Robust autoregression: Student-t innovations using variational Bayes

J. Christmas and R.M. Everson
Department of Computer Science
University of Exeter

***Abstract*—Autoregression (AR) is a tool commonly used to understand and predict time series data. Traditionally the excitation noise is modelled as a Gaussian. However, real-world data may not be Gaussian in nature, and it is known that Gaussian models are adversely affected by the presence of outliers. We introduce a Bayesian AR model in which the excitation noise is assumed to be Student-t distributed. Variational Bayesian approximations to the posterior distributions of the model parameters are used to overcome the intractable integrations inherent in the Bayesian model. Independent Automatic Relevance Determination (ARD) priors over each of the AR coefficients are used to estimate the model order.**

Using synthetic data we show that the Student-t model performs well against both Gaussian and leptokurtic data, in terms of parameter estimation (including the model order), and is much more robust to outliers than either Gaussian or finite mixtures of Gaussians models.

We apply the model to strongly leptokurtic EEG signals and show that the Student-t model makes more accurate one-step-ahead predictions than the Gaussian model and provides more consistent estimates of the AR coefficients over simultaneously recorded EEG channels.

***Index Terms*—Autoregressive processes, variational methods, Bayes procedures, Student-t distribution, robustness.**

I. Introduction

Autoregression is a tool commonly used to understand and predict time series data where observations taken closely in time are statistically dependent on one another. Each observation in the series is modelled as a linear combination of the previous p observations to which an element of excitation noise from a random innovations process is added. An autoregression model of order p is defined as

$$x_n = \sum_{i=1}^{p} \theta_i x_{n-i} + \epsilon_n \tag{1}$$

where x_n is the nth observation in the ordered time series data vector $\mathbf{x}$ (of length N), the θ_i are the autoregressive coefficients and ϵ_n is the excitation noise associated with this observation. Using (1) recursively to write x_n in terms of the innovations process shows that an AR model may also be viewed as a finite impulse response filter of the innovations.

Traditionally the excitation noise is presumed to be Gaussian distributed, which, due to the linearity of the AR model, means that the observations are also Gaussian distributed. However, in many real datasets observations are distributed with tails that decay more slowly than Gaussian. The presence of these observations distant from the mean adversely affects the robustness of the AR model with Gaussian excitations. Roberts and Penny [1] mitigated this problem by modelling the excitation noise with a finite mixture of Gaussians (referred to in this paper as the Gaussian Mixture Model (GMM)), thereby allowing leptokurtic distributions to be modelled. While this leads to improved performance, the tails still decay exponentially, like $\exp(-\epsilon_n^2)$, and the variance is always finite. Here we allow for very slow decay of the tails, and possibly infinite variances, by modelling the excitation process using an infinite mixture of Gaussians.

We introduce an autoregression model where the excitation noise is modelled by a Student-t distribution. With $\mathcal{S}()$ denoting the Student-t distribution, the likelihood of the noise is defined as

$$\begin{aligned} \mathrm{p}(\epsilon \,|\, 0, \lambda, d) &= \mathcal{S}(\epsilon \,|\, 0, \lambda, d) &(2)\\ &= \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \left(\frac{\lambda}{\pi d}\right)^{1/2} \left(1 + \frac{\lambda}{d}\epsilon^2\right)^{-(d+1)/2} &(3) \end{aligned}$$

where λ is known as the precision and d the degrees of freedom. As d becomes large the Student-t tends to a single Gaussian distribution. As d decreases the tails decay more slowly: when $d \le 2$ the variance is infinite, and when $d = 1$ the Student-t distribution is equivalent to the Cauchy distribution with tails that decay like ϵ^{-2}.

Denoting the Gaussian density with mean μ and variance σ^2 by $\mathcal{N}(x \,|\, \mu, \sigma^2)$ and the Gamma density by $\mathcal{G}$, defined as $\mathcal{G}(\tau \,|\, a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)$, the Student-t distribution can also be seen as an infinite mixture of Gaussians with common mean and variances scaled by the Gamma density:

$$\mathcal{S}(\epsilon \,|\, 0, \lambda, d) = \int_0^\infty \mathcal{N}(\epsilon \,|\, 0, (\lambda z)^{-1}) \mathcal{G}(z \,|\, d/2, d/2)\, dz. \tag{4}$$

Lange et al. [2] show the utility of using the Student-t distribution for robust statistical inference in a number of models, including linear and non-linear regression, and they show how the parametrisation of this distribution allows them to control the degree of downweighting of outliers to achieve more robust models. More recently Tipping and Lawrence have employed the Student-t distribution for robust Bayesian interpolation [3]; like them we use a variational approach for inference.

In order to construct a model for a particular set of data, we need to infer values for the parameters ($\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^{\mathrm{T}}$, λ and d) given the observations. Rather that just determining point values for them, we adopt a Bayesian approach so that we also obtain a measure of confidence in the inference and average over posterior distributions to reduce parameter uncertainty when making predictions.

Finding exact expressions for the estimated posterior distributions leads to intractable integrals. For Gaussian excitation sequences this problem has been tackled by Markov chain Monte Carlo sampling from the posterior distribution [4], [5], [6]. To avoid the computational expense of MCMC methods, here we use the variational Bayesian technique for finding approximations to the posterior distribution [7], [8], [9], [10]. This method minimises the Kullback-Leibler divergence between the approximate and actual posterior distributions to determine the optimal hyperparameter values for the approximations; for tutorials see [11], and [12, chapter 10].

Determining the model order p of an AR process from data can be problematic. One approach, adopted by Troughton and Godsill [13], is to integrate over all model orders which, however, requires a reversible jump MCMC sampler to accomplish the integration. Here we employ an automatic relevance determination (ARD) prior [14] over the auto-regressive coefficients θ_i, which has the effect of 'switching off' or setting to zero those coefficients for which there is no evidence in the data. In the image processing community ARD, Student-t noise and variational learning have been used for image deconvolution [15], [16]. We also draw attention to the work of Le et al. [17] who examine robust model selection for AR models by explicitly modelling the additive process that generates the outliers. They use a Bayesian approach, but with a 'robust likelihood', accomplishing inference with Laplace approximations for the integrals. In contrast, here we assume that the leptokurtic nature of the observations arises from the excitation sequence which permits us to model a range of observed distributions from Gaussian to very heavy-tailed.

We demonstrate with synthetic data that this method is able effectively to estimate values for both the hyperparameters of the posterior distributions and the order of the model and apply it to real EEG signals to demonstrate that it provides a better model than the standard autoregression with Gaussian excitation noise.

II. Bayesian autoregression

For a data set of observations written as a vector $\mathbf{x} = (x_1, \ldots, x_n)$, an alternative way of expressing the autoregression (AR) model shown in (1) is (following Ó Ruanaidh and Fitzgerald [18])

$$\mathbf{x} = \mathbf{L}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{5}$$

where $\mathbf{L}$ is the N by p matrix whose nth row contains the lags for element x_n, i.e. $(x_{n-1}, ..., x_{n-p})$. Combining this with the excitation noise distribution (2) allows the likelihood of the data to be written as

$$\mathrm{p}(\mathbf{x} \,|\, \boldsymbol{\theta}, \lambda, d) = \mathcal{S}(\mathbf{x} \,|\, \mathbf{L}\boldsymbol{\theta}, \lambda, d). \tag{6}$$

Using (4) this may be written as:

$$\mathrm{p}(\mathbf{x} \,|\, \boldsymbol{\theta}, \lambda, \mathbf{z}) = \mathcal{N}(\mathbf{x} \,|\, \mathbf{L}\boldsymbol{\theta}, (\lambda\, \mathrm{diag}(\mathbf{z}))^{-1}) \tag{7}$$

$$\mathrm{p}(z_n \,|\, d) = \mathcal{G}(z_n \,|\, d/2, d/2) \tag{8}$$

where the z_n are latent variables modifying the precision of the Gaussian mixture for each observation and $\mathrm{diag}(\mathbf{z})$ is the diagonal matrix with the z_n arranged along the diagonal.

