Estimation of Stationary Autoregressive Models with the Bayesian LASSO

Daniel F. Schmidt and Enes Makalic

Centre for MEGA Epidemiology

The University of Melbourne, Carlton, AUSTRALIA

{dschmidt,emakalic}@unimelb.edu.au

February 9, 2013

Abstract

This paper explores the problem of estimating stationary autoregressive models from observed data using the Bayesian LASSO. By characterising the model in terms of partial autocorrelations, rather than coefficients, it becomes straightforward to guarantee that the estimated models are stationary. The form of the negative log-likelihood is exploited to derive simple expressions for the conditional likelihood functions, leading to efficient algorithms for computing the posterior mode by coordinate-wise descent, and exploring the posterior distribution by Gibbs sampling. Both empirical Bayes and Bayesian methods are proposed for estimation of the LASSO hyper-parameter from the data. Simulations demonstrate that the Bayesian LASSO performs well in terms of prediction errors when compared to a standard autoregressive order selection method.

1 Introduction

Autoregressive models are a flexible class of linear time series models that find frequent use in signal processing. A k-th order autoregressive model with coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$ explains a time series by

$$y_t + \sum_{j=1}^{k} \phi_j y_{t-j} = \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$ are identically and independently distributed Gaussian innovations. In general, however, rather than being given a coefficient vector, one is only presented with an observed time series of n samples, $\mathbf{y} = (y_1, \ldots, y_n)'$. The problem is then to estimate the coefficients of a suitable autoregressive model on the basis of this data alone. A particular strength of the autoregressive model is its great flexibility: given a large enough k, an autoregressive model can exactly represent *any* stationary Gaussian process (Broersen, 2006). The problem is that the required order is generally *a priori* unknown, and if it is chosen to be too high the resulting flexibility of the model is outweighed by the increased variance in the estimated coefficients.

It is therefore common to not only attempt to estimate the coefficients on the basis of the data, but also to estimate a suitable order for the autoregressive model. The traditional method for fitting AR models is to use an information criterion, or model selection, approach and choose the $\hat{k}$-th order autoregressive model, with $\hat{k}$ given by

$$\hat{k} = \underset{k \in \{0,\ldots,q\}}{\arg\min} \left\{ \log 1/p(\mathbf{y}|\hat{\boldsymbol{\phi}}_k(\mathbf{y}), \hat{\sigma}_k^2(\mathbf{y})) + \alpha(k, n) \right\},$$

where $p(\cdot)$ is the likelihood function, $\hat{\boldsymbol{\phi}}_k(\mathbf{y})$ and $\hat{\sigma}_k^2(\mathbf{y})$ are the maximum likelihood estimates of the coefficients and innovation variance, respectively, and $\alpha(k, n)$ is the model complexity penalty. Different choices of $\alpha(k, n)$ yield different information criteria, i.e., $\alpha(k, n) = k$ for the Akaike Information Criterion (AIC) (Akaike, 1974), $\alpha(k, n) = 3k/2$ for the Symmetric Kullback Information Criterion (KIC) (Cavanaugh, 1999) and $\alpha(k, n) = (k/2) \log n$ for the Bayesian Information Criterion (BIC) (Schwarz, 1978).

The biggest problem with this model selection approach to fitting autoregressive models is that a nested structure is enforced on the various models, in the sense that only models in which the first k coefficients are non-zero are considered. Some attempts have been made to circumvent this problem by examining the problem in partial autocorrelation space (McLeod & Zhang, 2006). However, as with most forward-selection type model selection approaches, such procedures suffer from instability (Breiman, 1996), and in the case of all subsets-selection, quickly become computationally infeasible as k grows.

The Least Absolute Shrinkage and Selection Operator (LASSO) procedure (Tibshirani, 1996) was developed to overcome similar problems in the regular linear regression setting, and has recently been adapted to the autoregressive model setting. The existing LASSO procedures for autoregressive models operate in coefficient space, and are based on the conditional likelihood due to the complexity of the complete data likelihood in coefficient space (Wang et al., 2007; Hsu et al., 2008; Nardi & Rinaldo, 2008). For a chosen order k, the regular LASSO procedure

estimates an AR model by finding the coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$ that minimise the penalized sum-of-squares

$$\sum_{i=k+1}^{n} \left(y_t - \sum_{j=1}^{k} \phi_j y_{t-j} \right)^2 + \lambda \sum_{j=1}^{k} |\phi_j|, \tag{1}$$

where λ is a regularisation parameter that controls the complexity of the resulting model. As λ is increased, the resulting estimates are more greatly attenuated, or *shrunken*, in comparison to the unrestricted estimates ($\lambda = 0$), with all estimates being zero when $\lambda = \infty$. The great strength of the LASSO procedure is that for a finite λ, certain coefficients may be estimated as *exactly* zero, and thus the LASSO simultaneously performs model selection in addition to parameter estimation. Unfortunately, the autoregressive LASSO, as given by (1), has several problems:

1. use of the conditional likelihood means that the first k data points are discarded; for short time series this leads to increased variance in the parameter estimates;

2. there is no guarantee that for a particular λ the resulting estimates will represent a stationary process;

3. there are difficulties in finding robust and accurate standard errors for the estimated coefficients (Kyung et al., 2010);

4. it is difficult to choose a suitable λ as standard methods such as cross-validation are difficult to implement in the time series setting.

Problem (2) can be mitigated by ignoring those estimates in the LASSO path that do not represent stationary processes (using for example, a standard test for stationarity (Box et al., 1994)), while problem (4) can be tackled with a tool such as the stationary bootstrap (Politis & Romano, 1994). We propose a Bayesian approach to the LASSO based on the partial autocorrelation representation of autoregressive models that overcomes all these problems within a single, unified framework.

2 Bayesian LASSO for Autoregressive Models

For the remainder of this article we restrict our attention to stationary autoregressive models; an autoregressive model is stationary if and only if all roots of its associated characteristic polynomial $1 + \sum_{j=1}^{k} \phi_j z^{-j}$ lie entirely within the unit circle. Let Φ_k denote the set of coefficients that define all k-th order stationary autoregressive processes. There is a one-to-one correspondence between coefficients, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$, in the stationarity region

Φ_k, and partial autocorrelations, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)'$, where the j-th partial autocorrelation is defined as

$$\rho_j = \text{corr}(y_t, y_{t-j} | y_{t-1}, \ldots, y_{t-j+1}).$$

Shrinkage of partial autocorrelations was first suggested to the authors by K. Knight in a personal communication in 2008. Knight proposed to soft-threshold the empirical partial autocorrelations using their asymptotic distribution to guide the choice of thresholding parameter. Parameterizing the LASSO in terms of partial autocorrelations solves problems 1) and 2) associated with the coefficient-space conditional likelihood LASSO. An autoregressive model is stationary if and only if all partial autocorrelations are in the set

$$\left\{ \boldsymbol{\rho} \in \mathbb{R}^k : |\rho_i| < 1, i = 1, \ldots, k \right\},$$

making it simple to enforce stationarity when estimating models. Furthermore, the form of the likelihood function in partial autocorrelation space means that it is relatively straightforward to utilize the complete data likelihood in place of the conditional likelihood.

The LASSO may be implemented within the Bayesian framework by exploiting the fact that the sum-of-absolutes penalty implied by the LASSO is equivalent to using a double exponential, or Laplace distribution, as a prior distribution over the parameters. The regular LASSO specifies a single hyper-parameter which controls the scale of the Laplace distribution for all parameters. In this paper, we examine a slightly generalised case in which there is a hyper-parameter λ_j associated with each partial autocorrelation parameter ρ_j. It is clear that the regular LASSO may easily be obtained by setting all λ_j equal to a single global hyperparameter λ, that is, $\boldsymbol{\lambda} = \lambda \mathbf{1}_k$.

Let $p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)$ denote the likelihood of the observed data $\mathbf{y}$ given the autoregressive model parameters $(\boldsymbol{\rho}, \sigma^2)$. We may make inferences using the posterior distribution of the parameters $(\boldsymbol{\rho}, \sigma^2)$, conditional on the observed data $\mathbf{y}$ and the hyper-parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)'$, which is given by

$$p(\boldsymbol{\rho}, \sigma^2 | \mathbf{y}, \boldsymbol{\lambda}) \propto p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2) \pi_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\sigma^2, \boldsymbol{\lambda}) \pi_{\sigma^2}(\sigma^2), \tag{2}$$

where $\pi_{\boldsymbol{\rho}}(\cdot)$ and $\pi_{\sigma^2}(\cdot)$ denote the prior distributions over the parameters. These are given by

$$\pi_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\sigma^2, \boldsymbol{\lambda}) \quad \propto \quad \left(\frac{1}{2\sigma} \right)^k \prod_{j=1}^{k} \lambda_j \exp\left(-\frac{\lambda_j |\rho_j|}{\sigma} \right), \qquad \boldsymbol{\rho} \in (-1, 1)^k \tag{3}$$

$$\pi_{\sigma^2}(\sigma^2) \quad \propto \quad \left(\frac{1}{\sigma^2}\right)^{\nu}, \tag{4}$$

where and $\nu > 0$. This Bayesian formulation of the LASSO resolves issues 3) and 4) associated with the regular LASSO (as described by equation (1)): confidence intervals for the parameters arise naturally from the specification of the posterior distribution, and estimation of $\boldsymbol{\lambda}$ may be done in either an empirical Bayes fashion, or by extending the Bayesian hierarchy to include $\boldsymbol{\lambda}$ as a vector of hyper-parameters.

