Journal of Bioprocessing & Biotechniques

Liu and Fang, J Bioprocess Biotech 2018, 8:3
DOI: 10.4172/2155-9821.1000326

Research Article Open Access

Glutamate Fermentation Process Model Based on Gams with Double Penalty Approach

Chunbo Liu[1,2*] **and Xiangling Fang**[3]

[1] *Faculty of Science, School of Human Science, University of Western Australia, Stirling Highway Crawley, Australia*

[2] *Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Lihu Avenue, Wuxi, Jiangsu, China*

[3] *State Key Laboratory of Grassland Agro-ecosystems, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, China*

***Corresponding author:** Chunbo Liu, Faculty of Science, School of Human Science, University of Western Australia, Stirling Highway Crawley, Australia, Tel:+61893864234; E-mail: l.chunbo@yahoo.com.au

Received date: April 26, 2018; **Accepted date:** May 07, 2018; **Published date:** May 14, 2018

Abstract

There are many fermentation variables which can influence fermentation product during fermentation process, including fermentation time, temperature, pH, oxygen uptake rate (OUR), carbon dioxide evolution rate (CER), dissolved oxygen (DO) and stirring speed. For understanding the relationship between fermentation product and associated fermentation variables, one current interest is the investigation of significant fermentation variables to construct an interpretable and stable fermentation model. In this study, the significant variables in the fermentation process of glutamate were selected based on the generalized additive models through a shrinkage approach. The GAM2 which is based on less fermentation variables after variable selection has the same value of adjusted R-square (0.972) and deviance explained (97.6%) as GAM1 which includes all the measuring fermentation variables. Results showed that the proposed approach not only can identify the significant variables in the fermentation process of glutamate, but also can improve the performance of the fermentation model.

Keywords: Variables selection; Fermentation process; Glutamate; Generalized additive models

Introduction

Fermentation process is a complicated process that includes many variables such as fermentation time, temperature, pH, oxygen uptake rate (OUR), carbon dioxide evolution rate (CER), dissolved oxygen (DO) and stirring speed [1]. Such variables can influence the yield of fermentation product. For example, different DO levels influenced the yield of glutamate and lactate dehydrogenase activities in the fermentation process of glutamate [2]; the yield of glutamate was influenced by chosen fermentation variables such as DO, OUR and CER [3]; and a tight control on the fermentation temperature was needed to get an industrial standard production of glutamate [4]. Moreover, the fermentation time was suggested as an important variable among the process variables [1].

Due to the influences of different variables in fermentation process, it is essential to conduct variables selection to get the significant variables that influence the fermentation process first before constructing an effective model. From a pragmatic point of view, it aims at determining which variables have the strongest influences on the yield of fermentation production, whereas from a statistical perspective it represents a means to achieve a balance between goodness of fit and parsimony.

Several statistical methods have been used for variable selection, including interactive variable selection [5], uninformative variable elimination [6], interval partial least squares (PLS) [7], iterative PLS [8], genetic algorithms (GA) [9] and PLS-bootstrap [10]. However, these algorithms are developed for specific applications and are based on different principles that are impossible to know in advance which algorithm will be the best suited one for a particular data set. One problem for variable selection is the possibility of over fitting (i.e., removing an excessively high number of variables will cause the model to perform well on calibration but not on external validation). Another problem might be the inclusion of noisy and irrelevant variables. When one or both of these situations occurring, less robust models will be obtained [11,12]. The multi-way partial least-squares modeling approach can provide an accurate inference of the quality variables in the fermentation process that is difficult to measure online [13]. This approach enables the prediction of the final yield of product and the detection of faults that influence the fermentation productivity, but like artificial neural network models, it requires a huge amount of experimental data, which are often hard to define and access, and lack of data often leads to a reduction of the process behavior estimation.

This study applied the generalized additive models (GAM) [14-16] to select the significant variables in the fermentation process of glutamate. Within the class of GAM, the proposals avoid having to use nonparametric testing methods for which there is no general reliable distributional theory. In addition, the selection of significant fermentation variables and modeling in glutamate fermentation process can be carried out in one step as opposed to the selection procedures that involve an exhaustive search of all possible models.

