BAYESIAN METHODS FOR AUTOREGRESSIVE MODELS

W.D. Penny and S.J. Roberts

Department of Engineering Science,
Oxford University.
E-mail: wpenny,sjrob@robots.ox.ac.uk

Abstract. We describe a Variational Bayesian (VB) learning algorithm for parameter estimation and model order selection in autoregressive (AR) models. With uninformative priors on the precisions of the coefficient and noise distributions the VB framework is shown to be identical to the Bayesian Evidence framework. The VB model order selection criterion is compared with the Minimum Description Length (MDL) criterion on synthetic data and on EEG.

INTRODUCTION

In the autoregressive modeling of stationary stochastic processes there are two basic problems: estimation of the parameters and selection of the optimum model order. These problems are usually tackled by first estimating the model parameters using eg. Burg, Levinson-Durbin or Maximum-Likelihood algorithms [10] and then choosing the optimal model order using a variety of criteria eg. Akaikes Information Criterion (AIC), Final Prediction Error (FPE), Minimum Description Length (MDL) etc. (see [7, 4] for recent approaches).

In this paper we show that the Variational Bayesian (VB) methodology provides a single framework within which we can perform parameter estimation *and* model order selection. Whilst the Bayesian methodology has a long history the use of VB is relatively recent. A notable recent example is the application of VB to Principal Component Analysis [2].

In the first section of the paper we describe the autoregressive process as a probabilistic Bayesian model. We then describe the VB method for the estimation of AR parameters and discuss the VB model order selection criterion in the context of other methods.

AUTOREGRESSIVE MODELS

We define an autoregressive model as

$$y_n = \boldsymbol{x}_n \boldsymbol{w} + e_n \tag{1}$$

where y_n is the nth value of a time series, $\boldsymbol{w}$ is a column vector of AR coefficients (weights), $\boldsymbol{x}_n = [y_{n-1}, y_{n-2}, ..., y_{n-p}]$ are the p previous time series values and e_n is additive Gaussian noise with zero mean and precision (inverse variance) β. Given a data set $D = \{\boldsymbol{x}_n, y_n\}$ where $n = 1..N$ the likelihood of the data is given by

$$p(D \mid \boldsymbol{w}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp(-\beta E_D(\boldsymbol{w})) \tag{2}$$

where

$$\begin{aligned} E_D(\boldsymbol{w}) &= \frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n\boldsymbol{w})^2 \\ &= \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) \end{aligned} \tag{3}$$

and $\boldsymbol{Y}$ is a column vector with entries y_n and the nth row of the matrix $\boldsymbol{X}$ contains $\boldsymbol{x}_n$. The weights are drawn from a zero-mean Gaussian prior with an isotropic covariance having precision α

$$p(\boldsymbol{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{p/2} \exp(-\alpha E_W) \tag{4}$$

where

$$\begin{aligned} E_W &= \frac{1}{2}\sum_{i=1}^{p} w_i^2 \\ &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \end{aligned} \tag{5}$$

The weight and noise precisions are drawn from Gamma priors

$$\begin{aligned} p(\alpha) &= \Gamma(\alpha; b_\alpha, c_\alpha) \\ p(\beta) &= \Gamma(\beta; b_\beta, c_\beta) \end{aligned} \tag{6}$$

where

$$\Gamma(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{7}$$

We concatenate all the parameters of the model into the vector $\boldsymbol{\theta} = [\boldsymbol{w}, \alpha, \beta]$. The overall prior is therefore

$$p(\boldsymbol{\theta}) = p(\boldsymbol{w}|\alpha)p(\alpha)p(\beta) \tag{8}$$

and the likelihood is

$$p(D|\boldsymbol{\theta}) = p(D|\boldsymbol{w}, \beta) \tag{9}$$

VARIATIONAL BAYES

The 'evidence' or 'marginal likelihood' of a probabilistic model is the likelihood of the model after its parameters have been integrated out. The log evidence can be written as

$$\log p(D) \quad = \quad \log \int q(\boldsymbol{\theta}|D) \frac{p(D, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|D)} d\boldsymbol{\theta} \tag{10}$$

Using Jensen's inequality we have

$$\log p(D) \geq F(p) \tag{11}$$

where

$$F(p) = \int q(\boldsymbol{\theta}|D) \log \frac{p(D, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|D)} d\boldsymbol{\theta} \tag{12}$$

The quantity $F(p)$ is known as the variational free energy and is a lower bound on the log evidence, with equality if the approximating posterior is equal to the true posterior ie. if $q(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|D)$. The aim of Variational Bayesian (VB) learning or 'Ensemble learning' is to maximise this lower bound. Using $p(D, \boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ we can write

$$F(p) = L_{av} - KL(q||p) \tag{13}$$

where

$$L_{av} = \int q(\boldsymbol{\theta}|D) \log p(D|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{14}$$

and

$$KL(q||p) = \int q(\boldsymbol{\theta}|D) \log \frac{q(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{15}$$

where the first term is the average likelihood of the data L_{av} and the second term is the KL-divergence [6] between the approximating posteriors and the priors. This acts as a penalty term which penalizes more complex models.