The variational Bayesian methodology which we employ below allows the model order p to be estimated. This is, however, computationally expensive because in essence a solution has to be located for each feasible model order. Instead we seek a sparse solution in which only AR coefficients θ_i for which there is support in the data are non-zero. This is accomplished by placing an ARD prior [14] over each of the θ_i:

$$\mathrm{p}(\boldsymbol{\theta}) = \prod_{i=1}^{p} \mathcal{N}(\theta_i \,|\, 0, \delta_i) = \mathcal{N}(\boldsymbol{\theta} \,|\, \mathbf{0}, \mathrm{diag}(\boldsymbol{\delta})). \tag{9}$$

The precisions δ_i thus control the magnitude of the AR coefficients, so that if δ_i is large θ_i is effectively 'switched off'. Rather than learn point estimates for the δ_i in a type-II maximum likelihood scheme (e.g. [19]), we place a common Gamma prior over the precisions:

$$\mathrm{p}(\delta_i) = \mathcal{G}(\delta_i \,|\, a_\delta, b_\delta). \tag{10}$$

With this choice the effective prior on θ_i is seen to be a scale mixture of Gaussian; in particular when $a_\delta = b_\delta$ the effective prior is a Student-t density (cf. (4)). Tipping [19] presents a nice graphical illustration that the joint distribution $\mathrm{p}(\theta_1, \theta_2)$ of two Student-t densities concentrates probability mass close to zero values of θ_1 and θ_2 rather than in regions where both θ_1 and θ_2 are non-zero, thus encouraging sparse solutions.

We specify a Gamma prior for the precision λ:

$$\mathrm{p}(\lambda) = \mathcal{G}(\lambda \,|\, a_\lambda, b_\lambda). \tag{11}$$

Finally, specification of the model is completed by assigning a Gamma prior with hyperparameters a_d and b_d to the degrees of freedom d:

$$\mathrm{p}(d) = \mathcal{G}(d \,|\, a_d, b_d). \tag{12}$$

Figure 1 summarises the Bayesian AR model and the interdependencies between the model parameters. The joint probability of the data, latent variables, $\mathbf{z}$, and parameters $\Omega = \{\boldsymbol{\theta}, \lambda, d, \boldsymbol{\delta}\}$ may be factorised as

$$\begin{aligned} \mathrm{p}(\mathbf{x}, \mathbf{z}, \Omega) &= \mathrm{p}(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \lambda, d, \boldsymbol{\delta}) && (13) \\ &= \mathrm{p}(\mathbf{x} \,|\, \boldsymbol{\theta}, \lambda, \mathbf{z})\, \mathrm{p}(\boldsymbol{\theta} \,|\, \boldsymbol{\delta})\, \mathrm{p}(\lambda)\, \mathrm{p}(\mathbf{z} \,|\, d)\, \mathrm{p}(d)\, \mathrm{p}(\boldsymbol{\delta}). && (14) \end{aligned}$$

The structure of this model does not permit exact expressions for the posterior $\mathrm{p}(\Omega \,|\, \mathbf{x})$ to be found. Rather than resort to MCMC methods, which can be computationally expensive, we approximate the posterior using the variational Bayes method, which we now briefly describe.

III. Variational Bayes

A number of techniques are available for determining the posterior distribution in Bayesian inference. The chief obstacle is the integration required to find the normalising factor $\mathrm{p}(\mathbf{x}) = \int \mathrm{p}(\mathbf{x} \,|\, \Omega)\, \mathrm{p}(\Omega)\, d\Omega$ which appears in the denominator of Bayes' rule. Variational Bayes [7], [8], [9], [10], [11], however, approximates the posterior density $\mathrm{p}(\Omega \,|\, \mathbf{x})$ by a

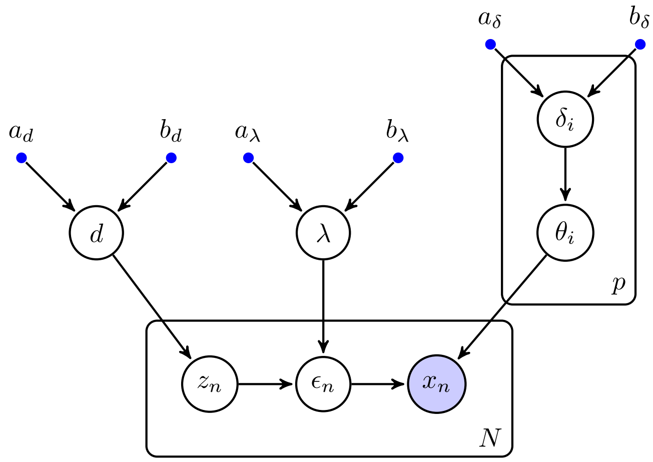


Figure 1: Graphical model showing model parameters and their interdependencies.

factorisation over groups of parameters Ω_i which are assumed to be independent when conditioned on $\mathbf{x}$; thus:

$$\mathrm{q}(\Omega\,|\,\mathbf{x}) = \prod_{i=1}^{G} \mathrm{q}_i(\Omega_i\,|\,\mathbf{x}) \tag{15}$$

The latent variables z_n are treated as additional parameters, which are random variables, so for notational simplicity we absorb them into $\Omega = \{\boldsymbol{\theta}, \mathbf{z}, \lambda, d, \boldsymbol{\delta}\}$ and approximate the posterior as:

$$\begin{aligned}
\mathrm{q}(\Omega\,|\,\mathbf{x}) &= \mathrm{q}(\boldsymbol{\theta}, \lambda, \mathbf{z}, d, \boldsymbol{\delta}\,|\,\mathbf{x}) &(16)\\
&= \mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x})\,\mathrm{q}(\lambda\,|\,\mathbf{x})\,\mathrm{q}(\mathbf{z}\,|\,\mathbf{x})\,\mathrm{q}(d\,|\,\mathbf{x})\,\mathrm{q}(\boldsymbol{\delta}\,|\,\mathbf{x})\\
&= \mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x})\,\mathrm{q}(\lambda\,|\,\mathbf{x}) \left[\prod_{n=1}^{N} \mathrm{q}(z_n\,|\,\mathbf{x})\right] \mathrm{q}(d\,|\,\mathbf{x})\,\mathrm{q}(\boldsymbol{\delta}\,|\,\mathbf{x}). &(17)
\end{aligned}$$

The log marginal probability of $\mathbf{x}$ may be written as

$$\begin{aligned}
\log(\mathrm{p}(\mathbf{x})) &= \overbrace{\int \mathrm{q}(\Omega\,|\,\mathbf{x}) \log\left(\frac{\mathrm{p}(\mathbf{x}, \Omega)}{\mathrm{q}(\Omega\,|\,\mathbf{x})}\right) d\Omega}^{\text{negative variational free energy}}\\
&\quad + \overbrace{\int \mathrm{q}(\Omega\,|\,\mathbf{x}) \log\left(\frac{\mathrm{q}(\Omega\,|\,\mathbf{x})}{\mathrm{p}(\mathbf{x}\,|\,\Omega)}\right) d\Omega}^{\text{KL divergence}} &(18)\\
&= \mathcal{F}(\mathrm{q}) + KL(\mathrm{q}(\Omega\,|\,\mathbf{x}) \parallel \mathrm{p}(\Omega\,|\,\mathbf{x})). &(19)
\end{aligned}$$

As indicated, the log marginal probability may be recognised as the sum of the Kullback-Leibler (KL) divergence between the approximate posterior and the true posterior, and the negative variational free energy. Since the KL divergence is non-negative (and zero if and only if $\mathrm{q}(\Omega\,|\,\mathbf{x})$ equals $\mathrm{p}(\Omega\,|\,\mathbf{x})$) the negative free energy is a lower bound on the log marginal probability and maximising $\mathcal{F}(\mathrm{q})$ by adjusting the approximate posterior $\mathrm{q}(\Omega\,|\,\mathbf{x})$ necessarily minimises $KL(\mathrm{q} \parallel \mathrm{p})$ so that q better approximates the posterior.

