Variational Bayes for Generalized Autoregressive Models

S.J. Roberts* & W. D. Penny
Robotics Research Group
Oxford University.
Oxford OX1 3PJ, U.K.
E-mail: sjrob,wpenny@robots.ox.ac.uk

May 22, 2002

Abstract

We describe a Variational Bayes (VB) learning algorithm for Generalized Autoregressive (GAR) models. The noise is modeled as a Mixture of Gaussians rather than the usual single Gaussian. This allows different data points to be associated with different noise levels and effectively provides robust estimation of AR coefficients. The VB framework is used to prevent overfitting and provides model order selection criteria both for AR order and noise model order. We show that for the special case of Gaussian noise and uninformative priors on the noise and weight precisions, the VB framework reduces to the Bayesian Evidence framework. The algorithm is applied to synthetic and real data with encouraging results.

1 Introduction

The standard autoregressive (AR) model assumes that the noise is Gaussian and therefore that the AR coefficients can be set by minimising a least squares cost function. Least squares, however, is known to be sensitive to outliers. Therefore, if the time series is even marginally contaminated by artifacts the resulting AR coefficient estimates will be seriously degraded. See Bishop ([1], page 209) and Press *et al.* ([2], page 700) for a general discussion of this issue and a number of proposed solutions.

This paper tackles this problem by modeling the noise with a Mixture of Gaussians (MoG) in which different data points can be associated with different noise levels. This provides a robust estimation of AR coefficients via a weighted least squares approach; data points associated with high noise levels are downweighted in the AR estimation step. This approach is thus well-suited to situations in which the signal is stationary (and may be modeled via an AR process) whilst the noise is non-Gaussian. The development of AR models with non-Gaussian excitation has a relatively long history. Work presented in [3], [4] and [5] utilizes essentially the same model as we discuss here. In these papers the issue of parameter inference is rightly tackled with a fully Bayesian approach using Markov Chain Monte Carlo (MCMC) sampling methods. Such schemes may be made efficient by exploiting tractable integration for part of the AR model [5, 6]. Recent use of reversible jump methods [7] applied to AR models [6] allows efficient exploration of model order in the Markov chain. Whilst these techniques are shown to be well-suited to AR model analysis, this paper considers an alternative formalism, variational Bayes (VB) learning, which offers a tractable (non-sampling) approach. Whilst the Bayesian methodology has a long history the use of VB is relatively new; the key idea of VB is to find a tractable approximation to the true posterior density which minimises the Kullback-Leibler (KL) divergence [8]. Notable recent applications are to Principal Component Analysis [9] and Independent Component Analysis [10]. We have also published short conference papers summarising some key results, for standard autoregressive models [11] and non-Gaussian AR models [12]. To our knowledge, VB has not previously been applied to such models.

Section 2 describes the autoregressive process as a probabilistic Bayesian model. We then describe the VB method (section 3) and show how it can be applied to generalized AR models (section 4). Section 5 describes the VB approach to the standard AR model and compares it to the evidence framework. Section 6 presents results on synthetic and real data and compares the difference between model evidence for VB and from a sampling step. Appendices, detailing some important results required elsewhere in the text, are included for completeness.

2 Generalized Autoregressive Models

We define an autoregressive model as

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

where y_n is the nth value of a time series, $\mathbf{w}$ is a column vector of AR coefficients (weights) and $\mathbf{x}_n = [y_{n-1}, y_{n-2}, ..., y_{n-p}]$ are the p previous time series values. The AR model has additive noise e_n which is usually modeled as a Gaussian. In this paper, however, we model the noise as a 1-dimensional Gaussian Mixture having m components. Component s has mixing coefficient π_s, mean μ_s and precision (inverse variance) β_s. We can write the parameters collectively as the vectors $\mathbf{w}$, $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_m]$, $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_m]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_m]$. The weights are drawn from a Gaussian prior (see later) with precision α. This choice of a Gaussian prior over the parameters is tantamount to a belief that the modeled data sequence has a smooth spectrum. For a more detailed discussion of this issue, and the choice of priors in AR models, the reader is referred to [13, 14].

We concatenate all the parameters into the overall parameter vector $\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{w}, \alpha\}$. We are given a set of data points $D = \{y_n, \mathbf{x}_n\}$ with $n = 1..N$. The likelihood of a data point is given by the mixture model

$$p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) = \sum_{s=1}^{m} p(s_n = s|\boldsymbol{\pi}) p(y_n|, \mathbf{x}_n, s_n, \beta_s, \mu_s, \mathbf{w}) \tag{2}$$

where s_n is an indicator variable indicating which component of the noise mixture model is selected at presentation of the n-th datum. These are chosen probabilistically according to

$$p(s_n = s|\boldsymbol{\pi}) = \pi_s. \tag{3}$$

The joint likelihood of a data point and indicator variable is

$$p(y_n, s_n|\mathbf{x}_n, \boldsymbol{\theta}) = p(s_n = s|\boldsymbol{\pi})p(y_n|\mathbf{x}_n, s_n, \beta_s, \mu_s, \mathbf{w}) \tag{4}$$

which, given that the noise samples are assumed independent and identically distributed (IID), gives

$$p(\mathbf{Y}, \mathbf{S}|\theta) = \prod_{n=1}^{N} p(s_n = s|\boldsymbol{\pi})p(y_n|\mathbf{x}_n, s_n, \beta_s, \mu_s, \mathbf{w}) \tag{5}$$

over the whole date set, where $\mathbf{Y} = [y_1, y_2, ..., y_N]^{\mathsf{T}}$ and $\mathbf{S} = [s_1, s_2, ..., s_N]^{\mathsf{T}}$.

Each component in the noise model is a Gaussian, hence:

$$p(y_n|\mathbf{x}_n, s_n, \beta_{s_n}, \mu_{s_n}, \mathbf{w}) = (2\pi)^{-1/2}\beta_{s_n}^{1/2} \exp\left(\frac{-\beta_{s_n}}{2}(y_n - [\mu_{s_n} + \mathbf{x}_n\mathbf{w}])^2\right) \tag{6}$$

2.1 Model priors

We choose the standard conjugate priors for each part of the model. These are either Normal, Gamma or Dirichlet densities which we define in the Appendix for completeness.

The prior on the model parameters is taken to factorise as:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_s p(\beta_s) \prod_s p(\mu_s)p(\mathbf{w}|\alpha)p(\alpha) \tag{7}$$

The prior over the mixing parameters of the MoG is a Dirichlet,

$$p(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\pi}, \boldsymbol{\lambda}) \tag{8}$$

where the hyperparameters are $\boldsymbol{\lambda} = [\lambda_0, \lambda_0, ..., \lambda_0]$ (ie. a symmetric Dirichlet). The prior over the precisions is a Gamma distribution

$$p(\beta_s) = \mathsf{Ga}(\beta_s; b_0, c_0), \tag{9}$$

the prior over the means a univariate Normal

$$p(\mu_s) = \mathsf{N}_1(\mu_s; m_0, v_0), \tag{10}$$

and the prior over the weights is a zero-mean Gaussian with an isotropic covariance having precision α

$$p(\mathbf{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{p/2} \exp(-\alpha E_W), \tag{11}$$

where

$$E_W = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{12}$$

Finally, the weight precision itself has a Gamma prior

$$p(\alpha) = \mathsf{Ga}(\alpha; b_\alpha, c_\alpha). \tag{13}$$

2.2 Gaussian noise models

The standard autoregressive model is recovered when the noise mixture consists of a single Gaussian component with zero mean and precision (inverse variance) β. The parameters now consist of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \mathbf{w}, \alpha\}$. The likelihood of the data set (from equation 2) now simplifies to:

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) \quad = \quad \left(\frac{\beta}{2\pi}\right)^{N/2} \exp(-\beta E_D(\mathbf{w})) \tag{14}$$

where

$$\begin{aligned} E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n\mathbf{w})^2 \\ &= \frac{1}{2}(\mathbf{Y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{Y} - \mathbf{X}\mathbf{w}) \end{aligned} \tag{15}$$

in which, as before, $\mathbf{Y}$ is a column vector with entries $(y_1, ...y_N)$ and the n-th row of the matrix $\mathbf{X}$ contains $\mathbf{x}_n$.

