Bayesian Sparse Linear Prediction with Pearson Type VII Distribution

Hongjie Wan¹* and Haiyun Zhang²
¹Information Engineering Dept, Beijing University of Chemical Technology, China
²College of Information Science and Engineering, Fujian University of Technology, China
*Corresponding author: Wanhj@mail.buct.edu.cn

Abstract. The speech signal can be modelled as AR models with an innovation noise model. The Pearson type VII distribution is used to model the real excitation. Variational Bayesian framework is used to estimate the posteriors of the AR coefficients and noise model parameters. The model is not conjugate, so MCMC is embed into VB framework to estimate the degree of freedom (DOF) parameter of Pearson type VII distribution. The model order selection is carried out by setting ARD priors on the coefficients. Simulation is carried out on synthetic and real data, the results show that the algorithm performs well for linear prediction both for synthetic data and speech signal, and the results are better than using least square method.

1. Introduction
For speech signal analysis, the autoregressive (AR) model is a typical model. The AR model is used widely for speech coding, speech recognition and speech synthesis. Traditionally, the innovation model of the AR model is assumed to be Gaussian model, and the usually methods for estimating the model parameters are least squares (LS), Durbin, Burg, and Lattice [1]. In this paper, the LS method be compared with.

Variational Bayesian (VB) is a common tool to estimate the posterior of the parameters for a model that is hard to compute the integral. It has great advantage than Maximum likelihood method and Maximum a posterior estimation other than merely point estimation. In addition, his framework can prevent over-fitting [2]. For the convenience of this method, Penny and Roberts provide a method based on VB to learn the AR parameter and the model order [3]. Their model takes Gaussian as the innovation noise model. The model order selection is based on the lower bound of the VB framework. Gaussian noise model is not robust to outliers, and it can’t model long tailed noise.

In this paper, a Bayesian model is built for AR model parameter estimation with Pearson type VII distribution [4][5] as the noise model. The VB based learning algorithm is derived to estimate the posterior of the parameters. ARD priors are set on each coefficient, other than cast isotropic priors on the coefficient as in [3]. The ARD prior can determine the model order by pruning the coefficients during VB iterations. When the coefficients are very near to zero, then they can be pruned automatically. This type of prior and pruning has been used in RVM [6], and the algorithm of this paper can be seen as the modified version of RVM. The EM based algorithm and the VB framework based algorithm has been given in [6] and [7] respectively. The VB framework requires computing the expected statistics of the parameters on the posteriors, then calculate these statistics round by round.
until convergence. To have this type of computation, the posteriors should have closed form. It is unfortunate that closed form distribution for the DOF parameter of Pearson type 7 distribution cannot be reached. To avoid this, the MCMC algorithm is embedded in VB to compute the expected statistics. As long as the sample number tends to large enough number, closely estimation of the expected statistics of the true posterior can be obtained. In VB framework, the lower bound is used to monitor the convergence of the algorithm. The lower bound is also derived to control the algorithm. In the embedded MCMC algorithm, a normal distribution is adopted to approximate the true posterior, for which the mean and covariance can be achieved. By this assumption, the approximated lower bound can be achieved and can be used to control the convergence.

This paper is organized as follows. In Section II, a Bayesian model is built for AR model using Pearson type 7 distribution and the ARD priors on every coefficients. In section III, the posteriors of the parameters except for the DOF parameter is firstly derived; Secondly, the embedded MCMC are described to approximate the posterior of the DOF parameter; in the last part of this section, the detailed approximated lower bound is given out. Experiments on synthetic and speech signal are performed in section IV. In section V, conclusions are given.

2. Bayesian AR Model

Given the signal with length \( T_s = (s_1, s_2, \ldots, s_T) \), and the known model order \( p \), the autoregressive model can be described as

\[
\begin{align*}
    s_t &= \sum_{i=1}^{p} w_i s_{t-i} + n_t
\end{align*}
\]

Where \((w_1, \ldots, w_p)\) are the coefficients of the AR model, \(n_t\) denotes the additive noise at time \(t\). Instead of using normal distribution as the additive noise model, the additive noise is modelled using the Pearson type VII distribution. The formulation of this distribution is as follows

\[
    p(n_t | 0, \lambda, \eta) = \frac{\sqrt{\lambda}}{B(\eta - 1/2, 1/2)} \left(1 + \frac{n_t^2}{\lambda}\right)^{-\eta}
\]

Where \(B(gg)\) is the Beta function. It has been proved that the Pearson type VII distribution can be formulated as the convolution of a normal and a Gamma distribution as following

\[
    p(n_t | 0, \lambda, \eta) = \int_0^\infty N(n_t | 0, \lambda^{-1}, x_t^{-1}) G(x_t | \eta - 1/2, 1/2) dx_t
\]