Materials and Methods

Strain and fermentation conditions

The strain *Corynebacterium glutamicum* S9114 was used in this study. The seed medium consisted of K_2HPO_4 1.5, glucose 25, $MnSO_4$ 0.005, $FeSO_4$ 0.005, $MgSO_4$ 0.6, corn slurry 25 and urea 2.5 (in g/L). The initial fermentation medium consisted of glucose 140, K_2HPO_4 1.0, $FeSO_4$ 0.002, $MgSO_4$ 0.6, $MnSO_4$ 0.002, thiamine 5.0×10^{-5}, corn slurry 15 and urea 3.0 (in g/L). Initial pH of all the above media was set at 7.0 to 7.2. During the fermentation process, 25% (w/w) ammonia

water was added to the liquid medium to maintain the pH at ~7.1 the added ammonia water also provided the nitrogen source required by glutamate synthesis during the fermentation process [3]. To ensure glucose concentration above 15 g/L during the fermentation process, it was added to the fermenter according to the requirement of substrate.

Generalized additive models with double penalty approach

A GAM is a generalized linear model (GLM) [17,18], with a linear predictor involving smooth functions of covariates:

$$g\{E(Y_i)\} = X_i^* \theta + \sum_J^J f_j(x_{ji}), i = 1, .., n \quad \text{.....(1)}$$

Where g is a specific link function, Yj is a univariate response that follows an exponential family distribution, X_i^* is the row of X* , which is the model matrix for any strictly parametric model components, with the corresponding parameter vector ɵ*, and fj being smooth functions of the covariates xj, which may be vector covariates (so xji denotes the Ith observation of the jth covariate). fj are subject to Identifiability constraints such as

$$\sum_i f_{j \geq}(x_{ji}) = 0 \vee j$$

Equation (1) can be estimated as a GLM, but to avoid over fitting it is necessary to estimate such a model by penalized maximum likelihood estimation, in which roughness measures are used to control over fit. In practice, the penalized likelihood is maximized by penalized iteratively reweighted least squares (P-IRLS), where the GAM is fitted by iterative minimization of the problem:

$$\left\| \sqrt{W^{[k]}} \left(Z^{[k]} - X\beta \right) \right\| + \sum \lambda_j \beta^T S_j \beta w.r.t.\beta \quad \text{.....(2)}$$

Where k is the iteration index, $z^{(k)} = X\beta^{[k]} + G^{[k]}\left(y - \mu^{[k]}\right), \mu_i^{[k]}$, is the current model estimate of $E(Y_i)$, $G^{[k]}$is a diagonal matrix such that $G_{ii}^{[k]} = g\left(\mu_i^k\right)$, $W^{[k]}$is a diagonal matrix given by $W_{ii}^k = \left[G_{ii}^{[k]2} V\left(\mu_i^{[k]}\right)\right]^{-1}$where $V\left(\mu_i^{[k]}\right)$gives the variance of Yi to within a response distribution scale parameter, ϕ, X includes the columns of X* and columns representing the spline bases for the fj, while β contains ɵ and all the smooth coefficient vectors, βj. Sj are matrices of known coefficients such that the terms in the summation measure the roughness of the smooth functions. λj are smoothing parameters that control the trade-off between fit and smoothness. The generic smoothing penalty matrix Sj associated with a smooth term of a GAM can be decomposed as:

$$W_j \wedge_j W_j^T \quad \text{.....(3)}$$

Where Wj is an eigenvector matrix associated with the jth smooth function, and $\wedge_j$ the corresponding diagonal eigenvalue matrix? The fact that a part of the spline basis space deals with the penalty null space implies that $\wedge_j$ contains zero eigenvalues. This may be problematic if variable selection has to be carried out. For instance, assuming that the jth smooth component is a nuisance function and that we use a penalty matrix as defined above during the model fitting process. Even if λj goes to infinity there will not be any guarantee that the smooth term will be suppressed completely (i.e., estimated as zero). In order to circumvent this difficulty, an extra penalty can be produced that penalizes only functions in the null space of the penalty so that a smooth component can be completely removed. Specifically, by decomposition of equation (3), an extra penalty can be formed as follows:

$$S_j^* = W_J^* W_J^{*T} \quad \text{.....(4)}$$

Where W_J^* is the matrix of eigenvectors corresponding to the zero eigenvalues of$\wedge_j$A GAM can be fitted subjecting each component function to a double penalty of the form:

$$\lambda_j \beta^T S_j \beta + \lambda_j^* \beta^T S_j^* \beta \quad \text{.....(5)}$$

Where both λ_j and λ_j will now have to be estimated. By introducing a penalty for the null space, smoothing parameter estimation (that is part of GAM fitting) can completely remove terms from the model.

Parameter	Parameter estimate	Standard error	T value	Pr >
Intercept	42.6837	0.2975	143.5	< 2e-16***
Source	Degrees of freedom estimate	Ref. df	F	P-value
s(Time)	5.8602	7.0166	274.119	< 2e-16***
s(Temp)	1.0663	1.5667	0.985	0.35854
s(pH)	0.4778	0.8159	0.076	0.72982
s(OUR)	4.5602	5.6267	7.057	2.29e-06***
s(CER)	1.9393	2.4429	3.665	0.02041 *
s(DO)	4.6654	5.6903	3.482	0.00358**
		R-square (adjusted)		Deviance explained
Value		0.972		97.60%

Table 1: Smoothing model 1 (GAM1) analysis using generalized additive models (GAMs) according to double penalty approach. *P<0.05; **P<0.01; ***P<0.001.

To reiterate the basic idea, any spline type smoother can be decomposed into two component functions: a component in the null space of the penalty, and a component in the range space of the penalty. The first term in equation (5) penalizes only function components in the range space, but can shrink these to zero, while the second term in equation (5) penalizes only function components in the null space, but can shrink these too to zero. For example, in the case of the usual cubic spline penalty, the second term in equation (5) would penalize straight line components to zero, while the first term would penalize (towards zero) function components representing departure from straight line behavior [19,20].

As a special case of GAMs, considering the continuous predicant Y and $X_1,\ldots,X_p$ covariates X_k, we formulate Y as a sum of unspecified smooth functions of the individual covariates by an additive model

$$Y=c+s(X_1,m_1)+s(X_2,m_2)+\ldots+s(X_p,m_p)+\mathcal{E} \quad (6)$$

Where Ɛ is assumed to be normally distributed random errors having constant feature and its mean value is zero; $s(X_i,m_i)$ (i=1,…,p) are smooth functions with efficient degree of freedom mi ≥ 1 to be estimated from data.

Results

This study performed variable selection in the fersmentation process of glutamate using the GAM approach. GAMs allow the relationship between the predicant and smooth functions of covariates to be modelled in an additive framework. Table 1 shows the model GAM1 according to double penalty approach in fermentation variables selection. From Table 1, the glutamate fermentation model GAM1 can be defined as

Glutamate=42.68+s(Time,5.86)+s(Temp,1.07)+s(pH,0.48)+s(OUR, 4.56)+ S(CER,1.94)+S(DO,4.67) (7)

To confirm the validity of GAM1 defined by equation (6), the diagnostic examination of the model was conducted (Figure 1). Results showed that the sampled data and residuals generated by GAM1 were close to normally distributed data (Figures 1a and 1b), hence GAM1 followed the generalized additive model assumption. The residuals randomly scattered around zero with no particular trend and pattern (Figure 1c), suggesting GAM1 can describe the in luences of different fermentation variables on glutamate production (Figure 1d). No serious in luential outliers existed between responses and itted values (Figure 1d).