To apply VB to AR models we approximate the posterior distribution over parameters with the factorised density

$$q(\boldsymbol{\theta}|D) = q(\boldsymbol{w}|D)q(\alpha|D)q(\beta|D) \tag{16}$$

Mackay [8] shows that maximising $F(p)$ with respect to each of $q(\boldsymbol{w}|D)$ and $q(\alpha|D)$ gives the following results. The weight posterior is a Normal density $q(\boldsymbol{w}|D) = N(\boldsymbol{w}; \hat{\boldsymbol{w}}, \hat{\boldsymbol{\Sigma}})$ and the weight precision posterior is a Gamma density $q(\alpha|D) = \Gamma(\alpha; b'_\alpha, c'_\alpha)$ where

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= (\hat{\beta}\boldsymbol{X}^T\boldsymbol{X} + \hat{\alpha}\boldsymbol{I})^{-1} \\ \hat{\boldsymbol{w}} &= \hat{\boldsymbol{\Sigma}}\boldsymbol{X}^T\hat{\beta}\boldsymbol{Y} \end{aligned} \tag{17}$$

and

$$\begin{aligned} 1/b'_\alpha &= \frac{1}{2}\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} + \frac{1}{2}Tr(\hat{\boldsymbol{\Sigma}}) + \frac{1}{b_\alpha} \\ c'_\alpha &= \frac{p}{2} + c_\alpha \\ \hat{\alpha} &= b'_\alpha c'_\alpha \end{aligned} \tag{18}$$

Extending Mackay's approach to the noise precision posterior shows that the optimal distribution is $q(\beta|D) = \Gamma(\beta; b'_\beta, c'_\beta)$ where

$$\begin{aligned} 1/b'_\beta &= \bar{E}_D(\hat{\boldsymbol{w}}) + \frac{1}{b_\beta} \\ c'_\beta &= \frac{N}{2} + c_\beta \\ \hat{\beta} &= b'_\beta c'_\beta \end{aligned} \tag{19}$$

and

$$\bar{E}_D(\hat{\boldsymbol{w}}) = E_D(\hat{\boldsymbol{w}}) + 0.5Tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{X}^T\boldsymbol{X}) \tag{20}$$

Practicalities

We use uninformative priors (see eg. [9]) by setting $b = 10^3$ and $c = 10^{-3}$ for each of the Gamma priors. The parameters of the VB posterior are initialised using the maximum likelihood solution (see later); $\hat{\boldsymbol{w}} = \boldsymbol{w}_{ML}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{ML}$. The VB equations can then be applied iteratively until a consistent solution is reached. Convergence can be measured by evaluating the negative free energy

$$F(p) = L_{av} - KL(\boldsymbol{w}) - KL(\alpha) - KL(\beta) \tag{21}$$

where

$$L_{av} = \frac{N}{2} < \log\beta > - \hat{\beta}\bar{E}_D - \frac{N}{2}\log 2\pi \tag{22}$$

and

$$< \log\beta > = \Psi(c'_\beta) + \log b'_\beta \tag{23}$$

and $\Psi()$ is the digamma function. The KL-divergence for Normal densities $q(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ and $p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ is

$$KL(q||p) = 0.5\log\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} + 0.5Trace(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + 0.5(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - \frac{d}{2} \tag{24}$$

where $|\mathbf{\Sigma}_p|$ denotes the determinant of the matrix $\mathbf{\Sigma}_p$. For Gamma densities $q(x) = \Gamma(\boldsymbol{x}; b_q, c_q)$ and $p(x) = \Gamma(\boldsymbol{x}; b_p, c_p)$ the KL-divergence is

$$\begin{aligned} KL(q||p) &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log\Gamma(c_q) \\ &+ \log\Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \tag{25}$$

We calculate $F(p)$ at each iteration and terminate training if the proportionate increase in $F(p)$ from one iteration to the next is less than 0.01 per cent.