2.3 Maximum Likelihood

The standard autoregressive model with Gaussian noise can be implemented using a Maximum Likelihood approach. The optimal AR coefficients, given by the maximum of equation 15 [15], are:

$$\mathbf{w}_{ML} \quad = \quad (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \tag{16}$$

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

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 \tag{17}$$

The covariance is then estimated as

$$\mathbf{C}_{ML} = \sigma^2_{ML} (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} \tag{18}$$

This ML solution provides a method for initializing the Bayesian analysis.

3 Variational Bayes Learning

The central quantity of interest in Bayesian learning is the posterior distribution $p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})$ which fully describes our knowledge regarding the parameters of the model. In non-linear or non-Gaussian models, however, the posterior is often difficult to estimate as although one may be able to provide values for the posterior for a particular $\boldsymbol{\theta}, \mathbf{S}$ the partition function, or normalization term, may involve an intractable integral. To circumvent this problem two approaches have been developed: the sampling framework and the parametric framework. In the sampling framework integration is performed via a stochastic sampling procedure such as Markov-Chain Monte-Carlo (MCMC). The latter, however, can be computationally intensive and assessment of convergence is often problematic. Alternatively, the posterior can be assumed to be of a particular parametric form. In the Laplace approximation, employed for example in the 'Evidence framework', the posterior is assumed to be Gaussian [16]. This procedure is quick but is often inaccurate. Recently an alternative parametric method has been proposed: Variational Bayes (VB) or 'Ensemble Learning'. A full tutorial on VB is given in [8]. In what follows we briefly describe the key features.

Given a probabilistic model of the data with AR order p and m MoG components in the noise model, the 'evidence' or 'marginal likelihood' is given by:

$$p(\mathbf{Y}|p, m) = \int\int p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}|p, m) d\boldsymbol{\theta} d\mathbf{S} \tag{19}$$

The log evidence can be written as (dropping p, m on the right hand side for brevity)

$$\log p(\mathbf{Y}) = \log \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} \tag{20}$$

where $q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})$ is, as we shall see, a hypothesized, or approximate, posterior density. This has been introduced in both denominator and numerator in the above Equation. Noting that $q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})$ is a density function (i.e. it integrates to unity) we may apply Jensen's inequality to obtain a strict bound on the true log posterior as

$$\log p(\mathbf{Y}) = \log \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} \geq \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S}. \tag{21}$$

To see this bound from another perspective, we may write

$$\begin{aligned} \log p(\mathbf{Y}) &= \int\int d\boldsymbol{\theta} d\mathbf{S} q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log p(\mathbf{Y}) \\ &= \int\int d\boldsymbol{\theta} d\mathbf{S} q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \left(p(\mathbf{Y}) \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})} \right) \\ &= \int\int d\boldsymbol{\theta} d\mathbf{S} q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} \\ &= \int\int d\boldsymbol{\theta} d\mathbf{S} q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \left(\log p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) + \log q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) - \log q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \right) \\ &= \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} + \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})}{p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} \end{aligned} \tag{22}$$

We may write the latter Equation as

$$\log p(\mathbf{Y}|p, m) = F(p, m) + KL_{post}(p, m) \tag{23}$$

where (re-introducing the dependence upon p, m)

$$F(p, m) \stackrel{\text{def}}{=} \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}|p, m)}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} \tag{24}$$

is known as the negative variational free energy and

$$KL_{post}(p, m) \stackrel{\text{def}}{=} \int\int q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) \log \frac{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})}{p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}, p, m)} d\boldsymbol{\theta} d\mathbf{S} \tag{25}$$

is the KL-divergence [17] between the approximate posterior and the true posterior.

Equation 23 is the fundamental equation of the VB-framework. Importantly, because the KL-divergence is always positive [17], $F(p, m)$ provides a strict *lower bound* on the model evidence. Moreover, because the KL-divergence is zero when the two densities are the same, $F(p, m)$ will become equal to the model evidence when the approximating posterior is equal to the true posterior ie. if $q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) = p(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})$.

The aim of VB-learning is therefore to maximise $F(p,m)$ and so make the approximate posterior as close as possible to the true posterior. To obtain a practical learning algorithm we must also ensure that the integrals in $F(p,m)$ are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of parameters (in physics this is known as the mean field approximation). Thus, following [18], we consider:

$$q(\boldsymbol{\theta},\mathbf{S}|\mathbf{Y}) = q(\mathbf{S}|\mathbf{Y})\prod_i q(\boldsymbol{\theta}_i|\mathbf{Y}) \tag{26}$$

where $\boldsymbol{\theta}_i$ is the ith group of parameters. The distributions which maximise the negative free energy can then be shown to be of the following form (see Appendix A), here shown for parameter group $\boldsymbol{\theta}_i$,

$$q(\boldsymbol{\theta}_i|\mathbf{Y}) = \frac{\exp[I(\boldsymbol{\theta}_i)]}{\int \exp[I(\boldsymbol{\theta}_i)]d\boldsymbol{\theta}_i}. \tag{27}$$

where

$$I(\boldsymbol{\theta}_i) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}^{\backslash i}|\mathbf{Y}) \log p(\mathbf{Y},\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}^{\backslash i} \tag{28}$$

and $\boldsymbol{\theta}^{\backslash i}$ denotes all the parameters *not* in the ith group. For models having suitable priors, the above equations are available in closed analytic form. This leads to a set of coupled update rules. Iterated application of these leads to the desired maximization.

3.1 Model Order Selection

By computing the evidence for models of different order (ie. p and m) we can estimate the posterior distribution over p and m using the negative free energy. We may thence use this to perform model order selection. The posterior over p,m is given via Bayes' theorem as

$$P(p,m|\mathbf{Y}) = \frac{P(\mathbf{Y}|p,m)P(m,p)}{\sum_{p',m'} P(\mathbf{Y}|p',m')P(m',p')} \tag{29}$$

where, for example, we may have uniform priors over model orders p and m.

The evidence is estimated using $F(p,m)$. To see that $F(p,m)$ is an intuitively suitable model order criterion we decompose it as follows. Using $p(\mathbf{Y},\mathbf{S},\boldsymbol{\theta}) = p(\mathbf{Y},\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and assuming that the approximate posterior factorizes as in Equation 26 i.e. $q(\boldsymbol{\theta},\mathbf{S}|\mathbf{Y}) = q(\boldsymbol{\theta}|\mathbf{Y})q(\mathbf{S}|\mathbf{Y})$ we can write

$$F(p,m) = L_{av}(p,m) - KL_{prior} \tag{30}$$

where

$$L_{av}(p,m) = \int\int q(\boldsymbol{\theta}|\mathbf{Y})q(\mathbf{S}|\mathbf{Y}) \log \frac{p(\mathbf{Y},\mathbf{S}|\boldsymbol{\theta})}{q(\mathbf{S}|\mathbf{Y})} d\boldsymbol{\theta} d\mathbf{S} \tag{31}$$

corresponds to the average likelihood and

$$KL_{prior} = \int q(\boldsymbol{\theta}|\mathbf{Y}) \log \frac{q(\boldsymbol{\theta}|\mathbf{Y})}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{32}$$

is the KL-divergence between the approximate posterior $q(\boldsymbol{\theta}|\mathbf{Y})$ and the prior $p(\boldsymbol{\theta})$. Now, because the KL term increases with the number of model parameters it acts as a penalty term in Equation 30 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 $N \to \infty$, $F(p,m)$ becomes equivalent to the Bayesian Information Criterion (BIC) [18, 19]

$$BIC(p,m) = \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) - \frac{N_{par}}{2} \log N \tag{33}$$

where N_{par} is the total number of adjustable parameters in the model. The BIC is itself equal to the negative of the Minimum Description Length (MDL) measure i.e. $BIC(p,m) = -MDL(p,m)$. These popular model order selection criteria can therefore be seen as limiting cases of the VB framework.

