ICIC Express Letters
Volume **13**, Number **5**, May **2019**

 ISSN 1881-803X
pp. 415–420

IMPROVED VARIATIONAL LINEAR REGRESSION WITH CONFIDENCE LEVEL

Jiajing Li[1], Tao Lv[1] and Zijian Dong[2,*]

[1]Information Center
The Second People's Hospital of Lianyungang
No. 41, Hailian East Road, Haizhou District, Lianyungang 222006, P. R. China

[2]School of Electronic Engineering
Huaihai Institute of Technology
No. 59, Cangwu Road, Haizhou District, Lianyungang 222005, P. R. China
*Corresponding author: dzjian@126.com

Received October 2018; accepted January 2019

Abstract. *High dimensional linear regression models are widely applied in engineering, economics, bioinformatics and other fields. Variational theory is an effective method to solve high-dimensional sparse linear regression problems. The essence of the variational method is to approximate the joint posterior distribution by the approximate method and obtain the analytical solution of the posterior distribution of the parameters to be inferred, thus greatly reducing the calculation cost of the sampling and iterative inference. However, when the problem scale is large and the number of samples is small, its inferring performance is still not satisfactory. By using the credibility information of variables in the iteration process, this paper proposes a method to dynamically delete the low confidence variables with no zero value in the iteration process. Simulation results show that the algorithm is simple and effective, and greatly improves the accuracy of inferences.*
Keywords: Linear regression models, Variational inference, Confidence level

1. **Introduction.** Linear regression is an analytical method using probability model to determine the quantitative relationship between two or more variables. It is widely used in engineering, economics, finance, bioinformatics and so on.

The linear regression model is first fitted by least square approximation, and then some fitting methods have been developed. However, on the whole, these fitting methods have high computational cost when dealing with large-scale linear regression models [1,2]. The Bayesian method can be used to infer the unknown variable (unknown weight) in the linear regression model [3]. The method is to specify a priori to constrain the model, and then to infer the uncertain estimation of the posterior distribution of the unknown variable. Because the posteriori distribution requires high dimensional integration of prior distribution and likelihood function, the computation complexity is very high. Markov Chain Monte Carlo (MCMC) algorithm is also used to solve linear regression models. The algorithm first generates a sample for the posterior distribution through Monte Carlo method, and then uses the Markov chain to sample the complex posterior distribution [4,5]. If we select the prior model carefully and set the unimportant variables to zero, we can also reduce the computational complexity [6-8]. However, in general, linear regression still has some problems in the application of high-dimensional sparse models, such as high computational cost and limited inference accuracy.

The variational method can also be used for solving linear regression models. The essence of the variational method is to simplify the computation of probability distribution through the factorization of the high dimensional joint posterior distribution [9-13]. The

DOI: 10.24507/icicel.13.05.415

advantage of the variational method is that it can obtain the analytical solution of the posterior distribution of the variable parameters, thus reducing the computational cost of iterative inference.

In practice, the variational method is accurate, if the scale of the problem is moderate and the sampling quantity is reasonable. However, when the sample size is large and the number of samples is small (the number of samples is far below the number of variables to be inferred), the inferring performance is unsatisfactory. This paper proposes a method of dynamically reducing the low confidence variable in the iterative process, and thus reduces the scale of the model and improves the inference accuracy. The simulation results show that the algorithm is simple and effective, and it can effectively improve the detection rate and false detection rate of the algorithm.

This paper is arranged as follows. The second section describes the variational theory and its algorithm. The third section proposes a modified variational method with dynamic reduction of variables. The fourth section is simulation, and the proposed algorithm is verified. Finally, a summary is made.

