (see *e.g.* [2]):

$$p(\sigma_a^2) = \mathrm{IG}\Big(\sigma_a^2 \mid \alpha_a, \beta_a\Big). \tag{16}$$

Similarly, the prior on the excitation variance is chosen to be:

$$p(\sigma_e^2) = \mathrm{IG}\Big(\sigma_e^2 \mid \alpha_e, \beta_e\Big) \tag{17}$$

which can closely approximate the uninformative Jeffreys' prior for scale parameters as $\alpha_e \to 0$ and $\beta_e \to 0$.

For the NAR indicators, we use a simple Bernoulli prior:

$$p(\boldsymbol{\beta}) = \prod_{i=1}^{n_b} \Big(\xi \beta_i + (1-\xi)(1-\beta_i)\Big) \tag{18}$$

where the prior probability of inclusion, ξ, is set by the experimenter. For the experiments described here, $\xi = 0.1$, but the results were found to be quite insensitive to this value unless extreme values are used. Again, it is straightforward to use a more informative prior, such as one with different prior probabilities for each term or with dependence between terms, should such knowledge be available.

Independent Gaussian priors are used for the NAR parameters. We expect *a priori* that the values of the NAR parameters of different degrees will be of different magnitudes, so we partition the parameter vector by degree. For the parameters associated with terms of degree m, the prior distribution

$$p(\boldsymbol{b}_{\{m\}} \mid \sigma_b^2) = \mathbf{N}\Big(\boldsymbol{b}_{\{m\}} \mid \mathbf{0}, \tfrac{\sigma_b^2}{s_{\{m\}}} \mathbf{I}_{n_{b_{\{m\}}}}\Big) \tag{19}$$

is used. The factor $s_{\{m\}} = E(|y_t^m|)$ scales the prior distribution to have a similar effect on parameters from different degrees, independent of the scaling of the signal. This prior can also be treated as a zero-mean multivariate Gaussian with an appropriate diagonal covariance matrix, $\mathbf{C}_{\boldsymbol{pb}}$. The hyperparameter, σ_b^2, is given an inverse Gamma prior, which is again conjugate.

Although $\boldsymbol{\beta}$ and $\boldsymbol{b}$ have been treated as independent *a priori* for convenience of notation, it will be seen in Section 4.1 that it is not necessary to specify a prior on those elements of $\boldsymbol{b}$ whose corresponding indicators in $\boldsymbol{\beta}$ are zero.

2.6 Bayesian hierarchy

We wish to reconstruct the signal, $\boldsymbol{x}$. Doing this using Eq. (7) requires knowledge of $\boldsymbol{\beta}$ and $\boldsymbol{b}_{\boldsymbol{\beta}}$, whose joint posterior is:

$$p(\boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}} \mid \boldsymbol{y}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)\, p(\boldsymbol{\beta})\, p(\boldsymbol{b}_{\boldsymbol{\beta}} \mid \sigma_b^2)\, p(\sigma_b^2). \tag{20}$$

This, however, is dependent on k and $\boldsymbol{a}^{(k)}$, which are also unknown, and have posterior:

$$p(k, \boldsymbol{a}^{(k)} \mid \boldsymbol{x}, \sigma_a^2, \sigma_e^2) \propto p(\boldsymbol{x} \mid k, \boldsymbol{a}^{(k)}, \sigma_e^2)\, p(k)\, p(\boldsymbol{a}^{(k)} \mid \sigma_a^2)\, p(\sigma_a^2) \tag{21}$$

which is dependent on $\boldsymbol{x}$.

3 Subset selection problem

Bayesian inference about the subset of nonlinear model terms is made on the basis of the marginal posterior probability distribution for the subsets:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \underset{\boldsymbol{b}, \boldsymbol{a}^{(k)}, \sigma_b, \sigma_a, \sigma_e}{\int \cdots \int} \sum_{k=0}^{k_{\max}} p(\boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_a^2, \sigma_e^2 \mid \boldsymbol{y})\, d\boldsymbol{b}_{\boldsymbol{\beta}}\, d\boldsymbol{a}^{(k)}\, d\sigma_b\, d\sigma_a\, d\sigma_e. \tag{22}$$

To find the most probable subset, it is necessary to maximise this expression. Since the ultimate aim is to estimate $\boldsymbol{x}$, if this distribution is multimodal it may be better to perform Bayesian model averaging, rather than simply to reconstruct $\boldsymbol{x}$ on the basis of a single estimate of $\boldsymbol{\beta}$. To do this, we need to sum over all 2^{n_b} possible subsets. Unless n_b is small, exploring all subsets is infeasible.

3.1 Stepwise search

Mercer [32] uses a similar model for the task of restoring distorted audio signals, but his approach is very different. He uses a stepwise regression procedure to choose model terms, with maximum likelihood parameter estimates. To reduce the computation required for each step of the algorithm, Mercer uses a coarse grid search followed by a local optimisation. After performing this selection procedure for several values of k, a model is chosen using the Akaike information criterion [1].

The problem with stepwise regression and other deterministic search algorithms is that they can get stuck at local minima, *i.e.* subsets which are better than all the neighbouring ones but not the best overall. This is a particular problem in nonlinear model selection problems, as the search space tends to be highly multimodal.

3.2 Bayesian variable selection

The task of subset selection in this nonlinear model is similar to that of variable selection in standard regression modelling.

George & McCulloch's *Stochastic Search Variable Selection* (SSVS) [12, 13] uses a different parameterisation from ours, in which all candidate model terms are included at all times, but the prior on each parameter, b_i, is dependent on the associated indicator, β_i: if $\beta_i = 1$, then a standard, fairly uninformative prior is used; if $\beta_i = 0$ then a narrow prior, centred on $b_i = 0$ is used, so that the term will have minimal effect on the model. $\boldsymbol{b}$ and $\boldsymbol{\beta}$ are sampled in separate, univariate Gibbs sampler moves. Choosing the width of the narrow prior is difficult, as it must be sufficiently narrow that unselected terms have negligible effect, to prevent overmodelling, but wide enough to be overridden by a strong likelihood, so that terms stand a chance of reinclusion.

Kuo & Mallick [27, 28] incorporate indicators, $\boldsymbol{\beta}$, into a regression model in a similar manner to Eq. (7), such that terms are completely disabled when the corresponding indicator is zero. They sample each element of $\boldsymbol{b}$ and $\boldsymbol{\beta}$ in separate Gibbs sampler steps. When $\beta_i = 0$, b_i is drawn from its prior, as the data provides no information.

Geweke [14] uses a different reparameterisation: the prior $p(b_i)$ is a mixture of a Gaussian and a δ-function at $b_i = 0$. The full conditional for b_i is sampled directly in a Gibbs sampler. He also allows for joint priors, $p(\boldsymbol{b})$.

3.3 Subset AR models

Chen [7] applies SSVS to subset linear AR models. Barnett et al. [3] use a sampling scheme similar to that of [14], but reparameterise the model in terms of partial autocorrelation coefficients in order to enforce stability. This reparameterisation is not applicable to NAR models.