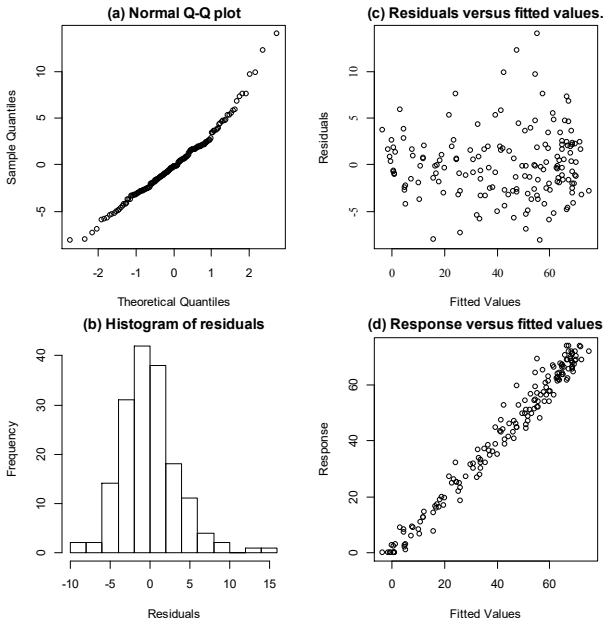

Figure 1: The diagnostics of GAM1 model: (a) normal Q-Q plot; (b) histogram of residuals. (c) residuals versus fitted values. (d) responses against fitted values.

To understand the individual influence of fermentation variables on the production of glutamate using GAM1, the influence of individual fermentation variable was predicted by conditioning the other variables constant at their mean values (Figure 2). Smooth function estimates were obtained by applying.

Parameter	Parameter estimate	Standard error	T value	Pr >
Intercept	42.6837	0.2996	142.5	<2e-16***
Source	Degrees of freedom estimate	Ref.df	F	P-value
s(Time)	5.843	7.01	556.237	< 2e-16***
s(OUR)	4.472	5.537	6.932	3.33e-06***
s(CER)	1.997	2.515	3.828	0.01613*
s(DO)	4.824	5.862	3.828	0.00154**
		R-square (adjusted)		Deviance explained
Value		0.972		97.60%

Table 2: Smoothing model 2 (GAM2) analysis using generalized additive models (GAMs) according to double penalty approach. *P<0.05; **P<0.01; ***P<0.001.

The double penalty approach with restricted maximum likelihood (REML) to fit a GAM on the glutamate fermentation dataset. The predicted results revealed the presence of non-linear relationships between the production of glutamate and the selected variable while which cannot be observed using a parametric approach. For example, a non-linear influence of time on the production of glutamate throughout the fermentation process was observed when the other fermentation variables are kept constant (Figure 2a). The production of glutamate increased rapidly during the first 20 h, and then increased slowly during the period of 20 to 34 h, and then became stable or even decreased (Figure 2a). The smoothest of time, OUR, CER and DO exhibited a strong non-linear behavior (Figures 2a, 2d, 2e and 2f), and hence these terms cannot be entered the model in a parametric manner.

While constructing the model GAM1, the influence of each fermentation variable was indicated as P-value. The P-values of Time, OUR, CER and DO were less than 0.05 and the P-values of the other variables were greater than 0.05 (Table 1). Thus, it was reasonable to exclude the non-significant variables to construct a more reliable fermentation mode [16]. The model GAM2 was constructed by only including the significant fermentation variables (Time, DO, OUR and CER) according to double penalty approach (Table 2), and can be defined as

$$Glutamate=42.68+s(Times,5.84)+s(OUR,4.47)+s(CER,2.00)+s(DO,4.82) \quad (7)$$

To confirm the validity of GAM2 defined by equation (7), the diagnostic examination was conducted (Figure 3). The quintile-to-quintile plot and the histogram of residuals showed the sampled data and residuals generated by GAM2 were close to normally distributed data (Figures 3a and 3b), hence GAM2 followed the GAM assumption. The residuals appeared randomly scattering around zero with no particular trend and pattern (Figure 3c), implying GAM2 has capability to describe the effects of different variables on glutamate production, and = no obvious influential outliers were observed between responses and fitted values (Figure 3d).

