Application of Gibbs Sampling in Modelling the Utilization Rate of Raw Materials for Drug Coating

Ge Song[1], Jiahui Yuan[2], Charlie Cheng-Jie Ji[2]

[1] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[2] Shanghai Hongpu Information Technology Co.,Ltd, Shanghai, China

Correspondence: Ge Song, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China. E-mail: 1154333780@qq.com

Abstract

Drug coating is one of the most important processes in the modern pharmaceutical industry. Improving the utilization rate of raw materials (URRM) in the drug coating process is thus important for cost saving and efficiency enhancement. There is little existing research on this topic in the literature of applied statistics. In this paper, motivated by a real dataset collected from a pharmaceutical company in China, we propose to use a novel predictive model that integrates a Bayesian framework with the Gibbs sampling algorithm to characterize the pattern of URRM. Based on certain prior distributional assumptions, the Gibbs sampling algorithm is then applied to sample the posterior distribution of the parameters to obtain more accurate and robust estimation results. By using the proposed method, the drugs can be properly separated into several categories with different patterns of URRM, and the pattern of each category can be properly recognized with selected covariates, which achieves the goals of clustering, variable selection, and regression simultaneously, and provides valuable insights into the improvement of the URRM for drug coating. Numerical studies show that the proposed method works well in practice.

Keywords: Bayesian analysis, clustering, Gibbs sampling, drug coating, utilization rate

1. Introduction

With the rapid development of modern pharmaceutical production, a wide variety of drugs are devised, and their production processes are now highly complicated. In the production process, drug coating is a process by which a thin continuous layer of solid is applied onto the surface of a tablet to protect drugs from dissolution or disintegration in the stomach (Ketterhagen et al., 2017). It is then of great importance to model and to improve the utilization rate of raw materials in the drug coating process for cost saving and efficiency enhancement. In this paper, we focus on the critical drug coating process, and apply modern computational statistical methods to assist in the improvement of the utilization rate of raw materials (URRM) used to coat tablets.

The process of drug coating consists of several sequential stages, including drying of core tablets, feeding of raw materials, preheating, spraying, drying of the coated tablets, cooling, and discharging. The raw materials for drug coating are first feed into a roller. After the aforementioned sequential steps with proper tuning parameter settings, tablets with enteric coating are produced, and the URRM of the tablets can be calculated and collected. The whole drug coating process involves many tuning parameters (e.g., roller rotation speed, air flow temperature) that need to be properly set up during the process. It is crucial to investigate statistically how the tuning parameters affect the final utilization rate, such that accurate and effective decisions and adjustments can be made in order to handle various scenarios.

Intuitively, in order to model the relationship between the URRM and the tuning parameters, a simple multiple linear regression model (Kutner et al., 2005) can be applied, i.e. $Y = X^T \beta + \varepsilon$, where Y is the URRM, X is the vector of tuning parameters/covariates, β is the corresponding vector of coefficients that can be estimated by using the famous least square method (Lai et al., 1978), and ε is the random error term. The major drawback of such approach is that it is based on an impractical assumption that different drugs follow the same regression model, i.e. share the same β. However, in practice, the relationship between the URRM and the tuning parameters varies from drug to drug. In such cases, using the same linear model is inaccurate, which may fail to properly describe such complicated relationship. On the other hand, it is also unrealistic to establish regression model for each type of drug, given the fact that the number of drug types is large, and the observations of certain drugs are collected in limited quantities, bringing efficiency loss and

the well-known "curse of dimensionality" problem. Based on the above discussion, it is desirable to develop a new approach to model the relationship between the URRM and the tuning parameters while considering the similarities and differences between different drugs. Ideally, drugs can be properly clustered into several groups, and the URRM pattern of each group can be characterized.