Attias [10] (see also [20], [21]) exploits the factorisation of the posterior (15) to find a general expression for the maximiser of the negative free energy in a mean-field sense. We seek to maximise the negative variational free energy, $\mathcal{F}(\mathrm{q}(\Omega\,|\,\mathbf{x}))$, with respect to all the $\mathrm{q}_i(\Omega_i\,|\,\mathbf{x})$. For readability Q_i represents $\mathrm{q}_i(\Omega_i\,|\,\mathbf{x})$:

$$\begin{aligned}
\mathcal{F}(\mathrm{q}) &= \int Q \log(\frac{\mathrm{p}(\mathbf{x}, \Omega)}{Q})\, d\Omega &(20)\\
&= \int \left(\prod_{i=1}^{G} Q_i\right) \log(\mathrm{p}(\mathbf{x}, \Omega))\, d\Omega_1, \ldots, d\Omega_G\\
&\quad - \int \left(\prod_{i=1}^{G} Q_i\right) \left(\sum_{i=1}^{G} \log(Q_i)\right)\, d\Omega_1, \ldots, d\Omega_G &(21)
\end{aligned}$$

Considering the integral with respect to Ω_j and keeping the remaining $Q_{i\neq j}$ fixed, the negative free energy can be written as

$$\begin{aligned}
\mathcal{F}(\mathrm{q}) &= \int Q_j \overbrace{\left[\int \log(\mathrm{p}(\mathbf{x}, \Omega)) \prod_{i\neq j} Q_i d\Omega_{i\neq j}\right]}^{(a)} d\Omega_j\\
&\quad - \int Q_j \log(Q_j)\, d\Omega_j + \mathit{const} &(22)
\end{aligned}$$

where terms that do not depend upon Q_j have been absorbed into the constant. The section of this expression marked (a) is the expectation of $\log(\mathrm{p}(\mathbf{x}, \Omega))$ with respect to each of the Q_j, where $i \neq j$. We denote this $\mathbb{E}_{i\neq j}[\log(\mathrm{p}(\mathbf{x}, \Omega))]$, and it may be recognised as the negative KL divergence between Q_j and $\mathbb{E}_{i\neq j}[\log(\mathrm{p}(\mathbf{x}, \Omega))]$; hence the maximum value is zero, which is obtained when

$$\log(Q_j) = \mathbb{E}_{i\neq j}[\log(\mathrm{p}(\mathbf{x}, \Omega))]. \tag{23}$$

If conjugate priors are chosen for each group, the approximate posterior turns out to have the same functional form as the prior [10], [22] and the variational approximations may thus be found by evaluating (23) for each group in turn. Of course, the hyperparameters of the posterior distribution for one group will generally depend upon the hyperparameters for other groups; consequently the parameters for each group are evaluated cyclically until convergence. Ghahramani and Beal [22] show that this scheme converges to a local maximum of $\mathcal{F}$, thus minimising $KL(\mathrm{q} \parallel \mathrm{p})$.

IV. Variational Bayesian autoregression

Here we use the factorised variational Bayes method to obtain approximate posterior distributions for the factorisation (17) using the joint probability (14). We consider each group in turn.

A. AR coefficients, $\boldsymbol{\theta}$

The approximate posterior for the AR coefficient that maximises $\mathcal{F}(\mathrm{q})$ is maximised when

$$\begin{aligned}
&\log(\mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x}))\\
&= \mathbb{E}_{/\boldsymbol{\theta}}[\log(\mathrm{p}(\mathbf{x}\,|\,\boldsymbol{\theta}, \lambda, \mathbf{z})\,\mathrm{p}(\boldsymbol{\theta}\,|\,\boldsymbol{\delta})\,\mathrm{p}(\lambda)\,\mathrm{p}(\mathbf{z}\,|\,d)\,\mathrm{p}(d)\,\mathrm{p}(\boldsymbol{\delta}))] &(24)
\end{aligned}$$

where $\mathbb{E}_{/a}[b]$ is the expectation of b taken with respect to the approximate posteriors of all variables except a. Expanding this and moving all terms not dependent on $\boldsymbol{\theta}$ into a single constant term we get

$$\begin{aligned}\log(\mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x})) &= \mathbb{E}_{/\boldsymbol{\theta}}[\log(\mathrm{p}(\mathbf{x}\,|\,\boldsymbol{\theta},\lambda,\mathbf{z}) + \log(\mathrm{p}(\boldsymbol{\theta}\,|\,\boldsymbol{\delta})] + const \qquad (25)\\ &= \mathbb{E}_{/\boldsymbol{\theta}}[\log(\mathcal{N}(\mathbf{x}\,|\,\mathbf{L}\boldsymbol{\theta},(\lambda\,\mathrm{diag}(\mathbf{z}))^{-1})\\ &\quad + \log(\mathcal{N}(\boldsymbol{\theta}\,|\,\mathbf{0},\mathrm{diag}(\boldsymbol{\delta})^{-1})] + const. \qquad (26)\end{aligned}$$

Again, expanding this and moving all terms not dependent on $\boldsymbol{\theta}$ into the constant term:

$$\begin{aligned}\log(\mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x})) &= \mathbb{E}_{/\boldsymbol{\theta}}[-\frac{\lambda}{2}(\mathbf{x}-\mathbf{L}\boldsymbol{\theta})^{\mathrm{T}}\,\mathrm{diag}(\mathbf{z})(\mathbf{x}-\mathbf{L}\boldsymbol{\theta})\\ &\quad - \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\,\mathrm{diag}(\boldsymbol{\delta})\boldsymbol{\theta}] + const \qquad (27)\\ &= -\frac{1}{2}\mathbb{E}_{\lambda}[\lambda](\mathbf{x}-\mathbf{L}\boldsymbol{\theta})^{\mathrm{T}}\,\mathrm{diag}(\mathbb{E}_{z}[\mathbf{z}])(\mathbf{x}-\mathbf{L}\boldsymbol{\theta})\\ &\quad - \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\,\mathrm{diag}(\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}])\boldsymbol{\theta} + const. \qquad (28)\end{aligned}$$

Since (28) is quadratic in $\boldsymbol{\theta}$, it can be seen that $\mathrm{q}(\boldsymbol{\theta}\,|\,\mathbf{x})$ is a Gaussian and we have

$$\mathrm{q}(\boldsymbol{\theta}\,|\,\mathrm{x}) = \mathcal{N}(\boldsymbol{\theta}\,|\,\boldsymbol{\mu}_{\theta},\boldsymbol{\Sigma}_{\theta}) \qquad (29)$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} &= \mathbb{E}_{\lambda}[\lambda]\mathbf{L}^{\mathrm{T}}\,\mathrm{diag}(\mathbb{E}_{\mathbf{z}}[\mathbf{z}])\mathbf{L} + \mathrm{diag}(\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}]) \qquad (30)\\ \boldsymbol{\mu}_{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbb{E}_{\lambda}[\lambda]\mathbf{L}^{\mathrm{T}}\,\mathrm{diag}(\mathbb{E}_{\mathbf{z}}[\mathbf{z}])\mathbf{x}. \qquad (31)\end{aligned}$$

B. Excitation noise precision, λ

Applying the same procedure for λ we obtain a Gamma distribution for the posterior noise precision:

$$\mathrm{q}(\lambda\,|\,\mathbf{x}) = \mathcal{G}(\lambda\,|\,\alpha_{\lambda},\beta_{\lambda}) \qquad (32)$$

where

$$\begin{aligned}\alpha_{\lambda} &= a_{\lambda} + \frac{N}{2} \qquad (33)\\ \beta_{\lambda} &= b_{\lambda} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\,\mathrm{diag}(\mathbb{E}_{\mathbf{z}}[\mathbf{z}])\,(\mathbf{x}-2\mathbf{L}\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}])\\ &\quad + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{z_n}[z_n]\mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{L}_n\boldsymbol{\theta})]. \qquad (34)\end{aligned}$$