4 Markov chain Monte Carlo

Since we cannot evaluate the required marginal distributions analytically, we take the MCMC approach (see *e.g.* [10, 15]), in which we generate samples from the joint posterior distribution of all the parameters, from which we can then make Monte Carlo estimates. We cannot sample from the joint distribution directly, but we can use Gibbs sampling [11] or the Metropolis-Hastings algorithm [25] in order to sample only from conditional distributions of subsets of the parameters, and yet generate a Markov chain which converges in the limit to produce samples from the correct joint distribution.

4.1 Gibbs moves for subset model

In plain Gibbs sampling, each parameter is sampled from its full conditional. Where there is strong interdependence between parameters, such as between b_i and β_i, the Markov chain will tend to be slow mixing [30, 39], so long runs will be required. We hence exploit the analytic structure of the model to marginalise b_i, allowing us to sample b_i and β_i jointly, in a similar manner to [14, 20, 22]:

$$\boldsymbol{\beta}_u, \boldsymbol{b}_{\boldsymbol{\beta}_u} \sim p(\boldsymbol{\beta}_u, \boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \tag{23}$$

which can be performed in two steps:

$$\boldsymbol{\beta}_u \sim p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \tag{24}$$

$$\boldsymbol{b}_{\boldsymbol{\beta}_u} \sim p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \tag{25}$$

where the subscript $(\cdot)_u$ denotes a partition containing the n_u element(s) we are updating, and $(\cdot)_f$ contains the remainder of the elements, which remain fixed during this step. Note that Eq. (24) is a discrete distribution, and independent of $\boldsymbol{b}_{\boldsymbol{\beta}_u}$, which is analytically marginalised from Eq. (23). We allow for the possibility of $n_u > 1$ in order to be able to sample several terms jointly, as in [22, 44, 46]. Both marginalisation and joint sampling have been shown to speed the convergence of this subset selection process [43].

We derive the (discrete) distribution for $\boldsymbol{\beta}_u$ from the likelihood as follows:

$$p(\boldsymbol{y}, \boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) = \overbrace{p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)}^{\text{likelihood}} \cdot \overbrace{p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \sigma_b^2)}^{\text{prior}} \tag{26}$$

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) = \int p(\boldsymbol{y}, \boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2)\, d\boldsymbol{b}_{\boldsymbol{\beta}_u} \tag{27}$$

$$p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \cdot \underbrace{p(\boldsymbol{\beta}_u)}_{\text{prior}} \tag{28}$$

which simplifies, as shown in Appendix B, to:

$$p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \propto p(\boldsymbol{\beta}_u) \sqrt{\frac{\left|\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}{\left|\mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}} \exp\Big(\tfrac{1}{2}\boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}^T \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big) \tag{29}$$

where

$$\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} = \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{C}_{\boldsymbol{x}}^{-1} \mathbf{Y}_{\boldsymbol{\beta}_u} + \mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} \tag{30}$$

$$\boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{C}_{\boldsymbol{x}}^{-1} (\boldsymbol{y}_1 - \mathbf{Y}_{\boldsymbol{\beta}_f} \boldsymbol{b}_{\boldsymbol{\beta}_f}) \tag{31}$$

and $\mathbf{Y}_{\boldsymbol{\beta}_u}$ contains those columns of $\mathbf{Y}$ which correspond to ones in $\boldsymbol{\beta}_u$.

Step (25) is a straightforward draw from a multivariate Gaussian:

$$p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \propto \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}, \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big). \tag{32}$$

There is no need to sample those elements of $\boldsymbol{b}_u$ for which $\beta_i = 0$, as their values are never used.

Since the parameters associated with different terms may well be correlated, but it is hard to anticipate which ones will have the strongest correlations, random groups of terms are sampled jointly. The value of n_u controls the trade-off between the speed of mixing of the Markov chain and the complexity of each iteration. Groups of size three have been found to be a good compromise.

It is straightforward to sample all the NAR parameters which are currently included in the model, $\boldsymbol{b}_{\boldsymbol{\beta}}$, jointly using Eq. (32), by partitioning such that $(\cdot)_u$ contains all the terms, leaving $(\cdot)_f$ empty. Making this move occasionally can further improve mixing.

4.2 Reversible-jump moves for the AR model order

The linear AR parameter vector $\boldsymbol{a}^{(k)}$ is of unknown dimension, as the value of k is unknown. We hence use reversible-jump MCMC [24], a generalisation of the Metropolis-Hastings algorithm which allows jumps between parameter spaces of different dimensionality.

Since the full conditional for the parameter vector $\boldsymbol{a}^{(k)}$ is available analytically, we can use it to propose efficient model moves [16, 45, 47]. This also results in the acceptance probability becoming independent of the parameter values.

For a jump from an AR model of order k to one of order k', the acceptance probability is derived in [45, Appendix A] as:

$$\begin{aligned}
&\alpha(k \to k' \mid \boldsymbol{x}, \sigma_a^2, \sigma_e^2) \\
&\quad = \min\left(1, \sigma_a^{k-k'} \sqrt{\frac{\left|\mathbf{C}_{\boldsymbol{ca}^{(k')}}\right|}{\left|\mathbf{C}_{\boldsymbol{ca}^{(k)}}\right|}} \frac{J(k' \to k)}{J(k \to k')} \frac{\exp\left(\frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{ca}^{(k')}}^T \mathbf{C}_{\boldsymbol{ca}^{(k')}}^{-1} \boldsymbol{\mu}_{\boldsymbol{ca}^{(k')}}\right)}{\exp\left(\frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{ca}^{(k)}}^T \mathbf{C}_{\boldsymbol{ca}^{(k)}}^{-1} \boldsymbol{\mu}_{\boldsymbol{ca}^{(k)}}\right)}\right)
\end{aligned} \quad (33)$$

where

$$\mathbf{C}_{\boldsymbol{ca}^{(k)}}^{-1} = \frac{1}{\sigma_e^2}\mathbf{X}^{(k)T}\mathbf{X}^{(k)} + \frac{1}{\sigma_a^2}\mathbf{I}_k \quad (34)$$

$$\boldsymbol{\mu}_{\boldsymbol{ca}^{(k)}} = \frac{1}{\sigma_e^2}\mathbf{C}_{\boldsymbol{ca}^{(k)}}^T \mathbf{X}^{(k)T} \boldsymbol{x}_1 \quad (35)$$

and $J(k \to k')$ is the probability of choosing to propose this change in order. To ensure good mixing, we want most proposed jumps to be small, but occasional large ones to occur too. We choose a discretised Laplacian density, centred on k:

$$J(k \to k') \propto \exp(-\lambda \left|k' - k\right|).$$

4.3 Gibbs moves for other parameters

The hyperparameters, σ_a^2, σ_b^2 and σ_e^2, can all be sampled in Gibbs moves from their full conditionals, which are inverse Gamma distributions [45].