4 VB for generalized AR models

To apply VB to Non-Gaussian Autoregressive models we approximate the posterior distribution over parameters with the factorized density

$$q(\boldsymbol{\theta}|\mathbf{Y}) = q(\boldsymbol{\pi}|\mathbf{Y})q(\boldsymbol{\beta}|\mathbf{Y})q(\boldsymbol{\mu}|\mathbf{Y})q(\mathbf{w}|\mathbf{Y})q(\alpha|\mathbf{Y}) \tag{34}$$

and the posterior distribution over hidden variables by $q(\mathbf{S}|\mathbf{Y})$. We then set each distribution so as to maximise $F(p,m)$ by applying equation 27, and substituting in the joint log-likelihood of the Generalized AR model

$$\log p(\mathbf{Y},\mathbf{S}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \pi_s + \frac{1}{2}\log \beta_s - \frac{1}{2}\beta_s(y_n - [\mu_s + \mathbf{x}_n \mathbf{w}])^2. \tag{35}$$

Given our choice of prior and model likelihood the optimal approximating densities take the following forms. For the precisions we have $q(\boldsymbol{\beta}|\mathbf{Y}) = \prod_s q(\beta_s|\mathbf{Y})$ and Gamma densities

$$q(\beta_s|\mathbf{Y}) = \mathsf{Ga}(\beta_s; b_s, c_s), \tag{36}$$

for the means, we have $q(\boldsymbol{\mu}) = \prod_s q(\mu_s)$ and univariate Normal densities

$$q(\mu_s|\mathbf{Y}) = \mathsf{N}_1(\mu_s; m_s, v_s), \tag{37}$$

for the mixing coefficients we have a Dirichlet

$$q(\boldsymbol{\pi}|\mathbf{Y}) \quad = \quad \mathsf{Dir}(\boldsymbol{\pi}; \lambda_1, ..., \lambda_m), \tag{38}$$

for the weights we have a multivariate Normal

$$q(\mathbf{w}|\mathbf{Y}) = \mathsf{N}(\mathbf{w}; \hat{\mathbf{w}}, \hat{\mathbf{C}}), \tag{39}$$

and for the weight precision a Gamma density

$$q(\alpha|\mathbf{Y}) = \mathsf{Ga}(\alpha; b'_\alpha, c'_\alpha) \tag{40}$$

We note finally that for the indicator posteriors we have $q(\mathbf{S}|\mathbf{Y}) = \prod_n q(s_n|\mathbf{Y})$.

Each distribution is updated as described below. Firstly, the responsibilities (i.e. the probabilities over the indicator variable set) of the Gaussian components in the noise model are updated, along with the hyperparameters which govern their distribution. Secondly, the component means and precisions are updated as well as the AR coefficients. All hyperparameters associated with the distributions over these variables are also updated during this second step. We detail the specific update equations for the variables in the following subsections.

4.1 Step 1

4.1.1 Component responsibilities

We first define some intermediate variables

$$\begin{aligned}
\log \tilde{\pi}_s &\stackrel{\text{def}}{=} \Psi(\lambda_s) - \Psi(\sum_{s'} \lambda_{s'}) \\
\log \tilde{\beta}_s &\stackrel{\text{def}}{=} \Psi(c_s) + \log b_s \\
\bar{\beta}_s &\stackrel{\text{def}}{=} b_s c_s
\end{aligned} \tag{41}$$

where $\Psi()$ is the digamma function [2] and

$$\begin{aligned} \tilde{\sigma}_s^2(n) &= \int q(\mathbf{w}|\mathbf{Y}) q(\mu_s|\mathbf{Y}) (y_n - [\mu_s + \mathbf{x}_n \mathbf{w}])^2 d\mathbf{w} d\mu_s \\ &= y_n^2 - 2m_s y_n - 2\hat{y}_n y_n + m_s^2 + v_s \\ &+ \; 2m_s \hat{y}_n + \mathbf{x}_n \hat{\mathbf{C}} \mathbf{x}_n^{\mathsf{T}} + \hat{y}_n^2 \end{aligned} \tag{42}$$

in which $\hat{y}_n = \mathbf{x}_n \hat{\mathbf{w}}$ is the mean AR prediction. If we define $\gamma_s^n \stackrel{\text{def}}{=} q(s_n|\mathbf{Y})$ to be the posterior of the indicator variable for the s–th component of the MoG (i.e. the probability that component s is responsible for data point y_n) then this step consists of updating the indicator posterior according to:

$$\tilde{\gamma}_s^n = \tilde{\pi}_s \tilde{\beta}_s^{1/2} \exp\left[-\frac{1}{2} \bar{\beta}_s \tilde{\sigma}_s^2(n)\right]. \tag{43}$$

These variables are then normalized to give

$$\gamma_s^n = \frac{\tilde{\gamma}_s^n}{\sum_{s'} \tilde{\gamma}_{s'}^n}. \tag{44}$$

4.2 Step 2

We now define some intermediate variables:

$$\begin{aligned} \bar{\pi}_s &\stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n \\ \bar{N}_s &\stackrel{\text{def}}{=} N \bar{\pi}_s \\ \bar{y}_s &\stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n y_n \\ \bar{\mathbf{x}}_s &\stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n \mathbf{x}_n \end{aligned} \tag{45}$$

i.e. $\bar{\pi}_s$ is the proportion of data associated with component s, $\bar{N}_s$ is the number of data points associated with component s and the quantities $\bar{\mathbf{x}}_s$ and $\bar{y}_s$ are weighted data values.

4.2.1 Component mixing fractions

The hyperparameters, $\{\lambda_s\}$ of the Dirichlet (Equation 38) are updated in the standard manner by adding the data counts to the prior counts, i.e.

$$\lambda_s = \bar{N}_s + \lambda_0 \tag{46}$$

4.2.2 Component precisions

If we define the expected variance of component s as

$$\tilde{\sigma}_s^2 = \frac{1}{N} \sum_{n=1}^{N} \gamma_s^n \tilde{\sigma}_s^2(n) \tag{47}$$

then the hyperparameters for the precisions (see Equation 36) are updated as

$$\begin{aligned} \frac{1}{b_s} &= \frac{N}{2} \tilde{\sigma}_s^2 + \frac{1}{b_0} \\ c_s &= \frac{\bar{N}_s}{2} + c_0 \end{aligned} \tag{48}$$

We can understand these equations by considering the corresponding mean variance (the inverse of the mean precision) which is given by $1/(b_s c_s)$. If we ignore terms involving the prior this comes out to be $\tilde{\sigma}_s^2 / \bar{\pi}_s$ which is the expected variance of that component re-weighted according to the number of examples that the component is responsible for.

4.2.3 Component means

In this step the hyperparameters governing the posterior distribution over the component means are updated (Equation 37). If we first define the means and precisions (i.e. the hyperparemeters) of the component means *as estimated from the data* as:

$$\begin{aligned} m_{data}(s) &= (\bar{\mathbf{x}}_s \hat{\mathbf{w}} - \bar{y}_s) / \bar{\pi}_s \\ \tau_{data}(s) &= \bar{N}_s \bar{\beta}_s \end{aligned} \tag{49}$$

then the posteriors of these hyperparameters are given by:

$$\begin{aligned} \tau_s &= \tau_0 + \tau_{data}(s) \\ m_s &= \frac{\tau_0}{\tau_s} m_0 + \frac{\tau_{data}(s)}{\tau_s} m_{data}(s) \end{aligned} \tag{50}$$

where $\tau_0 = 1/v_0$ is the prior precision and $\tau_s = 1/v_s$ is the posterior precision.