The individual influence of the significant fermentation variables (Time, OUR, CER and DO) on the production of glutamate using GAM2 was then determined (Figure 4). The influence of individual variable was estimated by conditioning the other fermentation variables constants at their mean values. There were obvious non-linear influences between the selected fermentation variables and the production of glutamate. For example, there were nonlinear influences of DO on the production of glutamate (Figure 4d). A minimal production of glutamate was observed when DO was maintained around 40%, indicating that keeping the DO at this level may result in a low production of glutamate, which is consistent with the previous report that glutamate production was at a low level when the DO concentration was low [3].

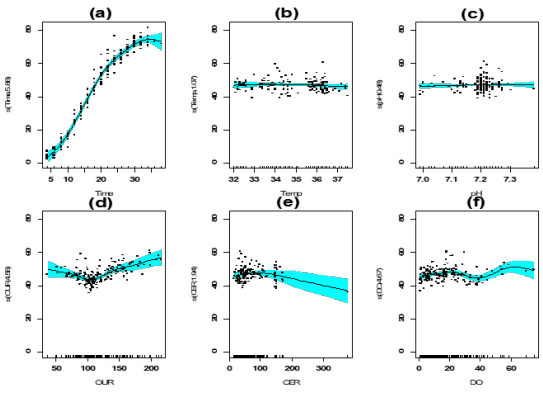

Figure 2: The individual influence of different fermentation variables on the production of glutamate predicted using GAM1. The predicted influence of (a) time, (b) temperature (Temp), (c) pH, (d) oxygen uptake rate (OUR), (e) carbon dioxide evolution rate (CER) and (f) dissolved oxygen (DO) on the production of glutamate when the other fermentation variables were kept constant at their mean values. Numbers in parentheses indicate the equivalent degrees of freedom for the smooth curves of fermentation variables. The black curve represents the predicted smooth functions of fermentation variables. The usual 95% Bayesian confidence intervals for glutamate production are shaded in green. The rug plot at the bottom of each graph indicates the value of fermentation variable.

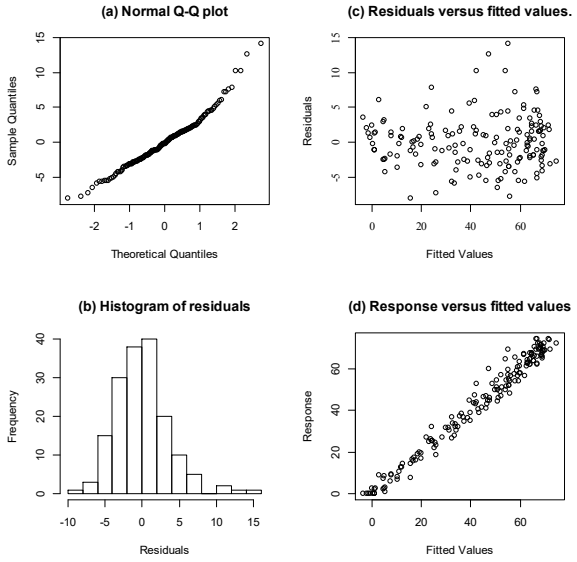

Figure 3: The diagnostics of GAM1 model: (a) normal Q-Q plot; (b) Histogram of residuals. (c) residuals versus fitted values. (d) responses against fitted values.