4.4 Sampling strategy

Since $\boldsymbol{x}$ is relatively expensive to compute, particularly when processing long signals (see Section 6.1), we separate our sampling moves into those which affect the linear AR model and those which affect the nonlinear model. In each iteration, these two groups are sampled in turn:

Choose initial values
for $i = \{1 \ldots$ number of iterations$\}$
  repeat
    Choose a group $(\cdot)_u$ of n_u nonlinear model terms to sample
    $\boldsymbol{\beta}_u \sim p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)$
    $\boldsymbol{b}_{\boldsymbol{\beta}_u} \sim p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)$
  until all nonlinear model terms have been sampled
  $\boldsymbol{b}_{\boldsymbol{\beta}} \sim p(\boldsymbol{b}_{\boldsymbol{\beta}} \mid \boldsymbol{y}, \boldsymbol{\beta}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)$
  $\sigma_b^2 \sim p(\sigma_b^2 \mid \boldsymbol{b}_{\boldsymbol{\beta}})$
  Reconstruct $\boldsymbol{x}$ from $\boldsymbol{y}$ using $\boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}$
  for $r = \{1 \ldots$ number of reversible-jump moves per iteration$\}$
    $k' \sim J(k' \mid k)$
    $z \sim U(0, 1)$
    if $z < \alpha(k \to k' \mid \boldsymbol{x}, \sigma_a^2, \sigma_e^2)$
      $k = k'$
    end if
    $\boldsymbol{a}^{(k)} \sim p(\boldsymbol{a}^{(k)} \mid \boldsymbol{x}, k, \sigma_a^2, \sigma_e^2)$
    $\sigma_a^2 \sim p(\sigma_a^2 \mid \boldsymbol{a}^{(k)})$
    $\sigma_e^2 \sim p(\sigma_e^2 \mid \boldsymbol{x}, k, \boldsymbol{a}^{(k)})$
  end for
end for

Sensible initial values for the parameters could be drawn from their prior distributions. As we are using quite uninformative priors, we choose to start with null models, *i.e.* $k = 0$ and $\boldsymbol{\beta} = \mathbf{0}$.

5 Experiments with single blocks

5.1 Synthetic data

To test the model estimation procedure, 8000 samples were generated from a synthetic AR-NAR process with AR order 6 and six nonlinear terms:

$$\begin{array}{ccc} b_{(2,2)} = -0.20 & b_{(4,1)} = 0.18 & b_{(5,4,1)} = -0.16 \\ b_{(5,5,3)} = 0.16 & b_{(5,5,5)} = 0.20 & b_{(2,2,2,1)} = -0.10. \end{array} \quad (36)$$

Fig. 2 shows the result of running the sampler for 2000 iterations with 82 candidate nonlinear terms (second and third degree to lag 6 and fourth degree to lag 2). Indicators were sampled in random triples, and eight reversible-jump moves were proposed each iteration. It was initialised with an empty model and arbitrary values for σ_e^2, σ_a^2 and σ_b^2. Because the initial values are atypical of the posterior distribution, as is usually the case with problems in

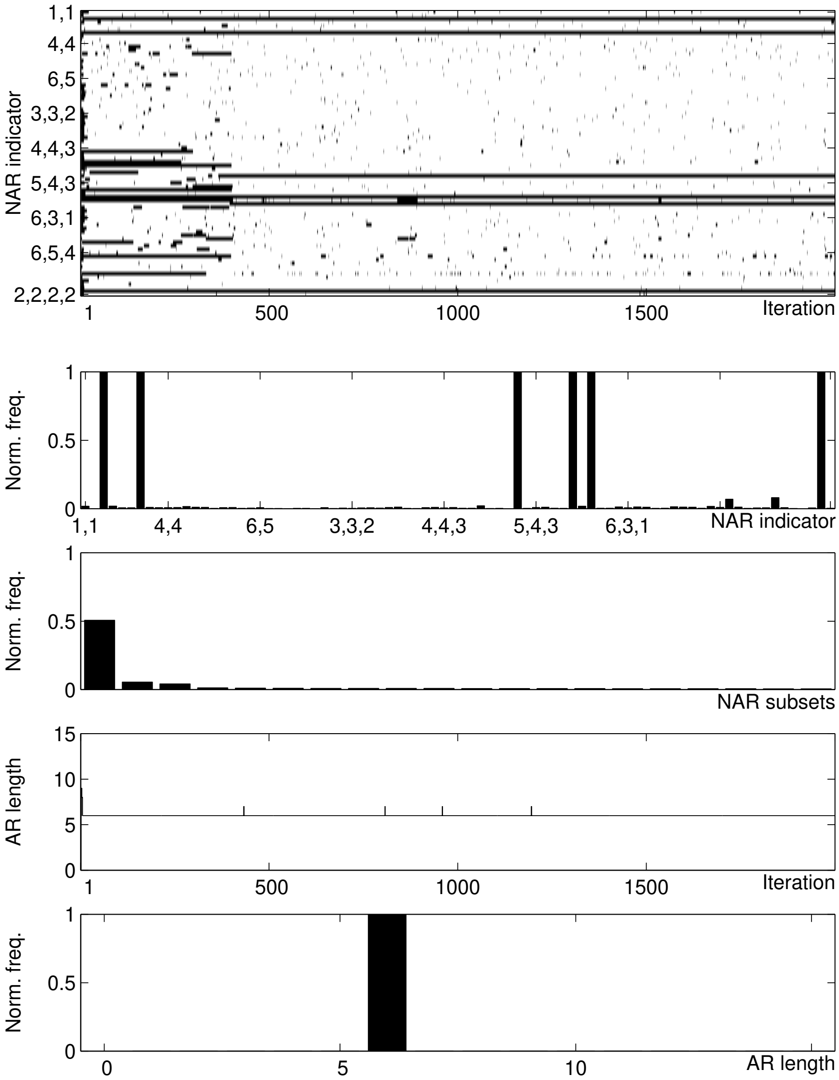

Fig. 2. Identifying synthetic AR-NAR data *(from top)*: (a) raw β values — black areas represent ones; (b) frequency of appearance of each NAR model term; (c) frequency of appearance of the most popular NAR subsets; (d) raw k values; (e) frequency of appearance of each k value.

high dimensions, the beginning of the run was discarded as 'burn-in' and only values from the final 1000 iterations were used for analysis.

Fig. 2.(d–e) shows that the sampler converged very quickly to the correct AR model order. The six nonlinear model terms which appear most frequently in the sampler output (Fig. 2.(a–b)) are correct; that subset accounts for over 50% of the iterations (Fig. 2.(c)).