2.1 Likelihood Function

The likelihood function plays a central role in Bayesian inference. The complete data likelihood function for an AR(k) model is given by a multivariate normal distribution with a special $(n \times n)$ covariance matrix. Following (McLeod & Zhang, 2006), when $n \geq 2k$, this may be compactly written as

$$-\log p(\mathbf{y}|\boldsymbol{\phi}, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log|\boldsymbol{\Gamma}(\boldsymbol{\phi})| + \frac{\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}}{2\sigma^2}, \tag{5}$$

where $\boldsymbol{\beta} = (1, \boldsymbol{\phi}')' \in \mathbb{R}^{k+1}$, $\mathbf{D} \in \mathbb{R}^{(k+1\times k+1)}$ is a matrix with the entries

$$D_{i,j} = \sum_{l=0}^{n-i-j+1} y_{l+i}y_{l+j},$$

and $\boldsymbol{\Gamma}(\boldsymbol{\phi}) \in \mathbb{R}^{(k\times k)}$ is the unit-variance process autocovariance matrix with entries $\Gamma_{i,j}(\boldsymbol{\phi}) = (1/\sigma^2)\mathrm{E}\left[y_{n-i}y_{n-j}\right]$ (Porat & Friedlander, 1986). This form of the likelihood function has several advantages: (i) once the matrix $\mathbf{D}$ has been formed, evaluation of the likelihood function requires only $O(k^2)$ operations, and (ii) the residual sum-of-squares (RSS) for a particular coefficient vector $\boldsymbol{\phi}$, is given by the simple formula

$$\mathrm{RSS}(\boldsymbol{\phi}) = \boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}.$$

In this work, Bayesian inference is being performed in partial autocorrelation space rather than the more conventional coefficient space. The Levinson–Durbin recurrence relations may be used to quickly transform a vector of partial autocorrelations, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)'$, to their corresponding coefficients $\boldsymbol{\phi}(\boldsymbol{\rho})$ when evaluating the likelihood

(5). Using these partial autocorrelations, the determinant term in (5) may be efficiently evaluated as (Kay, 1983)

$$|\boldsymbol{\Gamma}(\boldsymbol{\phi}(\boldsymbol{\rho}))| = \prod_{j=1}^{k} \frac{1}{(1-\rho_j^2)^j}. \tag{6}$$

This equation has the distinct advantage of being coordinate-wise independent in the sense that each partial autocorrelation appears only once in each factor of the product. This property may be exploited when sampling from the posterior distribution (2) to simplify the Gibbs sampling steps. Using (6) the negative log-likelihood in partial autocorrelation space, $l(\boldsymbol{\rho}, \sigma^2) \equiv -\log p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2)$, is given by

$$\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{j=1}^{p} j\log(1-\rho_j^2) + \frac{\boldsymbol{\beta}'(\boldsymbol{\rho})\mathbf{D}\boldsymbol{\beta}(\boldsymbol{\rho})}{2\sigma^2}, \tag{7}$$

where $\boldsymbol{\beta}(\boldsymbol{\rho}) = (1, \boldsymbol{\phi}'(\boldsymbol{\rho}))'$.

2.2 Conditional Negative-Log Likelihood Functions

Let $l_j(\cdot) \equiv l(\cdot; \boldsymbol{\rho}^{-j}, \sigma^2)$ denote the conditional negative log-likelihood, where $\boldsymbol{\rho}^{-j} = (\rho_1, \ldots, \rho_{j-1}, \rho_{j+1}, \ldots, \rho_k)'$ is a vector containing all parameters except for ρ_j. These functions play an important role in the sequel as they are used to efficiently find the posterior mode, and to efficiently sample from the posterior distribution (2). It is therefore desirable to find simplified expressions for these functions. Examining (7) reveals that the second term depends only on ρ_j when we consider the conditional negative log-likelihood $l_j(\rho_j)$, while the third term depends on the complete vector $\boldsymbol{\rho}$ through the transformation $\boldsymbol{\phi}(\boldsymbol{\rho})$, even for the conditional negative log-likelihood. However, the transformation $\boldsymbol{\phi}(\boldsymbol{\rho})$ is *multi-linear*, and thus the conditional transformation $\boldsymbol{\phi}(\rho_j; \boldsymbol{\rho}^{-j})$ is a linear function of ρ_j, which implies that the third term in the conditional negative log-likelihood $l_j(\rho_j)$ is *exactly* a quadratic function in ρ_j. Using this fact, we can represent the third term in the conditional negative log-likelihood by a quadratic expansion around the origin. This yields an alternative expression for the conditional negative log-likelihood which is given, up to constants, by

$$l_j(\rho_j) = g_j\rho_j + \frac{H_{j,j}\rho_j^2}{2} - \frac{j}{2}\log(1-\rho_j^2) + \text{const}, \tag{8}$$

where

$$g_j \quad = \quad \boldsymbol{\beta}'(\boldsymbol{\rho}^0)\left(\frac{\mathbf{D}}{\sigma^2}\right)\left(\frac{\partial\boldsymbol{\beta}(\bar{\boldsymbol{\rho}})}{\partial\bar{\rho}_j}\bigg|_{\bar{\boldsymbol{\rho}}=\boldsymbol{\rho}^0}\right),$$

$$H_{j,j} = \left(\frac{\partial \boldsymbol{\beta}(\bar{\boldsymbol{\rho}})}{\partial \bar{\rho}_j}\bigg|_{\bar{\boldsymbol{\rho}}=\boldsymbol{\rho}^0} \right)' \left(\frac{\mathbf{D}}{\sigma^2} \right) \left(\frac{\partial \boldsymbol{\beta}(\bar{\boldsymbol{\rho}})}{\partial \bar{\rho}_j}\bigg|_{\bar{\boldsymbol{\rho}}=\boldsymbol{\rho}^0} \right),$$

and $(\partial \boldsymbol{\beta}(\bar{\boldsymbol{\rho}})/\partial \bar{\rho}_j|_{\bar{\boldsymbol{\rho}}=\boldsymbol{\rho}^0})$ denotes the vector of derivatives of $\boldsymbol{\beta}(\bar{\boldsymbol{\rho}})$ with respect to component $\bar{\rho}_j$, evaluated at

$$\boldsymbol{\rho}^0 = (\rho_1, \ldots, \rho_{j-1}, 0, \rho_{j+1}, \ldots, \rho_k)'.$$

Details on computing the required vector of derivatives are presented in Appendix A.

3 Computing the Posterior Mode

Due to the form of the Laplace prior (3), the mode of the posterior distribution (2) can be interpreted as a solution to a LASSO-type equation with a fixed $\boldsymbol{\lambda}$. Although this paper details an MCMC approach to exploring the posterior and obtaining confidence intervals, a method for efficiently finding the posterior mode is also presented. Maximising the posterior is equivalent to minimising the negative log-posterior, $P(\boldsymbol{\rho}, \sigma^2) \equiv -\log p(\boldsymbol{\rho}, \sigma^2 | \mathbf{y}, \boldsymbol{\lambda})$, which is given by

$$P(\boldsymbol{\rho}, \sigma^2) = l(\boldsymbol{\rho}, \sigma^2) + \frac{1}{\sigma} \sum_{j=1}^{k} \lambda_j |\rho_j| + \left(\frac{k}{2} + \nu \right) \log \sigma^2 + \text{const.} \tag{9}$$

An efficient method for solving the standard LASSO problem is cyclical coordinate wise descent. We note that although the procedure is presented for the general LASSO in which each λ_j may be different, setting $\boldsymbol{\lambda} = \lambda \mathbf{1}_k$ allows the regular LASSO solution for a single, global, regularisation parameter to be computed.

3.1 Coordinate-Wise Descent Algorithm

Cyclic coordinate-wise descent (Friedman et al., 2007) is a simple and highly effective optimisation algorithm that works by iteratively optimising each coordinate of a multivariable function while keeping all the remaining coordinates fixed. The major strength of this approach is that each stage of the optimisation involves optimising a simpler function than the complete, joint function. Let $P_j(\cdot) \equiv P(\cdot | \boldsymbol{\rho}^{-j}, \sigma^2)$ denote the conditional negative log-posterior. Starting with some initial values for $\hat{\boldsymbol{\rho}}$ and $\hat{\sigma}^2$ (for example, the unrestricted maximum likelihood estimates), the cyclic coordinate-wise algorithm minimises (9) by the following steps:

1. Evaluate current negative log-posterior value

$$P_0 \leftarrow P(\hat{\boldsymbol{\rho}}, \hat{\sigma}^2)$$

2. Coordinate-wise updates for $\hat{\boldsymbol{\rho}}$; for $j = 1, \ldots, k$

$$\hat{\rho}_j \leftarrow \underset{\rho \in (-1,1)}{\arg\min} \{P_j(\rho)\} \tag{10}$$

   This is discussed in Section (3.2).