In this paper, motivated by a real dataset that comes from a real production line for drug coating of a pharmaceutical company in China, we introduce a novel Bayesian analytical framework for modelling the relationship between the URRM for drug coating and the tuning parameters. Specifically, based on prior distributional assumptions, the Gibbs sampling algorithm is applied to sample the posterior distributions of the tuning parameters to obtain more accurate and robust estimation results. The Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm applied for obtaining a sequence of observations which are approximated from multivariate conditional probability distributions, in cases when direct sampling is nontrivial. It was originally applied in the field of image processing (Geman and Geman, 1984). Gelfand and Smith (1990) showed how the Gibbs sampler can be used to solve various Bayesian inference problems. A detailed tutorial was given by Casella and George (1992). Nowadays, the Gibbs sampler has been widely applied in many important research frontiers including natural language processing (Chen et al., 2018), gene expression (Wang et al., 2019), etc. By using the proposed method, the drugs can be properly clustered into several categories, and the pattern of each category between the URRM and the tuning parameters can be effectively characterized with selected influential covariates, which provides valuable insights into the improvement of the URRM for drug coating.

The remainder of the paper is organized as follows. Section 2 gives the proposed Bayesian framework and the methods based on Gibbs sampling for the modelling of URRM pattern. Section 3 presents the real dataset. Section 4 present the numerical results for analyzing the real dataset. The conclusions are given in Section 5.

2. Statistical Methodology

2.1 Problem Formulation

Suppose we have real observations of M types of drugs. For each type of drug, there is at least one observation. The model of concern can then be written as

$$Y_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_i + \varepsilon_{ij}, \ i = 1, \dots, M, j = 1, \dots, n_i, \tag{1}$$

where Y_{ij} denotes the observed value of URRM for the jth observation of the ith drug, $\boldsymbol{X}_{ij}$ denotes the p-dimensional covariates after certain feature engineering, $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^T$ is the vector of p-dimensional coefficients for the ith drug, ε_{ij} is the random error term, and $n_i \geq 1$ is the number of observations for the ith drug. Noting that different drugs may share the same pattern (i.e. their coefficient vectors are equal), our goals are then to find the drugs with identical utilization rate pattern, to cluster the drugs into several groups, and to estimate the regression coefficients of each group. Unfortunately, such goals can hardly be achieved by using classic regression analysis methods such as least squares regression. The main reason is that $\boldsymbol{\beta}_i$s need to be estimated separately, and the estimation results are extremely unrobust or even invalid in cases when n_i is extremely small (e.g. $n_i = 1$ or $n_i \ll p$). For instance, in the real dataset described in Section 2.1, many drugs only have one observation. To solve this problem, in this paper we utilize a Bayesian modelling framework to learn the hidden relationship between different drugs. Then by using the Gibbs sampling algorithm, the drugs can be clustered properly, and the coefficients for each cluster can be estimated.

We now give the basic assumptions of the Bayesian framework as follows.

1. Each type of drug belongs to a specific category with unique URRM pattern. Denote by H_i the category of the ith drug. Its possible values could be $1, \dots, K$, where K is the number of categories. The prior probability of $H_i = k$ is $\pi_{ik} = 1/K, \ k = 1, \dots, K$.

2. The prior distribution of β_{ij} is $\beta_{ij} \sim N(\mu_{kj}, v)$, where k denotes the category to which the ith drug truly belongs.

3. Denote by Z_{kj} the status of μ_{kj}. $Z_{kj} = 1$ if μ_{kj} is influential and $Z_{kj} = 0$ otherwise. The prior distribution of μ_{kj} given Z_{kj} is

$$\mu_{kj}|Z_{kj} \sim I(Z_{kj} = 0)\delta_0(\mu_{kj}) + I(Z_{kj} = 1)N(0, \tau^2).$$

4. The prior distribution of Z_{kj} is $Z_{kj} \sim Bernoulli(\omega)$.

5. The prior distributions of ω, v, and τ^2 are respectively $\omega \sim Beta(a_0, b_0)$, $\tau^2 \sim InvGa(s_1, s_2)$, and $v \sim InvGa(\gamma_1, \gamma_2)$.