C. Latent variables, $\mathbf{z}$

Locating a joint distribution for $\mathbf{z}$ is not analytically tractable, so we examine the z_n individually, obtaining:

$$\begin{aligned}\log(\mathrm{q}(z_n\,|\,\mathbf{x})) &= \frac{1}{2}\log(z_n)\\ &\quad - \frac{1}{2}\mathbb{E}_{\lambda}[\lambda]\mathbb{E}_{\boldsymbol{\theta}}[(x_n-\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}z_n(x_n-\mathbf{L}_n\boldsymbol{\theta})]\\ &\quad + \mathbb{E}_{d}[\log(\mathcal{G}(\mathbf{z}\,|\,d/2,d/2))] \qquad (35)\\ &= \left(\frac{\mathbb{E}_d[d]+1}{2}-1\right)\log(z_n)\\ &\quad - \left(\frac{1}{2}\mathbb{E}_{\lambda}[\lambda]\mathbb{E}_{\boldsymbol{\theta}}[(x_n-\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(x_n-\mathbf{L}_n\boldsymbol{\theta})]\right.\\ &\quad \left. + \frac{\mathbb{E}_d[d]}{2}\right)z_n \qquad (36)\end{aligned}$$

where $\mathbf{L}_n$ is the nth row of $\mathbf{L}$. On inspection we see that $\mathrm{q}(z_n\,|\,\mathbf{x})$ is a Gamma distribution:

$$\mathrm{q}(z_n\,|\,\mathbf{x}) = \mathcal{G}(z_n\,|\,\alpha_{z_n},\beta_{z_n}) \qquad (37)$$

where

$$\begin{aligned}\alpha_z &= \frac{\mathbb{E}_d[d]+1}{2} \qquad (38)\\ \beta_{z_n} &= \frac{\mathbb{E}_d[d]}{2} + \frac{1}{2}\mathbb{E}_{\lambda}[\lambda]\mathbb{E}_{\boldsymbol{\theta}}[(x_n-\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(x_n-\mathbf{L}_n\boldsymbol{\theta})]. \qquad (39)\end{aligned}$$

As the expected value of d becomes large, so that the Student-t distribution describing the excitation noise approaches a Gaussian, the posterior expected value of z_n (i.e. α_{z_n}/β_{z_n}) tends to 1 and likelihood of $\mathbf{x}$ tends towards the Gaussian $\mathcal{N}(\mathbf{x}\,|\,\mathbf{L}\boldsymbol{\theta},\lambda^{-1})$.

D. Degrees of freedom, d

For $\mathrm{q}(d\,|\,\mathbf{x})$ (again dropping the constant term) we obtain:

$$\begin{aligned}\log(\mathrm{q}(d\,|\,\mathbf{x})) &= -N\log(\Gamma(\frac{d}{2})) + N\frac{d}{2}\log(\frac{d}{2})\\ &\quad + \frac{d}{2}\sum_{n=1}^{N}(\mathbb{E}_{z_n}[\log(z_n)] - \mathbb{E}_{z_n}[z_n])\\ &\quad + (a_d-1)\log(d) - b_d d \qquad (40)\end{aligned}$$

which does not correspond to any standard distribution. But using Stirling's approximation for $\log(\Gamma(d/2))$, shown in the square brackets below, we get:

$$\begin{aligned}\log(\mathrm{q}(d\,|\,\mathbf{x})) &= -N\left[-(\frac{d}{2}-\frac{1}{2})\log(\frac{d}{2}) + \frac{d}{2}\right] + N\frac{d}{2}\log(\frac{d}{2})\\ &\quad + \frac{1}{2}\sum_{n=1}^{N}(\mathbb{E}_{z_n}[\log(z_n)] - \mathbb{E}_{z_n}[z_n])\\ &\quad + (a_d-1)\log(d) - b_d d \qquad (41)\\ &= (a_d + \frac{N}{2} - 1)\log(d)\\ &\quad - \left(b_d - \frac{N}{2} - \frac{1}{2}\sum_{n=1}^{N}(\mathbb{E}_{z_n}[\log(z_n)] - \mathbb{E}_{z_n}[z_n])\right)d \qquad (42)\end{aligned}$$

which may be recognised as the log of a Gamma distribution. Therefore

$$\mathrm{q}(d\,|\,\mathbf{x}) = \mathcal{G}(d\,|\,\alpha_d,\beta_d) \qquad (43)$$

where

$$\begin{aligned}\alpha_d &= a_d + \frac{N}{2} \qquad (44)\\ \beta_d &= b_d - \frac{1}{2}\left(N + \sum_{n=1}^{N}(\mathbb{E}_{z_n}[\log(z_n)] - \mathbb{E}_{z_n}[z_n])\right). \qquad (45)\end{aligned}$$

E. ARD precisions, $\boldsymbol{\delta}$

Finally, the posterior distributions for the ARD precisions are found as:

$$\mathrm{q}(\delta_i\,|\,\mathbf{x}) = \mathcal{G}(\delta_i\,|\,\alpha_{\delta},\beta_{\delta_i}) \qquad (46)$$

where

$$\alpha_\delta = a_\delta + 1 \tag{47}$$

$$\beta_{\delta_i} = b_\delta + \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}[\theta_i^2]. \tag{48}$$

F. Summary

Summarising the results from (24)-(45) and writing expectations $\mathbb{E}_a[f(a)]$ as $\langle f(a) \rangle$ for readability:

$$\mathrm{q}(\boldsymbol{\theta} \,|\, \mathbf{x}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \tag{49}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = (\langle\lambda\rangle \mathbf{L}^{\mathrm{T}} \operatorname{diag}(\langle\mathbf{z}\rangle)\mathbf{L} + \operatorname{diag}(\langle\boldsymbol{\delta}\rangle))^{-1} \tag{50}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \langle\lambda\rangle \mathbf{L}^{\mathrm{T}} \operatorname{diag}(\langle\mathbf{z}\rangle)\mathbf{x} \tag{51}$$

$$\mathrm{q}(\lambda \,|\, \mathbf{x}) = \mathcal{G}(\lambda \,|\, \alpha_\lambda, \beta_\lambda) \tag{52}$$

$$\alpha_\lambda = a_\lambda + \frac{N}{2} \tag{53}$$

$$\begin{aligned}\beta_\lambda = b_\lambda &+ \frac{1}{2}\mathbf{x}^{\mathrm{T}} \operatorname{diag}(\langle\mathbf{z}\rangle)\,(\mathbf{x} - 2\mathbf{L}\langle\boldsymbol{\theta}\rangle) \\ &+ \frac{1}{2}\sum_{n=1}^{N} \langle\mathbf{z}_n\rangle \langle(\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{L}_n\boldsymbol{\theta})\rangle\end{aligned} \tag{54}$$

$$\mathrm{q}(z_n \,|\, \mathbf{x}) = \mathcal{G}(z_n \,|\, \alpha_z, \beta_{z_n}) \tag{55}$$

$$\alpha_z = \frac{\langle d\rangle + 1}{2} \tag{56}$$

$$\beta_{z_n} = \frac{\langle d\rangle}{2} + \frac{1}{2}\langle\lambda\rangle\langle(x_n - \mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(x_n - \mathbf{L}_n\boldsymbol{\theta})\rangle \tag{57}$$

$$\mathrm{q}(d \,|\, \mathbf{x}) = \mathcal{G}(d \,|\, \alpha_d, \beta_d) \tag{58}$$

$$\alpha_d = a_d + \frac{N}{2} \tag{59}$$

$$\beta_d = b_d - \frac{1}{2}\left(N + \sum_{n=1}^{N}\left[\langle\log(z_n)\rangle - \langle z_n\rangle\right]\right) \tag{60}$$

$$\mathrm{q}(\delta_i \,|\, \mathbf{x}) = \mathcal{G}(\delta_i \,|\, \alpha_\delta, \beta_{\delta_i}) \tag{61}$$

$$\alpha_\delta = a_\delta + 1 \tag{62}$$

$$\beta_{\delta_i} = b_\delta + \frac{1}{2}\langle\theta_i^2\rangle \tag{63}$$

As noted previously, each approximate posterior distribution is dependent on the expected values of one or more of the others, so closed-form algebraic solution cannot be obtained. However, we can arrive at a set of solutions by initialising the required expectations (perhaps based on the priors) and then iteratively updating the estimate for each hyperparameter based on the current estimates of the values on which it depends, until convergence. The required current estimates are obtained using the standard expressions

$$\langle\boldsymbol{\theta}\rangle = \boldsymbol{\mu}_\theta \tag{64}$$

$$\langle\lambda\rangle = \alpha_\lambda / \beta_\lambda \tag{65}$$

$$\langle z_n\rangle = \alpha_z / \beta_{z_n} \tag{66}$$

$$\langle\log(z_n)\rangle = \psi(\alpha_z) - \log(\beta_{z_n}) \tag{67}$$

$$\langle d\rangle = \alpha_d / \beta_d \tag{68}$$

$$\langle\delta_i\rangle = \alpha_\delta / \beta_{\delta_i} \tag{69}$$

where $\psi(\cdot)$ is the digamma function, together with the following expansions:

$$\langle(\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{L}_n\boldsymbol{\theta})\rangle = \operatorname{Tr}(\mathbf{L}_n\langle\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\rangle\mathbf{L}_n^{\mathrm{T}}) + (\mathbf{L}_n\langle\boldsymbol{\theta}\rangle)^{\mathrm{T}}(\mathbf{L}_n\langle\boldsymbol{\theta}\rangle) \tag{70}$$

$$\langle(x_n - \mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(x_n - \mathbf{L}_n\boldsymbol{\theta})\rangle = x_n^2 - 2x_n\mathbf{L}_n\langle\boldsymbol{\theta}\rangle + \langle(\mathbf{L}_n\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{L}_n\boldsymbol{\theta})\rangle \tag{71}$$

where $\operatorname{Tr}(\cdot)$ denotes the trace operator.

