Robust Bayesian Model Selection for Autoregressive Processes With Additive Outliers

Nhu D. Le; Adrian E. Raftery; R. Douglas Martin

Journal of the American Statistical Association, Vol. 91, No. 433 (Mar., 1996), 123-131.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199603%2991%3A433%3C123%3ARBMSFA%3E2.0.CO%3B2-O

Robust Bayesian Model Selection for Autoregressive Processes With Additive Outliers

Nhu D. LE, Adrian E. RAFTERY, and R. Douglas MARTIN

Autoregressive (AR) models of order k are often used for forecasting and control of time series, as well as for the estimation of functionals such as the spectrum. Here we propose a method that consists of calculating the posterior probabilities of the competing AR(k) models in a way that is robust to outliers, and then obtaining the predictive distributions of quantities of interest, such as future observations and the spectrum, as a weighted average of the predictive distributions conditional on each model. This method is based on the idea of *robust Bayes factors,* calculated by replacing the likelihood for the nominal model by a *robust likelihood.* It draws on and synthesizes several recent research advances, namely robust filtering and the Laplace method for integrals, modified to take account of the finite range of the parameters. The method performs well in simulation experiments and on real and artificial data. Software is available from StatLib.

KEY WORDS: Additive outlier; Laplace approximation; Posterior probability; Robust filtering; Robust likelihood.

1. INTRODUCTION

The AR(k), or autoregressive (AR) model of order k, for a time series $\{x_t\}$ is defined by

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_k x_{t-k} + \varepsilon_t, \qquad (1)$$

where $\{\varepsilon_1, \varepsilon_2, \cdots\}$ is a sequence of independent $N(0, \sigma_\varepsilon^2)$ random variables. This model is useful for the forecasting and control of time series, as well as for the estimation of functionals such as the spectrum or the amount of energy in a given frequency band.

To use this model, a value for k must be specified. Because there is rarely a direct physical motivation for the AR(k) model (1), this choice must be based on the data. There has been much work on ways of making this choice (e.g., de Gooijer, Abraham, Gould, and Robinson 1985), with particular emphasis on automatic model selection criteria such as the Akaike information criterion (AIC; Akaike 1973) and the Bayes information criterion (BIC; Schwarz 1978).

To get an idea of the order selection difficulties caused by outliers, let $y_t = x_t + v_t$, where v_t is an iid outlier-generating component with variance σ_V^2 and x_t has variance σ_X^2. Then the lag-l correlations of the processes $\{y_t\}$ and $\{x_t\}$, denoted by ρ_l^Y and ρ_l^X, satisfy

$$\rho_l^Y = \rho_l^X(1 - R), \qquad l = 1, 2, \ldots,$$

where $R = \sigma_V^2/(\sigma_X^2 + \sigma_V^2)$. Therefore, as σ_V^2 increases to ∞ (i.e., as R increases to 1), ρ_l^Y decreases to zero. Thus model selection based on the empirical autocorrelation and partial autocorrelation functions can be misleading in the presence of outliers and tends to underestimate model order.

Both AIC and BIC are monotone functions of the prediction error variance σ_ε^2, usually estimated by the maximum likelihood estimate (MLE) $\hat{\sigma}_\varepsilon^2$. The commonly used MLE-based AIC and BIC are affected by outliers in two ways: (a) they are distorted by grossly unreliable parameter estimates, and (b) they are greatly inflated by outliers. This was pointed out by Martin (1980), who proposed a bounded-influence robustified likelihood approach to model selection. Another approach, based on robust filtering and an associated robustified likelihood, also was proposed by Martin (1981).

Here we propose a new approach to the comparison of AR models that attempts to overcome the difficulties associated with outliers and with model uncertainty. This consists of calculating the posterior probabilities of the competing AR(k) models in a way that is robust to outliers and then obtaining the predictive distributions of quantities of interest, such as future observations or characteristics of the spectrum, as a weighted average of the conditional predictive distribution given each of the models.

To obtain the posterior probabilities, we calculate the *Bayes factors,* or ratios of posterior to prior odds, for each of a set of pairwise model comparisons. The basic idea is explained in Section 2 in the context of AR models without outliers. In Section 3 we introduce the idea of *robust Bayes factors,* obtained by replacing the likelihood for model (1) by a *robust likelihood* following Martin (1981). This robust likelihood has two key ingredients. The first is a robust predictor that provides robust location or centering for the predictive distribution, along with an associated robust scale. The robust predictor and associated scale are obtained using the robust filtering algorithm of Masreliez (1975) and Martin (1979). The second ingredient is a bounded and continuous likelihood-type loss function that replaces the nonrobust sum of squared residuals in the Gaussian likelihood. Here robustness refers to robustness of the Bayes factors

Nhu D. Le is Research Scientist, Biometry Section, British Columbia Cancer Agency, Vancouver, B.C., Canada. Adrian E. Raftery is Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195. R. Douglas Martin is Vice President of Technology, StatSci Division of MathSoft and Professor of Statistics, University of Washington, Seattle, WA 98195. This research was supported by Office of Naval Research Contract N-00014-91-J-1074. Raftery's research was also partially supported by the Ministère de la Recherche et de l'Espace, Paris, by the Université de Paris VI, and by INRIA, Rocquencourt, France. Le's research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank Chris Chatfield, the editor, the associate editor, and three anonymous reviewers for their helpful comments on two earlier versions of this article.

Journal of the American Statistical Association
March 1996, Vol. 91, No. 433, Theory and Methods

against the outliers in the observed data, not against the prior distribution as often referred to in the literature.

Computation of the Bayes factors using the robust likelihood requires integration over the parameter space. Because this is analytically difficult, we follow Raftery (1988) and use the Laplace method for integrals (Tierney and Kadane 1986). In doing so, we reparameterize the model (1) in terms of the partial autocorrelation coefficients and modify the Laplace method to take into account the finiteness of the parameter space. The method works well in simulation experiments and in a real example (Secs. 4 and 5). Software to implement it is available; see Section 6.