3. Update for $\hat{\sigma}^2$

$$\hat{\sigma}^2 \leftarrow \underset{\sigma^2 \in \mathbb{R}_+}{\arg\min} \left\{P(\sigma^2; \hat{\boldsymbol{\rho}})\right\} \tag{11}$$

   This is discussed in Section (3.3).

4. If $|P_0 - P(\hat{\boldsymbol{\rho}}, \hat{\sigma}^2)| < \delta$, for some small δ, terminate; otherwise, return to Step 1.

3.2 Optimising for $\hat{\rho}_j$

From (9) it is clear that although the conditional negative log-posterior functions $P_j(\rho)$ are continuous, they are not differentiable at the point $\rho = 0$, which causes some problems for standard gradient descent procedures. The functions may instead by treated in a piece-wise fashion, with one piece on either side of $\rho = 0$. Due to the nature of the Laplace prior, the minimiser of $P_j(\rho)$ will either have the same sign as the minimiser of the conditional negative log-likelihood function $l_j(\rho_j)$, which is given by $-\text{sgn}(g_j)$, or be exactly equal to zero. This means we can minimise the piecewise function in the orthant of ρ-space corresponding to the sign of $-g_j$, and then test this solution against the solution at $\rho = 0$. This may be done using the following steps: first, minimise the piecewise function in the orthant corresponding to the sign of $-g_j$ by solving the cubic

$$H_{j,j}\rho^3 + \left(\tilde{\lambda}_j + g_j\right)\rho^2 - (H_{j,j} + j)\rho - \tilde{\lambda}_j - g_j = 0,$$

for the single solution, say $\tilde{\rho}_j$, that lies in $(-1, 1)$, where

$$\tilde{\lambda}_j = -\frac{\text{sgn}(g_j)\lambda_j}{\hat{\sigma}}.$$

and g_j and $H_{j,j}$ are discussed in Section 2.2. The solution to the update equation (10) is then given by

$$\hat{\rho}_j \leftarrow \begin{cases} \tilde{\rho}_j & \text{if } \mathrm{sgn}(\tilde{\rho}_j) = -\mathrm{sgn}(g_j) \\ 0 & \text{otherwise} \end{cases},$$

which clearly demonstrates the ability of the Laplace prior to produce solutions with elements that are exactly zero.

3.3 Optimising for $\hat{\sigma}^2$

Updating $\hat{\sigma}^2$ involves optimising the function

$$P(\sigma^2; \hat{\boldsymbol{\rho}}) = d \log \sigma^2 + \frac{S_1}{\sigma^2} + \frac{S_2}{\sigma}$$

over $\sigma^2 \in \mathbb{R}_+$, where

$$\begin{aligned} d &= \left(\frac{n+k}{2}\right) + \nu, \\ S_1 &= \boldsymbol{\beta}'(\hat{\boldsymbol{\rho}})\mathbf{D}\boldsymbol{\beta}(\hat{\boldsymbol{\rho}})/2, \\ S_2 &= \sum_{j=1}^{k} \lambda_j |\hat{\rho}_j|. \end{aligned}$$

The update (11) is then given by

$$\hat{\sigma} \leftarrow \frac{\left(S_2^2 + 16\, d\, S_1\right)^{1/2} + S_2}{4d},$$

which has time complexity $O(k^2)$.

4 MCMC Sampling from the Posterior Distribution

The posterior mode offers a single point-estimate summary statistic of the posterior distribution. However, this is often considered unsatisfactory in that it fails to describe the uncertainty associated with the point estimate. The more tightly concentrated the posterior distribution is around the posterior mode, the less uncertainty there is about this point estimate. The normalising constant for the posterior distribution (2) cannot be determined analytically, and the indirect approach of exploring $p(\boldsymbol{\rho}, \sigma^2 | \mathbf{y}, \boldsymbol{\lambda})$ based on Markov-Chain Monte-Carlo (MCMC) sampling is adopted instead.

To sample from the posterior (2), with priors given by (3) and (4), one may use a Gibbs sampling approach. This is possible even though the partial autocorrelations $\boldsymbol{\rho}$ are constrained to a compact subset of $\mathbb{R}^k$. The results in Gelfand et al. (1992) show that Gibbs sampling under parameter constraints is applicable if the constraints themselves do not depend on any of the variables being sampled. This condition is clearly met in the case of autoregressive models. The conditional posterior distributions required for Gibbs sampling are

$$p(\rho_j|\boldsymbol{\rho}_j^{-j}, \sigma^2, \boldsymbol{\lambda}, \mathbf{y}) \quad \propto \quad \exp\left(-g_j\rho_j - \frac{H_{j,j}\rho_j^2}{2} - \frac{\lambda_j|\rho_j|}{\sigma}\right) \cdot \left(1-\rho_j^2\right)^{\frac{j}{2}}, \tag{12}$$

$$p(\sigma^2|\boldsymbol{\rho}, \boldsymbol{\lambda}, \nu, \mathbf{y}) \quad \propto \quad \left(\sigma^2\right)^{-d} \cdot \exp\left(-\frac{S_1}{\sigma^2} - \frac{S_2}{\sigma}\right), \tag{13}$$

where g_j and $H_{j,j}$ are given in Section 2.2 and d, S_1 and S_2 are given in Section 3.3. Unfortunately, neither of these distributions are of standard form. However, the conditional posterior for ρ_j is log-concave which implies that the adaptive rejection sampling (ARS) (Gilks & Wild, 1992) algorithm may be used to efficiently simulate from both distributions. To see this, note that the first factor in (12) is proportional to the posterior of a normal likelihood with a Laplace prior. The exact form of this posterior is known to be continuous and orthant-wise Gaussian (Hans, 2009), and thus log-concave. The second factor in (12) is log-concave, and as the product of two log-concave functions is also log-concave, it follows that ARS may be used, as long as a little care is taken to handle the non-differentiable point at $\rho = 0$. The conditional posterior for σ^2 is not log-concave, though it may be rendered so by taking the change of variables $v = \sqrt{1/\sigma^2}$, and ARS may also be used to simulate from this distribution. The details required for ARS sampling from these conditional posteriors are given in Appendix B.

4.1 Estimation of λ

The Gibbs sampling approach discussed in Section 4 simulates from the posterior of an autoregressive model with Laplace priors, given a fixed vector of hyperparameters $\boldsymbol{\lambda}$. Usually, suitable values of the hyperparameters are unknown *a priori*, and must be estimated from the data. There are a variety of ways to do this in the more conventional LASSO setting, such as using information criterion, cross-validation (Tibshirani, 1996) and the bootstrap (Hall et al., 2009). The latter approach can be applied to the autoregressive setting by the use of a method such as the stationary bootstrap (Politis & Romano, 1994). In the case of the Bayesian LASSO there are several alternative procedures available. This paper considers two such approaches: (i) an empirical Bayes procedure, which estimates the hyper-parameters using marginal maximum likelihood, and (ii) a fully Bayesian

approach in which a prior distribution is specified over the hyper-parameter λ, which is then included in the Gibbs sampling as an unknown parameter.

4.1.1 Empirical Bayes Estimation of λ

The empirical Bayes, or marginal maximum likelihood, approach to estimating the hyper-parameters advocates using the value of $\boldsymbol{\lambda}$ that maximises the marginal probability, that is,

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in \mathbb{R}_+^k}{\arg\max} \left\{ \int \int p(\mathbf{y}|\boldsymbol{\rho}, \sigma^2) \pi_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\sigma^2, \boldsymbol{\lambda}) \pi_{\sigma^2}(\sigma^2) d\sigma^2 \, d\boldsymbol{\rho} \right\}.$$

In the case of the Bayesian LASSO, the marginal distribution is difficult to compute; however, given that samples from the posterior distribution are available one may use Casella's empirical Bayes procedure based on MCMC (Casella, 2001) to obtain an approximate marginal maximum likelihood estimate. This procedure utilises the expectation-maximisation algorithm by viewing the model parameters as "missing data". We now give the details for the case of the regular Bayesian LASSO in which $\lambda_1 = \lambda_2 = \ldots = \lambda_k = \lambda$. Let $\lambda^{(i)}$ denote the i-th estimate of λ. The terms in the posterior that depend on λ are given by

$$-k \log \lambda + \frac{\lambda \sum_{j=1}^k |\rho_j|}{\sigma}. \tag{14}$$

To derive the empirical Bayes update, first take expectations of (14), conditional on the real data $\mathbf{y}$, and under $\lambda^{(i-1)}$; this involves replacing the terms σ and ρ_j with their conditional expectations $\mathrm{E}_{\lambda^{(i-1)}}[\sigma|\mathbf{y}]$ and $\mathrm{E}_{\lambda^{(i-1)}}[|\rho_j||\mathbf{y}]$. The empirical Bayes estimate is then found by maximising the expectation of (14) with respect to λ, yielding the update

$$\lambda^{(i)} \leftarrow \frac{k \, \mathrm{E}_{\lambda^{(i-1)}}[\sigma|\mathbf{y}]}{\sum_{j=1}^k \mathrm{E}_{\lambda^{(i-1)}}[|\rho_j||\mathbf{y}]}, \tag{15}$$

where the expectations are approximated by averaging the previous $m > 0$ samples of σ and $|\rho_1|, \ldots, |\rho_k|$ obtained from the Gibbs sampler. Adapting the proposal in Park & Casella (2008), the initial values of the hyperparameters were chosen to be

$$\lambda^{(1)} = \frac{2 \, k \, \hat{\sigma}_{\mathrm{ML}}}{\sum_{j=1}^k |\hat{\rho}_j^{\mathrm{ML}}|},$$

where $\hat{\sigma}^2_{\mathrm{ML}}$ is the maximum likelihood estimate of σ^2, and $\hat{\rho}_j^{\mathrm{ML}}$ are the maximum likelihood estimates of ρ_j. An efficient algorithm for computing the maximum likelihood estimates in the case of autoregressive models is detailed

in Schmidt & Makalic (2011). The above choice of initial value, along with the choice of $m = 100$, was found to work well in experiments, with the empirical Bayes estimator generally stabilising in under a thousand iterations of the Gibbs sampler.

