Non-homogeneous Markov Mixture of Periodic Autoregressions for the Analysis of Air Pollution in the Lagoon of Venice

Roberta Paroli[1], Silvia Pistollato[2], Maria Rosa[2], and Luigi Spezia[3]

[1] Istituto di Statistica
Università Cattolica S.C., Milano, Italy
(e-mail: roberta.paroli@unicatt.it)
[2] Dipartimento ARPAV Provinciale di Venezia, Mestre, Italy
(e-mail: spistollato@arpa.veneto.it, mrosa@arpa.veneto.it)
[3] Dipartimento di Scienze Economiche e Metodi Quantitativi
Università degli Studi del Piemonte Orientale, Novara, Italy
(e-mail: luigi.spezia@eco.unipmn.it)

Abstract. Markov mixtures of autoregressions (MMAR) have been recently used to analyse the behaviour of non-linear and non-Gaussian time series. A special MMAR model with periodic components and a non-homogeneous hidden Markov chain is proposed here: the transition probabilities of the hidden chain are time-varying, because they depend, through logistic functions, on the dynamics of exogenous variables. We perform a complete Metropolis-within-Gibbs algorithm associated to the random permutation sampling for model choice and variable selection and to constrained permutation sampling for the estimation of the unknown parameters and the latent data. An environmental application is developed on the series of sulphur dioxide and meteorological variables recorded by an air pollution testing station in the lagoon of Venice.

Keywords: Time-varying transition probabilities, exogenous variables, Metropolis-within-Gibbs, random and constrained permutation sampling, sulphur dioxide.

1 Introduction

Non-linear and non-normal time series can be modelled by autoregressive processes assuming that different autoregressions, each one depending on a latent regime, alternate according to the regime switching, which is driven by an unobserved Markov chain. When the chain is supposed homogeneous these models are widely known as Markov switching autoregressive models, introduced in the econometric literature by [Hamilton, 1994] to study economic and financial time series. When the Markov chain is non-homogeneous we have that the transition probabilities are time-varying and depend on exogenous variables. The class of non-homogeneous hidden Markov models depending on deterministic exogenous variables has been proposed by [Diebolt *et al.*, 1994] in the classical framework.

In this paper we propose the Bayesian analysis of Markov mixtures of autoregressions (MMAR) models with a periodic component and a non-homogeneous Markov chain defined on a general number of states, whose transition probabilities depend on deterministic exogenous variables through a logistic function. We introduce Metropolis-within-Gibbs algorithms for the estimation of the unknown parameters and for the computation of the marginal likelihood, needed for model comparison. In both cases we consider the problem of label switching, which recently has become one of the most interesting topics in the Bayesian analysis of independent and Markov-dependent mixture models. We tackle label switching through constrained permutation sampling algorithm in the case of parameter estimation and through random permutation sampling in the case of marginal likelihood computation.

In the applications these models can be efficient tools to analyse environmental time series, whose main characteristics are: *i*) different unobserved levels of pollutant mean concentrations, depending on the weather conditions, *ii*) serially correlated data, *iii*) periodicities, *iv*) missing values, *v*) availability of meteorological covariates. So we will apply our methodology to analyse a three year series of hourly mean concentrations of sulphour dioxide recorded in the lagoon of Venice.

2 The non-homogeneous Markov mixtures of periodic autoregressions

The non-homogeneous Markov mixtures of periodic autoregressionsof order $(m;p)$ (NHMMAR$(m;p)$) are discrete-time stochastic processes $\{Y_t; X_t\}$, such that $\{X_t\}$ is an unobservable non-homogeneous discrete-time Markov chain with a finite number of states, m, while $\{Y_t\}$, given $\{X_t\}$, is an observed autoregressive process of order p with a periodic component and depending on exogenous variables with the conditional distribution of Y_t depending on $\{X_t\}$ only through the contemporary X_t.