2. BAYES FACTORS FOR TIME SERIES

2.1 Bayes Factors and Accounting for Model Uncertainty

Fruitful approaches to statistical problems often involve postulating a class of probability models and comparing these models on the basis of how well they predict the observed data. The Bayesian approach to the problem of inference in the presence of several competing models is based on *posterior model probabilities.* If the class consists of the $(K+1)$ models $M_0, \ldots, M_K$, then the posterior probability of the model M_k given data D is

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=0}^{K} p(D|M_l)p(M_l)} . \qquad (2)$$

In Equation (2), $p(M_k)$ is the prior probability of model M_k and $p(D|M_k)$ is its *integrated likelihood,* defined by

$$p(D|M_k) = \int_{\Theta_k} p(D|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)\, d\boldsymbol{\theta}_k, \qquad (3)$$

where $\boldsymbol{\theta}_k$ is the (vector) parameter of model $M_k, p(\boldsymbol{\theta}_k|M_k)$ is its prior distribution, $p(D|\boldsymbol{\theta}_k, M_k)$ is the likelihood, and Θ_k is the parameter space.

Pairwise comparisons are based on the posterior odds ratio

$$\frac{p(M_k|D)}{p(M_l|D)} = \left[\frac{p(D|M_k)}{p(D|M_l)}\right]\left[\frac{p(M_k)}{p(M_l)}\right] = B_{kl}\lambda_{kl},$$

where B_{kl} is the Bayes factor for M_k against M_l and λ_{kl} is the corresponding prior odds. If M_k is nested within M_l, then the data D favor M_k if $B_{kl} > 1$, whereas they provide evidence for M_l if $B_{kl} < 1$. Jeffreys (1961, app. B) suggested that the evidence for the larger model be considered strong if $B_{kl} < 10^{-1}$ and conclusive if $B_{kl} < 10^{-2}$. The posterior probabilities can be recovered using the equation

$$p(M_k|D) = \left[B_{0k}\lambda_{0k}\left\{1 + \sum_{l=1}^{K} (B_{0l}\lambda_{0l})^{-1}\right\}\right]^{-1} .$$

This framework yields solutions to the estimation, prediction, and decision-making problems that take into account uncertainty about the order of the AR, unlike model *selection* methods that condition on a single selected model. If Δ is a quantity of interest, such as a property of the spectrum, the next observation, or the utility of a course of action, then its posterior distribution given the data D is evaluated by combining all models considered; that is,

$$p(\Delta|D) = \sum_{k=0}^{K} p(\Delta|M_k, D)p(M_k|D). \qquad (4)$$

This equation was first given by Leamer (1978, p. 117) and was proposed explicitly as a solution to the decision-making problem in the time series context in equation (5.1) of Poskitt (1988).

A simple approximation for $p(D|M_k)$, introduced by Schwarz (1978), is

$$\log p(D|M_k) \approx \log p(D|M_k, \hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\, d \log n, \qquad (5)$$

where $\hat{\boldsymbol{\theta}}_k$ is the MLE of $\boldsymbol{\theta}_k$, and d and n are the numbers of parameters and observations. We refer to Equation (5) as the BIC approximation; its error is $O(1)$ (Kass and Raftery 1995). Choosing the order that maximizes the right side of Equation (5) is the much-used BIC model selection procedure.

Akaike (1983) wrote that, asymptotically,

$$\log p(D|M_k) \approx \log p(D|M_k, \hat{\boldsymbol{\theta}}_k) - d, \qquad (6)$$

which we call the AIC approximation. This is true only if prior information increases at the same rate as the information in the data, which is unrealistic in most applications. Nevertheless, the procedure of choosing the order that maximizes the right side of Equation (6) has been much used, and so we include it in our comparison and examples. (For a review of Bayes factors, see Kass and Raftery 1995.)

2.2 Bayes Factors for Autoregressive Processes Without Outliers

We now apply the Bayesian framework to the model comparison problem where the data $y^T = (x_1, \ldots, x_T)$ are from a stationary Gaussian AR(k) process defined by (1); that is, the observations contain no outliers. Let M_k denote the Gaussian AR(k) model. To obtain the posterior probabilities, we need to evaluate the integrated likelihood, $p(y^T|M_k), k = 0, \ldots, K$, which is given by Equation (3). The log-likelihood function of the data given the model and its parameters is

$$\log p(y^T|\boldsymbol{\theta}_k, M_k) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum \log f_t^2 - \frac{1}{2}\sum\left(\frac{x_t - \hat{x}_t^{t-1}}{f_t}\right)^2,$$

where $\hat{x}_t^{t-1} = E(x_t|y^{t-1}), y^{t-1} = (x_{t-1}, x_{t-2}, \ldots, x_1)$, and $f_t^2 = E(x_t - \hat{x}_t^{t-1})^2$. Thus $\hat{x}_t^{t-1}$ is the conditional mean of x_t given the data up to time $(t-1)$, and f_t^2 is the corresponding conditional variance; both can be found using the Kalman filter (Harvey 1981).

There are several difficulties with the evaluation of the integral (3). The constraints on the parameters $(\phi_1, \ldots, \phi_k)$ that ensure stationarity are complicated. We avoid this difficulty by reparameterizing in terms of the first k partial autocorrelations $\boldsymbol{\theta}_k = (\pi_1, \ldots, \pi_k)$. The parameter space Θ_k

is then just the hypercube $(-1,1)^k$, and the mapping that transforms the $(\phi_1,\ldots,\phi_k)$ such that the process is stationary to $(\pi_1,\ldots,\pi_k)$ is one-to-one and onto Θ_k, and both it and its inverse are continuously differentiable (Barndorff-Nielsen and Schou 1973; Ramsey 1974).

