Variational Bayesian with Embedded Laplace Approximation for AR Model with Outliers

Hongjie Wan\textsuperscript{1,*} and Haiyun Zhang\textsuperscript{2}

\textsuperscript{1}Information Engineering Dept, Beijing University of Chemical Technology, China
\textsuperscript{2}College of Information Science and Engineering, Fujian University of Technology, Fuzhou, China

\*Corresponding author: Wanhj@mail.buct.edu.cn

Abstract. The student-\textit{t} distributed innovation is robust to Auto regressive (AR) observations with outliers. However, the characteristics of student-\textit{t} distribution lead to unclosed solution while using Variational Bayesian (VB) to estimate the model parameters. This paper proposed to embed the Laplace approximation in VB framework to get closed form solution. Markov Chain Monte Carlo (MCMC) is a typical method to unclosed form solutions, while it is known that MCMC has large computation costs. The degree of freedom parameter (DOF) changes greatly before and after the outlier points has been deleted, so the estimation can be improved. Experimental results show that, more accurate estimation coefficients estimation are obtained using the proposed methods compared with using the Gaussian innovation, and more accurate estimation is obtained after the additive outlier is deleted.

1. Introduction

Traditionally, the autoregressive (AR) model takes Gaussian noise as excitation which can’t model non Gaussian noise from the real world. Based on the Gaussian noise assumption, the parameters can be estimated by minimizing the least squares with known model order. However, this method has been proved to be sensitive to outliers, and the proper model order can’t be selected automatically.

To provide robust modelling, Roberts and Penny [1] model the noise using a finite mixture of Gaussians; and they compute the posteriors of the parameters using the VB framework to avoid over fitting and give automatic model order selection simultaneously. Christmas and Everson [2] take student-\textit{t} distributions as the noise model, which is natural for real world data and robust to outliers. But the prior of the degree of freedom (DOF) parameter is not conjugate, which leads to unclosed form posterior and this approach leads to not very accurate estimation of the noise model parameters. To eliminate the non-conjugate problem, Dahlin etc. [3] proposed a MCMC [4] method to achieve the posteriors of all the parameters of the mode which is computational costly.

This paper embeds the Laplace approximation in VB to eliminate the affection of non conjugacy for robust AR modelling with student-\textit{t} noise model. A hierarchal Bayesian model is built for the AR model. The model order selection is realized via ARD prior [5, 6], The Laplace approximation approach which has been studied in [7, 8] is adopted to estimate the DOF. The MCMC method is also embedded in VB for comparison.

The structure of the rest of the paper is as following. Section II describes the hierarchal Bayesian AR model. Section III give out the embedded Laplace approximation in VB and the parameter estimation. Experiments on are performed in section IV. Section V concludes this paper.
2. AR Modelling

AR model is used to describe the relationship between the time series. Assuming the observation at time $m$ is $y_m$, and then the AR model with known order $p$ can be defined as

$$y_m = \sum_{n=1}^{p} a_n y_{m-n} + n_m$$  \hspace{1cm} (1)

Where $(a_1, \ldots, a_p)$ are the model coefficients, $n_m$ is the additive noise. The student-$t$ distribution used for the additive noise with scale $\xi$ and degrees of freedom $\eta$ is as follows

$$p(x \mid 0, \xi, \eta) = \frac{\Gamma((\eta+1)/2)}{\Gamma(\eta/2)\sqrt{\pi\eta}} \left(1 + \frac{\xi^2}{\eta}x^2\right)^{-(\eta+1)/2}$$  \hspace{1cm} (2)

For the purpose of using ARD prior to determine the order of the model, (1) can be rewritten in terms of the following form

$$y = \Phi A + n$$  \hspace{1cm} (3)

Where the vector $y$ is a collection of $M$ given observations, and $n$ is the vector of noise; the coefficients vector is $A = (a_1, \ldots, a_p, \ldots, a_N)^T$, where $N$ is a large number compared to the real model order $p$. The matrix $\Phi$ is composed by the given observations whose row is built by the current value and the previous values.