4.1.2 Bayesian Estimation of λ

In contrast to the empirical Bayes procedure discussed in the previous section, Bayesian estimation of the hyper-parameter λ requires the specification of a prior distribution, $\pi_\lambda(\cdot)$. The advantage of Bayesian estimation, in comparison to the empirical Bayes approach, is that the resulting estimation procedure automatically takes into account the uncertainty associated with the point estimate of the hyperparameter, and facilitates the construction of suitable credible sets for λ if these are required. It is convenient to choose the prior to be a conjugate Gamma distribution with shape parameter α and rate (inverse-scale) parameter δ, i.e.,

$$\lambda|\alpha, \delta \sim \text{Ga}(\alpha, \delta).$$

With this choice of prior the conditional posterior distribution for λ is also a Gamma distribution, with a special rate and inverse-scale parameter,

$$\lambda|\boldsymbol{\rho}, \sigma^2, \mathbf{y}, \alpha, \delta \sim \text{Ga}\left(\alpha + k, \delta + \frac{1}{\sigma}\sum_{j=1}^{k}|\rho_j|\right). \tag{16}$$

A Bayesian estimate of λ may then be obtained by modifying the Gibbs sampler presented in Section 4 to include an extra step in which a sample of λ is drawn from (16).

If prior knowledge about the distribution of λ is available the hyperparameters α and δ may be chosen accordingly. However, this is in general unlikely to be the case, and a simple, empirical procedure for choosing α and δ is presented. In the case of the regular Bayesian LASSO for linear regression models, it is recommended (Park & Casella, 2008) that the prior be chosen so that it puts reasonable prior probability mass near the maximum likelihood estimates, while tailing off sufficiently fast as $\lambda \to \infty$. In contrast to the standard linear regression model, the assumption of stationarity ensures that the magnitude of the maximum likelihood estimates satisfies $|\rho_j^{\text{ML}}| < 1$; taking $\alpha = 1$, and $\delta = 1/(a\,\hat{\sigma}_{\text{ML}})$ yields a prior with a mean that is a times greater than the value of λ that would be expected *a priori* if the time series was generated by an autoregressive model with all partial autocorrelations close to the boundary of the stationarity region. The factor of $1/\hat{\sigma}_{\text{ML}}$ in this choice of δ ensures that the resulting inferences of

λ are not sensitive to scale-transformations of the time series. The estimates of λ obtained using this prior are not dissimilar to those obtained by the empirical Bayes procedure described in Section 4.1.1. To see this, note that the mode of the conditional posterior distribution (16) is given by

$$\frac{k\,\sigma}{\sum_{j=1}^{k} |\rho_j| + 1/a}. \tag{17}$$

Comparing (17) to (15) reveals a great similarity. Of course, in contrast to the empirical Bayes procedure, the Bayesian estimate is not solely characterised by the mode of the posterior. The form of the mode, however, serves to underline the similarity between the two procedures. One difference between the two procedures is the presence of the additional $1/a$ term in the denominator; larger values of a have less effect on the conditional posterior mode, but lead to prior distributions that more thinly spread the prior probability mass near values of λ that would be reasonable. The choice $a = 10$ results in a prior that puts sufficient probability mass on the range of values of λ that might reasonably be anticipated, irrespective of the order of the autoregressive model.

4.2 Point Estimation from the Posterior

The set of partial autocorrelations produced by the MCMC algorithm may be used to produce a suitable point estimate, or summary statistic, of the posterior distribution. There is no single best procedure for determining a point estimate from the posterior distribution, and we propose the following as plausible estimates of the partial autocorrelations:

- the coordinate-wise posterior medians (as per Park & Casella (2008));
- the posterior mode, as determined by the algorithm in Section 3 using the median of the λ samples generated by either of the procedures discussed in Section 4.1;
- a minimum Bayes risk estimator equipped with a suitable loss function, $L(\boldsymbol{\rho}, \boldsymbol{\varrho})$, i.e., by solving

$$\hat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\varrho}} \left\{ \sum_{i=1}^{s} L(\boldsymbol{\rho}^{(i)}, \boldsymbol{\varrho}) \right\}$$

  where s is the number of samples from the posterior distribution, and $\boldsymbol{\rho}^{(i)}$ is the i-th sampled vector of partial autocorrelations. Suitable loss functions include model error (Broersen, 1998) and Kullback–Leibler divergence (Kullback & Leibler, 1951).

In some instances it may be of more interest to make inferences about the coefficients, rather than the partial autocorrelations. This is particularly the case if the resulting model will be used to make predictions about future data. In this case, one option is to use the mean of the implied posterior for the coefficients. Let $\boldsymbol{\phi}(\boldsymbol{\rho})$ denote the coefficients corresponding to the partial autocorrelations $\boldsymbol{\rho}$. Given a set of s samples, $\boldsymbol{\rho}^{(i)}$, from the posterior over the partial autocorrelations, this can be approximated as

$$\mathrm{E}\left[\boldsymbol{\phi}|\mathbf{y}\right] \approx \frac{1}{s}\sum_{i=1}^{s}\boldsymbol{\phi}(\boldsymbol{\rho}^{(i)}). \tag{18}$$

This estimate is quick to compute and is guaranteed to be stationarity. This follows by noting that the stationarity region in coefficient space is convex, and that the all of the samples $\boldsymbol{\rho}^{(i)}$ represent stationary processes.

5 Discussion of the Priors

This work proposes to place truncated Laplace priors over the partial autocorrelation parameters, $\boldsymbol{\rho}$, of an autoregressive model; however, it is more conventional to parameterise an autoregressive model in terms of its coefficients, $\boldsymbol{\phi}$. It is therefore of some interest to explore the behaviour of the induced prior distribution in coefficient space. The transformation from partial autocorrelations to coefficients is complex and it is difficult to determine the exact form of the induced prior distribution $\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi})$. It is possible, however, to obtain some results regarding the first and second-order moments of the resulting prior, as established by the following theorem.

Theorem 1: *Let $\rho_1, \ldots, \rho_k$ be independent random variables satisfying $\mathrm{E}\left[\rho_j\right] = 0$, for all $j = 1, \ldots, k$. Then, the resulting coefficients $\phi_1, \ldots, \phi_k$ that correspond to the partial autocorrelations $\rho_1, \ldots, \rho_k$ satisfy*

$$\begin{aligned} \mathrm{E}\left[\phi_j\right] &= 0, \\ \mathrm{Cov}\left[\phi_i\phi_j\right] &= 0, \quad \text{for all } i \neq j. \end{aligned}$$

The proof of Theorem 1 is given in Appendix C. As the Laplace priors (3) satisfy the conditions of Theorem 1, it can be seen that the induced prior on $\boldsymbol{\phi}$ models the coefficients as zero meaned, uncorrelated, random variables. Furthermore, it is trivial to see from the equations presented in Appendix A that ϕ_k is distributed exactly as per $-\rho_k$.