The integral (3) cannot be evaluated analytically, and so we approximate it using the Laplace method for integrals (Tierney and Kadane 1986). The Laplace method was applied to Bayes factors by Raftery (1988) in the context of generalized linear models. The Laplace method is modified here to take into account the finiteness of the parameter space, as follows. Let $g(\boldsymbol{\theta})$ be a real-valued function from R^k to R, where $\boldsymbol{\theta}$ is a k-dimensional vector. A Taylor series expansion of $g(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ yields

$$g(\boldsymbol{\theta}) \approx g(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)'(\nabla g_{\boldsymbol{\theta}_0}) + \frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'(\nabla^2 g_{\boldsymbol{\theta}_0})(\boldsymbol{\theta}-\boldsymbol{\theta}_0),$$

where $(\nabla g_{\boldsymbol{\theta}_0})$ and $(\nabla^2 g_{\boldsymbol{\theta}_0})$ are the gradient and Hessian of $g(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$. Let $\boldsymbol{\theta}_0$ be the mode of $g(\boldsymbol{\theta})$. Then

$$\begin{aligned}\int_\Theta &\exp[g(\boldsymbol{\theta})]\,d\boldsymbol{\theta}\\ &\approx \int_\Theta \exp\left[g(\boldsymbol{\theta}_0) + \frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'(\nabla^2 g_{\boldsymbol{\theta}_0})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right] d\boldsymbol{\theta}\\ &= \exp[g(\boldsymbol{\theta}_0)] \int_\Theta \exp\left[\frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'(\nabla^2 g_{\boldsymbol{\theta}_0})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right] d\boldsymbol{\theta}\\ &= \exp[g(\boldsymbol{\theta}_0)]|-\nabla^2 g_{\boldsymbol{\theta}_0}|^{1/2}(2\pi)^{k/2}\int_\Theta \phi(\boldsymbol{\theta})\,d\boldsymbol{\theta}, \qquad (7)\end{aligned}$$

where $\phi(\boldsymbol{\theta})$ is the k-dimensional multivariate normal density with mean $\boldsymbol{\theta}_0$ and variance–covariance matrix $[-\nabla^2 g_{\boldsymbol{\theta}_0}]^{-1}$.

Applying the approximation (7) to the integral (3), with $\boldsymbol{\theta}_k = (\pi_1,\ldots,\pi_k)$ as the k-dimensional vector of partial autocorrelations and with

$$g(\boldsymbol{\theta}_k) = \log[p(y^T|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)],$$

yields

$$p(y^T|M_k) \approx (2\pi)^{k/2}|\nabla^2 g_{\boldsymbol{\theta}_k^*}|^{1/2}p(y^T|\boldsymbol{\theta}_k^*, M_k)p(\boldsymbol{\theta}_k^*|M_k) \times \int_{(-1,1)^k}\phi(\boldsymbol{\theta}_k)\,d\boldsymbol{\theta}_k, \qquad (8)$$

where $\boldsymbol{\theta}_k^*$ is the value of $\boldsymbol{\theta}_k$ that maximizes $g(\boldsymbol{\theta}_k)$. The integral on the right side of Equation (8) is evaluated by Monte Carlo integration. Arguments similar to those of Tierney and Kadane (1986) show that the error of the approximation (8) is $O(T^{-1})$.

Thus for a good approximation of the integrated likelihood $p(y^T|M_k)$, all we need are the posterior mode of $\boldsymbol{\theta}_k$ and the Hessian of the log-likelihood function, $\log p(y^T|\boldsymbol{\theta}_k, M_k)$, at that point. A natural parameterization for AR models is in terms of the partial autocorrelations, when the parameter space is the hypercube $[-1,1]^p$. When little prior information is available, a reasonable "noninformative" prior is uniform in the partial autocorrelations; this is proper, and so difficulties with improper priors do not arise. With this prior, the posterior mode is equal to the MLE.

3. ROBUST BAYES FACTORS FOR AUTOREGRESSIVE MODELS

3.1 A Robust Likelihood for Autoregressive Models

We now consider the model comparison problem for AR processes with additive outliers. Suppose that the data $y^T = \{y_1,\ldots,y_T\}$ are generated by the model

$$y_t = x_t + z_t w_t, \qquad (9)$$

where $\{x_t\}$ follows Equation (1), $\{w_t\}$ is a sequence of observations from a outlier-generating distribution whose variance is much larger than σ_ε^2, and $\{z_t\}$ is a $0-1$ process with $P[z_t = 1] = \gamma$ being the fraction of outliers in the data. When $z_t = 1, y_t$ is called an additive outlier.

Our approach to the comparison of different AR orders in the model (9) is to use a *robust likelihood* that approximates the likelihood of the (unobserved) series $\{x_t\}$. Following Martin (1981), this is defined as

$$\log\tilde{p}(y^T|M_k,\boldsymbol{\theta}_k) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log s_t^2 - \frac{1}{2}\sum_{t=1}^{T}\rho\left(\frac{y_t - \tilde{x}_t^{t-1}}{s_t}\right). \qquad (10)$$

In Equation (10), $\tilde{x}_t^{t-1}$ and s_t are robust estimates of the conditional mean and standard deviation of x_t given $x_1,\ldots,x_{t-1}$, found by robust filtering as described in Section 3.2.

The function ρ is chosen to be bounded and continuous so as to ensure that one observation does not have a large influence on the likelihood function and that small changes do not produce large changes in the likelihood function. Here we use the function

$$\begin{aligned}\rho(x) &= x^2 \quad \text{if } |x| \le a,\\ &= a^2 \quad \text{if } |x| > a.\end{aligned}$$

The basic idea is to downweight potential outliers, namely observations whose prediction residuals $y_t - \tilde{x}_t^{t-1}$ are large compared to their predictive standard deviations s_t. Here we use $a = 2.5$ as the tuning constant, so that an observation is censored once its prediction residual is more than 2.5 times its predictive standard deviation. A robust integrated likelihood, $\tilde{p}(y^T|M_k)$, is defined by replacing $p(y^T|M_k,\boldsymbol{\theta}_k)$ with $\tilde{p}(y^T|M_k,\boldsymbol{\theta}_k)$ in (8).

An exact Bayesian solution involves specifying a prior distribution for σ_ε^2 and integrating over this parameter. Here this is complicated, and instead we approximate the resulting integral by conditioning on a robust estimate of the innovations variance, as described in Section 3.3.

3.2 Robust Filtering

To calculate the robust Bayes factors, we need the prediction location and scale of the observations, $\hat{x}_{t|t-1}$ and s_t.

We obtain these using the robust filter of Maserliez (1975) and Martin (1979).