Where \(G(g)\) is the Gamma distribution and \(\eta\) is the degree of freedom parameter with the constraint condition \(\eta > \frac{1}{2}\). From above equation, after the complex form (2) is decomposed into two simple distributions, a hidden variable \(x_t\) is introduced at every time \(t\). All the hidden variables are collected into a vector \(x = (x_1, x_2, \ldots, x_T)\). To form Bayesian model, the Gamma distributions is cast on the parameters \(\lambda\) and \(\eta\) as priors

\[
    p(\lambda) = G(\lambda | a_\lambda, b_\lambda)
\]

For the DOF parameter \(\eta\), note that \(\eta - \frac{1}{2} > 0\), so \(\nu = \eta - \frac{1}{2}\) is used as the parameter and cast Gamma distributed prior on it

\[
    p(\nu) = G(\eta | a_\nu, b_\nu)
\]

Then similarly as sparse learning, the AR model is transformed as the following form

\[
    p(\lambda | a_\lambda, b_\lambda)
\]

\[
    p(\nu | a_\nu, b_\nu)
\]
This equation is given by writing all the observance in to one vector \( s \), and all additive noise into one vector \( n = (n_1, n_2, K, n_p) \), and \( w = (w_1, L, w_m) \) are the collection of all the coefficients. To give sparse learning and determine the model order in the VB iteration, the length of the coefficient vector \( w \) is changed as a large number \( M \), which is larger than the maximum probable model order \( p \). Then the coefficient vector is \( w = (w_1, L, w_m) \). The matrix \( \Phi \) is formed the observance that each row contains the current observance and its \( M-1 \) previous observed values, such as \( \Phi_{011}, \Phi_{s1}, \Phi_{sn} \). Then priors should be placed on the coefficients to form a Bayesian model. Here ARD priors are adopted, which can introduce automatic model selection. The prior on each coefficient is set with a normal distribution with zero mean and \( \gamma_m \) variance as follows

\[
p(w_m | \gamma_m) = N(w_m | 0, \gamma_m^{-1})
\]

Where \( \gamma_m \) determines the amplitude of the coefficient \( w_m \) in VB iteration. Then each \( \gamma_m \) is set with a Gamma distributed prior with the same shape and rate parameters as follows

\[
p(\gamma_m) = \mathcal{G}(\gamma_m | a, b)
\]

Also, all the \( \gamma_m \) parameters are written into one vector \( \gamma = (\gamma_1, \gamma_L, \gamma_p, L, \gamma_n) \).

3. Parameter Estimation using Variational Bayesian

All the parameters and hidden variables can be written into a set as \( \theta = \{w, \gamma, x, \lambda, \nu\} \). VB is used to approximate the true posterior \( p(\theta | s) \), thus reaching a closest approximation to \( p(s) \). The following factorized product is used to approximate the posterior of all the items of \( \theta \).

\[
q(\theta) = q(w) q(\gamma) q(x) q(\lambda) q(\nu) = \prod_{i=1}^{I} q(\theta_i)
\]

Where \( \theta_i \) is the \( i \)-th item of the parameter set \( \theta \), and \( I \) is the total number of parameter items, here \( I \) equals to 5. The posterior approximation \( q(\theta_i) \) can be obtained by Variational EM algorithm [8]

\[
q(\theta_i) \propto \exp \{ \log p(s, \theta_i) \}
\]

Where \( \langle \cdot \rangle_{\ast i} \) denotes the expected statistics with respect to distributions other than \( q(\theta_i) \). The VB framework is to compute the expected statistics of each \( \theta_i \) round by round until convergence. The joint probability of all the parameters and the observed values \( p(\theta | s) \) is

\[
p(s, \theta) = p(s | w, \lambda, x) p(w | \gamma) p(\gamma | x) p(\nu) p(\lambda)
\]

3.1. Closed form Posteriors

Applying (10) to \( w \), it can be seen that \( q(w) \) is a multivariate normal distribution, which has the covariance and mean as follows

\[
\Sigma = \langle \lambda \rangle \Phi^T \text{diag}(\langle x \rangle) \Phi + \text{diag}(\langle \gamma \rangle)^{-1}
\]

\[
\mu = \langle \lambda \rangle \Phi^T \text{diag}(\langle x \rangle) s
\]

Notice that the subscript denoting the expected statistics with respect to which distribution is omitted here and it will be omitted in the following chapters. Then taking the same procedure, it can be derived that each \( \gamma_m \) is Gamma distributed with shape and rate parameters as follows
With the same steps, each hidden variable $x_i$ is obtained as gamma distributed with the following shape and rate parameters

$$\alpha_i = a_i + \frac{1}{2}$$  \hspace{1cm} (13a)  

$$\beta_i = b_i + \frac{1}{2}w_i^2$$  \hspace{1cm} (13b)  