Fig. 3 shows Monte Carlo estimates of posterior distributions of parameter values, produced from those iterations which selected the most popular model. It can be seen that the estimated distributions have significant mass close to the known true values. The scatter plot in Fig. 4 shows that the AR parameters

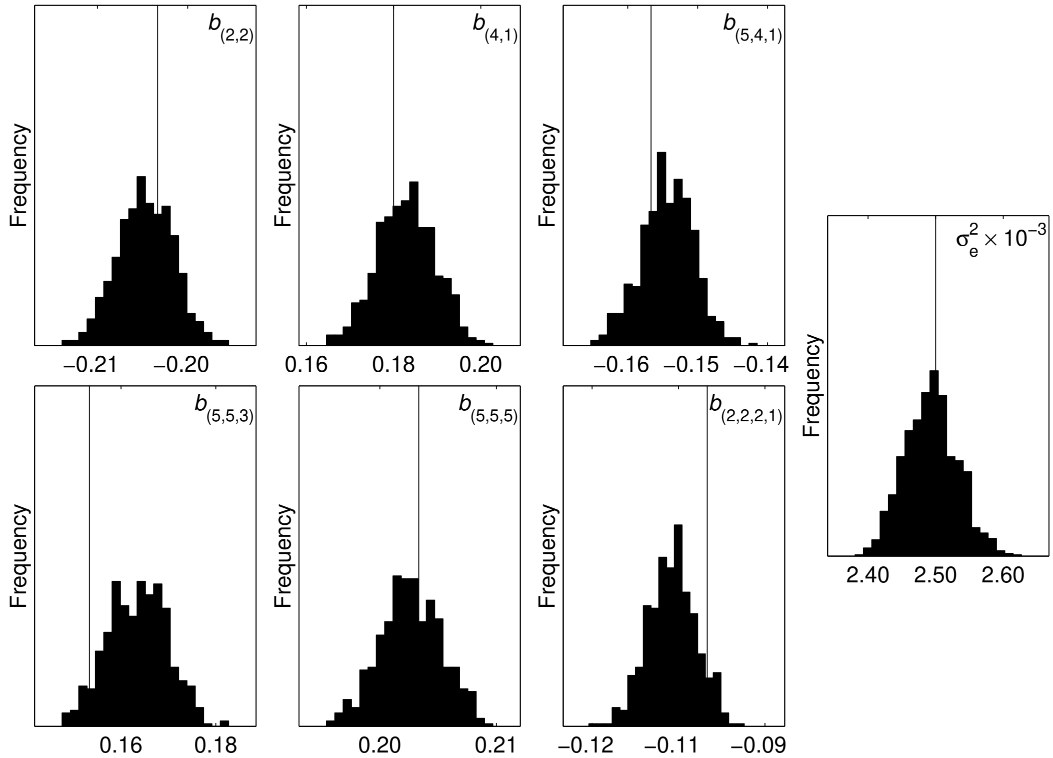

Fig. 3. Identifying synthetic AR-NAR data: histograms of σ_e^2 and NAR parameter values (true values marked by lines).

were also accurately estimated.

5.2 Synthetically distorted audio

1000 samples from a 44.1 kHz pop music recording were distorted by the following NAR filter:

$$y_t = x_t - 0.15y_{t-2}^2\, y_{t-1} + 0.2y_{t-2}\, y_{t-1}^3 - 0.08y_{t-2}^2\, y_{t-1}^3 + 0.005y_{t-2}\, y_{t-1}^4. \quad (37)$$

This introduced distortion 11 dB below the r.m.s. signal level. We call this a distortion-to-signal ratio (DSR) of -11 dB.

The sampling scheme was run for 4500 iterations, again from arbitrary initial values, with 58 candidate nonlinear terms (second and third degree up to lag 4, fourth degree up to lag 3 and fifth and sixth degree up to lag 2), and allowing AR models up to order 100. Fig. 5 shows the sampler output together with the DSR. It can be seen that after 1700 iterations the correct nonlinear terms have been chosen and the reconstruction is almost perfect — the DSR is improved to an average of around -69 dB. It is interesting that, even after as few as 100 iterations, the DSR is consistently around -34 dB, a useful

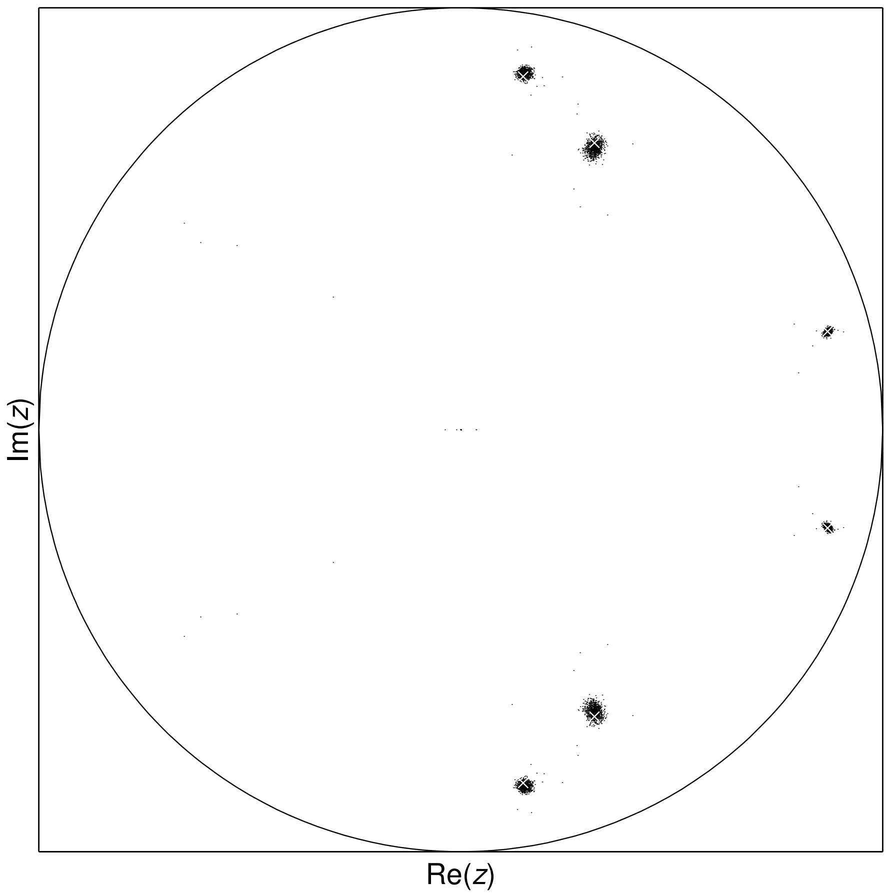

Fig. 4. Identifying synthetic AR-NAR data: sampled linear AR pole positions (true values marked by white crosses).

improvement. Fig. 6 shows the initial, distorted waveform, the reconstruction after 100 iterations and the reconstruction after 3000 iterations, each with the original, undistorted waveform for comparison.

6 Extension to long signals

A possible application of this method is the restoration of distorted audio signals. A conventional approach to speech and audio modelling is to break the signal into blocks which are sufficiently short that an assumption of stationarity is reasonable (see *e.g.* [23]). Typical block lengths are around 25 ms.