The model (1) and (9) can be written in state-space form as

$$\mathbf{x}_t = \mathbf{\Gamma x}_{t-1} + \boldsymbol{\varepsilon}_t$$

and

$$y_t = \mathbf{H x}_t + v_t, \tag{11}$$

where $v_t \equiv z_t w_t$ denotes the outlier-generating component, $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$ have dimension $k, \mathbf{\Gamma}$ is a $p \times p$ matrix, and $\mathbf{H}$ is a $1 \times p$ matrix, defined by

$$\mathbf{\Gamma} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{k-1} & \phi_k \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

$$\mathbf{x}_t^T = (x_t, x_{t-1}, \dots, x_{t-k+1})$$

$$\mathbf{H} = (1, 0, \dots, 0)$$

$$\boldsymbol{\varepsilon}_t^T = (\varepsilon_t, 0, \dots, 0).$$

We denote the state prediction density by $f(\mathbf{x}_t|y^{t-1})$; this is assumed to exist for $t \geq 1$. The observation prediction density is $f(y_t|y^{t-1})$. The conditional mean of $\mathbf{x}_t$ given y^t is denoted by $\hat{\mathbf{x}}_t = E[\mathbf{x}_t|y^t]$.

When ε_t and v_t in (11) are Gaussian, the computation of $\hat{\mathbf{x}}_t = E(\mathbf{x}_t|y^t)$ yields the Kalman filter recursion equations. Unfortunately, $\hat{\mathbf{x}}_t$ is hard to calculate exactly when v_t is non-Gaussian, except in a few special cases such as that of stable random variables (Stuck 1976). But there is a simplifying assumption that does allow calculation of $\hat{\mathbf{x}}_t$ (Masreliez 1975)—that the state predictor density is Gaussian, namely

$$f(\mathbf{x}_t|y^{t-1}) = N(\mathbf{x}_t; \hat{\mathbf{x}}_t^{t-1}, \mathbf{M}_t),$$

where $N(\cdot; \mu, \Sigma)$ denotes the multivariate normal density with mean μ and covariance matrix Σ and

$$\mathbf{M}_t = E\{(\mathbf{x}_t - \hat{\mathbf{x}}_t^{t-1})(\mathbf{x}_t - \hat{\mathbf{x}}_t^{t-1})^T|y^{t-1}\}$$

is the conditional covariance matrix for the state prediction error. Given this, $\hat{\mathbf{x}}_t$ satisfies the recursions

$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^{t-1} + \mathbf{M}_t \mathbf{H}^T \Psi_t(y_t), \tag{12}$$

$$\mathbf{M}_{t+1} = \mathbf{\Gamma P}_t \mathbf{\Gamma}^T + \mathbf{Q}, \tag{13}$$

and

$$\mathbf{P}_t = \mathbf{M}_t - \mathbf{M}_t \mathbf{H}^T \Psi_t'(y_t) \mathbf{H M}_t, \tag{14}$$

where

$$\Psi_t(y_t) = -\left(\frac{\partial}{\partial y_t}\right) \log f_Y(y_t|y^{t-1})$$

is the score function for the observation prediction density $f_Y(y_t|y^{t-1})$. The matrix $\mathbf{Q}$ is the covariance matrix of $\boldsymbol{\varepsilon}_t$ that is equal to σ_ε^2 at the (1, 1) position and to zero everywhere else, $\hat{\mathbf{x}}_t^{t-1} = \mathbf{\Gamma}\hat{\mathbf{x}}_{t-1}$, and

$$\Psi_t'(y_t) = -\left(\frac{\partial}{\partial y_t}\right) \Psi_t(y_t).$$

The density $f_Y(y_t|y^{t-1})$ is generally intractable when outliers are present. Thus it is difficult to obtain the Ψ function. But, as noted by Martin (1979), Ψ and Ψ' can be well approximated by appropriately chosen bounded continuous functions. Boundedness ensures that y_t does not have an unbounded influence on $\tilde{x}_t$, and continuity ensures that small changes in y_t do not produce large changes in $\tilde{x}_t$. The empirical study of Martin and Su (1985) showed that Hampel's two-part redescending function caused little bias in outlier-free situations while providing good robustness towards outliers. Thus here we use Hampel's two-part redescending function,

$$\begin{aligned} \psi(y) &= y, \quad |y| \leq \alpha c, \\ &= \alpha(c-y)/(1-\alpha), \quad \alpha c < y \leq c, \\ &= -\alpha(c+y)/(1-\alpha), \quad -c \leq y < -\alpha c, \\ &= 0, \quad |y| > c, \end{aligned}$$

with $\alpha c = 2.5$ and $c = 4.0$. That is, observations with prediction residuals (divided by their predictive standard deviations) in the interval (2.5, 4.0) are downweighted linearly, and those with prediction residuals greater than 4 are given zero weight. To ensure boundedness and continuity of Ψ', Martin and Su (1985) also recommended that Ψ' be replaced by the weight function $w(z) = \psi(z)/z$.

Let s_t^2 be the (1, 1) element of M_t. Then the recursions (12)–(14) may be replaced by

$$\tilde{\mathbf{x}}_t = \mathbf{\Gamma}\tilde{\mathbf{x}}_{t-1} + \frac{\mathbf{m}_t}{s_t^2} s_t \psi\left(\frac{r_t}{s_t}\right),$$

$$\mathbf{M}_{t+1} = \mathbf{\Gamma P}_t \mathbf{\Gamma}^T + \mathbf{Q},$$

and

$$\mathbf{P}_t = \mathbf{M}_t - w\left(\frac{r_t}{s_t}\right) \frac{\mathbf{m}_t \mathbf{m}_t^T}{s_t^2},$$

where $\mathbf{m}_t$ is the first column of M_t and r_t is the observation prediction residual

$$r_t = y_t - \mathbf{H}\tilde{\mathbf{x}}_t^{t-1}.$$

3.3 Implementation

We first find the values of the parameters that maximize the robust log posterior density, and we find its Hessian at the posterior mode. The parameters of an AR(k) model are the k partial autocorrelations $\Pi = (\pi_1, \dots, \pi_k)$ and the innovations variance σ_ε^2. In the nonrobust setup, it is easy to find the joint posterior mode of these $(k+1)$ parameters given the data $\{y_t\}$. But in the robust setup this is harder,