Let $\{X_t\}$ be a discrete-time, first-order, non-homogeneous Markov chain on a finite state-space S_X with cardinality m $(S_X = \{1, \dots, m\})$. For any $t = 2, \dots, T$, $\Gamma_t = \left[\gamma^t_{i,j}\right]$ is the $(m \times m)$ transition matrix, where $\gamma^t_{i,j} = P(X_t = j \mid X_{t-1} = i)$, for any $i, j \in S_X$; the initial distribution is the vector $\delta = (\delta_1, \dots, \delta_m)'$, where $\delta_i = P(X_1 = i)$, for any $i \in S_X$; $x^T = (x_1, \dots, x_T)'$ is the sequence of the states of the Markov chain and, for any $t = 1, \dots, T$, x_t has values in S_X. At any time $t = 2, \dots, T$, the transition probabilities $\gamma^t_{i,j}$ can be obtained as logistic functions of the vector z_t of exogenous deterministic variables, i.e.

$$\text{logit}(\gamma^t_{i,j}) = \ln\left(\gamma^t_{i,j} \,/ \gamma^t_{i,i}\right) = z'_t \alpha_{i,j} \qquad \text{for any } i, j \in S_X$$

$$\gamma^t_{i,j} = \left(\exp\left(z'_t \alpha_{i,j}\right)\right) \Big/ \left(1 + \textstyle\sum_{j \neq i} \exp\left(z'_t \alpha_{i,j}\right)\right) \quad \text{for any } i, j \in S_X$$

where $\alpha_{i,j}$ is an n-dimensional vector of parameters, $\alpha_{i,j} = (\alpha_{i,j,0}, \alpha_{i,j,1}, \ldots, \ \alpha_{i,j,n-1})'$, if $i \neq j$, and an n-dimensional vector of zeros, if $i = j$; z_t is an n-dimensional vector, $z_t = (1, z_{t,1}, \ldots, z_{t,n-1})'$, for any $t = 2, \ldots, T$. Instead of placing the first or the last entry of the transition matrix at the denominator of the logit as usual, we place there the diagonal entry because this statement allows us to perform constrained permutation sampling and random permutation sampling algorithms, as we shall see in Sections 3. Notice that when the last $n-1$ entries of z_t are equal to zero for any t, the Markov chain is homogeneous.

Hence, given the order-p dependence and the contemporary dependence conditions, the equation describing the NHMMAR model is

$$Y_{t(i)} = \mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)} + E_{t(i)}, \qquad (1)$$

where $Y_{t(i)}$ denotes the generic variable Y_t when $X_t = i$, for any $1 \leq t \leq T$ and for any $i \in S_X$; the autoregressive coefficients $\varphi_{\tau(i)}$, for any $\tau = 1, \ldots, p$ and for any $i \in S_X$, depend on the current state i of the Markov chain; $w_{t,j}$, for any $1 \leq t \leq T$, are the observations of the j-th exogenous deterministic variable, for any $j = 1, \ldots, q$, that are elements of the matrix W of dimension $(T \times q)$, weighted by the coefficients $\theta_{j(i)}$, for any $j = 1, \ldots, q$ and for any $i \in S_X$, that depend on the current state of the Markov chain. The term $\beta_{t(i)}$ is the harmonic component of periodicity $2s$, depending on the current state i of the Markov chain

$$\beta_{t(i)} = \sum_{j=1}^{s^*} \left(\beta_{1,j(i)} \cos(\pi j t / s) + \beta_{2,j(i)} \sin(\pi j t / s) \right),$$

where s^* is the number of significant harmonics $(s^* \leq s)$. $E_{t(i)}$ denotes the Gaussian random variable E_t when $X_t = i$, with zero mean and precision λ_i $\left(E_{t(i)} \mathrm{sim} \mathcal{N}(0; \lambda_i)\right)$, for any $i \in S_X$, with the discrete process $\{E_t\}$, given $\{X_t\}$, satisfying the conditional independence and the contemporary dependence conditions.

By these statements the conditional distribution of any variables $Y_{t(i)}$, given state i, is normal,

$$Y_{t(i)} \mathrm{sim} \mathcal{N} \left(\mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)}; \lambda_i \right),$$

for any $t = 1, \ldots, T$ and for any $i \in S_X$, while the marginal distribution of any variable Y_t is a mixture of m normals, whose mixing distribution is a row of the transition matrix Γ_t,

$$Y_t \mathrm{sim} \sum_{i=1}^{m} \gamma_{x_{t-1},i} \mathcal{N} \left(\mu_i + \sum_{\tau=1}^{p} \varphi_{\tau(i)} y_{t-\tau} + \sum_{j=1}^{q} \theta_{j(i)} w_{t,j} + \beta_{t(i)}; \lambda_i \right),$$

for any t.