The equivalent variational posteriors calculated for a *Gaussian* AR model are as follows:

$$\mathrm{q}(\boldsymbol{\theta} \,|\, \mathbf{x}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \tag{72}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = (\langle\lambda\rangle\mathbf{L}^{\mathrm{T}}\mathbf{L} + \operatorname{diag}(\langle\boldsymbol{\delta}\rangle))^{-1} \tag{73}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\langle\lambda\rangle\mathbf{L}^{\mathrm{T}}\mathbf{x} \tag{74}$$

$$\mathrm{q}(\lambda \,|\, \mathbf{x}) = \mathcal{G}(\lambda \,|\, \alpha_\lambda, \beta_\lambda) \tag{75}$$

$$\alpha_\lambda = a_\lambda + \frac{N}{2} \tag{76}$$

$$\beta_\lambda = b_\lambda + \frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}\left(\mathbf{x} - 2\mathbf{L}\langle\boldsymbol{\theta}\rangle\right) + \langle(\mathbf{L}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{L}\boldsymbol{\theta})\rangle\right) \tag{77}$$

$$\mathrm{q}(\delta_i \,|\, \mathbf{x}) = \mathcal{G}(\delta_i \,|\, \alpha_\delta, \beta_{\delta_i}) \tag{78}$$

$$\alpha_\delta = a_\delta + 1 \tag{79}$$

$$\beta_{\delta_i} = b_\delta + \frac{1}{2}\langle\theta_i^2\rangle \tag{80}$$

In the Student-t AR case as $d \to \infty$ and the excitation sequence becomes effectively Gaussian it can be seen that we recover from (49) to (63) the expressions (72) to (80).

V. Illustration: synthetic data

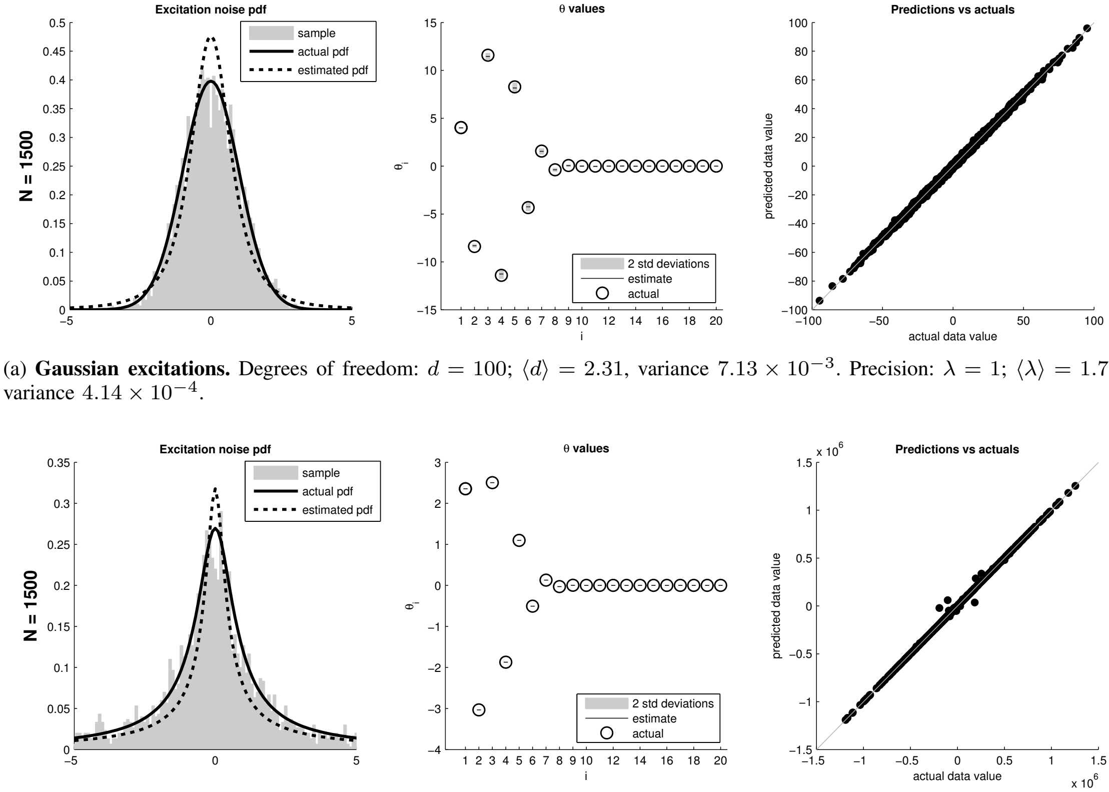
We demonstrate three different aspects of the efficacy of this Student-t AR model. Firstly, in section V-B, we show that it is able to make good estimates of the parameters of data which have been synthesised to fit the model. Secondly, in section V-C, we show that it is able to determine the correct model order. Lastly, where the excitation noise is Gaussian in nature, we show, in section V-D, that it is more robust to the addition of outliers than both the standard Gaussian AR model and the AR model with GMM excitation noise.

A. Synthesising data

We generate synthetic data with randomly chosen AR coefficients. With specified values for N, p, d, λ, we choose $\boldsymbol{\theta}$ to describe a stationary AR process as follows. If p is even, $p/2$ complex conjugate pairs $\eta_i, \bar{\eta}_i$ lying within the unit circle in the complex plane are drawn randomly (uniformly with respect to area); if p is odd, a single η_p is drawn on the real axis and $\lfloor p/2 \rfloor$ conjugate pairs are drawn for the remainder. The η_i and $\bar{\eta}_i$ are used as the roots of the auxiliary polynomial and $\boldsymbol{\theta}$ is the vector whose elements are its coefficients.

The data sequence is initialised by generating random values for the first p elements of $\mathbf{x}$. Every x_n, for n from $p+1$ to $2N$, is calculated from (1), where a random sample of excitation noise, ϵ_n, is selected from the Student-t distribution. Expression (4) indicates the way to sample from a Student-t distribution: first a value for z_n is randomly drawn from $\mathcal{G}(z_n \,|\, d/2, d/2)$, then this result is used in the random draw from the $\mathcal{N}(\epsilon_n \,|\, 0, (\lambda z_n)^{-1})$ distribution to generate one ϵ_n value. Finally, the first N elements of $\mathbf{x}$ are deleted to ensure that the whole sample conforms to the autoregression model.

Since we wish the model to be able to fit as wide a range of problems as possible, uninformative priors were selected. Each of λ, d and δ_i have Gamma priors, $\mathcal{G}(\lambda \,|\, a_\lambda, b_\lambda)$, $\mathcal{G}(d \,|\, a_d, b_d)$



(a) **Gaussian excitations.** Degrees of freedom: $d = 100$; $\langle d \rangle = 2.31$, variance 7.13×10^{-3}. Precision: $\lambda = 1$; $\langle \lambda \rangle = 1.76$, variance 4.14×10^{-4}.

(b) **Student-t excitations.** Degrees of freedom: $d = 0.5$; $\langle d \rangle = 0.25$, variance 8.45×10^{-5}. Precision: $\lambda = 1$; $\langle \lambda \rangle = 2.18$, variance 6.31×10^{-3}.