OTHER METHODS

Maximum Likelihood

We compare the VB solution to the maximum likelihood solution (see eg. [12]).

$$\boldsymbol{w}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{26}$$

From the AR predictions $\hat{y}_n = \boldsymbol{x}_n \boldsymbol{w}_{ML}$ the maximum likelihood noise variance, σ^2_{ML}, can be estimated as

$$\sigma^2_{ML} = \frac{1}{N-p}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 \tag{27}$$

The covariance is then estimated as

$$\mathbf{\Sigma}_{ML} = \sigma^2_{ML}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{28}$$

Bayesian Evidence

An alternative Bayesian approach is that offered by the Evidence framework [9]. Given that the observation noise is zero-mean Gaussian with precision β and that the weights are drawn from a prior distribution with zero mean and isotropic covariance having precision α, the posterior distribution is Gaussian with mean and covariance given by

$$\begin{aligned} \hat{\mathbf{\Sigma}} &= (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1} \\ \hat{\boldsymbol{w}} &= \hat{\mathbf{\Sigma}}\boldsymbol{X}^T\beta\boldsymbol{Y} \end{aligned} \tag{29}$$

In the evidence framework there are no priors over α or β. The evidence for the model is therefore given by (see eg. page 409 in [3])

$$p(D) = \int p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w} \tag{30}$$

The log of the evidence can then be written as

$$EV(p) = -\alpha E_W - \beta E_D(\hat{\boldsymbol{w}}) + 0.5 \log |\boldsymbol{\Sigma}| + \frac{p}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log 2\pi \quad (31)$$

The precision parameters are then set so as to maximise the evidence as follows

$$\alpha = \frac{\gamma}{2E_W} \quad (32)$$

$$\beta = \frac{N - \gamma}{2E_D(\hat{\boldsymbol{w}})} \quad (33)$$

where γ, the number of 'well-determined' coefficients, is given by

$$\gamma = p - \alpha Tr(\boldsymbol{\Sigma}) \quad (34)$$

where $Tr()$ denotes the trace of a matrix and γ is calculated using the 'old' value of α. The update for α is therefore an implicit equation. We can also write it as the explicit update

$$\alpha = \frac{p}{2E_W + Tr(\boldsymbol{\Sigma})} \quad (35)$$

This is equivalent to the update for $\hat{\alpha}$ from the VB framework (equation 19) if we choose uninformative priors on α. Similarly, if we choose uninformative priors for β then the VB update is (from equation 20)

$$\hat{\beta} = \frac{N}{2E_D(\hat{\boldsymbol{w}}) + Tr(\hat{\boldsymbol{\Sigma}} \boldsymbol{X}^T \boldsymbol{X})} \quad (36)$$

which is equivalent to equation 33 as, from an eigendecomposition of $\hat{\boldsymbol{\Sigma}}$, we can show that $Tr(\hat{\boldsymbol{\Sigma}} \boldsymbol{X}^T \boldsymbol{X}) = \gamma / \beta$.

Model Order Selection

The Variational Bayes (VB) framework maximises the negative free energy

$$F(p) = L_{av} - KL(Q(\boldsymbol{\theta}) || p(\boldsymbol{\theta})) \quad (37)$$

where the first term is the average likelihood of the data and the second term is the KL-divergence between the approximating posteriors and the priors. This acts as a penalty term which penalizes more complex models.

As the number of samples increases, the parameter posterior becomes sharply peaked about the most probable values (which are also the maximum likelihood values), $\hat{\boldsymbol{\theta}}$. It can then be shown that, in the large sample limit $N \to \infty$, $F(p)$ becomes equivalent to the Bayesian Information Criterion [1, 5]

$$BIC(p) = \log p(D|\hat{\boldsymbol{\theta}}) - \frac{p}{2} \log N \quad (38)$$

which is itself equal to the negative of the Minimum Description Length (MDL) ie. $BIC(p) = -MDL(p)$. These popular model order selection criteria can therefore be seen as a limiting case of the VB framework. For the AR model we have

$$BIC(p) = \frac{N}{2} \log \beta_{ML} - \frac{p}{2} \log N \tag{39}$$

where β_{ML} is the precision of the maximum likelihood model.