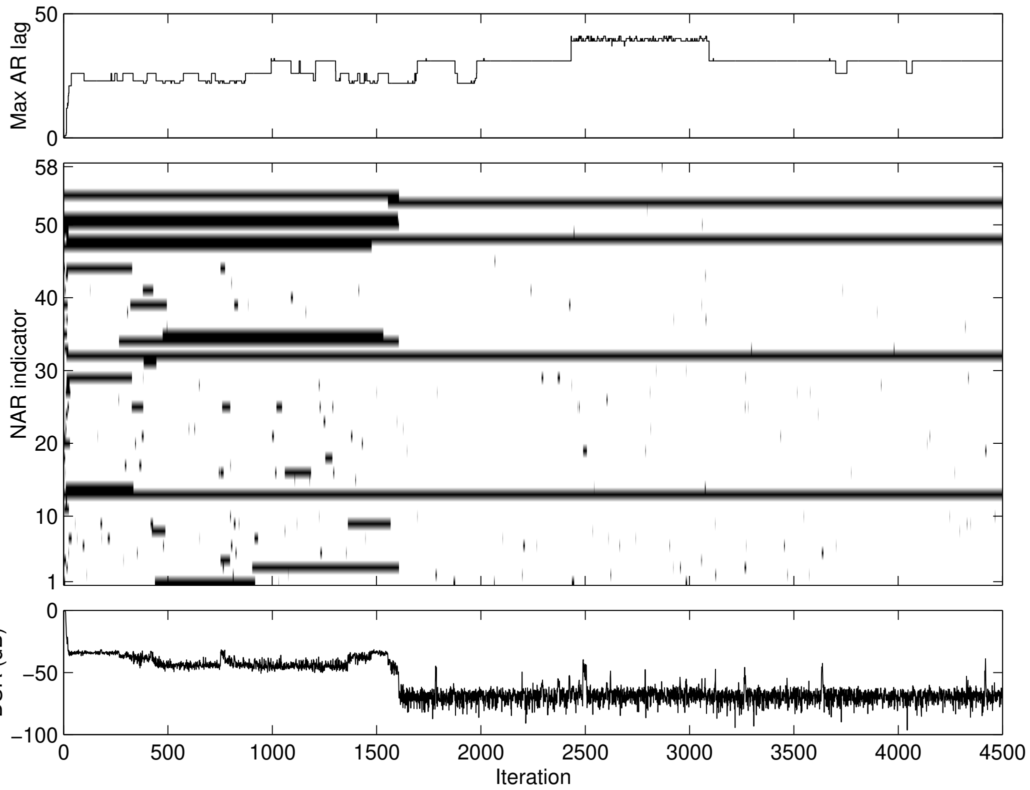

Fig. 5. Synthetically distorted pop music extract: *(from top)*: (a) AR model order, k; (b) NAR indicators, $\boldsymbol{\beta}$ (black pixels indicate $\beta_i = 1$); (c) distortion-to-signal ratio for the reconstructed signal.

6.1 Joint estimation over multiple blocks

A crude approach would be to process each block separately. With many distortion problems, however, we can expect the distortion process to remain unchanged for the duration of the signal, which could be many minutes. This can be exploited by estimating $\boldsymbol{b}$ and $\boldsymbol{\beta}$ over many (not necessarily contiguous) blocks of audio, $\boldsymbol{y}_{[r]},\ r \in \{1 \ldots R\}$, the source model for each of which has separate parameters $k_{[r]}$, $\boldsymbol{a}^{(k)}_{[r]}$ and $\sigma^2_{e\,[r]}$.

This changes the structure of $\mathbf{C}_{\boldsymbol{x}}^{-1}$ (Eq. 10), which is required when sampling the global parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}$. It can, however, be computed quickly as

$$\mathbf{C}_{\boldsymbol{x}}^{-1} = \sum_{r=1}^{R} \frac{1}{\sigma^2_{e\,[r]}} \tilde{\mathbf{A}}^{(k_{[r]})T}_{[r]} \tilde{\mathbf{A}}^{(k_{[r]})}_{[r]} \tag{38}$$

where $\tilde{\mathbf{A}}^{(k_{[r]})}_{[r]}$ is $\mathbf{A}^{(k_{[r]})}_{[r]}$ padded left and right with an appropriate number of columns of zeros. The computation required hence increases only linearly with the number of blocks, and much can be performed in parallel.

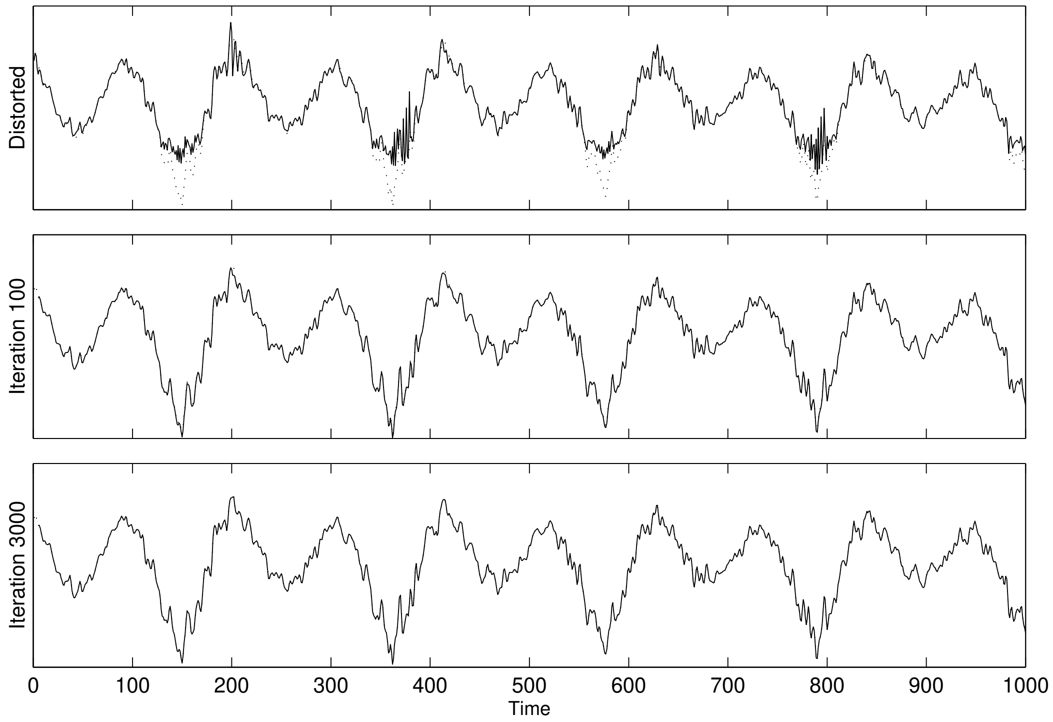

Fig. 6. Synthetically distorted pop music extract: original signal (dotted) compared with (solid) distorted signal, restoration after 100 iterations and restoration after 3000 iterations.

This multiple block technique is a powerful approach, as it allows large quantities of data to be used. As long as the signal does vary significantly over its duration, linear model terms can be permitted in the channel model without introducing ambiguity. This introduces the possibility of modelling linear phenomena such as reverberation as well as purely nonlinear effects.