First of all, Gaussian prior over each AR coefficient is set

$$p(a_i \mid \gamma_i) = N(a_i \mid 0, \gamma_i^{-1})$$  \hspace{1cm} (4)

Where $\gamma_i$ is a parameter controlling the strength of the prior, then a conjugate Gamma prior with shape $a_\gamma$ and rate $b_\gamma$ is placed over each parameter $\gamma_i$:

$$p(\gamma_i) = G(\gamma_i \mid a_\gamma, b_\gamma)$$  \hspace{1cm} (5)

The concept of conjugate prior will be described in section 3.2. For the student-$t$ noise model, it can be factorized as the convolution of a Gaussian and a Gamma distribution. The Gaussian is zero mean and variance $(\xi x_m)^{-1}$, where $x_m$ is a hidden Gamma distributed variable with equal shape and rate parameter $\eta/2$, and the hidden variables can be represented by a vector $x = (x_1, \ldots, x_M)^T$. Therefore, the noise model (2) can be represented by the following two distributions

$$p(n_m) = N(n_m \mid 0, (\xi x_m)^{-1})$$  \hspace{1cm} (6)

$$p(x_m) = G(x_m \mid \eta/2, \eta/2)$$  \hspace{1cm} (7)

Then, Gamma priors are placed on the $\xi$ and $\eta$ with shape and rate parameters $a_\xi$, $b_\xi$ and $a_\eta$, $b_\eta$ respectively. Very tiny values are set as $a_\xi = b_\xi = a_\eta = b_\eta = 1e-6$, thus non-informative priors are set for $\gamma$, $\xi$ and $\eta$ [4].

3. Variational Bayesian with Embedded Laplace Approximation

For a probabilistic model with known structure, let $y$ represent the given observations collection and $\theta$ represent the parameters and latent variables collection, the Bayesian inference approach is to compute the posterior $p(\theta \mid y) = p(y, \theta) / p(y)$. In VB, the posterior $p(\theta \mid y)$ can be approximated by a distribution $Q(\theta)$ when the maximum value of $L(Q) = \int Q(\theta) \log p(y, \theta) / Q(\theta) d\theta$ is obtained [9].

Assuming the vector $\theta$ can be divided into $I$ independent sets $\theta = \{\theta_i\}_{i=1}^I$, then the distribution $Q(\theta)$
can be decomposed as products of $I$ distributions $Q(\theta) = \prod_{i=1}^{I} Q_i(\theta_i)$. By taking the derivative of $L(Q)$ with respect to each $Q_i(\theta_i)$ equal to zero, that is \( \frac{\partial L(Q)}{\partial Q_i(\theta_i)} = 0 \), the local maxima of the $L(Q)$ can be achieved. At the same time, the posterior approximation $Q(\theta)$ is obtained for each set $\theta_i$,

$$Q(\theta_i) \propto \exp\{\log p(y, \theta_i)\}_{j\neq i}.$$  

(8)

Where \( \langle \cdot \rangle_{j\neq i} \) is an expectation operator, the subscript $j \neq i$ means all the distributions other than $Q(\theta_i)$ are required to compute the expectation, furthermore, only the variables involved in $Q(\theta_i)$ are desired to calculate the expectation.

3.1. Embedded Laplace approximation for Unclosed form Posterior

In VB framework, if the posterior approximation (8) has no closed form expression, it is difficult to calculate the expectations. One approach is the Laplace approximation [7, 8], which takes a Gaussian at point of the MAP estimate as the approximated posterior. Here, assuming $Q(\theta)$ doesn’t have a closed form, let

$$f(\theta) = \langle \log p(y, \theta) \rangle_{j\neq i}.$$  

(9)

Because the MAP estimation is obtained at the maximum value, the maximum value of $f(\theta)$ should be obtained firstly, and let $\tilde{\theta}$ denotes corresponding point. This can be done by setting the gradient of $f(\theta)$ equal to zero, and. Then, based on Taylor series, $f(\theta)$ can be expanded as follows

$$f(\theta) = f(\tilde{\theta}) + (\theta - \tilde{\theta})^T \nabla f(\theta) |_{\theta = \tilde{\theta}} + \frac{1}{2} (\theta - \tilde{\theta})^T \nabla^2 f(\theta) |_{\theta = \tilde{\theta}} (\theta - \tilde{\theta}).$$  