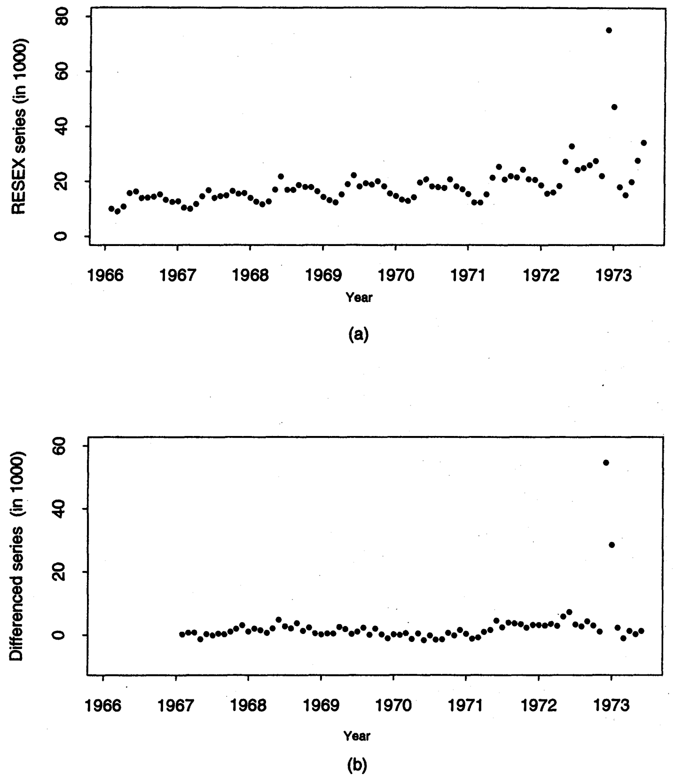

Figure 1. Monthly Inward Movement of Residence Extensions: RESEX Series. (a) The RESEX series; (b) the seasonally differenced RESEX series.

and we estimate the innovations variance in advance by substituting a highly robust estimate of σ_X into the equation

$$\sigma_\varepsilon^2 = \sigma_X^2 \prod_{i=1}^{k} (1 - \pi_i^2).$$

The highly robust scale estimate $\hat{\sigma}_X$ we use is the median of absolute deviations about the median (MADM) defined by

$$\hat{\sigma}_X = \frac{1}{.6745} \text{ median } (|y_t - \text{median } \{y_t\}|).$$

This estimate is highly robust with respect to bias (Martin and Zamar 1993). The robust likelihood is then maximized as a function of Π by the Newton–Raphson method.

The estimates are not sensitive to the initial value of Π, which affects the number of iterations required in the Newton–Raphson method. To reduce the number of iterations, we estimate the partial autocorrelations sequentially. We first estimate the lag-1 partial autocorrelation π_1 for the AR(1) model with starting value 0, then estimate (π_1, π_2) for the AR(2) model with starting values $(\hat{\pi}_1, 0)$, and so on. Convergence in this sequential method is quick, usually taking from two to eight iterations. In all of the examples that we have worked with, the estimates from the sequential method were very similar to those from the global method in which all the partial autocorrelations are estimated at once, and the global method always took longer than the sequential one. For any given data set, it is hard to know whether the objective function has a unique maximum, but in experience with data sets used in the examples and the simulation study, we have never had any convergence problems.

4. EXAMPLE: TELEPHONE EXTENSION SERIES

Figure 1(a) shows the inward movement of residential telephone extensions (RESEX series) in a fixed geographic area in which each of the 89 months from January 1966 to May 1973 (Martin, Samarov, and Vandaele 1983). There are two large values in November and December 1972. The first value is due to a November "bargain month" (i.e., free installation of residence extensions), and the second value is due to a spillover effect in December because not all of the November orders could be filled in that month. The plot also shows some seasonal behavior. The seasonally differenced series, defined as

$$y(t) = \text{RESEX}(t) - \text{RESEX}(t - 12),$$

is shown in Figure 1(b); it has two large values, and the seasonal structure has been removed.

The posterior model probabilities are obtained using the robust, nonrobust, BIC, and AIC methods. Specifically, the posterior probabilities defined in Equation (2) are calculated with the integrated likelihood (3) obtained with the nonrobust method through Equation (8) and the robust method through Equations (8) and (10). The BIC and AIC approximations are evaluated through Equations (5) and (6).

Table 1 shows that according to the robust method, the differenced series can be represented by an AR(2) model. The robust posterior probability of the process having order 2 is .84, which is much greater than the second largest posterior probability of .11 for order 3. Brubacher (1974) also identified the series as an AR(2) process, using an interpolation technique described by Brubacher and Wilson (1976). The method requires a priori identification of the outlier locations. Although it is a trivial task for this series, it is generally difficult to correctly specify a priori the outlier locations. Martin (1980) obtained the same result using an iterative procedure. Although his procedure seems to work well for this series, its convergence is not guaranteed in general.

The nonrobust method (and BIC) favor the AR(1) model. The AIC approach yields low posterior probabilities for all orders. As expected, the robust method does much better because it corrects for the outliers which are clear in this data set. Table 2 shows the predictive scores of the robust and nonrobust methods. The predictive score for a given

Table 1. Posterior Model Probabilities for the RESEX Series

	Order						
Method	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Robust	0	.01	.84	.11	.04	0	0
Nonrobust	0	.74	.20	.05	.01	0	0
BIC	.01	.75	.20	.04	0	0	0
AIC	0	.38	.32	.19	.07	.03	.01

Table 2. Predictive Scores for the Data

Method	Model	Posterior probability	Predictive score
Robust	AR(1)	.01	−631.71
	AR(2)	.84	−628.21
	AR(3)	.11	−628.25
	AR(4)	.04	−627.27
	Combined		−628.15
Nonrobust	AR(1)	.74	−725.69
	AR(2)	.20	−724.87
	AR(3)	.05	−724.40
	AR(4)	.01	−724.41
	Combined		−725.16

order k is

$$\sum_{t=7}^{77} \log p(y_t | y^{t-1}, M_k). \tag{15}$$

The predictive distribution that takes into account model uncertainty is derived from Equation (4). Its performance is measured by its predictive score, namely

$$\sum_{t=7}^{77} \log \left(\sum_{k=0}^{6} p(y_t | y^{t-1}, M_k) * p(M_k | y^T) \right). \tag{16}$$

Note that the last y^T should in principle be replaced by y^{t-1}, but in this example it makes almost no difference.