6.2 Long audio signal

A four second extract from a 44.1 kHz sampled orchestral recording was artificially distorted by a NAR filter with the following parameters:

$$\begin{gathered} b_{(1,4,6)} = -0.07 \quad b_{(2,2,3)} = -0.05 \quad b_{(3,6,8)} = -0.06 \\ b_{(4,7,7)} = -0.06 \quad b_{(8,9,9)} = -0.05. \end{gathered} \tag{39}$$

The resulting signal has a DSR of -10.5 dB, and the degradation is clearly audible.

Thirty randomly chosen, non-contiguous blocks of 1000 samples were used for analysis, and the sampler was run for 500 iterations with all 220 third degree nonlinear terms up to lag 10 as candidates. The model space is sufficiently large that no subset appears more than once in the sampler output. However,

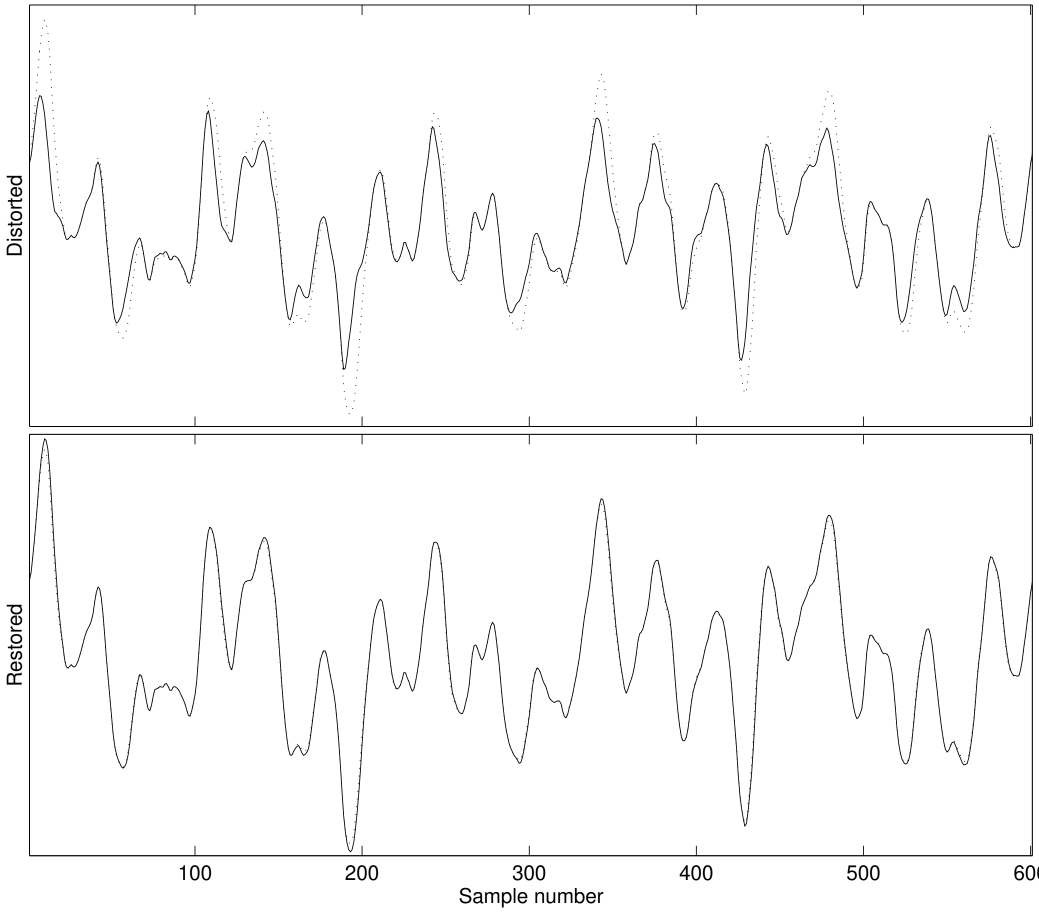

Fig. 7. Restoring a long audio signal: part of the original signal *(dotted)* with *(solid, from top)*: distorted signal $\boldsymbol{y}$; restored signal, $\hat{\mathbf{x}}$.

from the 50th iteration onwards, the DSR of the restoration is always better than -23.6 dB.

Allowing model mixing by making a Monte Carlo estimate of $\boldsymbol{x}$ directly from the final 250 iterations improves the DSR to -27.8 dB. Since the model is linear-in-the-parameters [35], this is equivalent to performing a restoration using a single NMA filter incorporating all the terms which appear in the sampler output, with the parameters averaged over all the iterations.

Fig. 7 compares the waveforms of part of the original signal with the distorted signal and the model-averaged restoration. It can be seen that the restored signal matches the original signal closely; the signals are hard to distinguish by listening.

7 Discussion

7.1 Real audio distortion

Physical distortion-producing processes, such as tape saturation, are often followed by linear processes, such are playback equalisation. Extremely large Volterra expansions can be required to model these adequately, so future work will consider modelling these explicitly as a third, either AR or MA, stage in the cascade model. The parameters of this extra stage can then be estimated jointly across the whole duration of the signal. The effects of bandlimiting could be similarly incorporated [36].

In practical applications, background noise and outliers caused by media defects are likely to be present. Robustness may therefore be improved by modelling these components explicitly. This can be done, using further indicator variables to mark outlier positions, within the same MCMC scheme [18–20,22].

7.2 Conclusions

The novel MCMC method presented here jointly estimates the structure and parameters of a cascade AR-NAR model, and allows for model mixing. Using concepts similar to our previous work [44–47], we have exploited the partially analytic structure of both the linear and nonlinear parts of the model to make joint moves, and efficient reversible-jump proposals, to speed the convergence of the Markov chain.

A Matrix definitions

The matrices $\mathbf{X}^{(k)}$ and $\mathbf{A}^{(k)}$ are rearrangements of $\boldsymbol{x}$ and $\boldsymbol{a}^{(k)}$, respectively [5]:

$$\mathbf{X}^{(k)} = \begin{bmatrix} x_k & x_{k-1} & \cdots & x_2 & x_1 \\ x_{k+1} & x_k & \cdots & x_3 & x_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n_x-2} & x_{n_x-3} & \cdots & x_{n_x-k} & x_{n_x-k-1} \\ x_{n_x-1} & x_{n_x-2} & \cdots & x_{n_x-k+1} & x_{n_x-k} \end{bmatrix} \tag{A.1}$$

$$\mathbf{A}^{(k)} = \begin{bmatrix} -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -a_k^{(k)} & \cdots & -a_1^{(k)} & 1 \end{bmatrix}. \tag{A.2}$$

B Derivation of marginal posterior for $\boldsymbol{\beta}$

In Section 4.1, it is necessary to sample from the distribution

$$\begin{aligned} &p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \\ &\qquad \propto p(\boldsymbol{\beta}_u) \int \underbrace{p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)}_{\text{Likelihood}} \, p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \sigma_b^2) \, d\boldsymbol{b}_{\boldsymbol{\beta}_u}. \end{aligned} \tag{B.1}$$