Figure 2: Gaussian and Student-t examples. Left: Comparison of the estimated posterior parameter distributions with those used to generate the observations. Centre: Estimated AR coefficients and those used to generate the observations. Right: Expected values of one-step-ahead predictions compared with the observations. $N = 1500$ and actual $p = 10$. Estimated values and variances are shown to 2 decimal places.

and $\mathcal{G}(\delta \,|\, a_\delta, b_\delta)$ respectively, where we choose $a_\lambda = b_\lambda = a_d = b_d = a_\delta = b_\delta = 10^{-3}$.

B. Parameter estimation

Figure 2 shows, for a dataset with $N = 1500$ and Gaussian excitation noise, a comparison of the expected values of the variational posterior parameter distributions with the actual values used to generate the samples for the dataset. The bottom row of figure 2 shows results for observations generated with a highly non-Gaussian $d = 0.5$ excitation sequence. In both cases is it is clear that the model accurately learns the coefficients and makes accurate predictions despite the vastly different natures of the excitation sequences. The actual $\boldsymbol{\theta}$ vectors used to generate the data were of length 10 (i.e. actual $p = 10$) but the model was trained with $p = 20$ to demonstrate the effect of ARD. This is clearly seen in the centre graphs where the θ_i values where $i > p$ have been “switched off”. The results in both examples are similar in that the variational posterior noise distribution is more compact than the actual, a tendency reported by a number of authors (for example [23], [24], [25]), the estimated $\boldsymbol{\theta}$ values are similar to the actuals, with a tendency to be underestimated, and the reconstructions of the data are good.

The over-compactness of the noise distribution and the underestimation of the $\boldsymbol{\theta}$ values warrants further investigation. To this end the model was trained against datasets which were created for every combination of λ and d between 0.01 to 10 in steps of 0.01, with $N = 1500$, $p = 10$ and a different, randomly-generated $\boldsymbol{\theta}$. The results (in figure 3) show that for λ there is an approximately linear relationship between predicted and actual, with the predicted value consistently over-estimated (and hence the variance is underestimated), while for d the results are non-linear and consistently underestimated. It is this combination of underestimated variance and degrees of freedom that lead to the more compact distribution compared with the actual. The predicted and actual $\boldsymbol{\theta}$ values are highly correlated, but the plot appears twisted clockwise with respect to the diagonal, indicating that the magnitudes of the coefficients are slightly underestimated, particularly for the smaller coefficients.

With a relatively small N it is highly unlikely that the sample will be truly representative of the distribution from which it is taken. This is particularly true of the Student-t distribution when the degrees of freedom are such that the variance is infinite (i.e. $d \leq 2$). This does not mean that the model is less able to represent the data, but it does mean that the estimated parameter values are less likely to reflect the

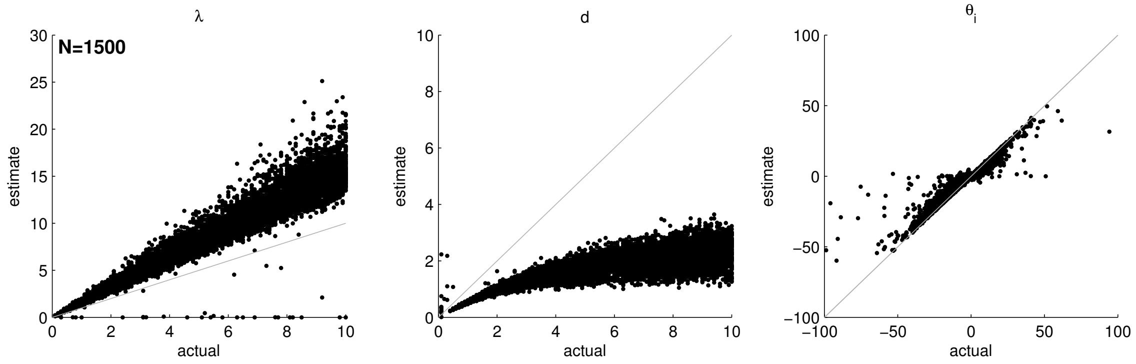


Figure 3: Comparison of estimate against actual λ (left), d (centre) and θ_i (right) for every combination of λ and d between 0.01 and 10 in steps of 0.01, with $N = 1500$, $p = 10$ and a different $\boldsymbol{\theta}$ each time. The solid gray lines mark estimate=actual.

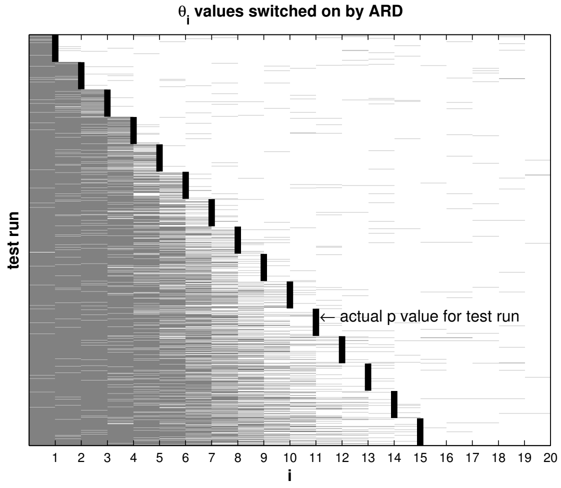


Figure 4: The model was trained against datasets generated for all λ and d values in the integer range 1 to 10 and every p from 1 to 15; 1500 datasets in all. Grey lines show which elements of $\boldsymbol{\theta}$ were switched on in each test. Heavy black lines show the actual p value for each test.

actuals.

C. Model selection

The model selection effects of ARD have been hinted at in the examples shown in section V-B. We demonstrate this now in more detail by training the model against synthetic datasets of 1500 observations each for every combination of p from 1 to 15, and every λ and d in the integer range 1 to 10. The prior for $\boldsymbol{\theta}$ is $\mathcal{N}(\boldsymbol{\theta} \,|\, \mathbf{0}, \mathrm{diag}(\boldsymbol{\delta})^{-1})$; if the estimated value of one of the θ_i is more than one standard deviation away from zero, i.e. $\theta_i^2 > 1/\delta_i$, then we deem it to be "switched on". Figure 4 shows, in grey, which θ_i are switched on in each of the 1500 test runs. The solid black lines indicate the actual model order in each case.

While the lower model orders are well estimated, the higher ones appear to be consistently underestimated. However, this is a consequence of the way in which the AR coefficients are generated: the scheme described in section V-A tends to produce θ_i which decay in magnitude with increasing i. This is illustrated in figure 5, which compares the actual and estimated $\boldsymbol{\theta}$ values for all the synthetic datasets where $p = 10$. The ARD mechanism is clearly suppressing θ_i when $i > p$, but, in addition, there is often insufficient support in the observations for the small θ_i (with $8 \lessapprox i \lessapprox 10$) so that they appear to be erroneously 'switched off' and the model order underestimated.

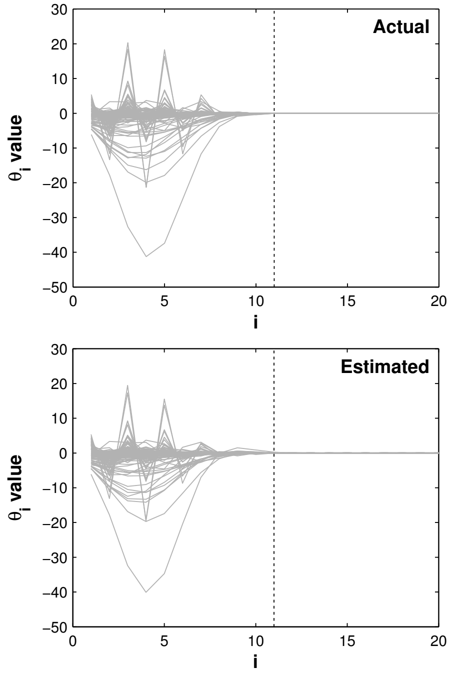


Figure 5: Plot of $\boldsymbol{\theta}$ for 100 synthetic datasets with $p = 10$. Estimated values where $i > p$ are constrained, by ARD, to be close to zero. Actual values where i is close to p tend to be small, so may be considered to be switched off.

