Practical Challenges in Bayesian Modeling and Elicitation of Probabilistic Information

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

In laboratories and industry, process information exists in a variety of ways. It may come from the specification of instruments, process operation conditions, expert knowledge etc. Such information is valuable for improving the quality of models, especially when data sets are high dimensional and with heterogeneous distributions and measurement noise. Traditional modeling methods, such as PCR, PLS, fail to incorporate such prior information in modeling. They also often implicitly assume Gaussian or uniform distributions, which may be far from the truth, and , may cause the model quality to be very poor. To make the best use of all the information, a Bayesian modeling method, Bayesian Latent Variable Regression (BLVR) (Nounou et al, 2002) is developed. This method is developed based on Bayes rule, which provides a rigorous way to incorporate data and prior information. In this method, prior distribution and the likelihood of data are combined to get a posterior distribution. It contains all the information available, the Bayesian estimate is obtained based on the posterior distribution and a chosen loss function. Since BLVR has the ability of using more information, it outperforms traditional methods in many cases.

However, there are some practical issues preventing the wide use of BLVR. The original BLVR solves an optimization problem by nonlinear programming (NLP), which is time consuming, particularly for a large number of variables. Also, this optimization-based BLVR can only provide a point estimate, and lacks the ability to readily provide uncertainty information. To overcome these problems, a sampling-based BLVR (Chen et al, 2006) is recently developed. It solves the optimization problem with Markov Chain Monte Carlo (MCMC) (Gamerman, 1997). This is a more practical Bayesian modeling method which is able to handle high dimensional data sets. BLVR usually assumes all variables to be stochastic. Since this assumption may not be valid for some discrete variables, especially those from designed experiments, a BLVR modeling procedure is also developed for hybrid data sets which contain both continuous and discrete variables. This modeling procedure can model the continuous and discrete variables with respective appropriate assumptions. With these advancements, Bayesian modeling methods are much easier to be applied to practical problems.

Nevertheless, applying traditional methods is still more convenient than applying Bayesian methods and the question remains: when should Bayesian

methods be used? That is, when is the extra effort of developing a Bayesian model instead of a conventional model being worth the extra effort? This presentation will demonstrate via theoretical and empirical arguments that if the amount of data available for modeling is small, then Bayesian modeling can perform better. This makes intuitive sense because with the incorporation of prior information, even with small amount of data, BLVR can still get good modeling results. In contrast, traditional methods often fail in this situation. When there are large amount of data available, the effect of prior information will be smaller. The signal to noise ratio also has an effect on the performance on BLVR. When the signal to noise ratio of output variable is much smaller than signal to noise ratio of input variable, BLVR has much better performance than traditional methods.

These effects are demonstrated by applying BLVR to a simulated example. Three types of prior is used for BLVR, uniform prior (u), historical Gaussian prior (h) and true Gaussian prior (t). Table 1 shows the MSE of testing data (200 observations) for PLS, MLPCR and BLVR with different priors normalized by the MSE of PCR when the number of training observations changes. Table 2 shows the MSE of testing data (200 observations) of those methods when the signal to noise ratio in output variables varies while table 3 shows the MSE when the signal to noise ration in input variables varies.

Table 1. Testing MSE of *Y* for different numbers of training data normalized by testing MSE of PCR, SNR in input variables and output variable are both 3, 50 realizations

Number of Training Data	PLS	MLPCR	BLVR(u)	BLVR(h)	BLVR(t)	Normalization Factor (MSE of PCR)
200	1.0090	0.9561	0.9587	0.8939	0.8671	277.7491
100	1.0884	0.9623	0.9663	0.8014	0.7896	306.5697
50	1.2237	0.9572	0.9592	0.6415	0.6413	421.8590
25	1.6583	0.9620	0.9427	0.4835	0.5084	618.3589
10	1.0000	0.9912	N/A	0.1730	0.1777	2467.5

Table 2. Testing MSE of *Y* for different SNR for *Y* normalized by testing MSE of PCR, 200 observations in training, SNR for *X* is 3, 50 realizations

SNR of *Y*	PLS	MLPCR	BLVR(u)	BLVR(h)	BLVR(t)	Normalization Factor (MSE of PCR)
27	0.9868	0.9581	0.9664	0.9201	0.9168	265.1168
9	0.9894	0.9567	0.9602	0.9181	0.9037	257.8655
3	1.0090	0.9561	0.9587	0.8939	0.8671	277.7491
1	1.0677	0.9522	0.9584	0.7798	0.7335	323.3444

Table 3. Testing MSE of *Y* for different SNR for *X* normalized by testing MSE of PCR, 200 observations in training, SNR for *Y* is 3, 50 realizations

SNR of *X*	PLS	MLPCR	BLVR(u)	BLVR(h)	BLVR(t)	Normalization Factor (MSE of PCR)
27	1.1949	0.9784	0.9772	0.6866	0.5976	53.8212
9	1.0678	0.9633	0.9615	0.8147	0.7573	120.1099
3	1.0090	0.9561	0.9587	0.8939	0.8671	277.7491
1	0.9466	0.9234	0.9428	0.9400	0.8804	594.6353

Another challenge in applying Bayesian modeling methods is in obtaining information about the prior and likelihood distributions. Such information often has to be obtained from experts, who may communicate it in a non-probabilistic manner. For example, the range of variation or values of first and second moments may be known a priori. Maximum Entropy (ME) (Jaynes, 1968) methods have been developed in other areas for the elicitation of prior distribution. These approaches can be adapted for getting prior and likelihood distributions for BLVR based on available information. These distributions can also be obtained via an empirical approach called empirical Bayes (Carlin and Louis, 2000). In this approach, even without extra information other than current data set itself, parameters for the prior and likelihood distributions can still be estimated. Noninformative prior can also be used, such as the well known Jeffreys prior (Jeffreys, 1961). With the help of those techniques, prior and likelihood distributions can be elicited in a rigorous manner.

This presentation will discuss a variety of practical case studies on Bayesian modeling including a simulated high dimensional data set, an industrial high throughput screening data set, and other modeling tasks based on laboratory data. Illustrative examples of elicitation of prior distribution will also be presented.

References:

Carlin B.P. and Louis T.A. (2000), Bayes and Empirical Bayes Methods for Data Analysis, Chapman & Hall/CRC

Chen H., Bakshi B.R. and Goel P.K. (2006), Sampling-based Bayesian Latent Variable Regression, Chemical Process Control 7, Alberta, Canada

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

Jaynes, E.T. (1968), Prior Probabilities, IEEE Transactions On Systems Science and Cybernetics, 4(3): 227-241

Jeffreys, H. (1961), Theory of Probability, Oxford University Press

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