2. Linear Regression Model and Its Variational Inference Algorithm.

2.1. Linear regression model.

The linear regression model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}_0 + \mathbf{e} \tag{1}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_N)^T$, y_i is the ith observation or dependent variable. $\mathbf{X}$ is the input matrix, and x_{jk} is the kth input in the jth test, $j = 1, 2, \ldots, N$, $k = 1, 2, \ldots, M$. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ is the weighted coefficient or the variable to be inferred. $\boldsymbol{\beta}_0$ is the intercept of $\boldsymbol{\beta}$. $\mathbf{e} = (e_1, e_2, \ldots, e_M)^T$ is the random observation noise. $\boldsymbol{\beta}_0$ can be eliminated in a certain way, so the linear regression model can be rewritten as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2}$$

2.2. Variational inference algorithm for linear regression model.

In many applications, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ is sparse, that is, most variables are 0 and non-zero variables account for a smaller proportion. So we can assume that $\boldsymbol{\beta}$ is a Gauss distribution, and $p(\boldsymbol{\beta}) \sim N\left(0, \frac{1}{\alpha}\mathbf{I}\right)$, $p(\alpha) \sim Gamma(a_0, b_0)$, where $Gamma$ represents gamma distribution and a_0, b_0 are two hyper parameters. Assuming that all observation noises are Independent and Identically Distributed (IID), $\mathbf{e}$ can be set as zero mean Gauss variables, $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$.

The likelihood function of the observed variables is

$$p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}) \sim \prod_{i=1}^{N} p(y_i|\boldsymbol{\beta}, \mathbf{X}_i) \tag{3}$$

where $\mathbf{X}_i$ is the ith row of $\mathbf{X}$, then

$$p(y_i|\boldsymbol{\beta}, \mathbf{X}_i) \sim N\left(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\right) \tag{4}$$

The joint distribution of $\mathbf{Y}$, $\boldsymbol{\beta}$, α is

$$p(\mathbf{Y}, \boldsymbol{\beta}, \alpha) = p(\mathbf{Y}|\boldsymbol{\beta}, \alpha)p(\boldsymbol{\beta}|\alpha)p(\alpha) \tag{5}$$

The variational technique is to find an approximate distribution for the posterior distribution of the target, which is more convenient in the inference process. This processing usually refers to a factorization of the analytical expression of the posterior distribution. According to the theory of factorization [14], we have

$$p(\boldsymbol{\beta}, \alpha|\mathbf{X}, \mathbf{Y}) \approx q(\boldsymbol{\beta}, \alpha) = q(\boldsymbol{\beta}|\alpha)q(\alpha) \tag{6}$$

The posteriori distribution of α is calculated first. According to variational approximation theory, $\ln q(\alpha)$ is a logarithmic joint distribution. Consider the expectation of all variables in $\boldsymbol{\beta}$,

$$\begin{aligned}\ln q(\alpha) &= E_{\boldsymbol{\beta}}[\ln p(\mathbf{Y}, \boldsymbol{\beta}, \alpha)] + const. \\ &= (a_0 + M/2 - 1)\ln\alpha - \alpha\left(b_0 + \frac{1}{2}E\|\boldsymbol{\beta}\|^2\right) + const.\end{aligned} \tag{7}$$

where $q(\alpha)$ is a gamma distribution, $q(\alpha) \sim Gamma(a_n, b_n)$, and the parameters are calculated as follows:

$$\begin{cases} a_n = a_0 + M/2 \\ b_n = b_0 + \dfrac{1}{2}E\|\boldsymbol{\beta}\|^2 \end{cases} \tag{8}$$

where $E\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\mu}_n\|^2 + tr(\mathbf{S}_n)$, $\boldsymbol{\mu}_n$ and $\mathbf{S}_n$ are the mean vector and variance matrix of the posteriori distribution of $\boldsymbol{\beta}$. The calculation of these two values will be discussed later.