The predictive densities in Equations (15) and (16) are calculated as follows. Assuming a contaminated normal distribution for residuals yields

$$y_t - \hat{x}_t^{t-1} \sim (1-\gamma) N(0, s_t^2) + \gamma N(0, \sigma_\omega^2),$$

where the prediction location $\hat{x}_t^{t-1}$ and scale s_t are defined and obtained as described in Section 3.2. The outlier-generating component variance σ_ω^2 is estimated using the following relationship [Equation (9)]:

$$\sigma_y^2 = \sigma_X^2 + \gamma \sigma_\omega^2.$$

The observation and state variances (σ_y^2, σ_X^2) are estimated by the sample variance and the robust estimator MADM (Sec. 3.3). The fraction of outliers γ is then estimated by the proportion of observations outside the range (median $\{y_t\} \pm 2.58\sigma_X$).

With the estimated values of $\gamma = .05, \sigma_X = 1{,}641.2$, and $\sigma_y = 6{,}976.8$, the estimated value of σ_ω is 30,370. The predictive density of $(y_t | y^{t-1})$ is then approximated by the density of $(y_t - \hat{x}_t^{t-1})$.

Our recommended procedure of robust estimation and model selection, with model averaging to take into account model uncertainty, has the best predictive performance of any of the methods considered here. Standard approaches would choose either an AR(1) or an AR(2) model and estimate them nonrobustly. Our approach improves on this considerably, by nearly 100 units of predictive score or nearly 200 units on the scale of twice the predictive score on which deviances and likelihood ratio test statistics are measured.

Almost all of the gain in predictive performance in this example is due to the fact that we have used a robust method. Model averaging has given some further improvement, but it is small. This is probably because there is little model uncertainty. In other situations where there is substantial model uncertainty, model averaging does yield meaningful improvements in out-of-sample predictive performance (Madigan and Raftery 1994; Raftery, Madigan, and Hoeting 1993; Raftery, Madigan, and Volinsky 1995).

5. SIMULATION

5.1 A Simulation Study

We now report a simulation study to compare the robust method of Section 3 with the nonrobust Gaussian method of Section 2.2 along with the BIC and AIC approximation methods, for both Gaussian AR time series and AR time series with additive outliers (AO's). The MLE-based BIC and AIC approximations are included to examine how well they approximate the integrated likelihood. Because we have not investigated the BIC and AIC based on the robust variance estimate, our results should be interpreted accordingly.

The results for AO series indicate how much is gained by using the robust method, whereas the results with Gaussian data indicate the loss from using the robust method unnecessarily. In each case we used 500 time series of 200 data points each, generated from Equations (1) and (9) with $k = 1$ and $\phi_1 = \pi_1 = .5$. The contamination fraction, γ, was set at .1, the innovations variance σ_ε^2 was chosen to be unity, and the value of outliers was fixed at $\pm\delta$, where δ is a positive constant. The outliers could be negative or positive with equal probabilities. We used $\delta = 0$, 4, and 7 to represent no outliers, "moderate" outliers, and "large" outliers. The standard deviation of x_t is 1.15 when there are no outliers. The posterior probabilities were calculated for orders zero to 5 using uniform priors for the partial autocorrelations and equal prior probabilities for the competing models. Table 3 shows the number of times that each order has the highest posterior probability.

Case 1: $\delta = 0$ (no outliers). The robust and nonrobust methods both performed very well. In most cases, the posterior probabilities of the process of order 1 were quite large,

Table 3. Percentage of Times That Each Order was Assigned With the Highest Posterior Probability

		Order					
δ	Method	0	1	2	3	4	5
0	Robust	0	96	3	1	0	0
	Nonrobust	0	96	3	1	0	0
	BIC	0	96	3	1	0	0
	AIC	0	75	12	6	5	2
4	Robust	0	85	13	2	0	0
	Nonrobust	21	71	8	0	0	0
	BIC	24	70	6	0	0	0
	AIC	8	60	20	6	3	3
7	Robust	0	92	6	2	0	0
	Nonrobust	76	23	1	0	0	0
	BIC	78	21	1	0	0	0
	AIC	43	35	10	5	5	2

NOTE: The true order was 1 in each case.

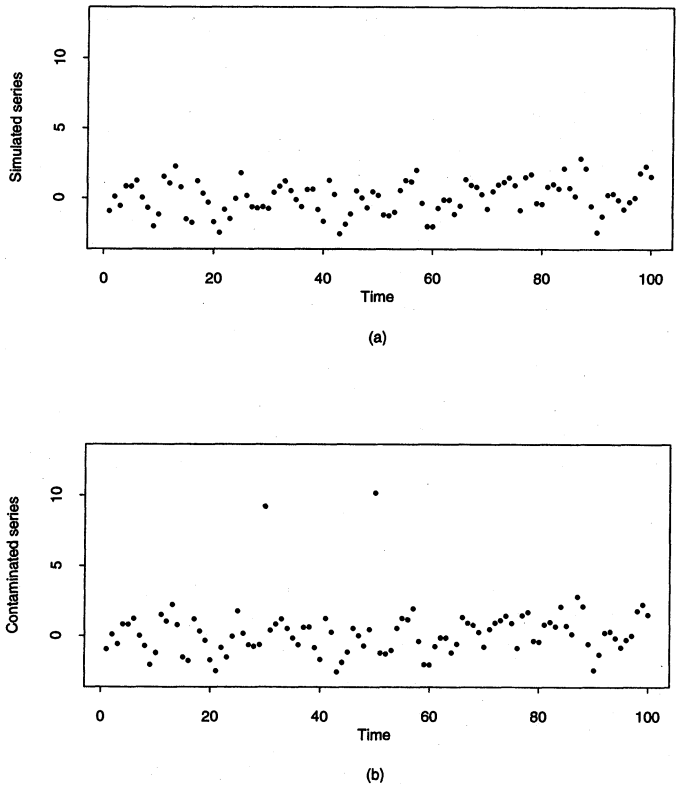

Figure 2. Simulated Series Used for the Spectrum Estimation Example. (a) The simulated series; (b) the simulated series with two additive outliers.

so that there is little to be lost by using the robust method when there are no outliers. The BIC method's performance was quite similar to that of the nonrobust method, as expected. The AIC method was inferior to the other three methods; it tended to favor higher-order models, in agreement with asymptotic results (Shibata 1976).