The approximate likelihood (Eq. 12) can be rewritten as:

$$p\Big(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2\Big) \approx \mathbf{N}\Big(\mathbf{A}^{(k)}(\boldsymbol{y}_1 - \mathbf{Y}_{\boldsymbol{\beta}_f}\boldsymbol{b}_{\boldsymbol{\beta}_f} - \mathbf{Y}_{\boldsymbol{\beta}_u}\boldsymbol{b}_{\boldsymbol{\beta}_u}) \mid \mathbf{0}, \sigma_e^2 \mathbf{I}_{n_e}\Big) \tag{B.2}$$

$$= \mathbf{N}\Big(\mathbf{A}^{(k)}\mathbf{Y}_{\boldsymbol{\beta}_u}\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \underbrace{\mathbf{A}^{(k)}(\boldsymbol{y}_1 - \mathbf{Y}_{\boldsymbol{\beta}_f}\boldsymbol{b}_{\boldsymbol{\beta}_f})}_{\triangleq\, \boldsymbol{e}_f}, \sigma_e^2 \mathbf{I}_{n_e}\Big) \tag{B.3}$$

such that $\boldsymbol{e}_f$ is the excitation signal which would need to be applied to the AR model if $\boldsymbol{x}$ was reconstructed from $\boldsymbol{y}$ using only the NAR model terms in the partition $(\cdot)_{\boldsymbol{\beta}_f}$, *i.e.* those corresponding to ones in $\boldsymbol{\beta}_f$. The partition $(\cdot)_{\boldsymbol{\beta}_u}$ is defined similarly.

After some manipulation, this Gaussian can be rearranged as:

$$p\Big(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2\Big) = \Upsilon_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} \, \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}, \mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big) \tag{B.4}$$

where

$$\mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \sigma_e^2 \left(\mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u}\right)^{-1} \tag{B.5}$$

$$\boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \left(\mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u}\right)^{-1} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \boldsymbol{e}_f \tag{B.6}$$

$$\Upsilon_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \frac{\exp\Big(-\frac{1}{2\sigma_e^2}\Big[\boldsymbol{e}_f^T \boldsymbol{e}_f - \boldsymbol{e}_f^T \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u} (\mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u})^{-1} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \boldsymbol{e}_f\Big]\Big)}{(\sqrt{2\pi}\sigma_e)^{n_e - n_{\boldsymbol{\beta}_u}} \sqrt{\left|\mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u}\right|}} \tag{B.7}$$

and $n_{\boldsymbol{\beta}_u}$ is the number of ones in $\boldsymbol{\beta}_u$.

Now that the likelihood is in the convenient form of a Gaussian in $\boldsymbol{b}_{\boldsymbol{\beta}_u}$, it can be multiplied by the prior on $\boldsymbol{b}_{\boldsymbol{\beta}_u}$ (Eq. 19) to form the integrand of Eq. (B.1):

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \\ &= p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}_{\boldsymbol{\beta}}, k, \boldsymbol{a}^{(k)}, \sigma_e^2)\, p(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \sigma_b^2) && \text{(B.8)} \\ &= \Upsilon_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} \, \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}, \mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big) \, \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \mathbf{0}, \mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big). && \text{(B.9)} \end{aligned}$$

Since the prior is conjugate, this product can be rearranged in the form of a single Gaussian:

$$= \Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \, \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}, \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big) \tag{B.10}$$

where

$$\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \left(\mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} + \mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1}\right)^{-1} \tag{B.11}$$

$$\boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} \boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} \tag{B.12}$$

$$\Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \frac{\Upsilon_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}}{\sqrt{2\pi}^{n_{\boldsymbol{\beta}_u}}} \sqrt{\frac{\left|\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}{\left|\mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right| \left|\mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}} \exp\Big(-\tfrac{1}{2}\Big[\boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}^T \mathbf{C}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} \boldsymbol{\mu}_{l\boldsymbol{b}_{\boldsymbol{\beta}_u}} - \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}^T \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1} \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big]\Big) \tag{B.13}$$

which simplify, after much cancellation, to:

$$\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \left(\tfrac{1}{\sigma_e^2} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u} + \mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}^{-1}\right)^{-1} \tag{B.14}$$

$$\boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \tfrac{1}{\sigma_e^2} \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \boldsymbol{e}_f \tag{B.15}$$

$$\Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} = \frac{\sqrt{\left|\mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}}{(\sqrt{2\pi}\sigma_e)^{n_e} \sqrt{\left|\mathbf{C}_{p\boldsymbol{b}_{\boldsymbol{\beta}_u}}\right|}} \exp\Big(-\tfrac{1}{2\sigma_e^2}\Big[\boldsymbol{e}_f^T \boldsymbol{e}_f - \tfrac{1}{\sigma_e^2} \boldsymbol{e}_f^T \mathbf{A}^{(k)} \mathbf{Y}_{\boldsymbol{\beta}_u} \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \mathbf{Y}_{\boldsymbol{\beta}_u}^T \mathbf{A}^{(k)T} \boldsymbol{e}_f\Big]\Big) \tag{B.16}$$

Now Eq. (B.1) can be written in the form

$$p(\boldsymbol{\beta}_u \mid \boldsymbol{y}, \boldsymbol{\beta}_f, \boldsymbol{b}_{\boldsymbol{\beta}_f}, k, \boldsymbol{a}^{(k)}, \sigma_b^2, \sigma_e^2) \\ \propto p(\boldsymbol{\beta}_u)\, \Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}} \int \mathbf{N}\Big(\boldsymbol{b}_{\boldsymbol{\beta}_u} \mid \boldsymbol{\mu}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}, \mathbf{C}_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}\Big)\, d\boldsymbol{b}_{\boldsymbol{\beta}_u} \tag{B.17}$$

but the unbounded integral is equal to unity, so

$$\propto p(\boldsymbol{\beta}_u)\, \Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}. \tag{B.18}$$

When sampling $\boldsymbol{\beta}_u$, those parts of $\Upsilon_{c\boldsymbol{b}_{\boldsymbol{\beta}_u}}$ which are independent of $\boldsymbol{\beta}_u$ can be neglected, resulting in Eq. (29).

Acknowledgements

The authors thank the anonymous reviewers for their helpful suggestions. This work was supported by the Engineering and Physical Sciences Research Council, UK.

References

[1] H. Akaike, A new look at statistical model identification, IEEE Transactions on Automatic Control AC-19 (1974) 716–723.

[2] C. Andrieu, P. M. Djurić, A. Doucet, Model selection by MCMC computation, Signal Processing (this issue).

[3] G. Barnett, R. Kohn, S. Sheather, Bayesian estimation of an autoregressive model using Markov chain Monte Carlo, Technical report, Australian Graduate School of Management, University of New South Wales (1994).

[4] S. A. Billings, S. Chen, M. J. Korenberg, Identification of MIMO non-linear systems using a forward-regression orthogonal estimator, International Journal of Control 49 (6) (1989) 2157–2189.

[5] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, Time Series Analysis: Forecasting and Control, Holden-Day, 1994, 3rd edition.

[6] J. E. Chance, K. Worden, G. R. Tomlinson, Frequency domain analysis of NARX neural networks, Journal of Sound and Vibration 213 (5) (1998) 915–941.