The model order selection criterion for the evidence framework (equation 31) can be rewritten in terms of the KL-divergence between the weight posterior and the weight prior, $KL(w)$. For the prior $p(\boldsymbol{w}) = N(\boldsymbol{w}; \mathbf{0}, (1/\alpha)\boldsymbol{I})$ and the posterior $q(\boldsymbol{w}) = N(\boldsymbol{w}; \hat{\boldsymbol{w}}, \hat{\boldsymbol{\Sigma}})$ we can show that

$$KL(w) = -0.5 \log |\boldsymbol{\Sigma}| - \frac{p}{2} \log \alpha + \alpha E_w - 0.5\gamma \tag{40}$$

Because $\beta E_d = 0.5(N - \gamma)$ (equation 33) we can combine the above equation with equation 31 to give

$$EV(p) = \frac{N}{2} \log \beta - KL(w) \tag{41}$$

where we have dropped the constant term $-0.5N(1 + \log 2\pi)$. The VB criterion is given by

$$F(p) = \frac{N}{2} < \log \beta > -KL(w) - KL(\alpha) - KL(\beta) \tag{42}$$

where, again, we have dropped the constant term $-0.5N(1 + \log 2\pi)$. The evidence and VB criteria are therefore identical except for the divergences on α and β. But as these don't scale with p we can ignore them. We also note that, in practice, the distribution over β is sufficiently peaked for the first expectation term not to make a significant difference ie. $< \log \beta > \approx \log \hat{\beta}$.

If we assume, a priori, that different model orders are equally likely then the posterior probability distribution over model order is given by

$$q(p) = \frac{\exp(F(p))}{\sum_{p'} \exp(F(p'))} \tag{43}$$

This can also be applied to BIC/MDL.

RESULTS

Synthetic Data

We generated multiple 3 second-blocks of 128Hz data from AR(6) models with varying Signal-to-Noise Ratios (SNRs) . We also generated data from AR(25) models, again with varying SNRs. We generated 10 data sets of each

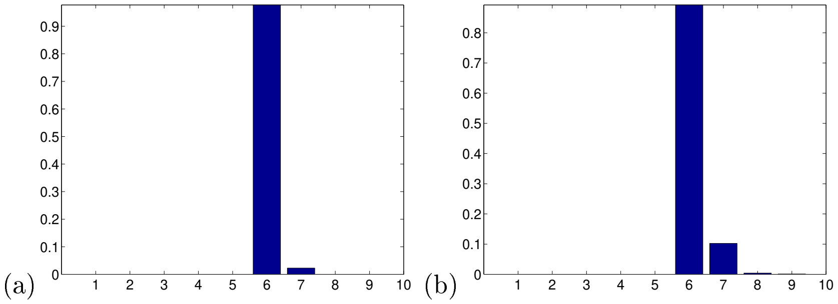

Figure 1: Model Order Selection for AR(6) data (SNR=2) with (a) BIC/MDL and (b) VB with 3-second blocks of data.

type (each with a different realisation of the noise process) and applied the VB and BIC/MDL model order selection methods. Typical results are shown in Figures 1 and 2 which are for $p = 6, SNR = 2$ and $p = 25, SNR = 20$ respectively. We have plotted the probability of each model order averaged over the 10 runs (the averaging was done on the probabilities rather than the log probabilities in order to show a greater spread). For both the AR(6) and AR(25) data, model order selection is more difficult at lower SNR. For the AR(6) data both methods essentially pick out the correct model order, with VB over-estimating it on a small proportion of runs. For the AR(25) data the BIC/MDL criterion significantly underestimates the model order. On longer blocks of data (eg. 10 seconds) both methods were seen to converge to the same estimate of model order (as predicted by theory).

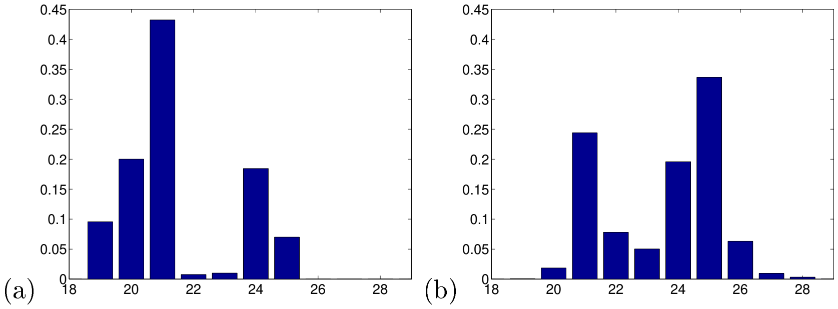

Figure 2: Model Order Selection for AR(25) data (SNR=20) with (a) BIC/MDL and (b) VB with 3-second blocks of data.

We also used the VB model order criterion with informative priors on α and β. These were set by running maximum-likelihood models on subsets of the data samples (typically we split the data into 3-sub blocks) and measuring the mean and variance of the precisions on the error and on the coefficients. The mean and variance of the gamma priors were then set so as to match these. The results differed from those of the evidence framework (as predicted by theory) and in many cases they were superior.