Where $\beta_i = (\beta_i, \phi, \beta_{\phi})$ is the collection of all rate parameters of $x_i$ and $o$ is the item by item product of two vectors; and $\langle ww^T \rangle = \Sigma + \langle w \rangle \langle w \rangle^T$. Then the posterior of $\lambda$ is also a gamma distribution, with the following parameters

$$\alpha = \frac{1}{2}a + \frac{T}{2}$$  \hspace{1cm} (15a)  

$$\beta = b + \frac{1}{2}(s_0s \cdot 2(\Phi \langle w \rangle) + \frac{1}{2}tr(diag(\langle x \rangle \Phi \langle ww^T \rangle \Phi^T))$$  \hspace{1cm} (15b)  

### 3.2. Embedded MCMC for Unclosed form Posterior

Based on (10), the posterior of the freedom parameter of Pearson type VII distribution will be

$$q(\upsilon) \propto \exp \left\{-T \log \Gamma(\upsilon) + \left[-T \log 2 + \sum_{i=1}^{T} \log x_i \right] \upsilon + (a_\upsilon - 1) \log \upsilon \right\}$$  \hspace{1cm} (16)  

The above equation contains Gamma function of the variable $\upsilon$, so it can’t be formed by a closed form distribution. From (14), it can seen that only the estimation of the mean of $q(\upsilon)$ should be obtained. Markov Chain Mote Carlo method can be used to get this estimation. By MCMC method, a sequence of samples can be got until the samples can be seen from a stationary distribution which is exactly the true distribution. Then the mean $\bar{\upsilon}$ of the samples can be reached. Metropolis-Hasting (MH) [9] sampler is used to get samples from the distribution $q(\upsilon)$. Normal distribution denoting $g$ is adopted to be the sampler. Let the sample to be generated be $\upsilon_i$ and the last sample be $\upsilon_{i-1}$. The MH acceptance probability is

$$\alpha = \min \left\{1, \frac{q(\upsilon_i) g(\upsilon_{i-1} | \upsilon_i)}{q(\upsilon_{i-1}) g(\upsilon_i | \upsilon_{i-1})} \right\}$$  \hspace{1cm} (17)  

Then a rand number $r$ is generated from a uniform distribution, if $r \leq \alpha$ the proposed sample is accepted, otherwise rejected. To avoid different initialization affections, burnin in the sampling process is considered. The variance $\sigma^2_\upsilon$ of the sampled samples can also be estimated. To form complete VB framework, $q(\upsilon)$ is assumed to be a normal distribution $N(\upsilon | \bar{\upsilon}, \sigma^2_\upsilon)$.  

In above equations, the mean of the variable with Gamma distribution, and the expected value of the logarithm of the variable with Gamma distribution should be computed, which can be the following equations

$$\langle x \rangle_{\text{post}} = \alpha / \beta$$  \hspace{1cm} (18)
\[
\langle \log x \rangle_{\text{pdf}} = \Psi(\alpha) - \log(\beta)
\]

Where \( x \) is the Gamma distributed variable, \( \alpha \) and \( \beta \) is the shape and rate parameters respectively.

4. Experiments

In this section, experiments are performed to evaluate the proposed algorithm. Firstly, the algorithm is run on simulation data to prove the efficiency of the algorithm. Then, the algorithm is applied to speech signal, and compares the result with the least square method. In the following, the proposed algorithm is denoted as VBMC and the traditional least square method as LS.

4.1. Experiment on Synthetic Data

In this experiment, synthesized data are generated by AR systems with non-Gaussian distributed innovations. Then the algorithm is run to estimate the parameters of AR coefficients and the innovation distributions. The AR coefficients are generated by polynomials whose roots are random selected within the unit circle of complex plane. The order of AR is set to be 10, and the maximum order in the algorithm to be 20. With known AR order and defined values \( \nu \) and \( \lambda \), the model can be defined. Then the observed sequences can be produced following (1) with 10 initial values. In the algorithm, \( \nu \) and \( \lambda \) are initialized randomly and the sample number is set to be 4000.

Firstly, the noise model parameters \( \nu \) and \( \lambda \) are set to be 1.5 and 1/3 respectively, and then the algorithm is verified whether it can estimate the noise model parameters. The AR coefficients are generated by the above methods, and the sample number is 800. It should be noted that the parameters of the noise distribution can’t be estimated by LS approach, and only the mean and variance of the normal distribution can be estimated by the prediction errors. The noise distribution is plotted in figure 1. From the figure, it can be seen that the proposed algorithm can accurately estimate the parameter of the noise model.

