PARAMETER ESTIMATION FOR NON-GAUSSIAN AUTOREGRESSIVE PROCESSES

Edward R. Beadle and Petar M. Djuric

Department of Electrical Engineering
State University of New York at Stony Brook, Stony Brook, N.Y. 11794

ABSTRACT

It is proposed to jointly estimate the parameters of non-Gaussian autoregressive (AR) processes in a Bayesian context using the Gibbs sampler. Using the Markov chains produced by the sampler an approximation to the vector MAP estimator is implemented. The results reported here used AR(4) models driven by noise sequences where each sample is iid as a two component Gaussian sum mixture. The results indicate that using the Gibbs sampler to approximate the vector MAP estimator provides estimates with precision that compares favorably with the CRLBs. Also briefly discussed are issues regarding the implementation of the Gibbs sampler for AR mixture models.

1. INTRODUCTION

Joint estimation of model parameters is of great interest in the statistical analysis of signals. For analytical work rational transfer function models are very popular choices. Within this class, autoregressive (AR) models are the most popular since most physical systems are well described by them. When the driving noise is Gaussian, many well known approaches such as least squares, Yule-Walker, and maximum likelihood have become standard methods to estimate the unknowns [1]. However, in applications such as the restoration of audio signals the adoption of a Gaussian noise model is inadequate [2]. In cases like these more complicated noise models are required to accurately represent the physical system.

An alternative to the Gaussian AR model is an AR model driven by a noise process **u** where the u_n's are iid as a Gaussian sum mixture. The general model form is,

$$x_n = \sum_{j=1}^{p} a_j x_{n-j} + u_n \tag{1}$$

and the finite Gaussian sum mixture is expressed as,

$$f(u) = \sum_{i=1}^{k} \lambda_i f_i(u) \tag{2}$$

where $0 < \lambda_i < 1$ and $\sum \lambda_i = 1$, and each f_i is a Gaussian pdf with parameters (μ_i, σ_i^2). A wide range of driving noise models can be accommodated by choosing the parameters of *f* appropriately. In this paper an AR(4) model driven by a two "component" Gaussian sum mixture is used which implies $p = 4$ and $k = 2$. Thus (2) takes the form,

$$f(u) = (1-\varepsilon) f_1(u) + \varepsilon f_2(u). \tag{3}$$

Further, we use a contaminated noise model, so both "component" distributions share a common mean, which is assumed known, and the variances satisfy $\sigma_1^2 << \sigma_2^2$.

In this work we investigate the use of a Markov Chain Monte Carlo (MCMC) method called the Gibbs sampler to jointly estimate the noise and process parameters [3]. MCMC methods are of growing interest to the signal processing community as they can be used where the likelihood or posterior pdfs are intractable for algorithms requiring either direct sampling or summary statistics of the distributions. The principle difficulty arises due to the complicated form of the likelihood and posterior distributions under non-Gaussian stimulation of a linear model. The Gibbs sampler greatly simplifies the joint estimation problem for non-Gaussian AR processes relative to other iterative techniques, such as the expectation-maximization algorithm and Newton-Raphson approaches, by directly generating samples from the posterior distribution $p(\underline{\theta} | \underline{x})$.

Also in the case of AR models driven by Gaussian sum mixtures, an "information paradox" is reached for classical Bayesian estimation approaches [5]. The paradox is, as more data are acquired the estimator performance improves, but at the same time can become impossible to evaluate despite the availability of a closed form expression. Using the Gibbs sampler and an alternative form of the "missing data" approach proposed in [2] and [4] these difficulties are mitigated and very good estimation performance can be achieved.

2. PROBLEM STATEMENT

The problem is one of parametric modeling in a Bayesian context using an observed data vector $\underline{\mathbf{x}}$ to jointly estimate all the model parameters, where only the process order and number of mixands are known.

It is assumed that the observed data were generated by a linear asymptotically stationary real AR process (1) of known order, driven by iid Gaussian mixture noise (3) with two "components". The assumption of asymptotic stationarity implies the poles of the process are inside the unit circle in the z-plane. This imposes constraints on the possible values of the process structure parameters for an AR(*p*) system. It is assumed without any loss in generality, that the means of the individual distributions of the mixture are equal to zero.

3. PRIOR SELECTION

Standard engineering assumptions in specifying the prior pdfs of the unknowns is used with the goal of being uninformative in the sense of Jeffreys' priors [4]. Independence between the noise pdf and the AR structure parameters is assumed. The noise model unknowns are the mixture parameter ε and the variances of f_1 and f_2. The mixture parameter ε is assumed independent of the variances with a uniform pdf for $\varepsilon \in (0,0.25)$. Assuming $\varepsilon \in (0,0.25)$ implies that less than 25 % of the noise samples are from the contaminating distribution f_2. When modeling physical systems ε<.1 are common.