Using the variational theory again, we calculate the expectation of the joint distribution of $p(\mathbf{Y}, \boldsymbol{\beta}, \alpha)$

$$\begin{aligned}\ln q(\boldsymbol{\beta}) &= E_{\alpha}[\ln p(\mathbf{Y}, \boldsymbol{\beta}, \alpha)] + const. \\ &= -\frac{1}{2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + E(\alpha)\mathbf{I}\right)\boldsymbol{\beta}\boldsymbol{\beta}^T - \frac{1}{2\sigma^2}\sum_i(-2\mathbf{X}_i y_i)^T\boldsymbol{\beta} + const.\end{aligned} \tag{9}$$

We can find that $q(\boldsymbol{\beta})$ is a Gaussian distribution, $q(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_n, \mathbf{S}_n)$, and

$$\begin{cases} \mathbf{S}_n = \left(\dfrac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \dfrac{a_n}{b_n}\mathbf{I}\right)^{-1} \\ \boldsymbol{\mu}_n = \dfrac{1}{\sigma^2}\mathbf{S}_n\mathbf{X}^T\mathbf{Y} \end{cases} \tag{10}$$

The forms of (9) and (10) constitute a complete iterative process. After a sufficient number of iterations, the iteration will arrive at a steady state. In practice, a relatively simple method can be used to judge whether the iteration process is stable, whether the steady state is achieved, and whether we can stop the iteration. We can calculate the difference of parameters in two successive iterations, for example, $S_n = \|\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n-1}\|^2$. If S_n is diminishing, the algorithm is stable. When S_n is small enough (depending on requirements), the iteration can be stopped.

3. Variational Inference Algorithm Based on Dynamic Reduced Variable Set.

When the number of variables in $\boldsymbol{\beta}$ is far more than the number of samples in $\mathbf{Y}$, that is, M is far greater than N, and the number of non-zero elements in $\boldsymbol{\beta}$ is relatively small, the inferential performance of the variational method is poor.

The reason is that the variational inference cannot infer most of the zero value "variables" to 0 by recursion, and these variables have been iterated around 0, so the parameter inference speed is slow and the inference precision is poor. For these parameters that are in the vicinity of 0, we can set these parameters to zero and delete them from the model to simplify the model dimension and reduce the computational complexity, thus speeding up the calculation and improving the calculation precision.

The idea is that we judge the parameters after the end of an iteration, and if the value of a variable is small enough, it can be judged to be a zero-value variable, and it can be removed from the model. The model after deleting continues to follow the next iteration.

We set a vector $\mathbf{T}_k$ in the kth iteration. If β_{ki} is zero, and will be deleted from model, then we set $t_{ki} = 0$; otherwise $t_{ki} = 1$. Let $C_k = \sum t_{ki}$. $\boldsymbol{\beta}_k$ is a $C_k \times 1$ vector from $\boldsymbol{\beta}_{k-1}$ by deleting some zero-value variable. $\mathbf{X}_k$ is obtained by deleting the column in $\mathbf{X}_{k-1}$ with $t_{ki} = 0$, and thus $\mathbf{X}_k$ is an $N \times C_k$ matrix, and is a subset of $\mathbf{X}_p$, $1 \le p < k$.

By model reduction, now the model $\mathbf{Y} = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{e}$, $\mathbf{X}_k$ is an $N \times C_k$ matrix, $\boldsymbol{\beta}_k$ is $C_k \times 1$ vector.

Now the algorithm is summarized as follows:

Input matrix $\mathbf{X}$, *input the observation vector* $\mathbf{Y}$; *set the initial super parameter* $a_0 = 1$, $b_0 = 1$. *Set a smaller number to* ξ *and* ψ.
Flag $= 1$;
While Flag (outside loop)
Repeat (inner loop):
1) calculate $\mathbf{S}_n$ *and* $\boldsymbol{\mu}_n$ *by (10)*;
2) calculate a_n *and* b_n *by (8)*;
3) calculate $S = \|\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n-1}\|^2$;
until $S < \xi$;
If $|\boldsymbol{\mu}_{ni}| < \psi$, *then set* $t_{ni} = 0$; *otherwise* $t_{ni} = 1$, *calculate* C_n.
Then, according to the method described above, get $\mathbf{X}_n$, $\boldsymbol{\beta}_n$ *from* $\mathbf{X}_{n-1}, \boldsymbol{\beta}_{n-1}$.
If $C_n = C_{n-1}$, *Flag* $= 0$;
End While
Output $\boldsymbol{\mu}_n$ *as the inferred result of* $\boldsymbol{\beta}$.