2.2 Parameter Estimation via Gibbs Sampling

Based on the problem formulation in Section 2.1, the parameters in the Bayesian framework that need to be estimated can be summarized as $\{\beta_{ij}, \mu_{kj}, Z_{kj}, \omega, v, \tau^2, \pi_{ik}\}$. In such a situation, the Gibbs sampling algorithm is often used to sample posterior distribution in a Bayesian framework. It is a MCMC algorithm that is applicable when the joint distribution of the parameters is unknown, but the conditional distributions of the parameters is easy to sample from (Geman and Geman, 1984). Specifically, the Gibbs sampler generates a sample from the distribution of each parameter in turn, conditional on the current values of the other parameters. It has been well demonstrated that the sequence of samples constitutes a Markov chain, whose stationary distribution is just the sought-after joint distribution (Gelman et al. 2013). Based on the above discussion, in this paper we utilize the Gibbs sampling algorithm to sample the posterior distributions of parameter to be estimated.

Before implementing the Gibbs sampling algorithm, the distributions of the parameters conditional on the other parameters must be given. After derivations, we have the following results:

$$Z_{kj}|Y, \beta, H, \omega, v, \tau^2 \sim Bernoulli\left(\frac{R_{kj}}{1 + R_{kj}}\right), k = 1, \dots, K, j = 1, \dots, p,$$

where $\log(R_{kj}) = \log\left(\frac{\omega}{1+\omega}\right) + \frac{1}{2}\log\left(\frac{v}{v+m_k\tau^2}\right) + \frac{m_k\tau^2(\sum_{\{i:\, H_i=k\}}\beta_{ij}/m_k)^2}{2v(v+m_k\tau^2)}$ and $m_k = \sum_{i=1}^{M} I(H_i = k)$,

$$\mu_{kj}|Z_{kj}, Y, \beta, H, \omega, v, \tau^2 \sim I(Z_{kj} = 0)\delta_0(\mu_{kj}) + I(Z_{kj} = 1)N\left(\frac{\tau^2}{v + m_k\tau^2}\sum_{\{i:\, H_i=k\}}\beta_{ij}, \frac{\tau^2 v}{v + m_k\tau^2}\right),$$

$$k = 1, \dots, K, j = 1, \dots, p,$$

$$\beta_{ij}|\mu_{kj}, Y, \beta, H, \omega, v, \tau^2 \sim N\left(\left(\frac{1}{v} + X_{\cdot j}^T X_{\cdot j}\right)^{-1}\left[\frac{1}{v}\mu_{kj} + X_{\cdot j}^T\left(Y_i - X_{[-j]}^T \boldsymbol{\beta}_{i[-j]}\right)\right], \left(\frac{1}{v} + X_{\cdot j}^T X_{\cdot j}\right)^{-1}\right),$$

$$i = 1, \dots, M, j = 1, \dots, p,$$

where $X_{\cdot j}$ is the jth column of the feature matrix $(X_{i1}, \dots, X_{in_i})^T$, $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$, $\boldsymbol{\beta}_{i[-j]}$ is $\boldsymbol{\beta}_i$ except that the jth element equals zero,

$$\pi_{ik}|Y, \beta, Z, v = \frac{e^{-\frac{\sum_{j=1}^{p}(\beta_{ij} - \mu_{kj})^2}{2v}}}{\sum_{k'=1}^{K} e^{-\frac{\sum_{j=1}^{p}(\beta_{ij} - \mu_{k'j})^2}{2v}}}, i = 1, \dots, M, k = 1, \dots, K$$

$$v|Y, \beta, \mu, H \sim InvGa\left(\gamma_1 + \frac{Mp}{2}, \gamma_2 + \frac{\sum_{i,j}(\beta_{ij} - \mu_{kj})^2}{2}\right),$$

$$\tau^2|Y, \beta, Z, \mu, H \sim InvGa\left(s_1 + \sum_{k,j}\frac{Z_{kj}}{2}, s_2 + \sum_{k,j}\mu_{kj}^2 Z_{kj}\right),$$

and

$$\omega|Y, Z, H \sim Beta\left(a_0 + \sum_{k,j} Z_{kj}, b_0 + \sum_{k,j}(1 - Z_{kj})\right).$$

Based on the above conditional distributions of the parameters, the Gibbs sampling algorithm can now be applied to sample the posterior distributions of the parameters, and the expected value of any parameter can be approximated by averaging over all the samples after steady state. The specific algorithm is summarized as follows.