A sufficient condition for the stationarity of the process (1) is that all the m sub-processes generated by the m states of the chain are stationary, that is, for any $i \in S_X$, the roots of the auxiliary equations are all inside the unit circle. To automatically satisfy the constraint on any $\varphi_i = \left(\varphi_{1(i)}, \ldots, \varphi_{p(i)}\right)'$, we can reparametrize φ_i in terms of the partial autocorrelations $r_i = \left(r_{1(i)}, \ldots, r_{p(i)}\right)'$ of any sub-process, for any $i \in S_X$, according to [Jones, 1987]. Our inference will be based on the logarithmic transformation $R_{j(i)} = \ln\left(\frac{1+r_{j(i)}}{1-r_{j(i)}}\right)$, which maps any partial autocorrelation $r_{j(i)}$ from $(-1;1)$ to $\Re$, for any $j = 1, \ldots, p$ and any $i \in S_X$.

In the framework of the mixture models the problem of identifiability concerns the invariance of the mixture under permutation of the indices of the components. In model (1) we have m states and we have $m!$ ways to label them; so different models are interchangeable by permuting their labeling. This is often called the "label switching" problem and it can be overcome by placing some identifiability constraints on some parameters with a data-driven procedure based on random permutation sampling algorithm [Frühwirth-Schnatter, 2001]. In this paper we shall introduce the random permutation sampling and the constrained permutation sampling algorithms.

Furthermore to be able to estimate the state-dependent seasonal component we need to assume the same hidden state for all the s times of any sub-period.

The unknown parameters and latent data of the NHMMAR to be estimated are: α the matrix of the vectors $\alpha_{i,j}$; μ the vector of the signals; λ the vector of the precisions; R the matrix of the coefficients $R_{j(i)}$; θ the matrix of the coefficients $\theta_{j(i)}$; β the matrix of the state-dependent harmonic coefficients; x^T the sequence of the hidden states; y^* the vector of all the missing observations. For our Bayesian inference, we place independent multivariate normal priors on each entry of matrix α; independent normal priors on each entry of vector μ; independent gamma priors on each entry of vector λ; independent multivariate normal priors of dimension p on each entry of the vector R_i; independent multivariate normal priors of dimension q on each vector θ_i; independent multivariate normal priors of dimension $2s^*$ on each vector β_i.

Let $y^T = (y_1, \ldots, y_T)'$ be the sequence of the observations; the posterior distribution of the parameter vector $\psi = (\alpha, \mu, \lambda, R, \theta, \beta, x^T, y^*)$ is

$$\pi\left(\psi \mid y^T, y^0, Z, W, V, \delta\right) = f(\alpha, \mu, \lambda, R, \theta, \beta, x^T, y^* \mid y^T, y^0, Z, W, V, \delta) \propto$$
$$\propto f\left(y^T, y^* \mid \mu, \lambda, R, \theta, \beta, W, V, x^T, y^0\right) f\left(x^T \mid \alpha, Z, \delta\right) p(\alpha)p(\mu)p(\lambda)p(R)p(\beta)p(\theta),$$

where $y^0 = (y_{-p+1}, \ldots, y_0)'$ are the initial values fixed for the p-dependence condition; Z is the matrix of dimension $(T \times n)$ of z_t, the exogenous variables of the Markov chain; V is a $(T \times 2s^*)$ matrix whose generic element on the t-th row of the j-th odd column is $\cos(\pi jt/s)$, while the generic element on the t-th row of the j-th even column is $\sin(\pi jt/s)$, for any $j = 1, 2, \ldots, s^*$.

3 Bayesian analysis

Bayesian approach to inference of mixture models is based on MCMC methods. We introduce a Metropolis-within-Gibbs procedure for model choice, variable selection and for parameter estimation.