The iteration consists of an inner loop and an outside loop. In the inner loop, the standard variation inference method is used to infer the parameters, and the mean square error is calculated. If the mean square error is lower than the threshold, the inner loop ends. In the outer loop, some variables are deleted from the model according to the method described above, thus reducing the dimension of the model. In the outer loop, the dimensions of the two adjacent models are compared. If the dimensions of the two models are the same, that means the model cannot shrink again, then output the results. The whole iteration stops.

4. **Simulations.** In order to verify the performance of the proposed algorithm, we conducted several simulations. The input parameters used in the linear regression model are randomly generated and the variables are randomly generated, but we control the non-zero value ratio less than 10% to embody the sparsity. In the simulation, we compare two sets of models, and compare the standard Variation technique for Linear Regression model (VLR) and our proposed reliability based variational method (VLR-RB).

Suppose that there are T non-zero variables in the original model, and T' is the number of non-zero variables deduced from an algorithm, which includes T_T, the number of non-zero variables in the original model, and T_F, the number of variables inferred as a non-zero variable, but being zero variable in the original model. $T' = T_F + T_T$.

We define two parameters to characterize the performance of the algorithm. One is Power of Detection (PD) to signify the ability of finding correct non-zero variable, $PD = T_T/T$, and one is False Detection Rate (FDR) to indicate the error ratio caused by false inference of the algorithm, defined as $FDR = T_F/T'$. If the infer result is completely correct, $T_F = 0$, $T' = T_T$, $PD = 1$, $FDR = 0$. Figure 1 is a simulation case where the parameter size is 1200, of which 120 are non-zero variables. The number of simulated observations is $200 : 20 : 500$, that is, the observed value is $200, 220, \ldots, 500$. Gauss white noise is added, and the signal to noise ratio is 10.

It can be seen that the proposed algorithm VLR-RB is superior to VLR in terms of detection rate, and PD is close to 100% when the number of observations is above 250, while the VLR algorithm needs more than 350 observations to reach 90%. In terms of FDR, the VLR-RB performance is also excellent. When the number of observations is over 250, the false detection rate is close to 0, while the FDR of VLR has been maintained at around 50%.

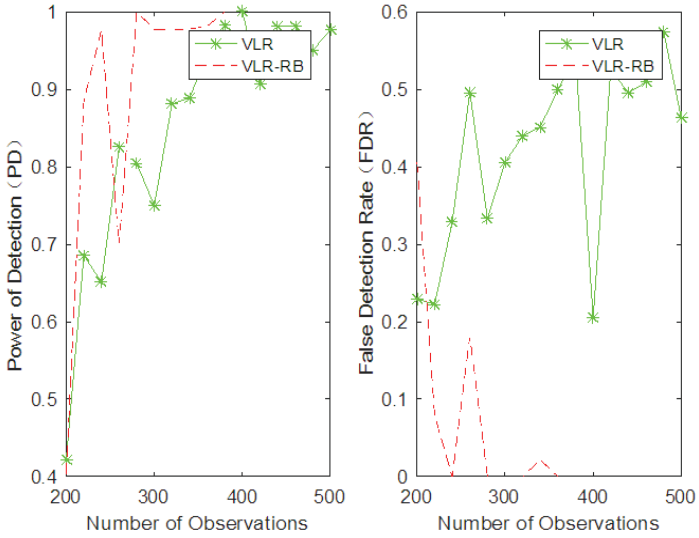

FIGURE 1. The number of parameters to be determined is 1200, SNR is 10, the left side is the detection rate, and the right side is the false detection rate.