Step 1. Set $t = 0$ and initial value $\boldsymbol{P}^{(t)} = (\beta_{ij}^{(t)}, \mu_{kj}^{(t)}, Z_{kj}^{(t)}, \omega^{(t)}, v^{(t)}, \tau^{2(t)}, \pi_{ik}^{(t)}) \triangleq (P_1^{(t)}, \dots, P_7^{(t)})$.

Step 2. Obtain the next sample $\boldsymbol{P}^{(t+1)}$. To do this, we sample each component of $\boldsymbol{P}^{(t+1)}$ from its distribution conditioned on all other components sampled so far. Specifically, to sample $P_i^{(t+1)}$, we update it based on the distribution specified by $P_i^{(t+1)}|P_1^{(t+1)}, \dots, P_{i-1}^{(t+1)}, P_{i+1}^{(t)}, \dots, P_7^{(t)}$.

Step 3. Repeat Step 2 B times.

Based on the stationary distributions of each parameter after Gibbs sampling, the URRM patterns of the drugs can be

properly divided into K groups, and the coefficients of each group can be estimated. It should be noted that K is a hyperparameter that can be selected by using the Bayesian information criterion (BIC). The BIC is a criterion for model selection among a finite set of possible models, and the model with the smallest BIC value is preferred. The BIC is typically defined as (Wit, Heuvel, & Romejin, 2012):

$$BIC = -2\ln(L) + k * \ln(n),$$

where L is the maximized likelihood function of the considered model, n is the total number of observations, and k is the number of parameters to be estimated.

3. Real Dataset

The real dataset used in the paper comes from a real production line for drug coating of a pharmaceutical company in China. The dataset was recorded from Nov 2017 to Mar 2018, and consists of $n = 809$ observations of $M = 85$ types of drugs. In each observation, the utilization rate of drug coating materials and the influential covariates are recorded. The number of observations for different drug varies, which is displayed in Figure 1. From Figure 1, it can be observed that some drugs only have limited observations, causing challendges for further statistical analysis. The fitted density function and the Q-Q plot of the utilization rate values after standardization are shown in Figure 2, from which we can observe that the normality assumption is significantly invalid. Based on the above discussions, it is unreasonable to build a single linear regressional model for all the observations, or to build separate linear regressional models for each type of drug. To solve this problem, in the following section, we apply the proposed Bayesion framework based on Gibbs sampling in Section 2 to achieve clustering and parameter estimation.
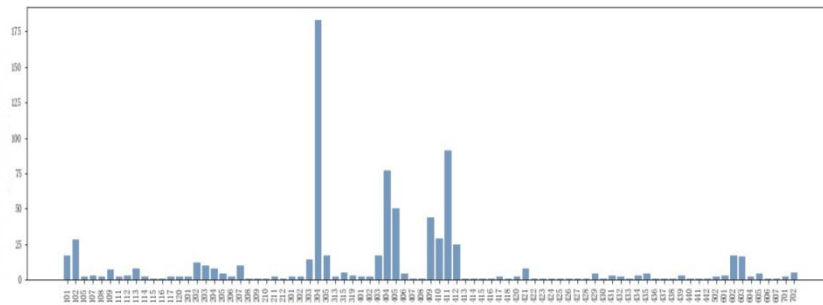