Secondly, with the noise parameters given above, AR observations with random initial values are generated. The true value of AR coefficients and the coefficients estimated by VBMC and LS are all plotted in figure 2. The figure shows that the VBMC approach can estimate the large coefficients accurately, but some of the small coefficients not very accurate. It can be also seen that the model order by VBMC is 10, which shows that it can estimate the model order automatically. Although the estimated coefficients are not exactly the same as the original, they are closer than that estimated by LS method.

![Figure 1. Original noise model, estimated by the proposed algorithm and by LS method.](image1)

![Figure 2. AR coefficients of the original, estimated by VBMC and LS method.](image2)
4.2. Experiment on speech signals

Speech signal can be modelled by AR model and several methods have been presented for AR coefficients estimation [1]. Here, experiment is used to prove that the speech signal is not generated by normal distributed noise model. If the noise model is normal distribution, then the observation should be normal distributed. A section of speech signal is selected with duration of 20ms and 16 KHz sampling rate, then plot the quantiles of the data versus the normal distribution in figure 3. If the speech signal can be modelled by AR model with Gaussian noise model, then the quantile will be a line. Figure 3 shows that it is not a line thus proved that speech signal is not generated by Gaussian noise model.

![Figure 3. Quantiles of speech signal vs. standard normal distribution](image1)

The original signal is shown in figure 4, which is a periodic signal and the sharp peak and valley shows that there are outliers in the signal. Note that the speech signal is normalized, so the values are small. For signal with outliers, the noise model with long tail is suitable for innovation model of AR modelling. The prediction error and the estimated noise model are plotted in figure 5. The result shows that there really exist outliers in prediction error which correspond to the innovation model of AR model. The estimated noise model has long tail, thus the model can process such outliers. While for Gaussian noise model, it doesn’t have long tail and can’t process such data points.

![Figure 4. The original speech signal](image2)

![Figure 5. The prediction error and the estimated noise model](image3)

The AR model order is set to be from 10 to 20, and the RMSE between the original signal and the one time prediction signal are computed. The results are listed in Table 1 and the AR coefficients
estimated by VBMC and that estimated by LS are plotted in figure 6. In table 1, the RMSE for order 17 to 20 is the same as 16, so only the results from 10 to 16 are listed. From table 1 and figure 6, it can be seen that LS can’t automatically decide which the best model order is and it can only give result under given model order. The model order is 16 and VBMC can automatically give out this model order by automatically setting the irrelevant coefficients to be zero or almost zero.

![Figure 6. The estimated coefficients of speech signal](image)

| Order | RMSE by LS | RMSE by VBMC |
|-------|------------|--------------|
| 10    | 0.0198     |              |
| 11    | 0.0196     |              |
| 12    | 0.0194     |              |
| 13    | 0.0194     |              |
| 14    | 0.0186     |              |
| 15    | 0.0187     | 15 0.0168    |
| 16    | 0.0184     |              |

5. Conclusion
In this paper, Bayesian method is proposed to estimate the AR coefficients of speech signal. Speech signal can be modelled by AR model and long tailed Pearson type 7 distributions. Variational Bayesian (VB) framework is used to get the posterior estimate of the coefficients and noise model parameters. To eliminate the non-conjugate problem, MCMC is embedded into the VB framework. The model order can be computed automatically by ARD. Experiments on synthetic data and speech signal are given for the proposed method and LS approach. Results show that the proposed algorithm is suitable for speech linear prediction.

References
[1] C Ben G, Nelson M and Dan E 2011 *Speech and Audio Signal Processing: Processing and Perception of Speech and Music, 2nd edition* (New York: Wiley–Interscience).
[2] 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.
[3] Penny W D and Roberts S J Bayesian methods for autoregressive models 2002 *Neural Networks for Signal Processing X, IEEE Signal Processing Society Workshop*.
[4] Pearson K Mathematical contributions to the theory of evolution, xix: Second supplement to a memoir on skew variation 1916 *Philos. Trans. Roy. Soc. Of London, Ser. A* 216 429–57.
[5] Sun J Ata K and Garibaldi J M Robust mixture clustering using pearson type vii distribution 2010 *Pattern Recognition Letters* 31(16) 2447-54.
[6] Tipping M E The relevance vector machine 2000 *Advances in Neural Information Processing Systems*.
[7] Bishop C M and Tipping M E Variational relevance vector machines 2000 *Proc. 16th Conf. Uncertainty in Artificial Intelligence* p 46–53.
[8] 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.
[9] Gilks W R, Richardson S and Spiegelhalter D J 1996 *Markov Chain Monte Carlo in Practice* (London: Chapman and Hall).