5.1 Beta Priors for Partial Autocorrelations

This work is not the first to consider Bayesian analysis of partial autocorrelations. Recent work (Daniels & Pourahmadi, 2009) has examined the use of Beta distributions as independent priors for partial autocorrelations in the general setting of covariance matrix estimation, i.e., $(\rho_j + 1)/2 \sim \text{Be}(\alpha, \beta)$, which leads to a density on ρ_j of the form

$$\rho_j|\alpha, \beta \propto (1+\rho)^{\alpha-1} (1-\rho)^{\beta-1}. \tag{19}$$

In the case that $\alpha = \beta > 1$, the prior (19) induces shrinkage behaviour on the *maximum a posteriori* estimates, as probability mass is concentrated near $\rho = 0$. The value of α controls the degree of shrinkage, with larger values resulting in greater levels of shrinkage. However, in contrast to the truncated Laplace priors (3) proposed in this work, the Beta priors do not lead to sparse estimates of $\boldsymbol{\rho}$. To see this, consider the conditional negative log-posterior, up to constants, for ρ_j when $(\rho_j + 1)/2 \sim \text{Be}(\alpha, \alpha)$:

$$l_j(\rho_j) - (\alpha - 1)\left(\log(1+\rho_j) + \log(1-\rho_j)\right), \tag{20}$$

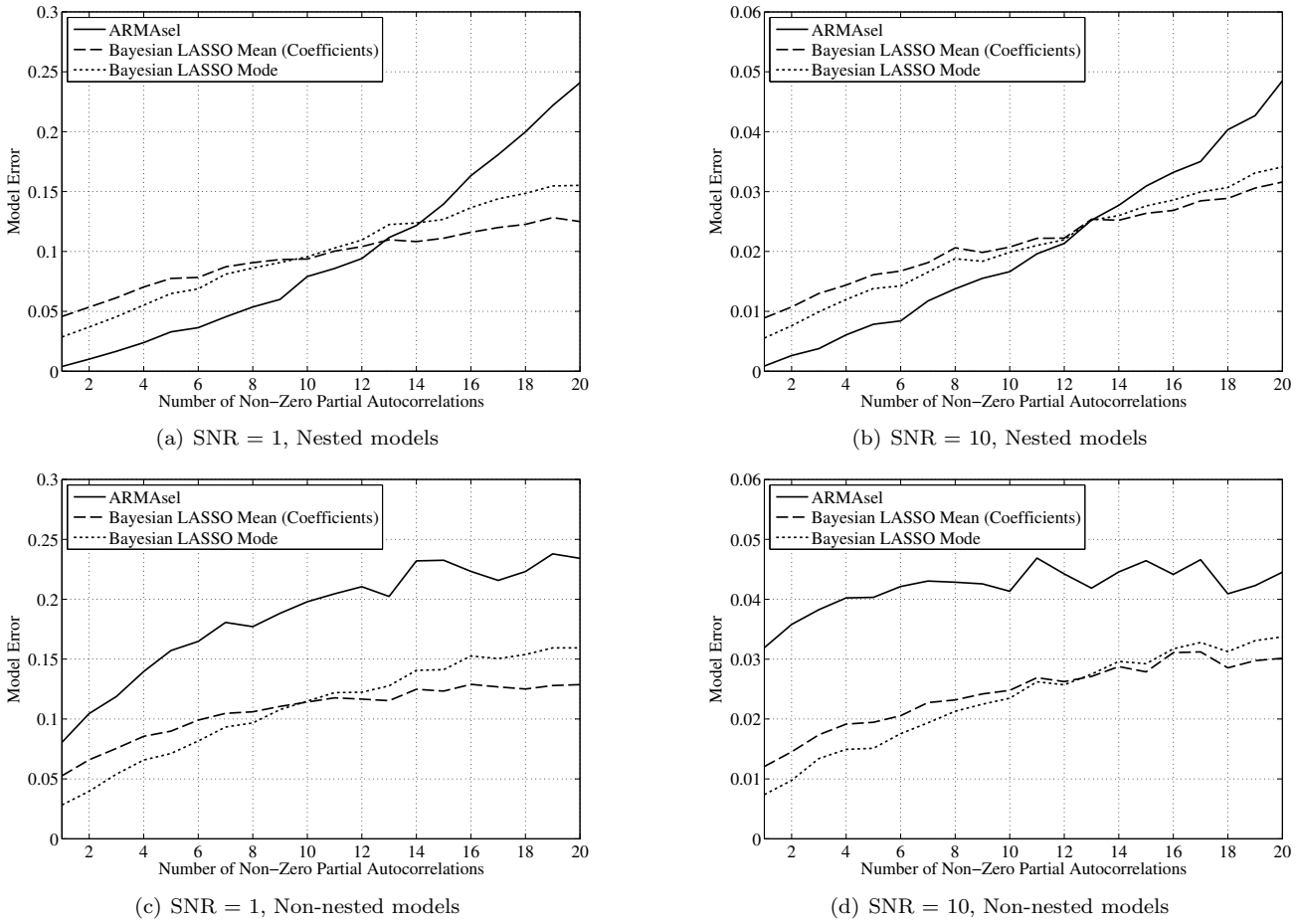
for $j = 1, \ldots, k$, where $l_j(\rho_j)$ is given in Section 2.2. The value of ρ_j that minimises (20) is a solution of the cubic equation

$$H_{j,j}\rho^3 + g_j\rho_j^2 - (2\alpha + j + H_{j,j} - 2)\rho - g_j = 0. \tag{21}$$

where g_j and $H_{j,j}$ are given in Section 2.2. Due to the fact that g_j is obtained from a quadratic expansion of $l_j(\rho_j)$ around the point $\rho_j = 0$, it is clear that $g_j = 0$ if and only if $\rho_j = 0$ minimises the conditional negative log-likelihood $l_j(\rho_j)$; this implies that $\rho_j = 0$ can never be a solution of (21) for finite α unless $\rho = 0$ is also a solution of the maximum likelihood estimator (i.e., no shrinkage is being done). It therefore follows that symmetric Beta priors cannot lead to sparse estimates of $\boldsymbol{\rho}$ for finite values of α.

6 Experiments

The behaviour of the Bayesian LASSO was assessed by comparing it against a well established model selection method for estimating autoregressive models. The ARMAsel (Broersen, 2006) toolbox has been developed over a series of papers, and uses a finite-sample information criterion (Broersen, 2000) in combination with the Burg estimator (Burg, 1967), to automatically select a suitable autoregressive model from a given time series.



(a) SNR = 1, Nested models

(b) SNR = 10, Nested models

(c) SNR = 1, Non-nested models

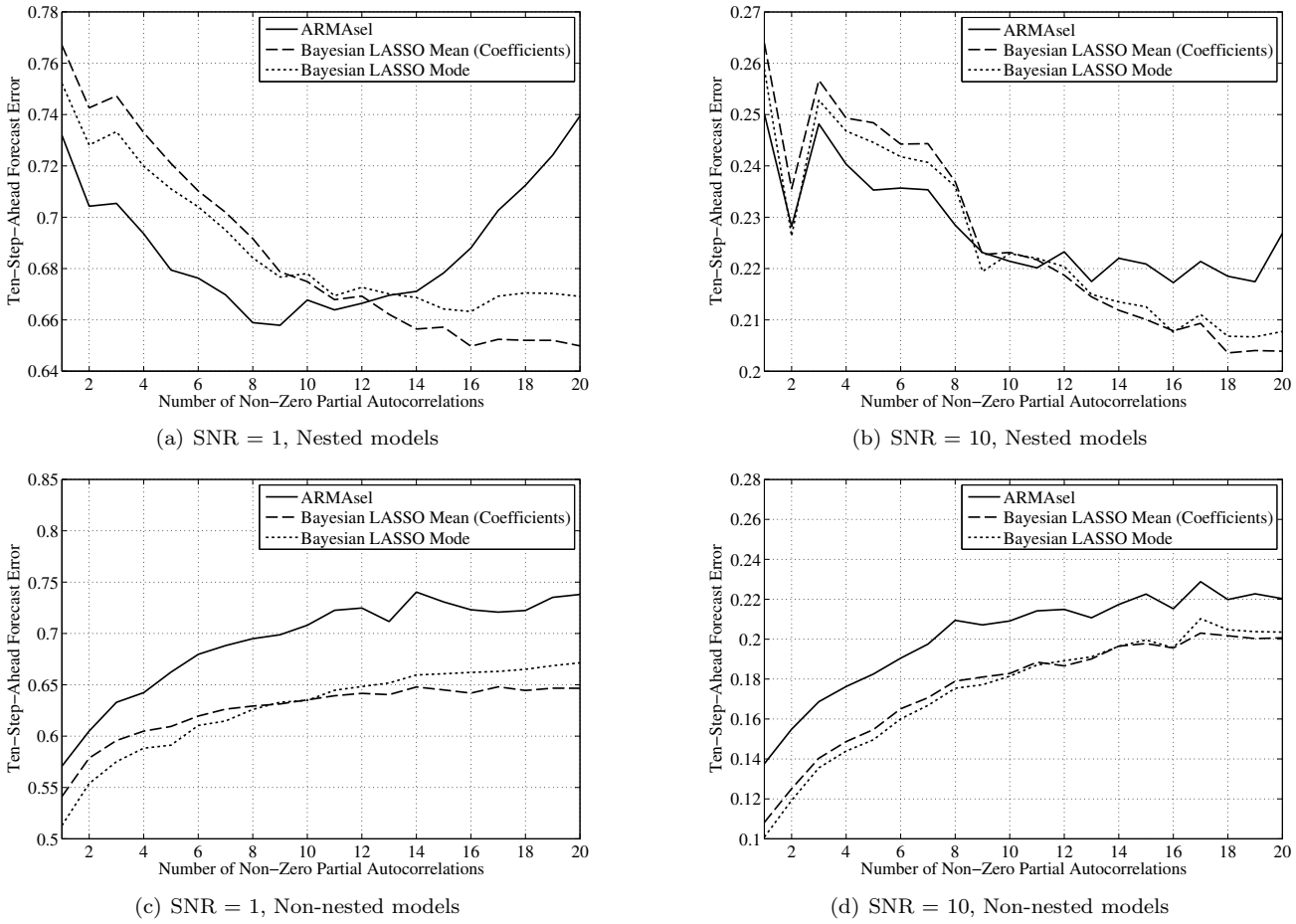
(d) SNR = 10, Non-nested models

Figure 1: Model errors for varying sparsity levels

6.1 Synthetic Data

Four experiments were undertaken to gauge the ability of the Bayesian LASSO methods to estimate autoregressive models under varying levels of signal-to-noise ratio (SNR) and model "sparsity" for both nested and non-nested model structures. In this setting, we defined sparsity as the number of non-zero partial autocorrelations in the underlying, true autoregressive model that generated the data. In all experiments, the number of partial autocorrelations estimated by the LASSO methods was $k = 20$. For each combination of sparsity level, $p = \{1, \ldots, 20\}$, signal-to-noise ratio SNR $= \{1, 10\}$, and model structure, 500 random autoregressive models were generated. The model structure determined the pattern of non-zero partial autocorrelations as follows:

- *Non-nested*: p of the first k partial autocorrelations were randomly chosen to be non-zero.