Model choice and variable selection can be performed by means of Bayes factors in which the marginal likelihoods of the competing models are computed according to [Chib, 1995] and [Chib and Jeliazkov, 2001] corrected by the random permutation sampling algorithm [Frühwirth-Schnatter, 2001]. For model choice we need to select the unknown cardinality of the state-space of the hidden Markov chain m and the autoregressive order p, while for variables selection we require to find the best subsets of explanatory variables Z and W among all the exogenous variables to be included in the final model. To encourage the moves between the $m!$ subspaces, we can use the random permutation sampling algorithm. So at the k-th iteration of the Metropolis-within-Gibbs algorithm we use to estimate the marginal likelihood, once $\psi^{(k)}$ has been drawn, we select randomly a permutation $(\rho(1), \ldots, \rho(m))'$ of the current labeling $(1, \ldots, m)'$ and then relabel the sequence of hidden states and the switching parameters.

We can estimate the unknown parameters of NHMMAR models via a Metropolis-within-Gibbs procedure, that we briefly discuss here.

To overcome label switching the Metropolis-within-Gibbs sampler is run on a subspace only, by placing some parameters in increasing or decreasing order. The identifiability constraint is chosen ex post after simulations by a data-driven procedure, based on random permutation sampling algorithm, so as to respect the geometry and the shape of the unconstrained posterior distribution; different identifiability constraints can be derived by different data sets. By plotting the couples of the outputs of the estimates, obtained via unconstrained Metropolis-within-Gibbs algorithm, performed associated with random permutation sampling, we can check if there are as many groups as the hidden states and if these groups can suggest special ordering in their labeling. Without loss of generality, and since for our data set the constraint is based on the precisions, we discuss our methodology assuming that the entries of λ must be in decreasing order ($\lambda_i > \lambda_j$, for $i < j$, $i, j \in S_X$), but the procedures can be easily adapted to any other type of constraint. If λ is not ordered, instead of rejecting the vector and going on sampling till an ordered vector is obtained, we adopt the constrained permutation sampling algorithm [Frühwirth-Schnatter, 2001]. At any k-th iteration of the MCMC sampler, after the generation of the sequence of the hidden states, we generate the vector of the precisions; so we have m couples $\left(i, \lambda_i^{(k)}\right)$. If the $\lambda_i^{(k)}$'s are unordered, we apply a permutation $\rho(\cdot)$ to order them; consequently also the corresponding i's must be permuted according to the permutation, $\{\rho(1), \ldots, \rho(m)\}$; then the permutation is extended to the sequence of states $x^{T(k)}$ just generated, and to the switching-parameters generated in the previous iteration, $\rho\left(\mu^{(k-1)}\right)$, $\rho\left(R^{(k-1)}\right)$, $\rho\left(\theta^{(k-1)}\right)$, $\rho\left(\beta^{(k-1)}\right)$, $\rho\left(\alpha^{(k-1)}\right)$; finally all the parameters and the missing observations are generated.

The iterative scheme of the Metropolis-within-Gibbs algorithm at the k-th iteration can be summarized as follows:
1) the sequence $x^{T(k)}$ of hidden states is generated by the forward filtering-backward sampling algorithm, [Carter and Kohn, 1994] and [Frühwirth-Schnatter, 1994];
2) the parameters $\lambda_i^{(k)}$, for any $i \in S_X$, are generated independently from gamma distributions; the entries of the vector $\lambda^{(k)}$ must be in decreasing order to satisfy the identifiability constraint. If $\lambda^{(k)}$ is not ordered, we apply the constrained permutation sampling algorithm;
3) the parameters $\mu_i^{(k)}$, for any $i \in S_X$, are generated independently from normal distributions;
4) the parameters $R_{j(i)}^{(k)}$, for any $j = 1, \ldots, p$ and any $i \in S_X$, are generated independently, by a Metropolis step, from the random walk $R_{j(i)}^{(k)} = R_{j(i)}^{(k-1)} + U_R$, where U_R is a Gaussian noise with zero mean and constant precision.
5) the parameters $\theta_i^{(k)}$, for any $i \in S_X$, are independently generated from normal distributions of dimension q;
6) the parameters $\beta_i^{(k)}$, for any $i \in S_X$, are independently generated from normal distributions of dimension $2s^*$;
7) the parameters $\alpha_{i,j}^{(k)}$, for any $i, j \in S_X$, with $i \neq j$, are generated independently, by a Metropolis step, from the random walk $\alpha_{i,j}^{(k)} = \alpha_{i,j}^{(k-1)} + U_A$, where U_A is a Gaussian noise with zero mean and constant precision matrix.
8) every missing observation y_t^* is generated from the conditional normal distribution.