4.2.4 AR coefficients

The distribution of AR coefficients is multivariate normal with hyperparameters governing the mean and covariance of the distribution (Equation 39). These hyperparameters are updated via:

$$\hat{\mathbf{C}} = \left(\sum_s \bar{\beta}_s \mathbf{X}^{\mathsf{T}} \mathbf{\Gamma}_s \mathbf{X} + \bar{\alpha} \mathbf{I} \right)^{-1} \tag{51}$$

$$\hat{\mathbf{w}} \;=\; \hat{\mathbf{C}} \sum_s \bar{\beta}_s \left(\mathbf{X}^{\mathsf{T}} \mathbf{\Gamma}_s \mathbf{Y} + m_s \mathbf{X}^{\mathsf{T}} \mathbf{\Gamma}_s \mathbf{1} \right)$$

where $\mathbf{\Gamma}_s = diag([\gamma_s^1, \gamma_s^2, ..., \gamma_s^N])$, the nth row of $\mathbf{X}$ contains $\mathbf{x}_n$, the nth entry in the column vector $\mathbf{Y}$ is y_n and $\mathbf{1}$ is a column vector of 1's of length N.

4.2.5 AR coefficient precisions

The update equations for the coefficient precision and associated hyper-parameters (Equation 40) are:

$$\begin{aligned} 1/b'_\alpha &= \frac{1}{2}\hat{\mathbf{w}}^{\mathsf{T}}\hat{\mathbf{w}} + \frac{1}{2}\mathsf{Tr}(\hat{\mathbf{C}}) + \frac{1}{b_\alpha} \\ c'_\alpha &= \frac{p}{2} + c_\alpha \\ \bar{\alpha} &= b'_\alpha c'_\alpha \end{aligned} \tag{52}$$

in which $\mathsf{Tr}(\cdot)$ denotes the trace of a matrix.

4.3 Co-Mean Noise Model

While it is possible to use a full Gaussian Mixture model for the noise (as described above), the numerical examples in this paper use a simplified model in which all the means are zero. The resulting model then resembles weighted least squares where the weights are given by the precisions and responsibilities of the noise components. This simplifies a number of update equations. For step 1, as above, we now have (cf. equation 42):

$$\tilde{\sigma}_s^2(n) = y_n^2 - 2\hat{y}_n y_n + \mathbf{x}_n \hat{\mathbf{C}} \mathbf{x}_n^{\mathsf{T}} + \hat{y}_n^2. \tag{53}$$

In this case equations 49 and 50 are dispensed with and the second line in equation 51 simplifies to

$$\hat{\mathbf{w}} = \hat{\mathbf{C}} \sum_s \bar{\beta}_s \mathbf{X}^{\mathsf{T}} \mathbf{\Gamma}_s \mathbf{Y} \tag{54}$$

4.4 Pseudo-code

The above steps may be conveniently implemented in the following algorithm, shown here in pseudo-code form in Table 1.

```
initialize;
WHILE (ΔF(p,m) < tolerance )
    update indicator posteriors, γ̃_s^n using Equations 43,44;
    update component means, precisions & associated hyperparamaters using Equations 46-50;
    update AR coefficients and precisions using Equations 51, 52;
    calculate negative free energy, F(p,m) using Equation 55;
    calculate ΔF(p,m) ≝ |F(p,m)^new − F(p,m)^old|;
END WHILE;
```

Table 1: Pseudo-code for VB-AR updates.

4.5 Practicalities

The prior distribution $p(\boldsymbol{\pi})$ is chosen to be uniform and we set $\lambda_0 = 5$. For the Gamma distributions we use vague priors (see eg. [16]); for $p(\beta_s)$ we set $b_0 = 10^3$, $c_0 = 10^{-3}$, and for $p(\alpha)$ we set $b_\alpha = 10^3$ and $c_\alpha = 10^{-3}$. It is noted that we do not find particular sensitivity to these values.

The posterior distributions $q(\boldsymbol{\pi}|\mathbf{Y})$, $q(\beta_s|\mathbf{Y})$, $q(\alpha|\mathbf{Y})$ and $q(\mathbf{w}|\mathbf{Y})$ have the parameters λ_s, b_s, c_s, b'_α, c'_α, $\hat{\mathbf{w}}$ and $\hat{\mathbf{C}}$ which are initialized as follows. The posterior for the AR coefficients is initialized using the Maximum Likelihood (ML) solutions of equations 16, 17 & 18. The posterior for the weight precision is then initialized using equation 52. The remaining posteriors are set as follows. We first calculate the errors $e(n)$ from the ML model and then define a new variable $z(n) = |e(n) - \bar{e}|$ which is the absolute deviation of each error from the mean error $\bar{e}$. We then apply k-means clustering [1] to $z(n)$ which results in mixing coefficients $\lambda_z(s)$ and means $m_z(s)$. We then set $\lambda_s = 100\lambda_z(s)$. The parameters b_s and c_s are then set so as to achieve means of $1/m_z(s)^2$ and variances of $\mathrm{var}(1/m_z)$ (the mean and variance of a Gamma density are bc and b^2c respectively).

The VB equations are then applied iteratively until a consistent solution is reached. Convergence is measured by evaluating the negative free energy

$$F(p,m) = L_{av} - KL(\mathbf{w}) - KL(\boldsymbol{\alpha}) - KL(\boldsymbol{\pi}) - KL(\boldsymbol{\beta}) \tag{55}$$

where we use the shorthand notation $KL(a) \stackrel{\text{def}}{=} KL(q(a|\mathbf{Y}) \parallel p(a))$. The first term of the above equation is given as:

$$L_{av} = H(q(\mathbf{S}|\mathbf{Y})) + \sum_{s=1}^{m} \bar{N}_s(\log \tilde{\pi}_s + \frac{1}{2}\log \tilde{\beta}_s) - \frac{1}{2}N\sum_{s=1}^{m} \bar{\beta}_s \tilde{\sigma}_s^2 - \frac{1}{2}N\log 2\pi \tag{56}$$

The entropy over the hidden variables is

$$H(q(\mathbf{S}|\mathbf{Y})) = -\sum_{n=1}^{N}\sum_{s=1}^{m} \gamma_s^n \log \gamma_s^n \tag{57}$$

The KL terms for Normal, Gamma and Dirichlet densities are given in the Appendix. As the update for the AR coefficients is computationally more intensive than the updates for the other parameters (as it involves a matrix inversion) we perform this step only once every W_t iterations. In our experiments we used $W_t = 5$. We evaluate $F(p, m)$ every W_t iterations (after the AR updates) and terminate optimization if the proportionate increase from one evaluation to the next is less than 0.01 per cent.

We note that the formalism adopted in this paper does not, unlike many AR modeling approaches, *guarantee* a stable model (i.e. the poles of the characteristic function do not lie outside the unit circle in the z-domain). Whilst it is possible to enforce stability by reflecting poles across the unit circle we do not regard this as an elegant solution. Placing stability priors on the *reflection coefficients* of the model is another possibility (as described in [20]) but this leaves analytically troublesome descriptions of the AR coefficients themselves for which (slow) sample-based approaches must thence be used. It is noted, however, that in all examples in this paper, and indeed all data analyzed thus far, no unstable models have been found.

5 The Gaussian Autoregressive model

The standard autoregressive model is recovered when the mixture consists of a single zero-mean Gaussian. Because of the factored form of the prior distribution the form of the approximate posterior which maximises the negative free energy is also of factorized form (we note that, unlike the general case in variational Bayes methods, we do not have to *assume* this), i.e.

$$q(\boldsymbol{\theta}|\mathbf{Y}) = q(\mathbf{w}|\mathbf{Y})q(\alpha|\mathbf{Y})q(\beta|\mathbf{Y}) \tag{58}$$

The weight posterior is a Normal density $q(\mathbf{w}|\mathbf{Y}) = \mathsf{N}(\mathbf{w}; \hat{\mathbf{w}}, \hat{\mathbf{C}})$. By inspection of equation 51, noting that there is now only a single value of β, that $m_1 = 0$ and that the weighting matrix Γ_s is an identity matrix we get

$$\hat{\mathbf{C}} = (\hat{\beta}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \hat{\alpha}\mathbf{I})^{-1} \tag{59}$$

$$\hat{\mathbf{w}} = \hat{\mathbf{C}}\hat{\beta}\mathbf{X}^{\mathsf{T}}\mathbf{Y}$$

The weight precision posterior is a Gamma density with parameters as before (equation 52). The noise precision posterior is $q(\beta|\mathbf{Y}) = \mathsf{Ga}(\beta; b'_\beta, c'_\beta)$ where

$$\begin{aligned} 1/b'_\beta &= \bar{E}_D(\hat{\mathbf{w}}) + \frac{1}{b_\beta} \\ c'_\beta &= \frac{N}{2} + c_\beta \\ \hat{\beta} &= b'_\beta c'_\beta \end{aligned} \tag{60}$$

and

$$\bar{E}_D(\hat{\mathbf{w}}) = E_D(\hat{\mathbf{w}}) + \frac{1}{2}\mathsf{Tr}(\hat{\mathbf{C}}\mathbf{X}^{\mathsf{T}}\mathbf{X}) \tag{61}$$

We note that in previous work MacKay has derived VB updates for the weight and weight precisions in a linear regression model [21], and that, re-assuringly, our equations (52,59) are identical to his. We note that these update rules can also be derived directly, i.e. without considering the model as a special case of GAR.