Case 2: $\delta = 4$ (moderate outliers). Table 3 shows the robust method to be better than the nonrobust method. The nonrobust posterior probabilities were larger and more dispersed than the robust posterior probabilities for order zero and were smaller and more dispersed for order 1. This is in line with the expectation that the nonrobust method would favor lower-order models when there are isolated additive outliers in the data. Again, the BIC method and the nonrobust method performed similarly. The AIC performed worse than the others; it tended to favor higher-order models.

Case 3: $\delta = 7$ (large outliers). Table 3 shows the robust method to be much better than the nonrobust method. The robust method generally gave high posterior probability to the correct order, whereas the nonrobust method gave high posterior probability to order zero in most cases, as did the BIC approximation. The nonrobust method did very poorly, favoring the correct model in only 23% of the cases, compared to 92% for the robust method. The nonrobust method chose order zero 76% of the time, which is not surprising because when $\delta = 7$, the theoretical lag correlations of the contaminated process are close to zero. Here the lag-1 autocorrelation of the observed process is .11, compared with .5 for the underlying uncontaminated process. The AIC method also performed poorly.

Overall, when there were no outliers in the data, the robust and nonrobust methods both worked well, indicating that there is little to be lost by using the robust method when the data are indeed Gaussian. When there were outliers in the data, the robust method performed better than the nonrobust method, particularly when the outliers were large. The performance of the nonrobust method deteriorated rapidly as the outliers got bigger.

The performance of the robust method was slightly better when there were either no outliers or large outliers than when the outliers were of moderate size. This is inherent in robust estimation, because it is easy to identify an outlier when it is clearly extreme but harder to identify moderate outliers that are too small to be identified with certainty but large enough to distort the results.

The performance of the BIC method was similar to that of the nonrobust method whether or not there were outliers in the data. The results indicate that the BIC provides a good approximation to the integrated likelihood, in line with the theoretical result of Kass and Wasserman (1995). This suggests that a robustified version of BIC, in which $\log \tilde{p}(y^T|\mathrm{M}_k, \boldsymbol{\theta}_k)$ is used instead of the maximized log-likelihood in Equation (5), may provide a good approximation to the robust integrated likelihood.

The AIC method performed poorly overall. It tended to favor higher-order models whether or not there were outliers in the data. The results indicate that the AIC method does not approximate the integrated likelihood well for autoregressive processes.

5.2 A Simulated Example: Spectrum Estimation

We estimate the power spectrum of a series of 100 observations generated from the AR(2) model,

$$y_t = .7y_{t-1} - .4y_{t-2} + \varepsilon_t,$$

where ε_t is from the standard normal distribution. We then reestimate the power spectrum after introducing two additive outliers to the series. The data are shown in Figure 2.

The power spectrum of an AR(k) model is

$$S(f) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{\left|1 + \sum_{j=1}^{k} \phi_j \exp(-ifj)\right|^2} .$$

Hence the power spectrum for an AR(k) model can be estimated by replacing ϕ_j by their estimates. To incorporate the model uncertainty, we estimate the power spectrum of the series by a weighted average of the estimated spectrum given each order, using the posterior model probabilities shown in Table 4 as weights.

The theoretical and estimated power spectra are displayed in Figure 3. As expected, for the clean series, both the robust and nonrobust estimates of the spectrum, including BIC and AIC, are reasonable in the sense that their peaks are close to the theoretical peak. But for the contaminated series, the

Table 4. Posterior Model Probabilities in the Spectrum Estimation Example

		Order						
Data	*Method*	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Gaussian	Robust	0	0	.79	.17	.03	.01	0
	Nonrobust	0	0	.77	.18	.04	.01	0
	BIC	0	0	.83	.14	.02	.01	0
	AIC	0	0	.43	.26	.15	.12	.04
Contaminated	Robust	0	.01	.79	.16	.03	.01	0
	Nonrobust	.67	.25	.07	.01	0	0	0
	BIC	.68	.24	.07	.01	0	0	0
	AIC	.26	.28	.27	.12	.04	.02	.01

robust method still produces a peak close to the theoretical one, whereas the nonrobust, BIC, and AIC methods fail to do so.

6. DISCUSSION

We have proposed a method for the order selection problem when the data are from a stationary Gaussian AR process of order k with additive outliers. This method works well in the presence of outliers and also with clean Gaussian data. It yields the posterior probability of each of the models considered, allowing one to take into account model uncertainty. It is designed to deal specifically with additive outliers; its effectiveness for other types of outlier such as innovations outliers or level shifts remains to be examined.

The method takes about 10 minutes of CPU time on a Sparc 2 Workstation to analyze a series of length 200, obtaining the posterior probabilities of order 0, 1, ..., 5. An S-plus function to implement the approach is available in StatLib and can be obtained free of charge by sending the message "send ar.robust from S" to statlib@stat.cmu.edu.

Although AR model selection in the presence of outliers has not been much studied, there are many approaches to parameter estimation, filtering, and prediction in the presence of outliers. These include explicit state-space modeling (West and Harrison 1990), the use of mixture transition distribution (MTD) models (Le, Martin, and Raftery 1990), and testing for outliers (Tsay 1986); other references have been listed by Martin and Raftery (1987). It is possible that these other approaches could also be extended to yield robust Bayes factors for AR model order.

Other authors have discussed AR order selection methods that yield a set of possible orders rather than a single one (see, for example, Rezayat and Anandalingam 1988). Our proposal here goes beyond this in yielding posterior probabilities of all the models initially considered and in formally taking into account model uncertainty.

[Received February 1993. Revised April 1995.]