The variances are dependent on each other to the extent that at each iteration we require $\sigma_1^2 < \sigma_2^2$ to avoid some of the identifiabilty issues encountered with mixtures [5]. To be somewhat noninformative we assume a joint prior on the variances proportional to $(const / \sigma_1^2\sigma_2^2)U(\sigma_2^2 - \sigma_1^2)$, where the range of values are bounded by some means, and the function *U* is the unit step function.

Lastly, given the model specifications the uninformative joint prior is defined as $p(\underline{a}) \propto const$ for all $\underline{a}$ yielding stable AR models. Since real AR models are assumed, the structure parameters $\underline{a}$ are dependent (related) on each other, and synthesizing samples from this distribution has required a novel approach using the Levinson recursion equations [6]. Therefore the prior used is of the form,

$$p(\underline{a},\varepsilon,\sigma_1^2,\sigma_2^2) \propto (const / \sigma_1^2\sigma_2^2)U(\sigma_2^2 - \sigma_1^2) \;\; \forall\, \underline{\vartheta} \in \Theta \quad (4)$$

where Θ is defined by the considerations given above.

4. BACKGROUND ON THE GIBBS SAMPLER

Let $\underline{\vartheta}$, be the vector of unknowns with components $\vartheta_1, \vartheta_2, \ldots, \vartheta_k$, and that the objective is to obtain summary inferences for the joint posterior $p(\underline{\vartheta}|\underline{x}) = p(\vartheta_1, \ldots, \vartheta_k|\underline{x})$. This problem can be recast into one of iterative sampling from appropriate distributions to produce a Markov chain. The distributions controlling the evolution of the Markov chain are called the *full conditional densities* and are given by $p(\vartheta_i|\underline{x}, \vartheta_j, j \neq i)$ for $i = 1, \ldots, k$. These functions are easily determined for each ϑ_i, using the form $p(\underline{\vartheta}|\underline{x}) \propto p(\underline{x}|\underline{\vartheta})\, p(\underline{\vartheta})$. Using Bayes' theorem and the chain rule for conditional densities the details of the derivation are straightforward in this case.

The Gibbs algorithm is seeded with a set of arbitrary values $\vartheta_1^{(0)}, \ldots, \vartheta_k^{(0)}$ obtained from the parameter space Θ, using the restrictions and prior pdfs described above. Then the following iterative procedure is executed,

draw $\vartheta_1^{(1)}$ from $p(\vartheta_1|\underline{x}, \vartheta_2^{(0)}, \ldots \vartheta_k^{(0)})$,
draw $\vartheta_2^{(1)}$ from $p(\vartheta_2|\underline{x}, \vartheta_1^{(1)}, \vartheta_3^{(0)}, \ldots \vartheta_k^{(0)})$,
$\vdots$
draw $\vartheta_k^{(1)}$ from $p(\vartheta_k|\underline{x}, \vartheta_1^{(1)}, \vartheta_2^{(1)}, \ldots \vartheta_{k-1}^{(1)})$,
draw $\vartheta_1^{(2)}$ from $p(\vartheta_1|\underline{x}, \vartheta_2^{(1)}, \ldots \vartheta_k^{(1)})$,
$\vdots$
draw $\vartheta_k^{(2)}$ from $p(\vartheta_k|\underline{x}, \vartheta_1^{(2)}, \vartheta_2^{(2)}, \ldots \vartheta_{k-1}^{(2)})$,
$\vdots$
draw $\vartheta_1^{(t)}$ from $p(\vartheta_1|\underline{x}, \vartheta_2^{(t)}, \ldots \vartheta_k^{(t)})$,
$\vdots$
draw $\vartheta_k^{(t)}$ from $p(\vartheta_k|\underline{x}, \vartheta_1^{(t)}, \vartheta_2^{(t)}, \ldots \vartheta_{k-1}^{(t)})$,

and so on, terminating at the last iteration t.

5. IMPLEMENTATION OF THE GIBBS SAMPLER

In each iteration the random draw can be performed using a variety of methods [7]. One straightforward method circumventing many of the difficulties encountered with random variate generation techniques is to use the weighted resampling technique [8]. It directly implements sampling from conditional densities, however it is very computationally intensive. An alternative approach is to adopt a missing data structure to build a hierarchical model and use "conjugate" priors to greatly simplify the sampling from the full conditional densities. However, although this method allows efficient implementation of the sampling using well known techniques it produced poor estimation

performance. Therefore a combination approach is proposed which uses a modified "missing data" approach for the AR coefficients and retains the weighted resampling approach for the noise model parameters.