EEG Data

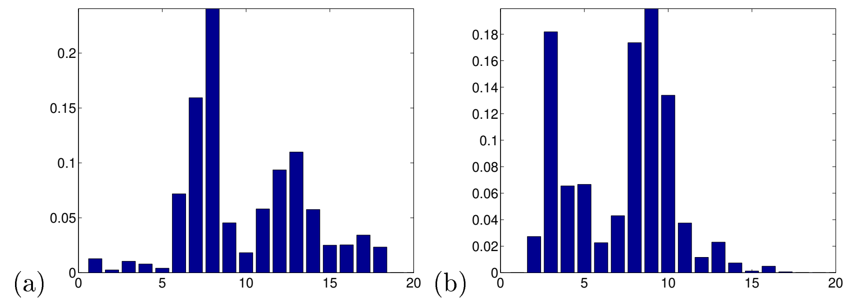

Figure 3: VB model order selection on EEG data from (a) awake subject and (b) anaesthetised subject.

We applied AR models to short blocks of EEG data recorded whilst a subject was awake or in an anaesthetised state. Fifty one-second blocks were selected randomly from each 30-minute recording and BIC/MDL and VB model order selection criteria were compared.

The results from VB in Figure 3 show that in the waking state there are modes at $p = 8$ and $p = 13$ and in the anaesthetised state the modes are at $p = 3$ and $p = 8/9$. The results clearly show a significant decrease in model order from the waking state to the anaesthetised state; comparing the means we see that the optimum model order decreases from 9.6 to 7.1 which, according to a t-test, is significant at the 0.0004 level. The results from BIC/MDL, however, showed no such discrimination; the difference of means (9.0 versus 10.4) was not significant; from a t-test we get a significance level of 0.062.

DISCUSSION

We have applied the Variational Bayesian framework to the autoregressive modeling of stationary stochastic processes. This allows for robust estimation of the AR parameters and of model order. Our experiments show that VB model order selection is superior to the BIC/MDL criterion.

We have shown that, for non-informative priors on the precision of the coefficients α and the precision of the noise β, the VB update rules are identical to those of the evidence framework. This is an extension of work by Mackay [8] who showed that, for a linear regression model with known β, the updates for α are the same. This link therefore provides further justification for the updating rules used in the evidence framework (as the VB update rules provably increase a lower bound on the log-evidence whereas no such convergence proof previously existed in the evidence framework). We have also shown that, apart from minor differences that have no practical impact, the VB and evidence model order selection criteria are identical (again, under the assumption of non-informative priors on α and β).

Whilst the VB and evidence approaches yield essentially the same algorithm when applied to the standard AR model, this is not generally the case. Indeed, for more complex models, VB is the preferred approach. This is because to apply the evidence framework we will generally need to compute the Hessian matrix, whereas this is not necessary in VB. This is demonstrated in a companion paper [11] where we apply VB to Non-Gaussian Autoregressive processes.

REFERENCES

[1] H. Attias, "A Variational Bayesian Framework for Graphical Models," in T. L. et al (ed.), **NIPS 12**, Cambridge, MA: MIT Press, 2000.

[2] C. Bishop, "Variational Principal Components," in **International Conference on Artificial Neural Networks**, 1999, pp. 509–514.

[3] C. M. Bishop, **Neural Networks for Pattern Recognition**, Oxford University Press, Oxford, 1995.

[4] P. Broersen and H. Wensink, "Autoregressive model order selection by a finite sample estimator for the Kullback-Leibler discrepancy," **IEEE Transactions on Signal Processing**, vol. 46, no. 7, pp. 2058–2061, 1998.

[5] D. Chickering and D. Heckerman, "Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables," Techn. Report MSR-TR-96-08, **Microsoft Research**, 1996.

[6] T. Cover and J. Thomas, **Elements of Information Theory**, John Wiley, 1991.

[7] P. Djuric and S. Kay, "Order Selection of Autoregressive Models," **IEEE Transactions on Signal Processing**, vol. 40, no. 11, 1992.

[8] D. Mackay, "Ensemble Learning and Evidence Maximization," Techn. report, **Cavendish Laboratory, University of Cambridge**, 1995.

[9] J. O'Ruaniaidh and W. Fitzgerald, **Numerical Bayesian Methods Applied to Signal Processing**, Springer, 1996.

[10] A. Papoulis, **Probability, Random Variables, and Stochastic Processes**, McGraw-Hill, 1991.

[11] W. Penny and S. Roberts, "Variational Bayes for Non-Gaussian Autoregressive Models," in **IEEE International Workshop on Neural Networks for Signal Processing (this volume)**, 2000.

[12] S. Weisberg, **Applied Linear Regression**, John Wiley, 1980.