For Gaussian noise models the negative free energy simplifies to

$$F(p) = L_{av} - KL(\mathbf{w}) - KL(\alpha) - KL(\beta) \tag{62}$$

where

$$L_{av} = \frac{N}{2}\langle \log \beta \rangle - \hat{\beta}\bar{E}_D - \frac{N}{2}\log 2\pi \tag{63}$$

and

$$\langle \log \beta \rangle = \Psi(c'_\beta) + \log b'_\beta \tag{64}$$

and $\Psi()$ is the digamma function.

An alternative Bayesian approach is that offered by the Evidence framework [16]. 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 also by equation 59. 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 [1])

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)d\mathbf{w} \tag{65}$$

The log of the evidence can then be written as

$$EV(p) = -\alpha E_W - \beta E_D(\hat{\mathbf{w}}) + \frac{1}{2}\log|\mathbf{C}| + \frac{p}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log 2\pi \tag{66}$$

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

$$\alpha = \frac{\gamma}{2E_W} \tag{67}$$

$$\beta = \frac{N-\gamma}{2E_D(\hat{\mathbf{w}})} \tag{68}$$

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

$$\gamma = p - \alpha \mathsf{Tr}(\mathbf{C}) \tag{69}$$

where 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

$$\alpha = \frac{p}{2E_W + \mathsf{Tr}(\mathbf{C})} \tag{70}$$

This is equivalent to the update for $\hat{\alpha}$ from the VB framework (equation 52) if we choose uninformative priors on α. Similarly, if we choose uninformative priors for β then the VB update is (from equation 60)

$$\hat{\beta} = \frac{N}{2E_D(\hat{\mathbf{w}}) + \mathsf{Tr}(\hat{\mathbf{C}}\mathbf{X}^{\mathsf{T}}\mathbf{X})} \tag{71}$$

which is equivalent to equation 68 as, from an eigendecomposition of $\hat{\mathbf{C}}$, it can be shown that $\mathsf{Tr}(\hat{\mathbf{C}}\mathbf{X}^{\mathsf{T}}\mathbf{X}) = \gamma/\beta$.

5.1 Model Order Selection

As mentioned section 3.1, the BIC model-selection criterion is a limiting case of the VB framework. For the AR model we have

$$BIC(p) = \frac{N}{2}\log\beta_{ML} - \frac{p}{2}\log N \tag{72}$$

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

The model order selection criterion for the evidence framework (equation 66) can be rewritten in terms of the KL-divergence between the weight posterior and the weight prior, $KL(\mathbf{w})$. For the prior $p(\mathbf{w}) = \mathsf{N}(\mathbf{w}; \mathbf{0}, (1/\alpha)\mathbf{I})$ and the posterior $q(\mathbf{w}) = \mathsf{N}(\mathbf{w}; \hat{\mathbf{w}}, \hat{\mathbf{C}})$ it can be shown that:

$$KL(\mathbf{w}) = -\frac{1}{2}\log|\mathbf{C}| - \frac{p}{2}\log\alpha + \alpha E_w - \frac{1}{2}\gamma. \tag{73}$$

Because $\beta E_d = \frac{1}{2}(N - \gamma)$ (equation 68) we can combine the above equation with equation 66 to give

$$EV(p) = \frac{N}{2} \log \beta - KL(\mathbf{w}) \tag{74}$$

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

$$F(p) = \frac{N}{2} \langle \log \beta \rangle - KL(\mathbf{w}) - KL(\alpha) - KL(\beta) \tag{75}$$

where, again, we have dropped the constant term $-\frac{1}{2}N(1 + \log 2\pi)$. The evidence and VB criteria are therefore identical except for the divergences of α and β. But as these do not scale with p we can ignore them. We also note that, for large data sets, the distribution over β is sufficiently peaked for the first expectation term not to make a significant difference ie. $\langle \log \beta \rangle \approx \log \hat{\beta}$. We may also thus see the evidence approach as a limiting case of the VB framework.

If we assume, a priori, that different model orders are equally likely then the posterior probability distribution over model order is given by Equation 29. This can also be applied to BIC/MDL.

6 Results

6.1 Synthetic Data: Gaussian noise model

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

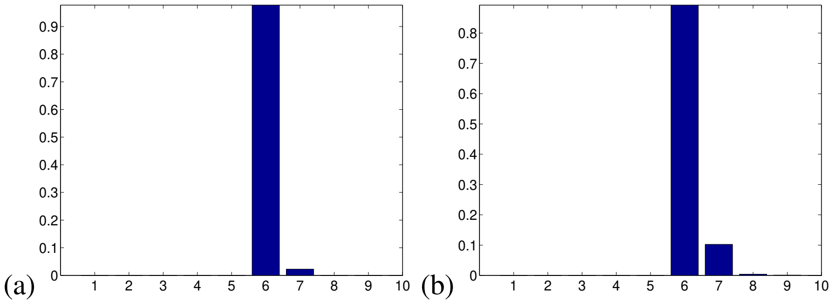

Figure 1: Model Order Selection for AR(6) data (SNR=2) with (a) BIC/MDL and (b) VB with 3-second blocks of data.

10 data sets of each 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[1]. We have plotted the probability of each

[1] The SNR was altered by adjusting the noise variance.

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).

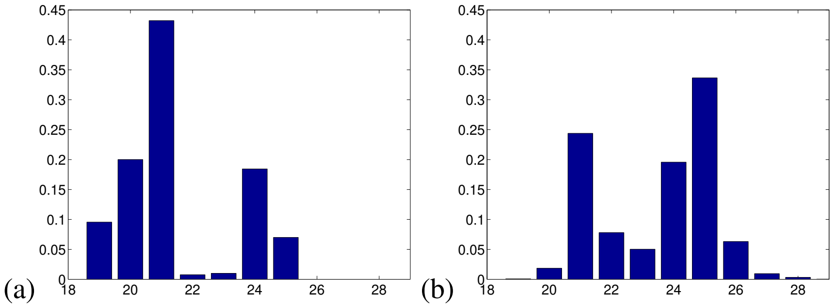

Figure 2: Model Order Selection for AR(25) data (SNR=20) with (a) BIC/MDL and (b) VB with 3-second blocks of data.

6.2 Synthetic Data: Non-Gaussian noise model

We generated three seconds of synthetic data from an AR(5) model with coefficients

$$\mathbf{w}_{true} = [-1.8517, 1.3741, 0.1421, -0.6852, 0.3506]^{\mathsf{T}} \tag{76}$$

and noise drawn from a 2-component Gaussian mixture model with mixing coefficients $\pi_1 = 0.9, \pi_2 = 0.1$ and variances $\sigma_1^2 = 1, \sigma_2^2 = 100$. We repeated the data generation process 10 times and used the negative free energy, $F(p, m)$ as a model order selection criterion. The results in Figures 3, 4 and 5 show

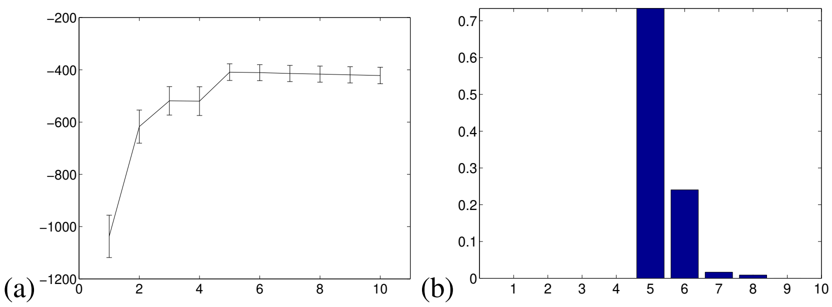

Figure 3: AR model order selection (a) $F(p, 2)$ versus p and (b) the corresponding probabilities, $P(p, 2)$ versus p.