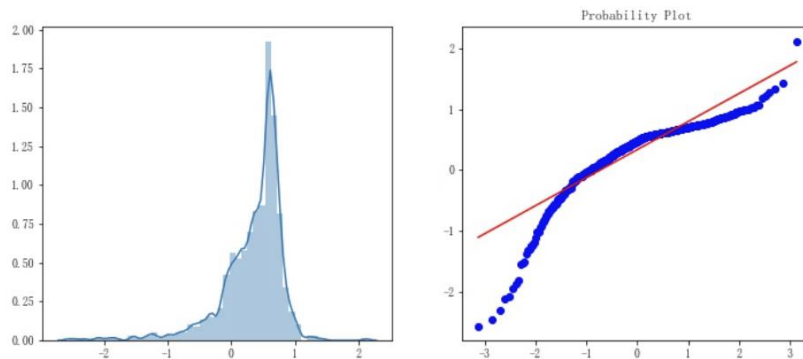Figure 1. The number of observations for each type of drug

Figure 2. The fitted density function and the Q-Q plot of the observed values of URRM

4. Results

In the current section, we apply the proposed method based on the Gibbs sampling algorithm to model the real dataset mentioned in Section 2. Originally, for each observation, there are 201 covariates. Based on exploratory analysis, we found that many covariates barely change among different drugs. To abandon such redundant covariates, we apply the idea of sure independence screening (SIS) method proposed by Fan and Lv (2008) for dimension reduction and feature selection. The basic idea of SIS is to select the most contributed covariates based on certain correlation criterion and to abandon the others. To implement SIS, we first calculate the Spearman's rank correlation coefficient between the response and each covariate:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where d_i is the difference between the two ranks of each observation and n is the number of observations. It should be noted that unlike Pearson's correlation, there is no requirement of normality for Spearman's correlation and hence it is nonparametric and more robust (Corder and Foreman, 2014). In this paper, we choose those covariates with $\rho \geq 0.1$, and 56 covariates that are significantly correlated with the utilization rate are remained for the following statistical analysis. The correlation matrix of the 56 covariates are displayed in Figure 3. From the plot, it can be concluded that certain weak correlation structure exists and that the effect of multicollinearity can be neglected.
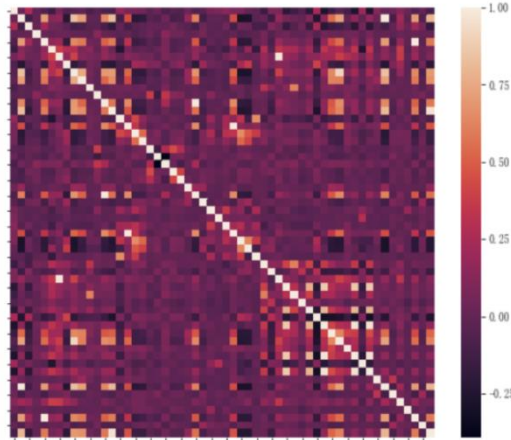

Figure 3. The correlation matrix of the 56 covariates

Now, we apply the proposed method to analyze the real dataset. Under model (1) and the Bayesian framework in Section 2.1, we apply the Gibbs sampling algorithm presented in Section 2.2 to estimate the model parameters. The mean square error (MSE) curves of the Gibbs sampling algorithm with various choices of K are displayed in Figure 4. From the plots, it can be observed that as the number of iterations increases, the MSE curves converge to very low levels and tend to be steady, indicating that the proposed procedure is effective. We then compute the BIC values of the model with various choices of K, and the results are summarized in Table 1. From the table, we can observe that the model with $K = 3$ has the smallest BIC value, which indicates that the drugs can be properly divided into 3 different groups with significantly different URRM patterns.
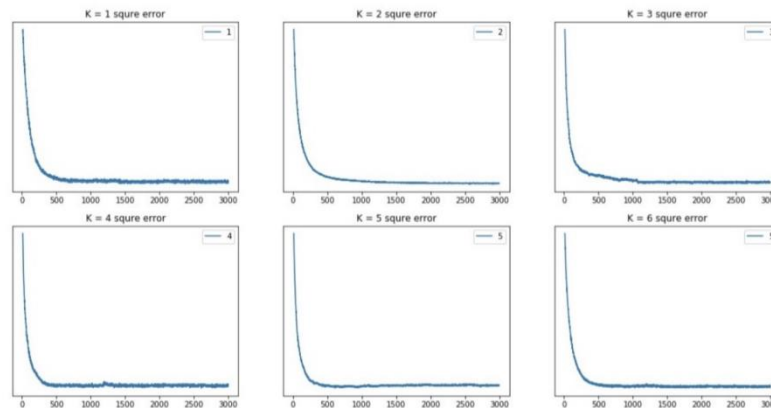