(10)

The second term of the above equation can be omitted as the gradient is zero, which gives

$$f(\theta) = f(\tilde{\theta}) + \frac{1}{2} (\theta - \tilde{\theta})^T H(\tilde{\theta})(\theta - \tilde{\theta}).$$  

(11)

Where $H$ denotes the hessian matrix. Substituting (11) into (8), after normalization, a local normal distribution approximation to $Q(\theta)$ is achieved

$$Q(\theta) \sim N(\tilde{\theta}, -H(\tilde{\theta})^{-1}).$$  

(12)

This method needs to compute a Hessian matrix and its inverse, which, for high dimensional parameter, will cost a large amount of computation. Fortunately, the parameter we deal with is one dimension in this paper, so it will not cause much computation. Another method is the MCMC method, but this method will cost relatively large computation.

3.2. Posterior Approximation of the parameters

According to the model built in section 2, the parameter set $\theta$ is divided into five sets as $\theta = \{A, \gamma, x, \xi, \eta\}$. The joint probability of the observations and the parameter set can be factorized into products of six items

$$p(y, \theta) = p(y | A, \xi, x) p(A | \gamma) p(\xi | \eta) p(x | \eta) p(\gamma) p(\eta).$$  

(13)

Take the VB update equation (8), we can have posterior approximations over the five parameter sets, $Q(A), Q(\gamma), Q(x), Q(\xi)$ and $Q(\eta)$. Firstly, keep only the $Q$ distributions related with $A$, gives

$$Q(A) \propto \exp\left\{ \log N(y | \Phi A, (\xi \text{diag}(x))^{-1}) + \log \prod_{n=1}^{N} p(\alpha_n | \gamma_x) \right\}_{Q(\xi)Q(x)Q(\eta)}.$$  

(14)
After mathematical derivation, \( Q(A) \) is a normal distribution over \( A \), its mean and variance are as follows:
\[
\Sigma = \left( \xi (x\gamma) \Phi^T \text{diag}(x\gamma) \Phi + \text{diag}(\gamma(y)) \right)^{-1}
\]
\[
\mu = \Sigma (\xi (x\gamma) \Phi^T \text{diag}(x\gamma) )y
\]
(15a)
(15b)

Applying the same procedure, the posterior \( Q(\gamma) \) can be similarly obtained
\[
Q(\gamma) \propto \exp \left\{ a_{\gamma} + \frac{1}{2} \sum_{n=1}^{N} \log \gamma_n - \sum_{n=1}^{N} \gamma_n \left( b_{\gamma} + \frac{1}{2} \{ a_{\gamma} \}^2 \right) \right\}
\]
(16)

It can be seen that \( Q(\gamma) \) is the product of \( N \) gamma distributions, then each \( \gamma_n \) is a gamma distributed variable, the shape and rate parameter can be written specifically as following:
\[
a_{\gamma} = a_{\gamma} + \frac{1}{2}
\]
(17a)
\[
b_{\gamma} = b_{\gamma} + \frac{1}{2} \{ a_{\gamma} \}
\]
(17b)

Note that for the posteriors of all \( \gamma_n \), they have the same shape parameter, and individual rate parameters. Taking similar steps over \( Q(x) \), we have
\[
Q(x) \propto \exp \left\{ \frac{\eta}{2} - \frac{1}{2} \sum_{m=1}^{M} \log x_m - \frac{\varepsilon}{2} \sum_{m=1}^{M} x_m \left( \gamma_m - \Phi_m A \right)^T (\gamma_m - \Phi_m A) \left( \gamma_m - \Phi_m A \right) \right\}
\]
(18)

Also, \( Q(x) \) can be factorized into products of \( M \) gamma distributions, the shape and rate parameter for each \( x_m \) are listed below:
\[
a_{x_m} = \frac{\eta}{2} + \frac{1}{2}
\]
(19a)
\[
b_{x_m} = \frac{\eta}{2} + \frac{1}{2} \left( \frac{\varepsilon}{2} \right)
\]
(19b)