Now, at the end of the k-th iteration of the MCMC sampler, the vector $\psi^{(k)}$ has been approximately simulated from $\pi(\psi \mid y^T, y^0)$, if k is large enough. We shall repeat these steps till we have an N-dimensional sample. This sample will be used to estimate each entry of ψ by means of posterior means, apart from the sequence of states, estimated thought posterior modes.

4 Application to air pollution in the lagoon of Venice

Air quality control includes the study of data sets recorded by air pollution testing stations. We are interested both in the analysis of the dynamics of the hourly mean concentrations of sulphur dioxide (SO2), in micrograms per cubic meter $\left(\mu g/m^3\right)$, recorded by an air pollution testing station in the lagoon of Venice (Italy), and in investigating its relationships with the daily meteorological variables. The series of the SO2 in the log scale from the 1st of January 2001 to the 31st of December 2003 (26280 observations) is plotted in Figure 1a and it can be noticed that some observations are missing. This happens either because sometimes the station must be stopped for automatic calibration or because of occasional mechanical failure, ordinary maintenance, or data quality inspections. Plotting the histogram of the values we can guess the presence of hidden states by noticing an asymmetric distribution.

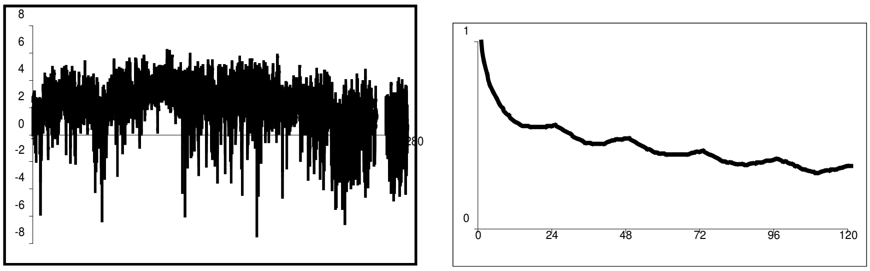

Fig. 1. (a) Series of the SO2 hourly log-concentrations; (b) 120 hours autocorrelations.

Just by looking at the series of observations we can notice a daily periodicity ($s = 24$) with 1 peak a-days ($s^* = 1$); the daily periodicity is confirmed by the correlogram of five days (Figure 1b). Atmospheric concentrations of the SO2 are influenced by many meteorological variables that are recorded together with the pollutant by the same station; we consider the following covariates: wind speed, temperature, atmospheric pressure, humidity, rainfall and solar radiation. Some of these variables will be included in the matrix W of the exogenous variables influencing the observed process and in the matrix Z of the covariates influencing the non-homogeneous Markov chain.

We develop our empirical analysis in three steps: *i*) model and variables selection, *ii*) constraint identification, *iii*) parameter estimation.

i) Model selection is performed for $m = 1, 2, 3, 4$ and $p = 0, 1, 2, 3, 4, 5, 6$ and the NHMMAR(3,1), i.e. a model with 3 hidden states and an autoregression component of order 1, is the best among all the competing models. Also variable selection is based on the values of the marginal likelihoods of all the models we analysed. The results show that temperature, humidity and wind are the variables to be included in the final model. They will be included both in the matrix W and in the matrix Z.

ii) In the second step of our analysis we have to select the identifiability constraint, which must respect the geometry and the shape of the unconstrained posterior distribution. Graphically analysing the outputs of the unconstrained NHMMAR(3;1) model, we chose the constraint on the precisions: $\lambda_1 > \lambda_2 > \lambda_3$ (Figures 2a) because the decreasing ordering is evident in the graph. Decreasing precisions is a reasonable constraint for these data, because when the low hidden state occurs, the variability of SO2 data depending on it is low and the concentrations of pollution are also low; by contrast when the high hidden state occurs, the variability of SO2 data depending on it is high and the concentrations of pollution are also high.

iii) Now we run constrained permutation sampling for the NHMMAR(3;1) model to estimate its parameters. The dynamics of the fitted values can be observed in Figure 2b: if we compare it with the dynamics of the actual data (in Figure 1a), we can see that these simulated values correctly follow the series according to the dynamics of the twenty-four hours. By this graph and by the values of the descriptive statistics we calculated to assess the fitting accuracy of the estimated model, we can argue that the fitting ability of the model is satisfactory.