(a) SNR = 1, Nested models

(b) SNR = 10, Nested models

(c) SNR = 1, Non-nested models

(d) SNR = 10, Non-nested models

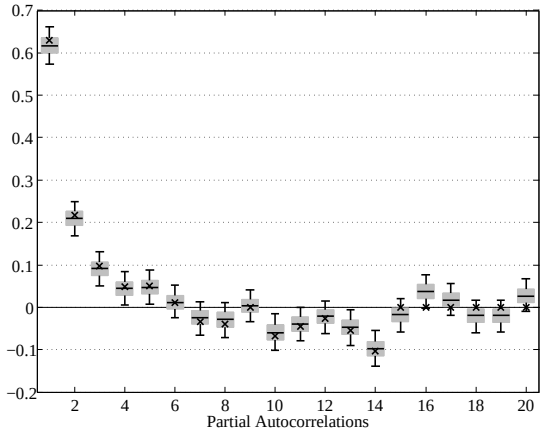
Figure 2: Ten-step-ahead forecast errors for varying sparsity levels

- *Nested*: the first p partial autocorrelations were always chosen to be non-zero.

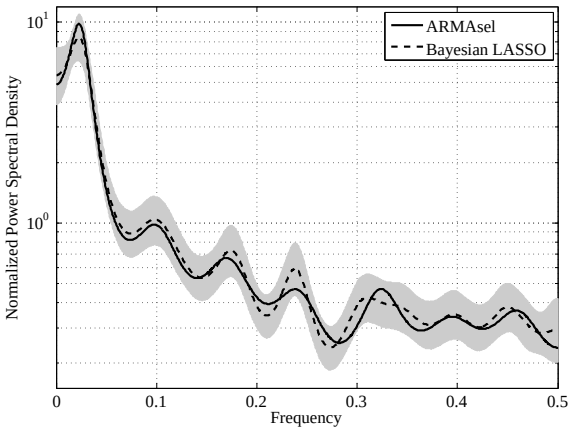
In both cases, the values of the non-zero partial autocorrelations were randomly sampled uniformly on $(-1, 1)$, and subsequently normalised to control the signal-to-noise ratio of the resulting autoregressive model. Letting $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_k)$ denote the randomly sampled partial autocorrelations, the normalisation is done by solving

$$\prod_{i=1}^{k} \frac{1}{(1 - (\kappa\varrho_i)^2)} - 1 = \text{SNR}$$

for κ, and using $\boldsymbol{\rho}^* = \kappa\boldsymbol{\varrho}$ as the partial autocorrelations of the true model. For each true model, a time series of $n = 100$ samples was randomly generated and ARMAsel and the Bayesian LASSO were used to estimate the partial autocorrelations. Two methods were used to obtain point estimates for the Bayesian LASSO: (i) the posterior mean



(a) Partial autocorrelation estimates from the regular Bayesian LASSO; whiskers cover the 95% credible intervals, crosses are the Burg estimates



(b) Spectral densities estimated by the Bayesian LASSO and ARMAsel

Figure 3: Analysis of the Southern Oscillation Index Data

of the samples in coefficient space, given by (18), and (ii) the posterior mode of the partial autocorrelations, using the median of the sequence of $\lambda^{(i)}$ samples produced by the Bayesian procedure (described in Section 4.1.2) as an estimate of λ. The accuracy of the three estimated models was assessed using two measures: (i) the normalised model error, and (ii) the normalised K-step-ahead forecast error, with $K = 10$. The model error between a true autoregressive model with coefficients $\boldsymbol{\phi}^*$, and an estimated autoregressive model with coefficients $\hat{\boldsymbol{\phi}}$ is given by

$$\mathrm{ME}(\hat{\boldsymbol{\phi}}, \boldsymbol{\phi}^*) = \left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right)' \boldsymbol{\Gamma}^* \left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right) / \gamma_0^*, \tag{22}$$

where $\boldsymbol{\Gamma}^*$ is a Toeplitz autocovariance matrix with entries

$$\Gamma^*_{i,j} = \gamma^*_{|i-j|}$$

and $\gamma_j^* = \mathrm{E}\left[y_t y_{t-j}\right]$ is the k-th autocovariance of data generated under the true model $\boldsymbol{\phi}^*$. The forecast error is computed as the mean squared-deviation between forecasts of the next K samples made by the estimated models, and the next K actual realisations of the time series, i.e.,

$$\mathrm{FE}_K(\hat{\boldsymbol{\phi}}, \mathbf{y}^*) = \left(\frac{1}{K\gamma_0^*}\right) \sum_{i=k+1}^{k+K-1} (\hat{y}_i(\hat{\boldsymbol{\phi}}|y_1, \ldots, y_k) - y_i^*)^2, \tag{23}$$

where $\hat{y}_i(\hat{\boldsymbol{\phi}}|y_1, \ldots, y_k)$ are forecasting predictions made using the model $\hat{\boldsymbol{\phi}}$, conditional on the previous k samples, and y_i^* are realisations of the generating process (Brockwell & Davis, 1987). The forecast errors were then averaged over 1,000 realisations of the time series generated from the true model $\boldsymbol{\phi}^*$. It should be noted that the model error measure given by (22) is equivalent to the expectation of the K-step-ahead forecast error, (23), with $K = 1$.

In all cases the Gibbs sampler was run for $4,000$ iterations, discarding the first $1,000$ samples as "burn-in". With this quantity of samples there appeared to be no convergence issues, and the adaptive rejection sampling attained acceptance rates in the order of $90\% - 95\%$ (i.e., around 1.1 samples per call to the ARS algorithm). The complete set of experiments took approximately one hundred hours to complete on a standard laptop.

The results, in terms of model error, are presented in Figures 1(a) through 1(d). The curves show the median model errors obtained by the ARMAsel procedure and the Bayesian LASSO estimates based on the posterior mean of the coefficients, and the posterior mode of the partial autocorrelations. There are several points of interest regarding the behaviour of the three methods with respect to each other and the model structure. The performance of the two Bayesian LASSO procedures is essentially unaffected by the choice of model structure for a given p and SNR. This is in contrast to the ARMAsel procedure for which the performance differs greatly depending on whether the underlying models are nested or non-nested. This is not unexpected, given that the procedure is an order selection method that makes the explicit assumption that the generating model is of a nested structure. Interestingly, the two Bayesian LASSO based approaches appear robust to this assumption. Furthermore, even in the case of nested models, the two Bayesian LASSO based approaches appear to be outperforming the ARMAsel procedure when the number of non-zero partial autocorrelations becomes large.

The posterior mode estimates outperform the posterior mean estimates when the number of non-zero partial autocorrelations is small; this is expected as the posterior mode is capable of estimating partial autocorrelations as exactly zero which is beneficial when the underlying model is sparse. In contrast, when the model is dense, the posterior mean of the coefficients performs better than the posterior mode. The results, in terms of ten-step-ahead forecast error are presented in Figures 2(a) through 2(d). The general trends are essentially the same as in the above discussion. Once again, the ARMAsel procedure performs worse than the Bayesian LASSO based methods for less sparse models, and when the underlying structure is non-nested. In contrast, the two Bayesian LASSO based approaches once again appear robust to the choice of model structure.

As a general conclusion, the results suggest that the methods based on the Bayesian LASSO offer a viable alternative to regular order selection procedures such as ARMAsel for estimating autoregressive processes, and are expected to perform well regardless of the underlying structure of the generating model. Further, unless we believe

the underlying structure to be quite dense, the results suggest that the posterior mode, with its ability to produce sparse, parsimonious partial autocorrelation vectors, offers an excellent alternative to the posterior mean estimates.

6.2 Analysis of the Southern Oscillation Index

The difference in behaviour between the Bayesian LASSO and the ARMAsel procedure was examined by applying both methods to a real dataset. The Southern Oscillation Index (SOI) data measures the monthly fluctuations in air pressure difference between Tahiti and Darwin, and is used in studying and predicting *El Niño* phenomena. The dataset analysed contained $n = 1,619$ monthly measurements from January, 1876 through to December, 2010, and was obtained from the Australian Government Bureau of Meteorology website.