D. Robustness to outliers

A useful characteristic of the Student-t distribution is its ability to handle outlying observations in a principled way. Here we illustrate the resilience to outliers of the Student-t

AR model, a resilience which is not shared by the (Bayesian) Gaussian AR model, which is closely related to the traditional AR model whose parameters are estimated by least-squares fitting.

Using a similar method as described previously, a synthetic dataset of 500 observations was generated with Gaussian excitation noise ($\lambda = 10$) and $p = 10$. Student-t, Gaussian and GMM AR models (the latter as per [1], with a mixture of 5 Gaussians) were each trained against it, and, as figure 6a shows, all three models accurately make one-step-ahead reconstructions of the data.

Three outliers were then added to the dataset, each as positive values (i.e. in the same direction), with values of 10 times the maximum size of the remaining observations. The Student-t, Gaussian and GMM AR models were each trained against this amended set; one-step-ahead predictions are compared with actual values in figure 6b. It is not surprising that the p values immediately following each outlier are poorly predicted by the models, so these are omitted from the plots in figure 6b. By comparing these with the corresponding graphs in figure 6a it is clear that the Gaussian and GMM models have been significantly affected by the presence of the outliers, while the Student-t model is robust to them. It is noticeable that the Gaussian and GMM AR predictions are worse than before (the points are spread away from the diagonal), the estimated mean has moved away from the actual and the noise variance has been underestimated (the plot appears twisted clockwise with respect to the diagonal). For the Student-t model the predictions do not appear to have deteriorated and the estimated mean has moved only slightly away from the true mean.

Outliers are observations which lie further from the true mean than would be likely given the true distribution. The Gaussian AR model is forced to accomodate them within the single Gaussian distribution it fits to the excitation noise. This causes the mean of the estimated distribution to move away from the actual and/or the variance to be overestimated; both of these effects are demonstrated here. In contrast, while all of the distributions in the Student-t mixture of Gaussians have the same mean, their range of variances allows the overall distribution to accomodate the outliers.

VI. Results: real data

EEG signals are often thought of as an example of data whose noise is heavier-tailed than Gaussian. Here we examine an EEG signal comprising 1150 observations. If we regard the data as having been generated by an underlying autoregressive model with Gaussian distributed excitation noise, then we expect the observations themselves also to be Gaussian distributed. Figure 7 shows the Normal probability plot (the sample quantiles of the observations versus theoretical quantiles from a normal distribution) for the selected data. The variation from the straight line shows that these data are significantly non-Gaussian in nature.

Both the Student-t and Gaussian AR models were evaluated against this selected sensor sample of 1150 observations. Where the degrees of freedom for a Student-t distribution is less than or equal to 2 the variance is, effectively, infinite, which makes direct comparison of the confidence it has in its predictions with the Gaussian AR model impossible. Instead, for each observation, Monte Carlo sampling of 1000 predictions was used to generate an 80% credibility interval. This was repeated for the Gaussian AR results to enable direct comparison. Figure 8 shows a subset of results for 50 observations. For each observation the Student-t confidence interval is noticeably tighter and the actual observed value falls within it. In fact for all 1150 observations the actual values lie within the 80% credibility intervals of the Student-t AR model. This is not the case for the Gaussian AR model.

The Student-t model estimates the AR coefficients with low variance and a model order of approximately 12. We also find that it identifies rather similar values for all 58 EEG channels that comprise a single observation set for a subject, whereas the Gaussian model does not; this is demonstrated in figure 9. An important consequence of this is that power spectral densities calculated from the AR coefficients estimated with Student-t excitations are considerably more consistent across a subject than estimates using Gaussian excitations.

VII. Conclusions

The standard AR model is based upon an assumption of Gaussian excitation noise. We have shown that a Bayesian model based on a Student-t assumption is more robust to outliers and is able to model data whose excitation noise is Gaussian distributed or heavier-tailed than the Gaussian distribution. The new model is shown to be a generalisation of the Gaussian model.

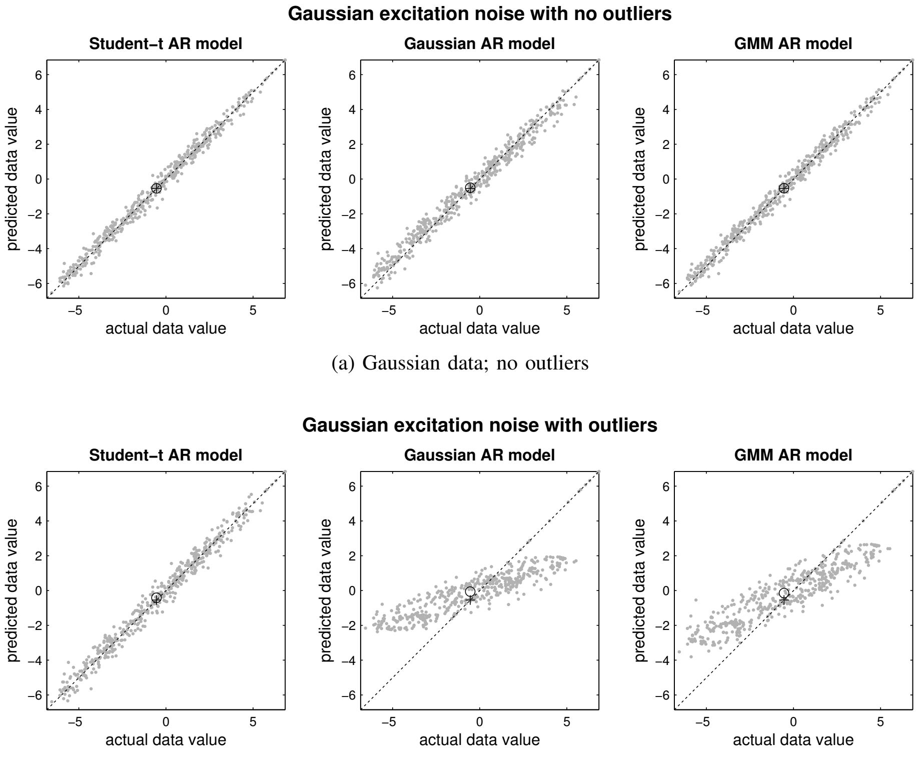
The Student-t model leads to intractable integrals in the calculation of the posterior densities for the model parameters. We have shown that the factorised variational Bayes technique provides good approximations and is computationally efficient, but it tends to underestimate the excitation noise variance, the degrees of freedom and the magnitudes of the model coefficients.

The Student-t model incorporates ARD priors over the AR coefficients and this has been shown to result in sparse solutions that accurately predict the model order.

For real EEG data that is heavier-tailed than Gaussian, we have shown that the Student-t model makes more accurate one-step-ahead predictions, with smaller variances, than the Gaussian model. The connection between the AR coefficients and the power spectrum of the observations has long been recognised [26] and exploited for the estimation of power spectra, but the variability of these estimates has been pointed out [27]. The consistency of the coefficients estimated by this Student-t AR model across EEG channels lead to much more consistent estimates of the power spectra.

VIII. Acknowledgments

Many thanks to Dr Aureliu Lavric at the School of Psychology, University of Exeter, for providing the EEG data.



(a) Gaussian data; no outliers

(b) Gaussian data; 3 outliers

Figure 6: One-step-ahead predictions plotted against actual values for (a) Gaussian data and (b) the same with the addition of 3 outliers. The black dotted diagonals indicate prediction = actual. The actual mean is marked with a black cross; the estimated mean with a black circle. The Student-t AR model (left) is largely unaffected by the outliers; the predictions are still very good and the estimated mean is very close to the actual. The Gaussian (centre) and GMM AR (right) models are noticeably less accurate than before and are overestimating the excitation variance (the plot appears twisted clockwise with respect to the diagonal). The estimated mean has moved away from the actual.