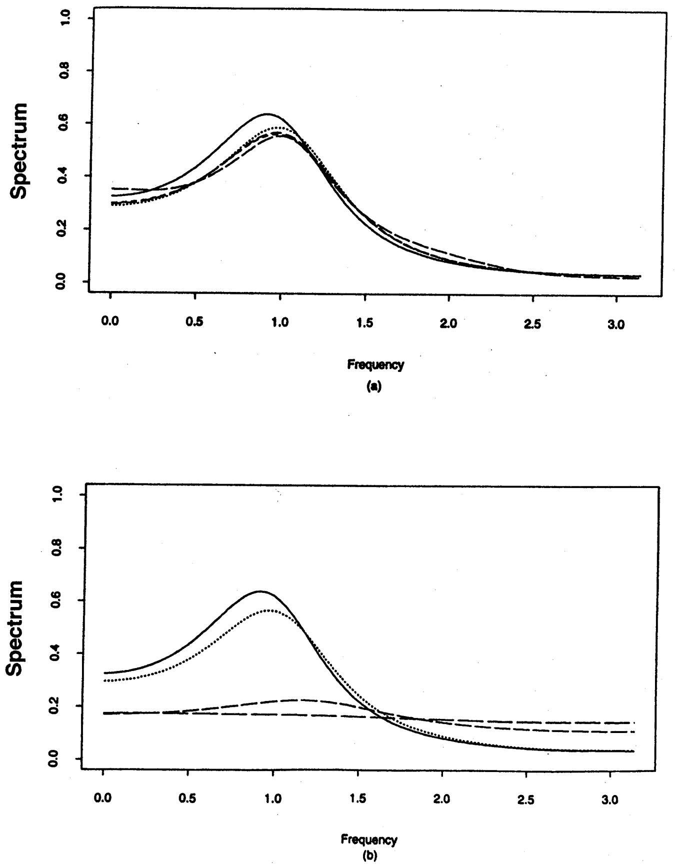

Figure 3. Theoretical and Estimated Power Spectra (a) Theoretical and Estimated Spectrum for Cleaned Simulated Series; (b) Theoretical and Estimated Spectrum for Simulated Series With Outliers. —— true; ··· robust; - - - nonrobust; — — — BIC; —— —— AIC.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *The Second International Symposium on Information Theory*, eds. by B. N. Petrov and F. Csake, Akademiai Kiado, Hungary, pp. 267–281.

——— (1983), "Information Measures and Model Selection," *Bulletin of the International Statistical Institute*, 50, 277–290.

Barndorff-Nielsen, O., and Schou, G. (1973), "On the Parameterization of Autoregressive Models by Partial Autocorrelations," *Journal of Multivariate Analysis*, 3, 408–419.

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis Forecasting and Control* (2nd ed.), San Francisco: Holden-Day.

Brubacher, S. R. (1974), "Time Series Outlier Detection and Modelling With Interpolation," Bell Laboratories technical memo.

Brubacher, S. R., and Wilson, G. T. (1976), "Interpolating Time Series With Application to the Estimation of Holiday Effect on Electricity Demand," *Journal of the Royal Statistical Society*, Ser. C, 25, 107–116.

De Gooijer, J. G., Abraham, B., Gould, A., and Robinson, L. (1985), "Methods for Determining the Order of an Autoregressive-Moving Average Process: A Survey," *International Statistical Review*, 53, 301–329.

Harvey, A. C. (1981), *Time Series Models*, New York: Halsted Press.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses in Large Samples," *Journal of the American Statistical Association*, 90, 928–934.

Le, N. D., Martin, R. D., and Raftery, A. E. (1990), "Modeling Outliers, Bursts and Flat Stretches in Time Series Using Mixture Transition Distribution (MTD) Models," Technical Report 194, University of Washington, Dept. of Statistics.

Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference With Nonexperimental Data*, New York: John Wiley.

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1335–1346.

Martin, R. D. (1979), "Approximate Conditional-Mean Type Smoothers and Interpolators," in *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Berlin: Springer, pp. 117–143.

——— (1980), "Robust Estimation of Autoregressive Models" (with discussion), in *Directions in Time Series*, eds. D. R. Brillinger and G. C. Tiao, Haywood, CA: Institute of Mathematical Statistical, pp. 228–254.

——— (1981), "Robust Methods for Time Series," in *Applied Time Series II*, ed. D. F. Findley, New York: Academic Press, pp. 683–759.

Martin, R. D., and Raftery, A. E. (1987), Comments on "Non-Gaussian State-Space Modeling of Nonstationary Time Series," by Genshiro Kitagawa, *Journal of the American Statistical Association*, 82, 1044–1050.

Martin, R. D., Samarov, A., and Vandaele, W. (1983), "Robust Methods for ARIMA Models," in *Applied Time Series Analysis of Economic Data*, ed. A. Zellner, Economic Resource Report ER-t, Washington, DC: Bureau of the Census.

Martin, R. D., and Su, K. Y. (1985), "Robust Filters and Smoothers: Definition and Design," Technical Report 58, University of Washington, Dept. of Statistics.

Martin, R. D., and Zamar, R. H. (1993), "Bias Robust Estimation of Scale," *The Annals of Statistics*, 21, 991–1017.

Masreliez, C. J. (1975), "Approximate Non-Gaussian Filtering With Linear State and Observation Relation," *IEEE Transactions on Automatic Control*, AC-20, 107–110.

Poskitt, D. S. (1988), "Decision-Theoretic Approach to Order Determination and Structure Selection in Vector Linear Processes," in *Bayesian Analysis of Time Series and Dynamic Models*, ed. J. C. Spall, New York: Marcel Dekker, pp. 53–74.

Raftery, A. E. (1988), "Approximate Bayes factors for generalized linear models," Technical Report 121, University of Washington, Dept. of Statistics.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1993), "Accounting for model uncertainty in survival analysis improves predictive performance" (with discussion), University of Washington, Dept. of Statistics.

Raftery, A. E., Madigan, D., and Volinsky, C. (1995), "Model Uncertainty in Survival Analysis and Nonparametric Smoothing" (with discussion), in *Bayesian Statistics 5*, eds. J. M. Bernardo et al., New York: Oxford University Press, to appear.

Rezayat, F., and Anandalingam, G. (1988), "Using Instrumental Variables for Selecting the Order of ARMA Models," *Communications in Statistics, Part A—Theory and Methods*, 17, 3029–3065.

Ramsey, R. L. (1974), "Characterization of the Partial Autocorrelation Function," *The Annals of Statistics*, 2, 1296–1301.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shibata, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117–128.

Stuck, B. W. (1976), "Minimum Error Dispersion Linear Filtering of Symmetric Stable Processes," *IEEE Transactions on Automatic Control*, AC-17, 507–509.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tsay, R. S. (1986), "Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 81, 132–141.

West, M., and Harrison, J. (1990), *Bayesian Forecasting and Dynamic Models*, New York: Springer.