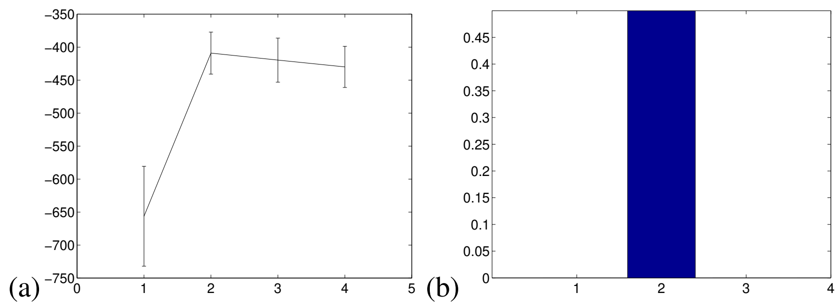

Figure 4: Noise model order selection (a) $F(5, m)$ versus m and (b) the corresponding probabilities, $P(5, m)$ versus m.

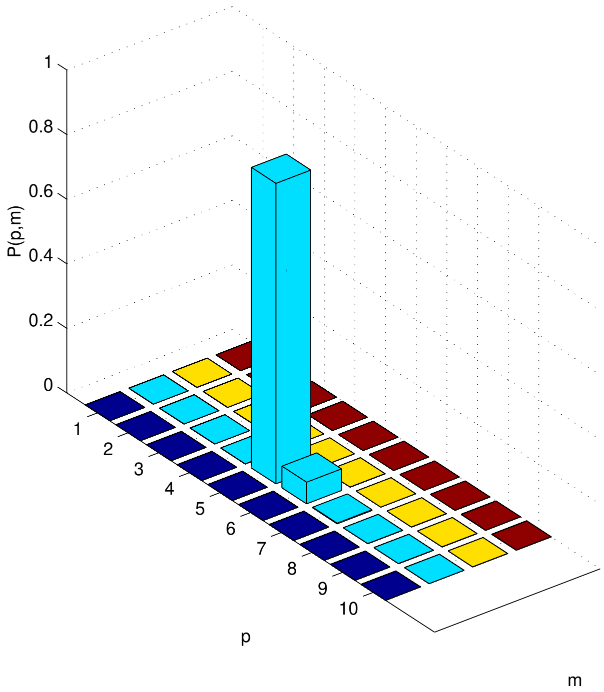

Figure 5: Noise and AR model order selection. The bar plot shows $P(p, m)$ against p, m.

that the VB model order selection criterion selects both the correct AR order and the correct noise model order. For more on VB model order selection, including a comparison with the Minimum Description Length (MDL) criterion, see [11].

We also calculated the error with which the AR coefficients were estimated using the measure $\| \hat{\mathbf{w}} - \mathbf{w}_{true} \|$. Over the ten runs, the errors from the Non-Gaussian AR model with two components were less than those from the Gaussian AR model by an average factor of six.

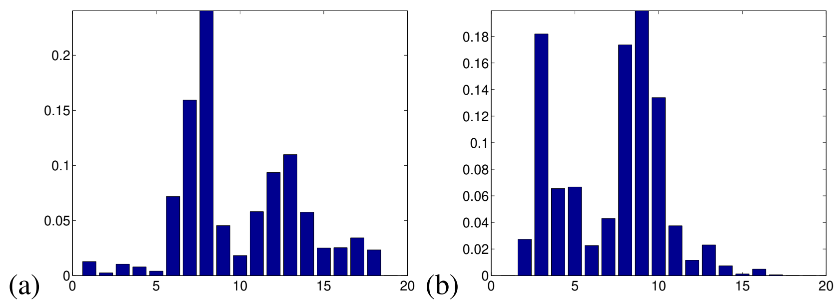

Figure 6: VB model order selection on EEG data from (a) awake subject and (b) anesthetized subject.

6.3 EEG Data: Gaussian noise model

We applied AR models to short blocks of EEG data recorded while a subject was awake or in an anesthetized 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 6 show that in the waking state there are modes at $p = 8$ and $p = 13$ and in the anesthetized state the modes are at $p = 3$ and $p = 8/9$. The results clearly show a decrease in model order from the waking state to the anesthetized state; according to a t-test, this decrease is significant at the 0.0004 level. The results from BIC/MDL, however, showed no such discrimination: the difference of means was not significant; from a t-test we get a significance level of 0.062.

6.4 EEG Data: Non-Gaussian noise model

We applied the model to a short sample of EEG data shown in Figure 7, the middle section of which is contaminated by an eye-movement artifact. For a Gaussian noise model ($m = 1$) the optimum AR model order was at $p = 27$ having $F_{27,1} = -2681$. For $m = 2$ and $m = 3$ the best models were also at $p = 27$ (this is not always the case) with $F_{27,2} = -2661$ and $F_{27,3} = -2671$. This shows that the non-Gaussian AR process with two noise components is the preferred model. This model took 25 iterations to converge. The final mixing coefficients in the noise model were $\pi_1 = 0.78$ and $\pi_2 = 0.22$ and the variances (inverse precisions) were $\sigma_1^2 = 0.014$ and $\sigma_2^2 = 0.066$, i.e. a low and high variance component. Figure 7 shows which data points 'belong' to which noise component. The first component corresponds to the majority of the data (78% of it) and the second component picks out the outliers which are mostly in the middle section. These points are then effectively downweighted in the corresponding

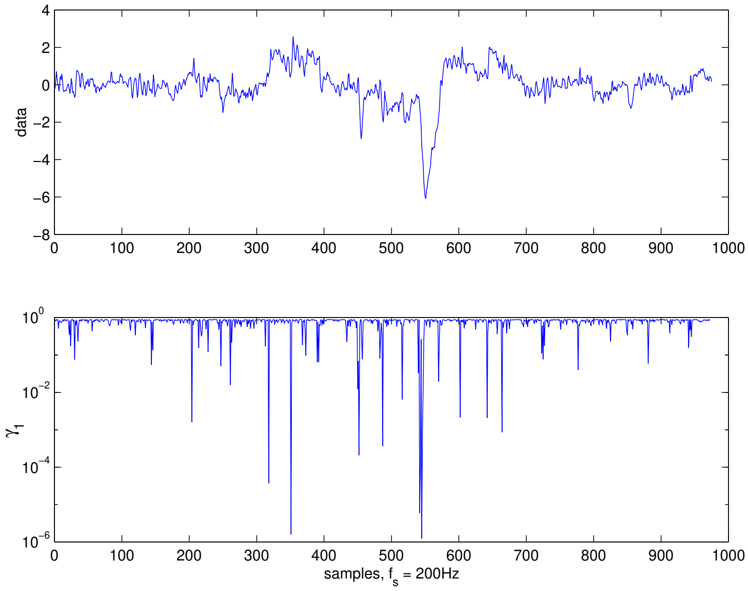

Figure 7: Top plot: Original EEG trace, Bottom plot: The log-probabilitity that the data belongs to the first (low-variance) noise component, $\log \gamma_1$ (calculated from equation 44). A high proportion of data in the middle section therefore clearly belongs to the second (high-variance) noise component. These points are downweighted in the estimation of the AR coefficients.

weighted least squares regression. The outliers therefore have a minimal influence on the estimation of the AR coefficients; this can be seen from equation 54 where $\bar{\beta}_s$ and $\mathbf{\Gamma}_s$ weight the samples accordingly.