The Bayesian LASSO was run with $k = 20$, for 10,000 iterations, discarding the first 3,000 iterations as burn-in. A box plot of the partial autocorrelations for is given in Figure 3(a). The ARMAsel procedure selected an AR(14) model, and the corresponding Burg estimates of the partial autocorrelations are shown as crosses in the same figure. Despite the large sample size, the results of both procedures exhibit considerable differences. It is clear that the posterior median estimates of the partial autocorrelations for the Bayesian LASSO are uniformly smaller in absolute magnitude than the corresponding (non-zero) Burg estimates. The regular Bayesian LASSO has not shrunken any of the partial autocorrelations very close to zero, while the ARMAsel procedure has removed partial autocorrelations from 15 onwards.

The normalized power-spectral densities (PSD) of the AR models estimated by the Bayesian LASSO and the ARMAsel procedure are plotted against normalized discrete frequency in Figure 3(b). The median PSDs for the LASSO method is similar to the ARMAsel PSD; however, there are two points of interest: (i) although the positions of the peaks and troughs generally coincide between all methods, the Bayesian LASSO has estimated the trough near $f = 0.27$ at a slightly lower frequency than the ARMAsel procedure, and (ii) the PSD for the ARMAsel is flatter than that estimated by the Bayesian LASSO, and this is attributed to the fact that six of the partial autocorrelations have been completely regularised to zero by the ARMAsel procedure, resulting in a less complex model.

7 Extensions and Future Work

7.1 Estimation of λ by Information Criteria

In the original paper from Tibshirani (Tibshirani, 1996) the LASSO was interpreted as a penalised maximum likelihood parameter estimator. This interpretation remains equally valid in the case of the autoregressive LASSO proposed in this paper, and suggests that one may employ information criteria to estimate a suitable value of λ instead of using the procedures detailed in Section 4.1. While exact expressions for the degrees-of-freedom of the LASSO estimator for autoregressive models are currently unknown, we can appeal to the asymptotic equivalence of autoregressive models and linear regressions to use the results in Zou et al. (2007). If $\hat{\boldsymbol{\rho}}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ are the posterior mode estimates obtained by minimising the negative log-posterior (9), we can estimate λ using an information criteria by solving

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}_+}{\arg\min} \left\{ -\log p(\mathbf{y}|\hat{\boldsymbol{\rho}}(\lambda), \hat{\sigma}^2(\lambda)) + \alpha(\hat{k}(\lambda), n) \right\}, \tag{24}$$

where

$$\hat{k}(\lambda) = ||\hat{\boldsymbol{\rho}}(\lambda)||_0$$

is the ℓ_0 norm, that is, the number of non-zero entries of $\hat{\boldsymbol{\rho}}(\lambda)$, and $\alpha(k, n)$ is a suitable penalty term. For example, $\alpha(k, n) = k$ yields the Akaike information criterion, $\alpha(k, n) = (k/2) \log n$ yields the Bayesian information criterion, and so on. Once an estimate $\hat{\lambda}$ has been obtained, the MCMC procedure detailed in Section 4 may be used to compute "standard errors" for this particular λ, circumventing the need to use the bootstrap, which can be problematic for LASSO-type estimates (Kyung et al., 2010).

7.2 Alternative Penalization Schemes

It is reasonably straightforward to adapt the Gibbs sampler presented in Section 4, and associated conditionals, to several different penalization schemes. The most straightforward is the so-called "ridge" penalty, in which the coefficients are given a multivariate normal prior distribution

$$\boldsymbol{\rho}|\sigma^2, \lambda \sim N\left(\mathbf{0}_k, \left(\frac{\sigma^2}{\lambda}\right)\mathbf{I}_k\right),$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix. The hierarchy in Section 4 may easily accommodate this type of prior by noting that the conditional posterior density remains log-concave, and the ARS step requires only a small modification. A slight modification to the empirical Bayes and Bayesian estimates discussed in Section 4.1 is also required to estimate the ridge parameter λ. This procedure may also be extended to the normal-gamma penalty (Griffin & Brown, 2010) as well as the generalized beta penalty scheme (Armagan et al., 2011).

A further extension of the penalization scheme is the "adaptive" LASSO (or potentially, ridge), in which a hyperparameter λ_i is associated with each partial autocorrelation ρ_i. It is clear that both the algorithm for computing the posterior mode, given in Section 3, and the Gibbs sampling hierarchy given in Section 4, automatically handle this level of flexibility. The only extension required in this case is for a modification of the procedures used to estimate the hyperparameters. It is straightforward to alter both the empirical Bayes and fully Bayesian procedures discussed in Section 4.1 to allow for estimation of each individual λ_i. The performance of an adaptive LASSO/ridge procedure would be an interesting topic for future research.

7.3 LASSO Estimation of ARMA Models

A straightforward extension to the autoregressive Bayesian LASSO is to apply it to the more general autoregressive moving-average (ARMA) model, which explains a time series by

$$y_t + \sum_{j=1}^{k} \phi_j y_{t-j} = \varepsilon_t + \sum_{j=1}^{l} \theta_j \varepsilon_{t-j},$$

where $\boldsymbol{\phi}$ are the autoregressive coefficients and $\boldsymbol{\theta}$ are the moving average coefficients. It is generally a requirement that an estimated ARMA model be both stationary and invertible, the latter condition being dependent on all the zeros of the characteristic polynomial of the moving average component being completely within the unit circle. An obvious way to ensure both these conditions are met would be to parameterise both the autoregressive component and the moving average component in terms of partial autocorrelations. Unfortunately, the negative-log likelihood for an ARMA model is significantly more complex than in the case of a pure autoregression and such an extension would require considerable work to efficiently sample from the posterior distribution. The authors are currently exploring this as a future topic of research.

Appendices

Appendix A

Given a vector $\boldsymbol{\rho} \in (-1,1)^k$ of partial autocorrelations, the corresponding coefficients $\boldsymbol{\phi}$ are found by the following recurrence relations:

$$\boldsymbol{\phi}^{(1)} = -\rho_1 \tag{25}$$

$$\boldsymbol{\phi}^{(i)} = \left(\boldsymbol{\phi}^{(i-1)} - \rho_i \tilde{I}\boldsymbol{\phi}^{(i-1)}, -\rho_i\right), \quad (i > 1) \tag{26}$$

where $\boldsymbol{\phi}^{(i)}$ denotes the i-dimensional coefficient vector formed at the i-th iteration of the recurrence relations and $\tilde{I}$ denotes the permutation matrix that reverses the order of the elements of a vector. For notational simplicity we define $\boldsymbol{\phi} \equiv \left(\boldsymbol{\phi}^{(k)}\right)'$, i.e., the final coefficient vector is the vector formed at the k-th iteration of the recurrence relations. The gradient and Hessian of the negative log-likelihood, given by (5), both depend on the vectors $\partial\boldsymbol{\phi}/\partial\rho_j$. From (25) and (26), it is straightforward to see that the recurrence relations required to calculate the gradient vectors are given by

$$\partial\boldsymbol{\phi}^{(1)}/\partial\rho_j = \begin{cases} -\rho_1 & \text{for } j > 1 \\ -1 & \text{for } j = 1 \end{cases}$$

$$\partial\boldsymbol{\phi}^{(i)}/\partial\rho_j = \begin{cases} \left(\left[\partial\boldsymbol{\phi}^{(i-1)}/\partial\rho_j\right] - \rho_i\tilde{I}\left[\partial\boldsymbol{\phi}^{(i-1)}/\partial\rho_j\right], -\rho_i\right) & \text{for } j > i \\ \left(-\tilde{I}\left[\partial\boldsymbol{\phi}^{(i-1)}/\partial\rho_j\right], -1\right) & \text{for } j = i \\ \left(\left[\partial\boldsymbol{\phi}^{(i-1)}/\partial\rho_j\right] - \rho_i\tilde{I}\left[\partial\boldsymbol{\phi}^{(i-1)}/\partial\rho_j\right], 0\right) & \text{for } j < i \end{cases}$$

with $\partial\boldsymbol{\beta}/\partial\rho_j = (0, \partial\boldsymbol{\phi}^{(k)}/\partial\rho_j)'$.