The missing data approach proposed in [2] and [4] first estimates the noise sequence samples using current AR process coefficient estimates, and then attempts to probabilistically classify each sample to only one component of the mixture based on likelihoods. The result of the classification produces a multivariate normal pdf of order *p* which can easily be sampled. The covariance matrix is diagonal, with the values on the diagonal being the current estimates of the component distribution variances. However, this approach tends to underestimate membership into the contaminating mixture component which in turn biases the noise parameter model estimates.

The proposed approach also produces a multivariate normal of order *p*, with a diagonal covariance matrix. But now the diagonal elements are formed by "blending" the individual variance estimates in proportion to the likelihoods. This method coupled with the weighted resampling on the noise parameters was able to approach the CRLBs while avoiding the introduction of bias. In the case of vector (joint) sampling of parameters, the scalar draws shown in Section 4 become vector draws [2].

As the number of iterations *t* grows, the resulting parameter vector $\underline{\vartheta}^{(t)}$ in each chain is a sample from a distribution asymptotically approaching the joint conditional pdf of the unknown parameters given the data, namely $p(\underline{\vartheta}|x)$. The exact point in where this occurs is subject to some interpretation and depends on the concepts of convergence of the Markov chains. A plethora of convergence diagnostics have been proposed in the literature [9].

In the present case the vector MAP estimator has been chosen, and as such convergence of the Markov chains are not required. However, convergence will in general indicate that the main mode(s) have been found. The convergence diagnostic chosen for this application is based on normal theory and is outlined in [10]. Basically this diagnostic compares the variances of each individual Markov chain being simulated for a particular experiment, and compares them to the cross chain variances via a scale reduction factor. When the factor is suitably small convergence is detected.

Using multiple independent chains in the parameter space, a random search type algorithm is implemented. In the simulations performed, 5 chains each with 100 iterations were used. The 100 iteration limit was set to match conditions imposed in [11]. The approximation to the (global) vector MAP was implemented by searching across all the chains for the sample with the highest posterior density value. To help the search for a global MAP one of the chains is seeded with the Burg estimates of the coefficients.

6. PERFORMANCE RESULTS

As stated, the vector MAP was approximated using 5 independent Markov chains of 100 iterations in each experiment. The true (global) vector MAP value is approximated by the simulated value across all the chains with the highest posterior density value. An experiment is defined as generating an observed data sequence $\underline{\mathbf{x}}$ of 1000 samples using the given model parameters, and running the Gibbs sampler as specified. The model parameters for the Gaussian mixture are $\varepsilon = .1$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 100$. The noise sequences generated are used to excite two different AR(4) models. The first is a narrow band system, and the second is a wide band system. To gather sufficient data on the performance of the algorithm proposed for both the wide band and narrow band AR models, 250 experiments have been performed. The estimator performance is shown in the tables below.

For the narrow band case (Figure 1) the results in Table 1 show the performance of the estimator for 250 experiments. The sample variance of the approximate vector MAP is comparable to the CRLBs and some bias in the variance estimates may be evident. With the limited number of experiments available to date it is possible that the difference between the estimates and true values is not due to bias, but to statistical fluctuation in the estimates. In addition the estimator precision for a_4 and ε are somewhat disappointing relative to the other parameters. The high variance relative to the CRLBs is currently attributed to the sensitivity of these estimates to small perturbations in the other model parameters. However, increasing the number of iterations in each chain may remedy these problems.

The performance in the wide band case (Figure 2) is shown in Table 2. Here the sample variance of the estimator is again comparable to the CRLBs, and possibly some bias in the variance estimates is evident. The increase in the estimator variances from the narrow band case is currently attributed to the wide band nature of the system. That is the noise spikes from the contaminating mixture are dissipated quickly and do not provide the ringing attributed with providing the higher precision estimates in the narrow and case [11]. Again more experiments are necessary to resolve whether bias is present.

The performance of the Gibbs approach presented here is far superior to least squares approaches and competitive

to the MLE approach presented in [11] for both wide and narrow band cases.