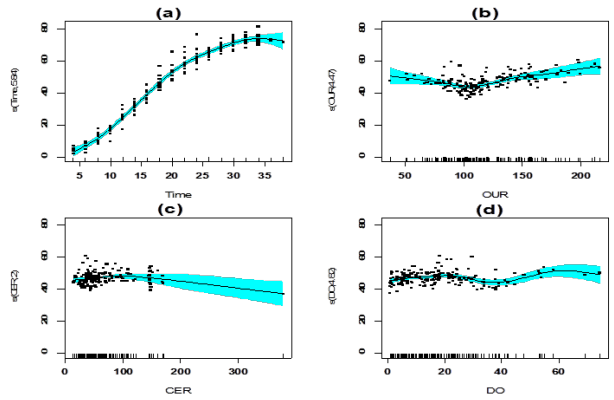

Figure 4: The individual influence of different fermentation variables on the production of glutamate predicted using GAM2. The predicted influence of (a) time, (b) temperature (Temp), (c) pH, (d) oxygen uptake rate (OUR), (e) carbon dioxide evolution rate (CER) and (f) dissolved oxygen (DO) on the production of glutamate when the other fermentation variables were kept constant at their mean values. Numbers in parentheses indicate the equivalent degrees of freedom for the smooth curves of fermentation variables. The black curve represents the predicted smooth functions of fermentation variables. The usual 95% Bayesian confidence intervals for glutamate production is shaded in green. The rug plot at the bottom of each graph indicates the value of fermentation variable.

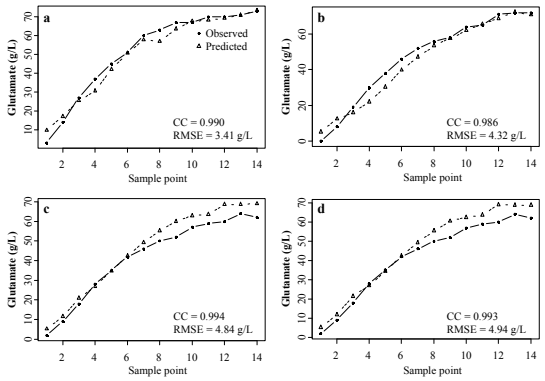

Figure 5: The observed and predicted production of glutamate from two different fermentation batches. Predicted production of glutamate based on (a) GAM2 and (b) GAM1 based on GAM2 and GAM1 using online recorded data of fermentation variables from the first fermentation batch. Predicted production of glutamate based on (c) GAM2 and (d) GAM1 using online recorded data of fermentation variables from the second fermentation batch.

Discussion

The comparison between the two models showed that GAM2 had the same value of adjusted R-square (0.972) and deviance explained (97.6%) as GAM1 (Tables 1 and 2), even GAM2 was based on less fermentation variables after variable selection. The comparison of two model check results showed that GAM2 had the same modeling capacity as GAM1 (Figures 1-4). To further check the performance of GAM2 constructed by only including the significant fermentation variables, following comparisons between the observed production and predicted production of glutamate using GAM1 and GAM2 were conducted by using online recorded data of fermentation variables from two different fermentation batches (Figure 5).

For the first fermentation batch, there was a highly significant correlation ($P<0.01$) between the observed production and predicted production of glutamate based on GAM2, with a correlation coefficient (CC) of 0.990 and a root-mean-square error (RMSE) of 3.41 g/L (Figure 5a). The correlation between the observed production and predicted production of glutamate based on GAM1 was also highly significant ($P<0.01$), with a CC of 0.986 and a RMSE of 4.32 g/L (Figure 5b). In terms of the second batch, the observed production and predicted production of glutamate based on GAM2 was highly ($P<0.01$) correlated, with a CC of 0.994 and a RMSE of 4.84 g/L (Figure 5c). There was also a highly significant correlation ($P<0.01$) between the observed production and predicted production of glutamate based on GAM1, with a CC of 0.993 and a RMSE of 4.94 g/L (Figure 5d).

For both fermentation batches, the observed production and predicted production of glutamate based on GAM2 exhibited a higher CC and also a smaller RMSE than that based on GAM1. And thus, the performance of the GAM2 was better than GAM1, suggesting the proposed approach not only can identify the significant variables in the fermentation process of glutamate, but also can improve the performance of the model.