Where \( \Phi_m \) is the \( m \)-th row of matrix \( \Phi \), this is different to RVM and compressive sensing. For convenience, all \( b_{x_m} \) can be collected into a vector as \( \beta = (\beta_1, \ldots, \beta_M)^T \), then
\[
\beta_i = \frac{1}{2} \left( \frac{\eta}{2} \right) + \frac{1}{2} \left( \frac{\varepsilon}{2} \right)
\]
(20)

Where \( \langle AA^T \rangle_{(Q(A))} = \Sigma + \langle A \rangle_{(Q(A))} \langle A \rangle_{(Q(A))^T}, \ \odot \) denotes the element wise product of two vectors, and \( \text{diag}() \) represents an operation to form a column vector by extracting the diagonal elements of a matrix.

The posterior of the freedom of student-t can be similarly obtained
\[
Q(\eta) \propto \exp \left\{ -M \log \Gamma (\eta) + M \eta \log(\eta) + \frac{\eta}{2} \sum_{m=1}^{M} \left( \log x_m - \gamma_m \right)_{(Q(A))} + (a_{\eta} - 1) \log \eta - b_{\eta} \eta \right\}
\]
(21)

It is obvious that the above equation can’t be formulized as a closed form distribution, as there is log gamma in it. Let
\[
f(\eta) = -M \log \Gamma (\eta) + M \eta \log(\eta) + \frac{\eta}{2} \sum_{m=1}^{M} \left( \log x_m - \gamma_m \right)_{(Q(A))} + (a_{\eta} - 1) \log \eta - b_{\eta} \eta
\]
(22)
The first order derivative is

\[
\frac{d}{d\eta} f(\eta) = \frac{M}{2} \psi(\eta) \left(1 - \frac{2}{\eta^2}\right) + \frac{M}{2} \log(\eta) + \frac{1}{2} \sum_{k=1}^{K} \frac{1}{a_k - \eta} - b_{\eta}
\]

Where \( \psi(\cdot) \) denotes the digamma function. Setting it equal to zero, we can achieve the maximum of \( f(\eta) \) at the point \( \tilde{\eta} \). The root of \( f'(\eta) \) can be obtained by Newton method, or fzero function in matlab. Giving the second order derivative \( f''(\eta) \), the posterior of \( \eta \) can be approximated by a normal distribution as

\[
\mathcal{N}(\eta | \tilde{\eta}, -f''(\tilde{\eta})^{-1})
\]

The second order derivative is

\[
f''(\eta) = \frac{M}{4} \psi'(\eta) + \frac{M}{2\eta} - \frac{a - 1}{\eta^2}
\]

Where \( \psi'(\cdot) \) denote the derivative of the digamma function.

Finally, the posterior approximation of \( \xi \) can similarly obtained as a gamma distribution, its shape and rate parameters are as following

\[
\alpha_{\xi} = a_{\xi} + \frac{M}{2}
\]

\[
\beta_{\xi} = b_{\xi} + \frac{1}{2} \left( y \cdot \Phi(x) \right)^{T} (y - 2 \Phi(A) x) + \frac{1}{2} \text{tr} \left( \text{diag}(y \Phi(x) \Phi A^T A) \Phi^T \right)
\]

Where \( \text{tr}(\cdot) \) denotes the trace operation over a matrix. Statistics of \( Q(A), Q(\gamma), Q(x), Q(\xi) \) and \( Q(\eta) \) can be updated co-ordinately by rounds of iteration until convergence, since the statistics of them depend on each other.

4. Experiments

This section performs experiment on synthetic data with outliers, and results are given to prove the performance of the proposed method. The proposed method (denoted as LA-VB) is compared with the model using Gaussian distributed innovation (denoted as Gau) and the embedded MCMC method (denoted as MC-VB).

4.1. Data Preparation

In the experiment, the actual order of the AR coefficients is set to be 10, and the maximum order \( N \) for automatic order determination is set to be 40. Under a given order \( p \), given parameters \( \xi \) and \( \eta \) for student-t distribution, after randomly generating \( p \) initial values, observed sequences can be produced following (1).