6.5 How good is the variational bound?

From Equation 23 we saw that the negative free energy term, $F(p, m)$, formed a strict lower bound to the true posterior (the KL term is strictly non-negative). How tight this bound is will clearly affect the confidence we have in the variational approach. We consider here a simple experimental validation of the bound and present results on the non-Gaussian noise example in Section 6.

Consider the marginal of Equation 19, once more dropping, for notational convenience, dependence upon p, m,

$$p(\mathbf{Y}) = \int\!\!\int p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}) d\mathbf{S} d\boldsymbol{\theta} \tag{77}$$

We may write the above as

$$p(\mathbf{Y}) = \int\!\!\int \left(\frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} \right) q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) d\mathbf{S} d\boldsymbol{\theta} = \int\!\!\int w(\boldsymbol{\theta}, \mathbf{S}) q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y}) d\mathbf{S} d\boldsymbol{\theta} \tag{78}$$

in which we define the weight function

$$w(\boldsymbol{\theta}, \mathbf{S}) \stackrel{\text{def}}{=} \frac{p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})} \tag{79}$$

As noticed by others [22, chap. 4] this enables the variational approximation, $q(\boldsymbol{\theta}, \mathbf{S}|\mathbf{Y})$ to be used as the proposal in an *importance sampling* step. We do not cover the details of importance sampling, as they may be found in standard texts (e.g. [23], chapter 5). Suffice to say, the true posterior may be estimated via importance weights which are evaluated as an expectation under the proposal, i.e Equation 78 above may be written as

$$p(\mathbf{Y}) \approx \mathbb{E}[w(\boldsymbol{\theta}, \mathbf{S})]_q \tag{80}$$

We applied this procedure to the non-Gaussian AR($p = 5, m = 2$) data analysed in Section 6. A total of 3s of data (384 samples) was generated from the true model as before. The VB-AR algorithm was applied to this data and then, using the variational inference densities as proposals, 1000 samples were evaluated using importance sampling to estimate the true model posterior[2]. This was repeated over all model orders from $p = 1$ to 10 and with the number of Gaussians in the noise model from $m = 1$ to 5. Figure 8 shows the relative increase in model evidence after the sampling procedure over the p, m grid. We note that the relative difference between the 'true' and approximate posteriors is of order one part in 10^5 at most (for these models). We note that the maximum model evidence appeared at the true model order in both cases. We note that, as expected from theory, the variational approximation is always a strict lower bound. It appears that the variational bound is very tight for low model orders but rises with increased numbers of model parameters, both in terms of the AR model order (p) but also, importantly, in terms of the number of mixture components in the noise model (m). This is to be expected due to the factored nature of the variational approximation which gets worse with increasing numbers of parameters. It is encouraging that the evidence 'peaks' at the correct order in both cases, indicating that the variational free energy may offer a realistic model-selection criterion. More importantly, the use of the variational approximation as a proposal for importance sampling makes the latter very fast and this approach is to be commended in cases where evaluation of the bound is crucial.

[2] As the dimensionality of the model increases, so does the tendancy fof the importance weights to be dominated by a small number of large weights. In the experiments reported here this did not occur, though we note this tendancy for high p models.

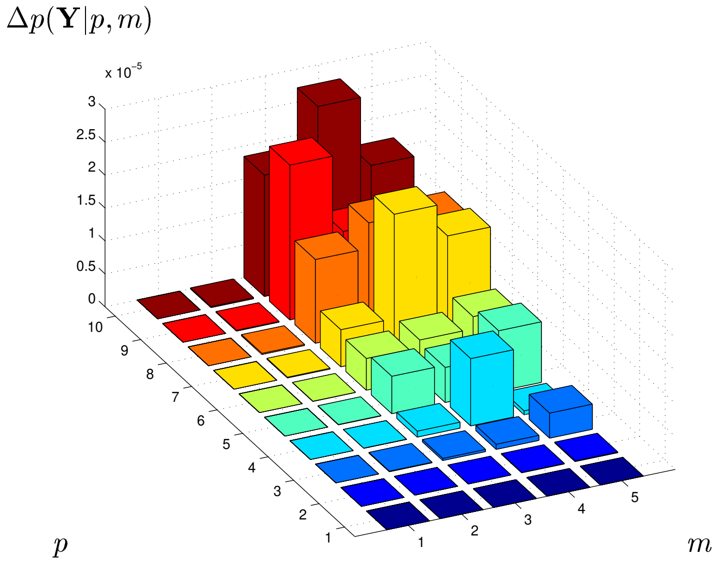

Figure 8: Relative increase in evidence versus model order after importance sampling using the variational posterior as proposal.

7 Discussion

We have proposed a non-Gaussian AR model for the modeling of stationary stochastic processes where the noise is modeled with a Mixture of Gaussians. The main application is as a model for stationary signals contaminated by non-Gaussian and/or non-stationary noise processes. The VB framework prevents overfitting and provides a model order selection criterion both for the AR order and the noise model order. In the experiments presented, we restricted the means in the noise mixture to be identically zero. The resulting model resembles weighted least squares and hence provides robust estimation of AR coefficients.

Our model provides an alternative to cumulant methods for Non-Gaussian AR modeling (see eg. [24, 25]), for which model order estimation is problematic.

We envisage that a main application of the model is to EEG data; in an EEG recording up to 30% [26] of data blocks are corrupted by muscle, eye-movement or other artifact and, traditionally, these data blocks are discarded. As we have shown, however, the non-Gaussian AR model is able to downweight the contaminated section of each block, thus allowing more reliable AR (and consequently spectral) estimation. We also envisage using the non-Gaussian AR algorithm to model the sources in an Independent Component Analysis model; this would extend the work in [10] by allowing for dynamic sources at the same time as using VB for model order estimation. It is noted that the model we use, in which the noise model has no dynamic structure, is equivalent to a 0-th order Hidden Markov Model (HMM) for the noise process with a number of hidden states equal to the number of components in the MoG. In the eye-movement artifact example the high variance activity occurred sporadically and the HMM(0) model is appropriate. It is arguable, however, that in some cases, where the high-variance noise occurs in temporally-correlated bursts, that a higher-order HMM is appropriate (a HMM(1), for example). We have not considered this here, though it is an important area of future research.

For Gaussian AR models, 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 [21] 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).

Acknowledgements

We would like to thank Ron Daniel and Peter Sykacek for interesting discussions and helpful comments and criticism of this work. This work was funded in part by grant EESD PRO264 from the UK Engineering & Physical Science Research Council (EPSRC) to whom we are most grateful. The comments of the anonymous referees were most valuable in the final revision of this paper.

References

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

[2] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C.* Cambridge, 1992.

[3] G. Barnett, R. Kohn, and S. Sheather. Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Journal of Econometrics*, 74(2):237–254, 1996.

[4] S. Godshill. Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes. *Statistical Review*, 65(1):1–21, 1996.

[5] S. Godshill and P. Rayner. Statistical reconstruction and analysis of autoregressive signals in impulsive noise. *IEEE Transactions on Speech & Audio Processing*, 6(4):352–372, July 1998.

[6] P Troughton and S. Godshill. A reversible jump sampler for Autoregressive time series, employing full conditionals to achieve efficient model space moves. Technical Report CUED/F-INFENG/TR.304, Department of Engineering, University of Cambridge, 1997.

[7] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society (Series B)*, 59(4):731–758, 1997.

[8] H. Lappalainen and J.W. Miskin. Ensemble learning. chapter in Advances in Independent Components Analysis, M. Girolami (Ed). Kluwer Academic Publishers, 2000.

[9] C.M. Bishop. Variational Principal Components. In *International Conference on Artificial Neural Networks*, pages 509–514, 1999.

[10] R. Choudrey, W.D. Penny, and S.J. Roberts. An Ensemble Learning approach to Independent Component Analysis. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[11] W.D. Penny and S.J. Roberts. Bayesian Methods for Autoregressive Models. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[12] W.D. Penny and S.J. Roberts. Variational Bayes for Non-Gaussian Autoregressive Models. In *IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia*, 2000.