[7] C. W. S. Chen, On the selection of best subset autoregressive time series models, Technical report, Department of Statistics, Feng-Chia University, Taiwan (October 1996).

[8] S. Chen, S. A. Billings, Modelling and analysis of non-linear time series, International Journal of Control 50 (6) (1989) 2151–2171.

[9] J. R. Deller, J. G. Proakis, J. H. L. Hansen, Discrete-time processing of speech signals, Macmillan, 1993.

[10] W. J. Fitzgerald, Signal processing applications of Markov chain Monte Carlo methods, Signal Processing (this issue).

[11] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6 (6) (1984) 721–741.

[12] E. I. George, R. E. McCulloch, Variable selection via Gibbs sampling, Journal of the American Statistical Association 88 (423) (1993) 881–889.

[13] E. I. George, R. E. McCulloch, Stochastic search variable selection, in: Gilks et al. [15], pp. 203–214.

[14] J. Geweke, Variable selection and model comparison in regression, in: J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith (eds.), Bayesian Statistics 5, Oxford University Press, 1996, pp. 609–620.

[15] W. R. Gilks, S. Richardson, D. J. Spiegelhalter (eds.), Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics, Chapman & Hall, 1996.

[16] S. J. Godsill, Some new relationships between MCMC model uncertainty methods, Technical Report CUED/F-INFENG/TR.305, Department of Engineering, University of Cambridge (November 1997).

[17] S. J. Godsill, P. J. W. Rayner, A Bayesian approach to the restoration of degraded audio signals, IEEE Transactions on Speech and Audio Processing 3 (4) (1995) 267–278.

[18] S. J. Godsill, P. J. W. Rayner, Robust noise modelling with application to audio restoration, in: Maher [31].

[19] S. J. Godsill, P. J. W. Rayner, Robust noise reduction for speech and audio signals, Proceedings of IEEE ICASSP-96 II (1996) 625–628.

[20] S. J. Godsill, P. J. W. Rayner, Robust treatment of impulsive noise in speech and audio signals, in: J. O. Berger, B. Betro, E. Moreno, L. R. Pericchi, F. Ruggeri, G. Salinetti, L. Wasserman (eds.), Bayesian Robustness, IMS Lecture Notes – Monograph Series, 1996, volume 29, pp. 331–342.

[21] S. J. Godsill, P. J. W. Rayner, Digital Audio Restoration: A Statistical Model Based Approach, Springer-Verlag, 1998.

[22] S. J. Godsill, P. J. W. Rayner, Statistical reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler, IEEE Transactions on Speech and Audio Processing 6 (4) (1998) 352–372.

[23] S. J. Godsill, P. J. W. Rayner, O. Cappé, Digital audio restoration, in: K. Brandenburg, M. Kahrs (eds.), Applications of Digital Signal Processing to Audio and Acoustics, Kluwer Academic Publishers, 1996, chapter 4, pp. 133–194.

[24] P. J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732.

[25] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1) (1970) 97–109.

[26] W. Klippel, Adaptive adjustment of nonlinear filters used for loudspeaker equalization, Presented at the *104th Convention of the Audio Engineering Society*, preprint 4646 (May 1998).

[27] L. Kuo, B. Mallick, Variable selection for regression models, Technical Report 94-26, Department of Statistics, University of Connecticut (1994).

[28] L. Kuo, B. Mallick, Bayesian semiparametric inference for the accelerated failure-time model, Canadian Journal of Statistics 25 (4) (1997) 457–472.

[29] G. P. Liu, V. Kadirkamanathan, S. A. Billings, On-line identification of nonlinear systems using Volterra polynomial basis function neural networks, Neural Networks 11 (9) (1998) 1645–1657.

[30] J. S. Liu, W. H. Wong, A. Kong, Covariance structure and convergence rate of the Gibbs sampler with various scans, Journal of the Royal Statistical Society B 57 (1) (1995) 157–169.

[31] R. C. Maher (ed.), IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995.

[32] K. J. Mercer, Identification of Distortion Models, Ph.D. thesis, University of Cambridge (1993).

[33] I. J. Morrison, The Application of Volterra Series to Signal Detection and Estimation, Ph.D. thesis, University of Cambridge (November 1990).

[34] J. J. K. Ó Ruanaidh, W. J. Fitzgerald, Interpolation of missing samples for audio restoration, Electronics Letters 30 (1994) 622–623.

[35] K. J. Pope, Time Series Analysis, Ph.D. thesis, University of Cambridge (October 1993).

[36] D. Preis, H. Polchlopek, Restoration of nonlinearly distorted magnetic recordings, Journal of the Audio Engineering Society 32 (1) (1984) 26–30.

[37] J. J. Rajan, P. J. W. Rayner, Unsupervised time-series classification, Signal Processing 46 (1) (1995) 57–74.

[38] M. J. Reed, M. O. Hawksford, Practical modelling of nonlinear audio systems using the Volterra series, 100th Convention of the Audio Engineering Society, preprint 4264 (May 1996).

[39] G. O. Roberts, S. K. Sahu, Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler, Journal of the Royal Statistical Society B 59 (2) (1997) 291–317.

[40] M. Schetzen, The Volterra and Wiener Theories of Nonlinear Systems, John Wiley & Sons, 1980.

[41] H. Schurer, C. H. Slump, O. E. Herrmann, Second order Volterra inverses for compensation of loudspeaker nonlinearity, in: Maher [31].

[42] H. Tong, Non-Linear Time Series: A Dynamical System Approach, Oxford Statistical Science Series, Oxford University Press, 1990.

[43] P. T. Troughton, Simulation Methods for Linear and Nonlinear Time Series Models with Application to Distorted Audio Signals, Ph.D. thesis, University of Cambridge (May 1999).

[44] P. T. Troughton, S. J. Godsill, Bayesian model selection for time series using Markov chain Monte Carlo, Proceedings of IEEE ICASSP-97 V (1997) 3733–3736.

[45] P. T. Troughton, S. J. Godsill, A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves, Technical Report CUED/F-INFENG/TR.304, Department of Engineering, University of Cambridge (November 1997).

[46] P. T. Troughton, S. J. Godsill, Bayesian model selection for linear and non-linear time series using the Gibbs sampler, in: J. G. McWhirter (ed.), Mathematics in Signal Processing IV, Oxford University Press, 1998, pp. 249–261.

[47] P. T. Troughton, S. J. Godsill, A reversible jump sampler for autoregressive time series, Proceedings of IEEE ICASSP-98 IV (1998) 2257–2260.

[48] S. V. Vaseghi, R. Frayling-Cork, Restoration of old gramophone recordings, Journal of the Audio Engineering Society 40 (10) (1992) 791–801.

[49] M. West, J. Harrison, Bayesian Forecasting and Dynamic Models, Springer-Verlag, 1997, 2nd edition.

[50] S. A. White, Restoration of nonlinearly distorted audio by histogram equalization, Journal of the Audio Engineering Society 30 (11) (1982) 828–832.