Figure 4. The MSE curves of the Gibbs sampling algorithm with various choices of K

Table 1. The BIC values for various choices of K

K	BIC
1	1128.15
2	1035.23
3	**978.76**
4	1061.35
5	1083.02
6	1197.11

Based on the steady states of the parameters after using the Gibbs sampling algorithm, we can obtain the specific pattern for each type of drug when $K=3$. Specifically, the first pattern includes 32 types of drugs and 291 observatons in total, the second pattern includes 20 types of drugs and 93 observations in total, and the third pattern includes 33 types of drugs and 425 observations in total. The histograms and fitted density functions of the utilization rate of the three groups are displayed in Figure 5. From the plots, it can be observed that the three patterns differ significantly, and the utilization rate of each pattern approximately follows normal distribution. The number of selected covariates of each group are shown in Table 2. It can be seen from the table that after the Gibbs sampling, the goal of high-dimensional variable selection is also achieved. For demonstration, we choose five covariates, and the corresponding estimated coefficients are shown in Table 3. Based on the estimated coefficients, the pattern of how the covariates affect the utilization rate is clear.
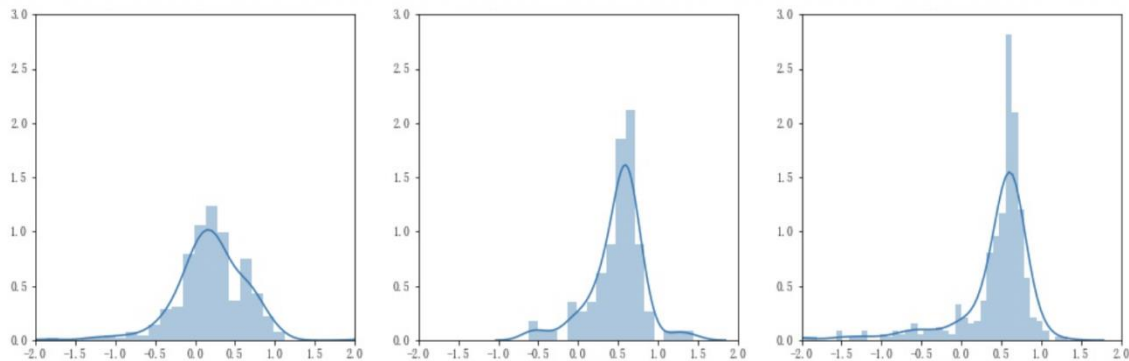

Figure 5. The histograms and fitted density functions of the utilization rate of the three groups

Table 2. Number of selected covariates of each group

	Group I	Group II	Group III
Number of selected covariates	14	7	8

Table 3. Estimated coefficients of some covariates. "-" means that the covariate is not selected

Covariate	Group I	Group II	Group III
Cooling time	-	-	-0.08
Peristaltic pump speed when cooling	0.029	0.078	-
Spray time	0.042	-	0.083
Motor frequency	0.073	-	-
Peristaltic pump speed when discharging	-0.043	-	-

Now, we compare the proposed method with the classic multiple linear regression method. In other words, it is assumed in such a situation that different drugs share the same relationship between the URRM and the tuning parameters. The model can then be written as $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n$. Specifically, we apply the 10-fold cross validation approach, in which the original dataset is randomly partitioned into 10 equal size subset. Of the 10 subsets, a single subset is the validation data for evaluating the fitted model, and the remaining 9 subsets are used as training data to train the model. The MSE values by using the 10-fold cross validation are summarized in Table 4. From the table, we can observe that the results are very unstable for many subsets, indicating that the basic assumption of linear regression fails, and that the drugs do not follow the same distribution. Based on the three groups by using Gibbs sampling, the fitting residuals of the three groups of the two approaches are shown in Figure 6. From the plots, it can be clearly observed that the residuals of linear regression tend to be much larger than those of Gibbs sampling. Finally, we can conclude that, instead of directly applying linear regression to all the data, it is more appropriate and effective to utilize the proposed Gibbs sampling approach.