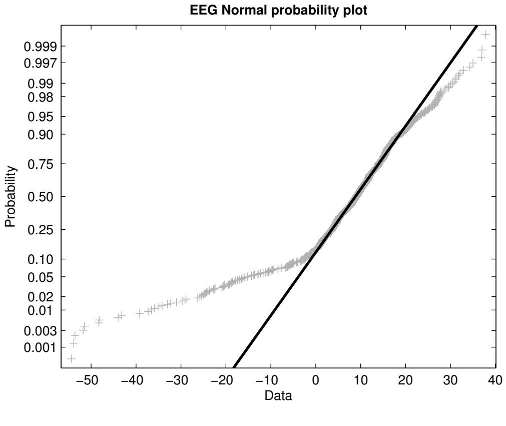


Figure 7: Normal probability plot for the selected EEG sensor sample. The variation from the black line shows that the data are significantly non-Gaussian in nature.

References

[1] S. Roberts and W. Penny, “Variational bayes for generalized autoregressive models,” *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, September 2002.

[2] K. Lange, R. Little, and J. Taylor, “Robust Statistical Modeling Using the t Distribution,” *Journal of the American Statistical Association*, vol. 84, pp. 881–896, 1989.

[3] M. Tipping and N. Lawrence, “Variational inference for student-t models: robust bayesian interpolation and generalised component analysis,” *Neurocomputing*, vol. 69, pp. 123–141, 2005.

[4] G. Barnett, R. Kohn, and S. Sheather, “Bayesian estimation of anautoregressive model using Markov chain Monte Carlo,” *Journal of Econometrics*, vol. 74, no. 2, pp. 237–254, 1996.

[5] S. Godshill, “Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes,” *International Statistical Review*, vol. 65, no. 1, pp. 1–21, 1996.

[6] S. Godshill and P. Rayner, “Statistical reconstruction and analysis of autoregressive signals in impulsive noise,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 352–372, 1998.

[7] D. Mackay, “Ensemble learning and evidence maximisation,” Cavendish Laboratory, University of Cambridge, Tech. Rep., 1995.

[8] ——, “Ensemble learning for hidden Markov models,” Cavendish Laboratory, University of Cambridge, Tech. Rep., 1997.

[9] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, “An introduction to variational methods for graphical models,” *Machine Learning*, vol. 37, no. 2, p. 183, 1999.

[10] H. Attias, “A variational Bayesian framework for graphical models,” *Advances in Neural Information Processing Systems*, vol. 12, pp. 209–215, 2000.

[11] H. Lappalainen and J. Miskin, *Advances in Independent Component Analysis*. Berlin: Springer-Verlag, 2000, ch. Ensemble Learning, pp. 75–92.

[12] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[13] P. Troughton and S. Godsill, “A reversible jump sampler for autoregressive time series,” in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 98)*, 1998.

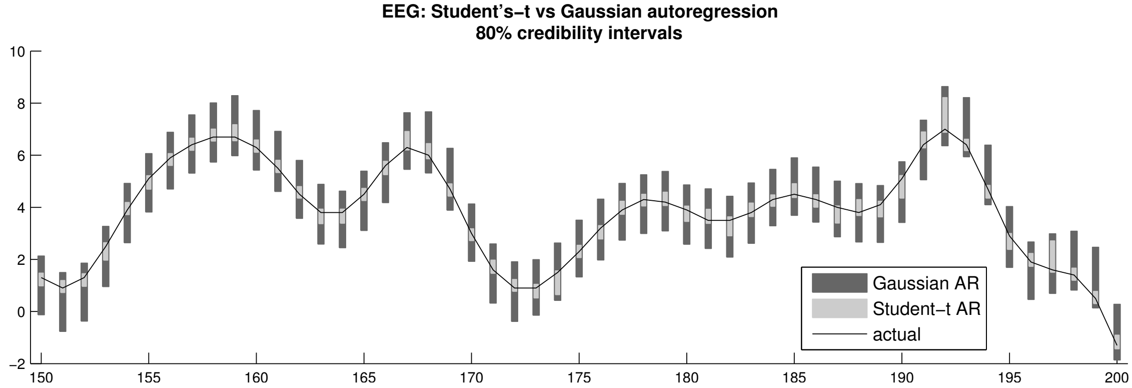


Figure 8: Comparison of 80% credibility intervals for the reconstructions of a 50-point extract of the EEG data using an AR model with Student-t excitations (light shading) and one with Gaussian excitations (dark). The line indicates the observed signal. Credibility intervals are symmetric about the mean predictions, which for clarity are not shown. The Student-t AR model shows higher confidence than the Gaussian.

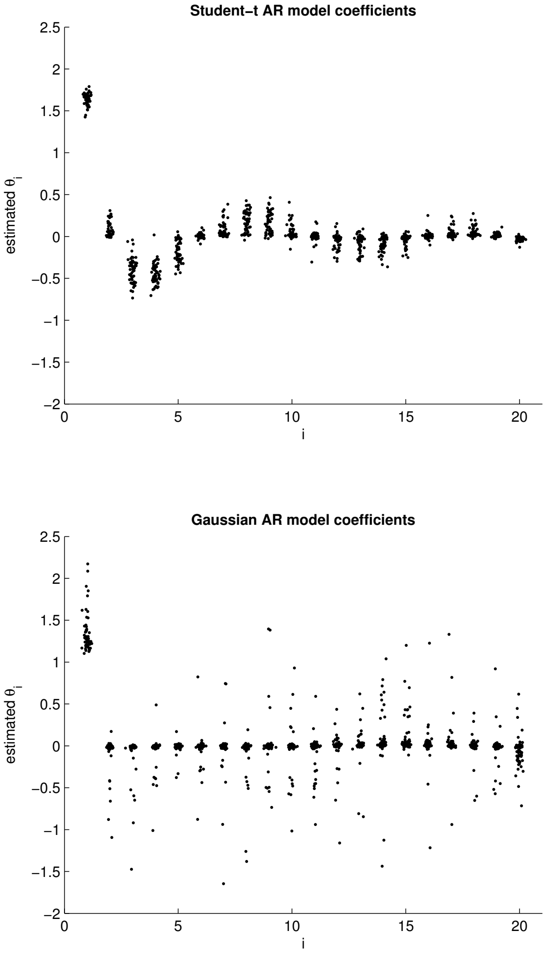


Figure 9: AR coefficients ($\boldsymbol{\theta}$) calculated for all 58 EEG channels in a single observation set using (top) the Student-t model and (bottom) the Gaussian model. Each point has been offset laterally by a small random amount to make the patterns clearer.

[14] D. Mackay, "Bayesian non-linear modelling for the prediction competition," *ASHRAE Transactions*, vol. 100, no. 2, pp. 1053–1062, 1994.

[15] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders, "Variational bayesian image restoration based on a product of t-distributions image prior," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1795–1805, 2008.

[16] D. Tzikas, A. Likas, and N. Galatsanos, "Variational bayesian sparse kernel-based blind image deconvolution with student's-t priors," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 753–764, 2009.

[17] N. D. Le, A. E. Raftery, and R. D. Martin, "Robust bayesian model selection for autoregressive processes with additive outliers," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 123–131, 1996. [Online]. Available: http://www.jstor.org/stable/2291388

[18] J. Ó Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer, 1996.

[19] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[20] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics*, vol. 7. Oxford University Press, 2002.

[21] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, 2003.

[22] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2001, pp. 507–513.

[23] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2003.

[24] B. Wang and D. Titterington, "Inadequacy of interval estimates corresponding to variational Bayesian approximations," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 373–380.

[25] G. Consonni and J.-M. Marin, "Mean-field variational approximate Bayesian inference for latent variable models," *Computational Statistics and Data Analysis*, vol. 52, pp. 790–798, 2007.

[26] H. Akaike, "Power spectrum estimation through autoregressive model fitting," *Annals of the Institute of Statistical Mathematics*, vol. 21, no. 1, pp. 407–419, December 1969. [Online]. Available: http://ideas.repec.org/a/spr/aistmt/v21y1969i1p407-419.html

[27] D. J. Christini, A. Kulkarni, S. Rao, E. R. Stutman, F. M. Bennett, J. M. Hausdorff, N. Oriol, and K. R. Lutchen, "Influence of autoregressive model parameter uncertainty on spectral estimates of heart rate dynamics," *Annals of Biomedical Engineering*, vol. 23, no. 2, pp. 127–134, 1995.