Appendix B: Sampling ρ_j and σ^2

Adaptive Rejection Sampling for ρ_j

The adaptive rejection sampling (ARS) algorithm presented in Gilks & Wild (1992) requires the derivative of the conditional log-posterior distributions. In the case of the conditional for ρ_j, given by (12), this is simply

$$\frac{\partial \log p(\rho_j|\boldsymbol{\rho}_j^{-j}, \sigma^2, \boldsymbol{\lambda}, \mathbf{y})}{\partial\rho_j} = -g_j - H_{j,j}\rho - \frac{\text{sgn}(\rho)\lambda}{\sigma} - \frac{j\rho}{1-\rho^2}.$$

The curvature of the conditional negative log-posterior around the mode is used to aid in the selection of the starting knots used in ARS algorithm. This is given by

$$c_j = H_{j,j} + \frac{j(\hat{\rho}_j^2 + 1)}{(\hat{\rho}_j^2 - 1)^2},$$

where $\hat{\rho}_j$ denotes the minimum of the conditional negative log-posterior, which may be found using the equations described in Section 3.2. Finally, the following six points are used as initial knots for the ARS algorithm:

$$\begin{aligned}
x_1 &\leftarrow \max\left\{\frac{\hat{\rho}_j - 1}{2},\ \hat{\rho}_j - \frac{2}{\sqrt{c_j}}\right\},\\
x_6 &\leftarrow \min\left\{\frac{\hat{\rho}_j + 1}{2},\ \hat{\rho}_j + \frac{2}{\sqrt{c_j}}\right\},\\
x_2 &\leftarrow \frac{\hat{\rho}_j + x_1}{2},\\
x_5 &\leftarrow \frac{\hat{\rho}_j + x_6}{2},\\
x_3 &\leftarrow \frac{\hat{\rho}_j + x_2}{2},\\
x_4 &\leftarrow \frac{\hat{\rho}_j + x_5}{2},
\end{aligned}$$

where the points are chosen so that $x_1 < x_2 < \ldots < x_6$. Finally, to avoid issues with the non-differentiable point at $\rho_j = 0$, simply remove any x_j if they happen to be exactly equal to zero.

Adaptive Rejection Sampling for σ^2

Although the conditional log-posterior for σ^2, given by (13) is unimodal, it is not a concave function of σ^2. However, taking the change-of-variables $v = \sqrt{1/\sigma^2}$ yields a transformed density of the form

$$p(v|\boldsymbol{\rho}, \boldsymbol{\lambda}, \nu, \mathbf{y}) \propto v^{2d-3} \cdot \exp\left(-S_1 v^2 - S_2 v\right). \tag{27}$$

From Section 3.3 we see that $S_1 > 0$, $S_2 > 0$ and $d = (n+k)/2 + \nu$, implying that (27) is log-concave as long as $(n + k + 2\nu \geq 3)$, which holds for all datasets if we take $\nu \geq 1/2$. The derivative of the conditional log-posterior for ν is given by

$$\frac{\partial \log p(v|\boldsymbol{\rho}, \boldsymbol{\lambda}, \nu, \mathbf{y})}{\partial v} = \frac{2d-3}{v} - 2S_1 v - S_2.$$

To select the initial knots we adapt the procedure discussed in Appendix 1 of Hans (2009), and use the mode of the conditional posterior, along with the curvature of the conditional negative log-posterior to determine the position of the initial knots for the ARS algorithm. The mode of (27) is given by

$$\hat{v} = \frac{\left(S_2^2 + (16\,d - 24)S_1\right)^{\frac{1}{2}} - S_2}{4\,S_1},$$

and the curvature of the conditional negative log-posterior at the mode is

$$c_v = \frac{2d-3}{\hat{v}^2} + 2S_1.$$

Let $s_v = \sqrt{1/c_v}$, and let $K \geq 1$. If $\hat{v} - s_v/2 > 0$, we choose the initial points to be

$$\mathbf{x} = (\hat{v} - Ks_v, \ldots, \hat{v} - s_v, \hat{v} - s_v/2, \hat{v} + s_v/2, \hat{v} + s_v, \ldots, \hat{v} + Ks_v),$$

discarding any points x_i such that $x_i < 0$. Otherwise, if $\hat{v} - s_v/2 \leq 0$, choose the initial points to be

$$\mathbf{x} = (\hat{v}/2, \hat{v} + s_v/2, \hat{v} + s_v, \ldots, \hat{v} + Ks_v).$$

Once a sample has been drawn from (27) it can be transformed into a sample from (13) by taking $\sigma^2 = 1/v^2$.

Appendix C: Proof of Theorem 1

The transformation from $\boldsymbol{\rho}$ to $\boldsymbol{\phi}$ is described by the recurrence relation (26), with the base case given by (25). We begin by proving that $\mathrm{E}\left[\phi_j\right] = 0$ for all $j = 1, \ldots, k$. Recall that $\mathrm{E}\left[\rho_j\right] = 0$ for all $j = 1, \ldots, k$. Taking expectations of $\boldsymbol{\phi} \equiv \boldsymbol{\phi}^{(k)}$ and noting that $\boldsymbol{\phi}^{(j)}$ does not contain any ρ_i such that $i > j$, yields

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{\phi}^{(k)}\right] &= \left(\mathrm{E}\left[\boldsymbol{\phi}^{(k-1)}\right] - \mathrm{E}\left[\rho_k\right]\,\mathrm{E}\left[\tilde{\mathbf{I}}\,\boldsymbol{\phi}^{(k-1)}\right], -\mathrm{E}\left[\rho_k\right]\right) \\ &= \left(\mathrm{E}\left[\boldsymbol{\phi}^{(k-1)}\right], 0\right). \end{aligned} \tag{28}$$

The expectation $\mathrm{E}\left[\boldsymbol{\phi}^{(k-1)}\right]$ can then be calculated by using (28); repeating this process until we arrive at $\mathrm{E}\left[\boldsymbol{\phi}^{(1)}\right] = 0$ completes the proof.

To prove that $\mathrm{Cov}\left[\phi_i, \phi_j\right] = 0$ for all $i \neq j$, we first note that due to the fact that $\mathrm{E}\left[\phi_j\right] = 0$ we may write the

covariance as

$$\mathrm{Cov}\left[\phi_i, \phi_j\right] = \mathrm{E}\left[\phi_i \phi_j\right]. \tag{29}$$

As all the ϕ_j are multi-linear functions of $\boldsymbol{\rho}$, the expectation (29) will be non-zero if and only if ϕ_i and ϕ_j have at least one term in common. To prove that ϕ_i and ϕ_j have no terms in common when $i \neq j$, rewrite (26) as

$$\boldsymbol{\phi}^{(k)} = \left(\phi_1^{(k-1)} - \rho_k \phi_{k-1}^{(k-1)},\ \phi_2^{(k-1)} - \rho_k \phi_{k-2}^{(k-1)},\ \ldots, \phi_{k-1}^{(k-1)} - \rho_k \phi_1^{(k-1)},\ -\rho_k\right). \tag{30}$$

From (30), and the fact that the entries in $\boldsymbol{\phi}^{(k-1)}$ contain only those ρ_i such that $i \leq (k-1)$, it is clear that if the entries of $\boldsymbol{\phi}^{(k-1)}$ have no terms in common with each other, then the entries of $\boldsymbol{\phi}^{(k)}$ cannot have any terms in common with each other. As the base case of the recurrence relation is simply $\boldsymbol{\phi}^{(1)} = -\rho_1$, it follows that the entries of $\boldsymbol{\phi} \equiv \boldsymbol{\phi}^{(k)}$ have no terms in common with each other, which in turn implies that $\mathrm{E}\left[\phi_i \phi_j\right] = 0$ for $i \neq j$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Armagan, A., Dunson, D. B. & Clyde, M. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira & K. Weinberger, eds. 523–531. ArXiv:1107.4976v1.

Box, G., Jenkins, G. & Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. Prentice- Hall, (3rd ed.) ed.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

Brockwell, P. J. & Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer–Verlag.

Broersen, P. M. (2006). *Automatic Autocorrelation and Spectral Analysis*. Springer, 1st ed.

Broersen, P. M. T. (1998). The quality of models for ARMA processes. *IEEE Transactions on Signal Processing* **46**, 1749–1752.

Broersen, P. M. T. (2000). Finite sample criteria for autoregressive order selection. *IEEE Transactions on Signal Processing* **48**, 3550–3558.

Burg, J. P. (1967). Maximum entropy spectral analysis. In *Proc. 37th Meet. Soc. Explorational Geophys.* Oklahoma City, OK.

Casella, G. (2001). Empirical bayes Gibbs sampling. *Biostatistics* **2**, 485–500.

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters* **42**, 333–343.

Daniels, M. & Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis* **100**, 2352–2363.

Friedman, J., Hastie, T., Höfling, H. & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

Gelfand, A. E., Smith, A. F. M. & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**, 523–532.

Gilks, W. R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **41**, 337–348.

Griffin, J. E. & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.

Hall, P., Lee, E. R. & Park, B. U. (2009). Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica* **19**, 449–471.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika* **96**, 835–845.

Hsu, N.-J., Hung, H.-L. & Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis* **52**, 3645–3657.

Kay, S. M. (1983). Recursive maximum likelihood estimation of autoregressive processes. *IEEE Transactions on Acoustics, Speech and Signal Processing* **31**, 56–65.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5**, 369–412.

McLeod, A. I. & Zhang, Y. (2006). Partial autocorrelation parameterization for subset autoregression. *Journal of Time Series Analysis* **27**, 599–612.

Nardi, Y. & Rinaldo, A. (2008). Autoregressive process modeling via the Lasso procedure.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

Politis, D. N. & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Porat, B. & Friedlander, B. (1986). Computation of the exact information matrix of Gaussian time series with stationary random components. *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**, 118–130.

Schmidt, D. F. & Makalic, E. (2011). Estimating the order of an autoregressive model using normalized maximum likelihood. *IEEE Transactions on Signal Processing* **59**, 479–487.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B)* **58**, 267–288.

Wang, H., Li, G. & Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 63–78.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics* **35**, 2173–2192.