Table 4. The MSE values of the linear regression method by using the 10-fold cross validation approach

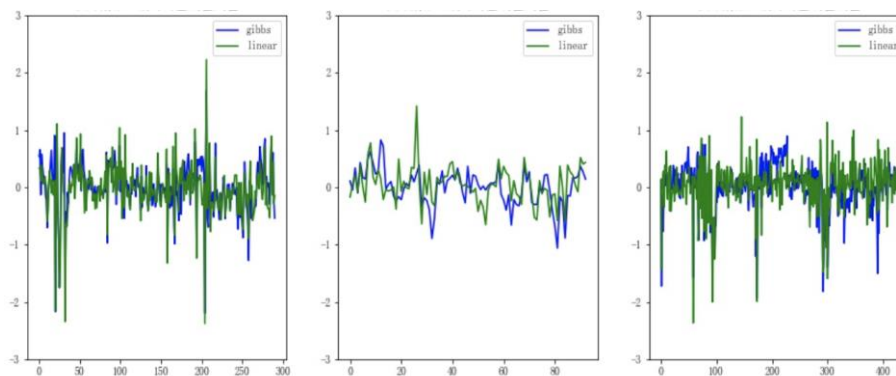| ID	1	2	3	4	5	6	7	8	9	10
MSE	0.57	0.46	0.18	0.29	0.25	0.31	0.17	0.08	0.31	0.23

Figure 6. The fitting residuals of the three groups by using Gibbs sampling and linear regression

5. Conclusions

In this paper, in order to model the functional relationship between the utilization rate of raw materials for drug coating and the tuning parameters during the coating process, we propose a Bayesian framework that integrates the Gibbs sampling algorithm. Based on certain prior distributional assumption, Gibbs sampling is applied to sample the posterior distributions of the parameters for obtaining accurate and robust parameter estimation results. Applying the proposed method on a real dataset, the drugs are clustered into three groups, and the patterns of utilization rate of the groups distinct from each other. Also, the influential covariates for each group are selected with estimated coefficients, by which engineers can properly recognize how the covariates affect the utilization rate.

Some important issues remain that need to be solved. First, the proposed approach is based on the assumption of normality. In certain cases, such assumption may be invalid. In future research, the application of the Gibbs sampler for other distributions (e.g. uniform) should be investigated. Second, convergence diagnostics for the Gibbs sampling when applying in the current problem should be investigated systematically in the future.

References

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174. https://doi.org/10.1080/00031305.1992.10475878

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Gelfand, A. E., & Smith, A. P. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press. https://doi.org/10.1201/b16018

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741. https://doi.org/10.1109/TPAMI.1984.4767596

Ketterhagen, W., Aliseda, A., Am Ende, M., Berchielli, A., Doshi, P., Freireich, B., & Prpich, A. (2017). Modeling tablet film-coating processes. In *Predictive Modeling of Pharmaceutical Unit Operations* (pp. 273-316). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100154-7.00010-7

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models.* New York: McGraw-Hill Irwin.

Lai, T. L., Robbins, H., & Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *PNAS, 75*(7), 3034–3036. https://doi.org/10.1073/pnas.75.7.3034

Su, J., Xu, J., Qiu, X., & Huang, X. (2018). Incorporating discriminator in sentence generation: A Gibbs sampling method. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/11990

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., ... Li, B. (2019). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nature neuroscience*, *22*(5), 691-699. https://doi.org/10.1038/s41593-019-0382-7

Wit, E., Heuvel, E. V. D., & Romeijn, J. W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217-236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

Copyrights