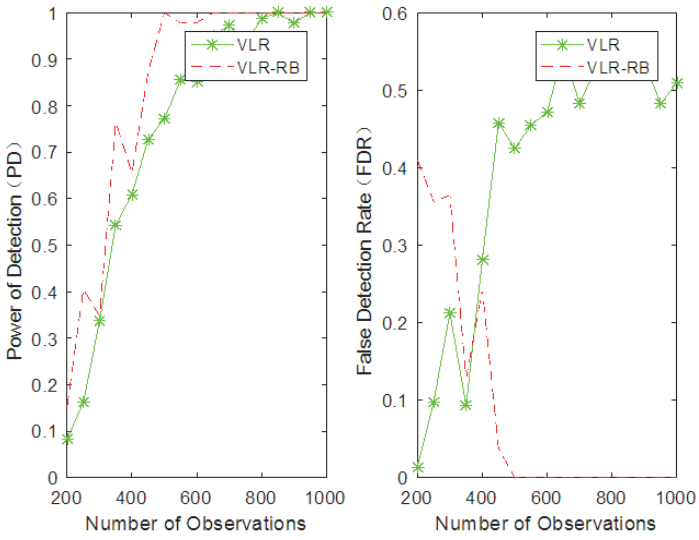

FIGURE 2. The number of parameters to be determined is 2000, SNR is 10, the left side is the detection rate, and the right side is the false detection rate.

Figure 2 is a simulation case where the parameter size is 2000, of which 200 are non-zero variables. The number of simulated observations is $200 : 50 : 1000$. The signal to noise ratio is 10. It can be seen that VLR-RB performance is still superior to VLR in terms of false detection rate. In terms of detection rate, the VLR-RB performance is better than VLR in the range of $450 \sim 800$.

5. **Conclusion.** The variational method can give the analytical solution directly due to its approximate factorization of the posterior distribution, thus avoiding the sampling process of the posterior distribution, and greatly simplifying the inferring cost. Based on the reliability of non-zero variables, this paper proposes a method of dynamic reduction of non-zero variables, thus improving the performance of the algorithm, especially in the false detection rate, and its performance is far superior to the original variational

inference algorithm. The simulation results also show the reliability of the algorithm. Some measures can be used to further improve the accuracy of inference and reduce the calculation time, such as the adoption of dynamic reduction strategy, and improved iterative end judgment.

REFERENCES

[1] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, 1973.
[2] M. N. Williams, C. A. Gomez Grajales and D. Kurkiewicz, Assumptions of multiple regression: Correcting two misconceptions, *Practical Assessment Research & Evaluation*, vol.18, no.11, pp.1-14, 2013.
[3] S. Geisser, Bayesian estimation in multivariate analysis, *Annals of Mathematical Statistics*, vol.36, no.1, pp.150-159, 1965.
[4] P. Dellaportas, J. J. Forster and I. Ntzoufras, On Bayesian model and variable selection using MCMC, *Statistics and Computing*, vol.12, no.1, pp.27-36, 2002.
[5] E. I. George and R. E. McCulloch, Variable selection via Gibbs sampling, *Publications of the American Statistical Association*, vol.88, no.423, pp.881-889, 1993.
[6] S. Xu, Estimating polygenic effects using markers of the entire genome, *Genetics*, vol.163, no.2, pp.789-801, 2003.
[7] X. Cai, Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping, *BMC Bioinformatics*, vol.12, no.1, 2011.
[8] A. Faul and J. J. T. Avenuse, Fast marginal likelihood maximisation for sparse Bayesian models, *Proc. of the 9th International Workshop on Artificial Intelligence & Statistics*, pp.3-6, 2003.
[9] R. B. O'Hara and M. J. Sillanpää, A review of Bayesian variable selection methods: What, how and which, *Bayesian Analysis*, vol.4, no.4, pp.85-117, 2009.
[10] M. W. Seeger and D. P. Wipf, Variational Bayesian inference techniques, *IEEE Signal Processing Magazine*, vol.27, no.6, pp.81-91, 2010.
[11] M. S. Sadough and M. Modarresi, Improved iterative joint detection and estimation through variational Bayesian inference, *AEUE – International Journal of Electronics and Communications*, vol.66, no.5, pp.380-383, 2012.
[12] X. Huang, J. Wang and F. Liang, A variational algorithm for Bayesian variable selection, *arXiv:1602.07640*, 2016.
[13] Z. Dong and Z. Wang, Variational inference of linear regression with non-zero prior means, *Communications in Statistics – Simulation and Computation*, vol.45, no.7, pp.2241-2248, 2014.
[14] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Springer, 2006.