Missing observations are simulated as extra latent variables; Figure 2c shows how simulated values fill the series according to the dynamics of the observed data. The dynamics of the hidden states, representing the three different levels of pollution occured during the analysed period, can be observed in Figure 2d, where we have depicted the sequence of the posterior modes of all generated states. State 3 underlies the observations with the highest level of pollution, while state 1 underlies those with the lowest level of pollution.

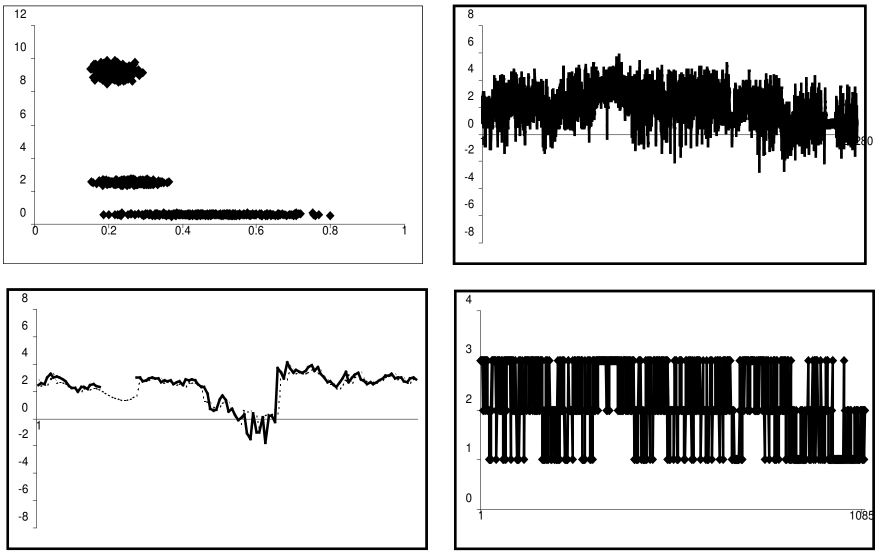

Fig. 2. (a) Couples of outputs of means and precisions of unconstrained algorithm with random permutations for m=3; (b) dynamics of the fitted values; (c) a subserie of actual (solid line) and fitted (dashes); (d) the sequence of the hidden states

5 Conclusions

We recurred to Bayesian non-homogeneous Markov mixtures of periodic autoregressions to analyse a time series about the hourly mean concentrations of sulphur dioxide, whose dynamics is characterized by cyclicity, non-normality and non-linearity. Model choice, exogenous variable selection and inference have been performed through Metropolis-within-Gibbs algorithms, considering the label switching problem, which has been efficiently tackled by permutation sampling.

References

[Carter and Kohn, 1994]C.K. Carter and R. Kohn. On gibbs sampling for state space models. *Biometrika*, pages 541–553, 1994.

[Chib and Jeliazkov, 2001]S. Chib and I. Jeliazkov. Marginal likelihoods from the metropolis-hastings output. *Journal of the American Statistical Association*, pages 270–281, 2001.

[Chib, 1995]S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, pages 1313–1321, 1995.

[Diebolt *et al.*, 1994]F.X. Diebolt, J.H. Lee, and G.C. Weinbach. Regime switching with time varying transition probabilities. In C.P. Hargreaves, editor, *Nonstationary Time Series Analysis and Cointegration*, pages 283–302, 1994.

[Frühwirth-Schnatter, 1994]S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, pages 183–202, 1994.

[Frühwirth-Schnatter, 2001]S. Frühwirth-Schnatter. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, pages 194–209, 2001.

[Hamilton, 1994]J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[Jones, 1987]M.C. Jones. Randomly choosing parameters from the stationarity and invertible region of autoregressive-moving average models. *Applied Statistics*, pages 134–138, 1987.