In conclusion, this study focused on fermentation variable selections in the fermentation process of glutamate based on generalized additive models. The model conducted after variable selection by including only the significant fermentation variables exhibited better performance compared with the model constructed by including all the fermentation variables. The proposed approach not only can identify the significant variables in the fermentation process, but also can improve the performance of the model. In fact, this approach can be extended to other processes, following Equation (6), because the smooth function s(Xi,mi) gives the ability to examine the relationship between covariate Xi and the predicant Y, the "data-driven" estimated s(Xi,mi) is therefore most helpful to describe unknown relationship between the covariate Xi and the predicant Y when there is no prior knowledge. So this approach is appealing since it has the properties of stability and prediction, and variable selection and modeling can be carried out in one single step as opposed to the selection procedures that involve an exhaustive search of all possible models. Furthermore, it avoids having to use testing methods for which there is no general distributional theory.

Financial Support

This research was funded by the National High Technology Development 863 Program grant numbers 2006AA020301-11, China.

Acknowledgments

We thank Dr Yun Li at Mathematics, Informatics and Statistics Leeuwin Centre, CSIRO, Australia for his constructive suggestions on this manuscript.

References

1. Ibanoglu S, Ibanoglu E (2001) Modelling of natural fermentation in cowpeas using response surface methodology. J Food Eng 48: 277-281.
2. Gao P, Lu JB, Duan ZY, Mao ZG, Shi ZP (2005) Effects of dissolved oxygen on the key enzyme in glutamate fermentation. Food and Fermentation Industries 31: 72.
3. Zhang C, Shi Z, Gao P, Duan Z, Mao Z (2005) On-line prediction of products concentrations in glutamate fermentation using metabolic network model and linear programming. Biochem Eng J 25: 99-108.
4. Uy D, Delaunay S, Goergen JL, Engasser JM (2005) Dynamics of glutamate synthesis and excretion fluxes in batch and continuous cultures of temperature-triggered *Corynebacterium glutamicum*. Bioproc Biosyst Eng 27: 153-162.
5. Fredrick L, Paul G, Stefan R, Svante W (1994) Interactive variable selection (IVS) for PLS (I): theory and algorithms. J Chemometr 8: 349-363.
6. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, et al. (1996) Elimination of uninformative variables for multivariate calibration. Anal Chem 68: 3851-3858.
7. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, et al. (2000) Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl Spectrosc 54: 413-419.
8. Osborne SD, Künnemeyer R, Jordan RB (1997) Method of wavelength selection for partial least squares. Analyst 122: 1531-1537.
9. Leardi R (2000) Application of genetic algorithm-PLS for feature selection in spectral data sets. J Chemometr 14: 643-655.
10. Lazraq A, Cléroux R, Gauchi JP (2003) Selecting both latent and explanatory variables in the PLS1 regression model. Chemometr Intell Lab 66: 117-126.
11. Barman I, Kong CR, Dingari NC, Dasari RR, Feld MS (2010) Development of robust calibration models using support vector machines for spectroscopic monitoring of blood glucose. Anal Chem 82: 9719-9726.
12. Swierenga H, Wülfert F, De Noord OE, De Weijer AP, Smilde AK, et al. (2000) Development of robust calibration models in near infra-red spectrometric applications. Anal Chim Acta 411: 121-135.
13. Zhang H, Lennox B (2004) Integrated condition monitoring and control of fed-batch fermentation processes. J Process Contr 14: 41-50.
14. Wood SN (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc 99: 673-686.
15. Wood SN (2006) On confidence intervals for generalized additive models based on penalized regression splines. Aust NZ J Stat 48: 445-464.
16. Wood SN (2012) On p-values for smooth components of an extended generalized additive model. Biometrika 100: 221-228.
17. Cordeiro GM, McCullagh P (1991) Bias correction in generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological) 53: 629-643.
18. Hastie T, Tibshirani R (1987) Generalized additive models: some applications. J Am Stat Assoc 82: 371-386.
19. Marra G, Wood SN (2011) Practical variable selection for generalized additive models. Comput Stat Data An 55: 2372-2387.
20. Wood SN, Pya N, Safken B (2016) Smoothing Parameter and Model Selection for General Smooth Models. J Am Stat Assoc 111: 1548-1563.