In the VB iteration procedure, the initial value of \( \xi \) and \( \eta \) each is set to be a random number between 0 and 1; the expectation of the latent variable vector \( x \) is initialized randomly. The iteration procedure can be regarded as converged when the relative change of \( L(Q) \) between two consecutive iteration steps is lower than a tolerance, e.g. 1e-5. For the MC-VB approach, the sample number is 4000.

4.2. Experiment on AR observations with outliers

The following experiment is performed on data sequences with single outlier. Gaussian noise model is used as innovations and outliers are added to the observations. The data number \( M \) is from 200 to 900 with step size 100, the true order \( p \) is 10 and the variance of the noise model is 1. The additive outliers (A-outliers) are generated by adding a value larger than 5 times of the standard deviation of the observations; the innovation outliers (I-outliers) are generated by adding a value larger than 5 times of the maximum value of the past \( N \) observations, \( N \) is set as 20. The experiment is performed 100 times and the average results are shown in Figure1. The point with the largest reconstruction error is selected as the outlier point.
Figure 1 shows the $\Delta \eta$ value obtained after deleting the A-outlier; correspondingly, $\Delta \eta$ value obtained after deleting a randomly selected non outlier data point is provided in Figure 1 (b). Note that when deleting A-outlier, the subsequent $N$-1 data points are also deleted. The values in Figure 1 (a) is much larger than that in Figure 1 (b). In Figure 1(a), LA-VB gets the most stable values; the values of MC-VB changes slowly as the data length increases and the value are large for all data length. Although MC-VB gets greater and more slowly changed $\Delta \eta$ values when data length changed, computational cost is a shortage of the method. Figure 1(c) and (d) shows the results of data with I-outlier similar to (a) and (b). The experiment results discussed above suggest that $\Delta \eta$ changes greatly after outlier points have been deleted.

The coefficients estimation error obtained by LA-VB, MC-VB before and after deleting outlier is displayed in Fig. 2. The results of these approaches are compared with that obtained by Gaussian in [6]. Figure 1(a) (b) showed the results with and without A-outlier respectively. It is clear that after deleting A-outlier, the performance has a great improvement. Figure 2(c) (d) showed the results with and without I-outlier respectively. There are slightly improvements after the outlier is deleted for LA-VB and MC-VB, whereas Gaussian gets greater improvement. In summary, the proposed method is robust to A-outlier and I-outlier; the estimation is affected by A-outlier, but better performance is obtained after deleting A-outlier.

5. Conclusion
The student-t noise model is robust to AR observations with outliers when being used as the innovation of AR models. To get closed form solution while estimating the parameters of the model using VB framework, the Laplace approximation is embedded into the VB framework. Results have demonstrated that more accurate parameter estimation can be obtained using the proposed method, and the deletion of the additive outliers can further improve the estimation accuracy.

References
[1] Penny W D and Roberts S J Bayesian methods for autoregressive models 2002 Neural Networks for Signal Processing X, IEEE Signal Processing Society Workshop.
[2] Christmas J and Everson R Robust autoregression: Student-t innovations using Variational Bayes 2011 IEEE Transactions on Signal Processing 59(1) 48-57.
[3] Dahlin J, Lindsten F, Schön T B and Wills A Hierarchical Bayesian approaches for robust inference in ARX models 2012 Proc. 16th IFAC Symposium on System Identification
[4] Gilks W R, Richardson S and Spiegelhalter D J 1996 *Markov Chain Monte Carlo in Practice* (London: Chapman and Hall).

[5] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Springer).

[6] Tzikas D G, Likas A C and Galatsanos N P The variational approximation for Bayesian inference 2008 *IEEE Signal Processing Magazine* 35(6) 131–46.

[7] Mathew J B Variational Algorithms for Approximate Bayesian Inference 2003 PhD thesis (UK: University of Cambridge).

[8] Chong W, David M B Variational inference in nonconjugate models 2013 The Journal of Machine Learning Research archive 14(1).

[9] Wan H, Ma X and Li X Variational Bayesian learning for removal of sparse impulsive noise from speech signal 2018 Digital Signal Processing 73 106–16.