Bayesian Latent Variable Regression of High Dimensional Data with Applications to Process Identification

Hongshu Chen, Bhavik R. Bakshi
Department of Chemical and Biomolecular Engineering, The Ohio State University
Prem K. Goel
Department of Statistics, The Ohio State Univeristy

With the development of modern experimental and analytical technology, it is increasingly common to encounter high dimensional data sets. These data sets may contain large number of variables or samples or both. Traditional modeling methods usually rely on simplifying assumptions of Gaussian noise and prior, and may fail to make the best use of the available data. Meanwhile, since experimenters often have some knowledge about the data set and a likely model, it will be extremely helpful if we can make use of these information in modeling. Bayesian statistics provides a rigorous way to combine the prior information and the likelihood of data. By using Bayes rule, we can get the posterior distribution from prior distribution and likelihood of data. The posterior distribution contains all the information available, thus, the model based on the posterior distribution would capture all the available knowledge and is expected to be better than the model get from traditional methods. This makes Bayesian modeling method a natural choice for modeling complex high dimensional data sets.

A Bayesian modeling method called Bayesian Latent Variable Regression (BLVR)(Nounou et al, 2002) has already been available for some time. It is a linear regression method which can incorporate the prior information, deal with measurement noise in both input variables and output variables, and handle collinearity of input variables. It assumes Gaussian measurement noise for observations and Gaussian prior distribution for observations and model parameters. This method is optimization based. It gets the Maximum A Posteriori (MAP) estimate by optimization routines, i.e., the estimate is the mode of the posterior distribution. This method is most suitable when the dimension of the data set is not very large. When the dimension is large, to solve such a constrained optimization problem with lots of parameters is extremely computationally expensive. As is well known, solving this kind of optimization problem is also problematic because of local minima and convergence issues. Furthermore, since the optimization based BLVR only provides the point estimate, we will lose other information from the posterior distribution, which makes it difficult to provide the confidence interval of our estimate.

To avoid the above problems of optimization, and make BLVR applicable for complex high dimensional data set, a sampling based approach was developed and will be described in this presentation. Instead of solving

optimization problem, this approach uses Monte Carlo approximation to obtain estimates from the sampled posterior distribution. This method uses Markov Chain Monte Carlo (MCMC) (Gamerman, 1997) to draw samples of parameters from the posterior distribution. MCMC is well known in Bayesian statistics community and widely used for Bayesian computing. However, *existing methods have not focused on latent variable regression methods* , which are popular for modeling of process and chemometric data. As long as we know the posterior distribution or the posterior density up to a constant, we can use MCMC to draw samples of this posterior distribution. There are two types of MCMC, Metropolis-Hastings sampling and Gibbs sampling. Gibbs sampling is very useful for high dimensional distribution because it draws samples of each dimension of the parameter vector in sequence. Hence, we use Gibbs sampler in our method. Based on these samples, we obtain the approximate posterior mean, mode and other statistics. This sampling based method is relatively computationally inexpensive and the results are more reliable than the optimization based BLVR. Also it is very easy to provide confidence interval of the estimate and other moments. In principle, this sampling based BLVR can handle any kind of distribution for likelihood and prior, yet the Gaussian assumption could greatly reduce the computation load. Hence, two programs of this sampling based BLVR were developed. One still assumes Gaussian likelihood and prior, since this is often reasonable in many situations and it runs more efficiently. The second approach to be developed in our work does not make any assumptions about Gaussian distributions, and uses Adaptive Rejection Metropolis Sampling (ARMS) (Gilks et al, 1995) method to facilitate the Gibbs sampling.

The complex high dimensional chemical and biological data sets often encountered in high throughput screening applications consists of both continuous and discrete variables. The discrete variables may represent some category and could be without measurement noise. This violates the Gaussian measurement noise assumption made in BLVR, hence, a procedure is developed to separately deal with continuous and discrete variables in Bayesian modeling.

This sampling based BLVR method has been applied to both simulated data set and industrial data set. Table 1 shows the results of a simulated example where there are 15 input variables and the true rank is 10, both the input and output variables are contaminated by measurement noises, the signal to noise ratio is 3. The results for the optimization based BLVR are based on 15 realizations and other results are based on 100 realizations. Other applications include system identification of an industrial distillation column and high throughput screening, which will be described in the presentation.

Table 1. Results for Simulated High Dimensional Data Set

MSE	Y(training)		Y(testing)		X(training)		X(testing)		CPU TIME (s)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
PCR	253.19	23.294	281.24	29.302	3.1897	0.11187	3.1679	0.11667	0.0171	0.004
PLS	242.19	22.17	280.17	31.429	4.5003	0.17886	4.553	0.21065	0.0089	0.003
BLVR-opt(u)	190.68	28.069	353.23	70.397	2.997	0.12417	2.9136	0.10151	194.37	16.96
BLVR-opt(h)	172.52	52.774	284.5	123.21	2.772	1.1854	2.7017	0.87297	1999.8	187.9
BLVR-mcmc(u)	161.74	14.193	269.58	28.178	3.0986	0.11251	3.0104	0.11796	16.172	0.072
BLVR-mcmc(h)	148	14.137	247.36	28.564	2.3404	0.08291	2.4304	0.09328	131.04	0.655

References:

Gamerman D. (1997), Markov Chain Monte Carlo, Chapman & Hall.

Gilks W.R., Best N.G. and Tan K.K.C. (1995), Adaptive Rejection Metropolis Sampling within Gibbs Sampling, Applied Statistics, 44(4):455-472.

Nounou M.N., Bakshi B.R., Goel P.K. and Shen X. (2002), Process Modeling By Bayesian Latent Variable Regression, AICHE Journal, 48(8):1775-1793