[13] G. Kitagawa and W. Gersch. A smoothness priors time-varying ar-coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control*, 30:45–56, 1985.

[14] G. Kitagawa and W. Gersch. A smoothness priors long ar model method for spectral estimation. *IEEE Transactions on Automatic Control*, 30:57–65, 1985.

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

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

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

[18] H. Attias. A Variational Bayesian Framework for Graphical Models. In T. Leen et al, editor, *NIPS 12*, Cambridge, MA, 2000. MIT Press.

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

[20] P. Sykacek, I. Rezek, and S. Roberts. Markov chain Monte Carlo methods for Bayesian sensor fusion. Technical report, Department of Engineering Science, University of Oxford, 2000. Technical Report PARG-00-10, available from http://www.robots.ox.ac.uk/∼parg/.

[21] D.J.C. Mackay. Ensemble Learning and Evidence Maximization. Technical report, Cavendish Laboratory, University of Cambridge, 1995.

[22] J. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, Cavendish Laboratory, 2000.

[23] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley, 1994.

[24] S. Minfen, S. Lisha, and P.J. Beadle. Parametric Bispectral Estimation of EEG signals in different functional states of brain. In *Advances in Medical Signal and Information Processing*, pages 66–72. IEE, 2000.

[25] C.L. Nikias. *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.

[26] A. Gevins, M.E. Smith, and D. Yu. High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7(4):374–385, 1997.

A Derivation of Variational Learning

We start by restating Equation 27 which conjectures that the optimal form of the proposal (q) distribution that maximizes the negative free energy (model evidence measure), with respect to parameter group $\boldsymbol{\theta}_i$, is

$$q(\boldsymbol{\theta}_i|\mathbf{Y}) = \frac{1}{Z_i}\exp[I(\boldsymbol{\theta}_i)] \tag{81}$$

where Z_i is the partition function (normalising constant) and where, as before, we define

$$I(\boldsymbol{\theta}_i) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}^{\backslash i}|\mathbf{Y}) \log p(\mathbf{Y}, \mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}^{\backslash i}. \tag{82}$$

This maximisation is equivalent to stating that the negative KL-divergence between $q(\boldsymbol{\theta}_i|\mathbf{Y})$ and $\exp[I(\boldsymbol{\theta}_i)]/Z_i$ is maximized (this is strictly non-positive and reaches its maximum when Equation 81 holds). The negative KL-divergence is given as

$$-KL(q(\boldsymbol{\theta}_i|\mathbf{Y}), \exp[I(\boldsymbol{\theta}_i)]/Z_i) = \int d\boldsymbol{\theta}_i q(\boldsymbol{\theta}_i|\mathbf{Y}) \left(\log\exp[I(\boldsymbol{\theta}_i)] - \log q(\boldsymbol{\theta}_i|\mathbf{Y}) - \log Z_i\right). \tag{83}$$

Subsituting from Equation 82 we obtain

$$\begin{aligned}
-KL(q(\boldsymbol{\theta}_i|\mathbf{Y}), \exp[I(\boldsymbol{\theta}_i)]/Z_i) &= \int d\boldsymbol{\theta}_i q(\boldsymbol{\theta}_i|\mathbf{Y}) \log q(\boldsymbol{\theta}_i|\mathbf{Y})^{-1} \int d\boldsymbol{\theta}^{\backslash i} q(\boldsymbol{\theta}^{\backslash i}|\mathbf{Y}) \log[p(\mathbf{Y}, \mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})] - \log Z_i \\
&= \iint d\boldsymbol{\theta}_i d\boldsymbol{\theta}^{\backslash i} q(\boldsymbol{\theta}_i|\mathbf{Y}) q(\boldsymbol{\theta}^{\backslash i}|\mathbf{Y}) \log\left[\frac{p(\mathbf{Y}, \mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_i|\mathbf{Y})}\right] - \log Z_i \\
&= \int d\boldsymbol{\theta} q(\boldsymbol{\theta}|\mathbf{Y}) \left[\frac{p(\mathbf{Y}, \mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_i|\mathbf{Y})}\right] - \log Z_i
\end{aligned} \tag{84}$$

where the last step follows from the factored representation of the proposal, i.e. Equation 26. Note that this is (to within an additive constant) the contribution to the free energy of Equation 24 of the parameter group $\boldsymbol{\theta}_i$. As the free energy is strictly non-negative it is maximized by maximization of each contribution, e.g. maximization of Equation 84. This is achieved when Equation 81 holds and hence we obtain a simple methodology to obtain densities for each component of q.

B Densities and Divergences

For completeness we include definitions of the Normal, Gamma and Dirichlet densities, their entropies and KL-divergencies.

B.1 The Univariate Normal density

For univariate normal densities $q(x) = \mathsf{N}_1(x; \mu_q, \sigma_q^2)$ and $p(x) = \mathsf{N}_1(x; \mu_p, \sigma_p^2)$ the KL-divergence is

$$KL(q \parallel p) = \frac{1}{2} \log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 + \sigma_q^2 - 2\mu_q\mu_p}{2\sigma_p^2} - \frac{1}{2}. \tag{85}$$

B.2 The Multivariate Normal density

The multivariate Normal density is given by

$$\mathsf{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \tag{86}$$

The KL-divergence for Normal densities $q(\mathbf{x}) = \mathsf{N}(\mathbf{x}; \boldsymbol{\mu}_q, \mathbf{C}_q)$ and $p(\mathbf{x}) = \mathsf{N}(\mathbf{x}; \boldsymbol{\mu}_p, \mathbf{C}_p)$ is

$$KL(q \parallel p) = \frac{1}{2} \log \frac{|\mathbf{C}_p|}{|\mathbf{C}_q|} + \frac{1}{2} \mathsf{Tr}(\mathbf{C}_p^{-1} \mathbf{C}_q) + \frac{1}{2} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^{\mathsf{T}} \mathbf{C}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - \frac{d}{2}. \tag{87}$$

where $|\mathbf{C}_p|$ denotes the determinant of the matrix $\mathbf{C}_p$.

B.3 The Gamma Density

The Gamma density is defined as

$$\mathsf{Ga}(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b} \right). \tag{88}$$

For Gamma densities $q(x) = \mathsf{Ga}(\mathbf{x}; b_q, c_q)$ and $p(x) = \mathsf{Ga}(\mathbf{x}; b_p, c_p)$ the KL-divergence is

$$\begin{aligned} KL(q \parallel p) &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log \Gamma(c_q) \\ &+ \log \Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \tag{89}$$

where $\Psi()$ is, as before, the digamma function [2].

B.4 The Dirichlet Density

The Dirichlet density is given by

$$\mathsf{Dir}(\boldsymbol{\pi};\boldsymbol{\lambda}) = \frac{\Gamma(\sum_{s=1}^{m}\lambda_s)}{\prod_{s=1}^{m}\Gamma(\lambda_s)}\prod_{s=1}^{m}\pi_s^{\lambda_s-1} \tag{90}$$

where λ_s is the sth element of $\boldsymbol{\lambda}$ and $\Gamma(x)$ is the Gamma function [2].

For the m-state Dirichlet density with $q(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\pi};\boldsymbol{\lambda}_q)$, $p(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\pi};\boldsymbol{\lambda}_p)$ and $\lambda_{qt} = \sum_{s=1}^{m}\lambda_q(s)$, $\lambda_{pt} = \sum_{s=1}^{m}\lambda_p(s)$ the KL-divergence is

$$\begin{aligned} KL(q \parallel p) &= \log\frac{\Gamma(\lambda_{qt})}{\Gamma(\lambda_{pt})} + \sum_{s=1}^{m}(\lambda_q(s)-\lambda_p(s))(\Psi(\lambda_q(s)) - \Psi(\lambda_{qt}) \\ &\quad + \sum_{s=1}^{m}\log\frac{\Gamma(\lambda_p(s))}{\Gamma(\lambda_q(s))}. \end{aligned} \tag{91}$$