7. CONCLUSIONS

The simulations shown here indicate that the Gibbs methodology can be an effective tool for joint estimation of model parameters of non-Gaussian AR systems in a Bayesian context. However, additional research work is required to improve the estimator performance to produce precision closer to the CRLBs. Currently, it is felt that more significant exploration of the posterior parameter space will yield more precise results. Thus increasing the number of chains, iterations per chain, or the number of samples used in the weighted bootstrap sampling procedure can help in this regard. This will increase the computational load, but constantly increasing computing power lessens this concern. Lastly, this method would also be of interest in the challenging problem of joint estimation for non-Gaussian moving average models.

	True Value	MAP Sample Mean	CRLB	MAP Sample Variance
a_1	2.7600	2.7602	1.6278e-5	1.9299e-5
a_2	-3.8090	-3.8091	8.0163e-5	10.666e-5
a_3	2.6540	2.6539	8.0163e-5	11.610e-5
a_4	-0.9240	-0.9238	1.6278e-5	2.7204e-5
ε	0.1000	0.1002	30.869e-6	72.486e-6
σ_1^2	1.0000	0.9853	2.8944e-3	3.5897e-3
σ_2^2	100.00	98.125	237.47	250.94

Table 1: Estimator performance for narrow band system.

	TRUE Value	MAP Sample Mean	CRLB	MAP Sample Variance
a_1	1.3520	1.3498	1.0491e-4	1.3015e-4
a_2	-1.3380	-1.3354	2.5961e-4	3.5653e-4
a_3	0.6620	0.6587	2.5961e-4	3.1529e-4
a_4	-0.2400	-0.2378	1.0491e-4	1.4966e-4
ε	0.1000	0.1010	30.869e-6	64.250e-6
σ_1^2	1.0000	0.9890	2.8944e-3	3.7579e-3
σ_2^2	100.00	101.92	237.47	337.78

Table 2: Estimator performance for wide band system.

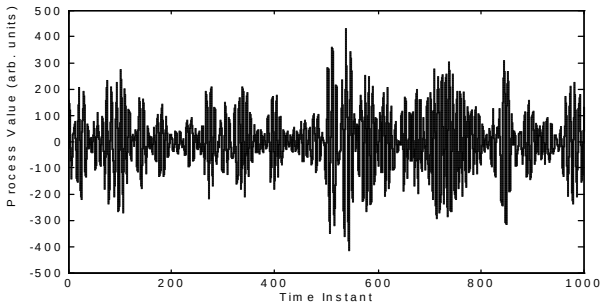

Figure 1: Typical observed data for the narrow band model

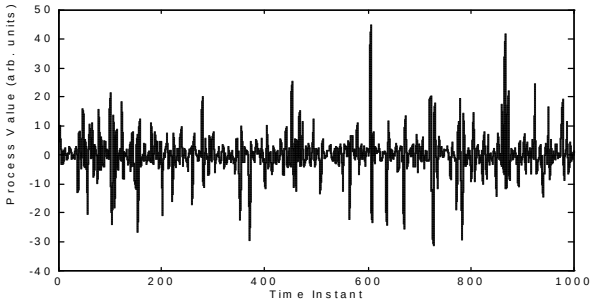

Figure 2: Typical observed data for the wide band model.

8. REFERENCES

[1] M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press:New York, 1981.

[2] S.J. Godsill and P.J. Rayner, "Robust noise reduction for speech and audio signals," I*CASSP Proc*., pp. 625-628, 1996.

[3] G. Casella and E. George, "Explaining the Gibbs Sampler," *The American Statistician*, vol 46, no 2, pp. 167-174, Aug. 1992.

[4] C. Robert, *The Bayesian Choice*, Springer:New York,1996.

[5] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley:New York, 1973.

[6] E.R. Beadle and P.M. Djuric, "Uniform random generation of stable minimum phase ARMA(p,q) Processes," submitted for publication in *IEEE Sig. Proc. Letters*.

[7] W. Gilks, et. al., *Markov Chain Monte Carlo Methods in Practice*, Chapman-Hall:New York, 1996.

[8] A.F.M. Smith and A.E. Gelfand, "Bayesian statistics without tears: A sampling-resampling perspective," *The American Statistician*, vol 46, no 2, pp. 84-88, May 1992.

[9] M. Cowles and B. Carlin, "Markov chain Monte Carlo convergence diagnostics: A comparative review," *J. Amer. Statis. Soc*., vol 91, no 434, pp. 883-904, June 1996.

[10] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol 7, no 4, pp. 457-511, 1992.

[11] D. Sengupta and S. Kay, "Efficient estimation of parameters for non-Gaussian autoregressive processes," *IEEE Trans. ASSP*, vol 37, no 6, pp. 785- 794, June 1989.