Chemometrics and Intelligent Laboratory Systems 128 (2013) 144–152

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Target-oriented overall process optimization (TOPO) for reducing variability in the quality of herbal medicine products

Bing Xu [a], Zhaozhou Lin [a], Zhisheng Wu [a], Xinyuan Shi [a,b,*], Yanjiang Qiao [a,b,*], Gan Luo [a]

[a] Research Center of TCM Information Engineering, Beijing University of Chinese Medicine, Beijing 100029, PR China
[b] Engineering Research Center of Key Technologies for Chinese Medicine Production and New Drug Development, Ministry of Education of People's Republic of China, Beijing 100029, PR China

ARTICLE INFO

Article history:
Received 25 April 2013
Received in revised form 23 July 2013
Accepted 16 August 2013
Available online 30 August 2013

Keywords:
Target-oriented overall process optimization (TOPO)
Quality variability
PLS regression
Bayesian approach
Probability trajectory
Herbal medicine

ABSTRACT

This paper presents a new strategy, target-oriented overall process optimization (TOPO), which can be used to assure the consistent quality in herbal medicine products. The methodology of TOPO includes four parts, target definition, data pretreatment, process modeling and overall process optimization. The Bayesian approach is integrated into the optimization step. The mechanism of TOPO involves optimizing multiple units of the production system step by step, giving each unit optimal operating conditions consistent with the quality target. The effects of TOPO were assessed using the descriptive statistics of the Bayesian posterior predictive distribution and the final target achievement. The probability trajectory was adjusted to monitor and optimize the production process. The proposed TOPO strategy was successfully applied to a seven-unit manufacturing process used to produce Lonicerae Japonicae extract. Results demonstrated that TOPO could keep the production process in line with the predefined target and reduce the variability of the final products.

1. Introduction

Herbal medicines and their derivatives have been used extensively for thousands of years in many Asian countries, such as China, Japan and Korea. For example, in China's National Essential Drug List (2012 Edition), 203 out of 520 total recommended drugs were herbal medicine preparations. In Europe and North America, herbal medicines have seen increasing use over the past few decades, largely in the form of dietary supplements, functional foods or health products. It is estimated that nearly 80% of the world's people still rely on herbal medicines for health related benefits [1]. According to the World Health Organization (WHO), the global market for herbal remedies and supplements was about U.S. $83 billion in 2008, and it continues to grow exponentially [2].

These trends lead to the imperative requirements of quality control of herbal medicine products, because the quality of herbal products is linked directly to their efficacy and safety [3]. Generally, quality control of herbal products involves identification of the starting material, details of the manufacturing process, and standards for the finished product. Currently, there are many technologies that can be used to control the quality of these products, including chemical profile methods, biological methods, on-line analytical tools, integrated evaluation approaches, etc. [4–7]. However, these methods mainly refer to the analytical aspects of quality control, and very few studies have evaluated the engineering aspects.

The major difficulties and challenges in the quality control of herbal medicines lie in the variability of the herbal material, the degree of which depends on factors such as the location of growing, the time of harvest, preprocessing methods and storage conditions [8–10]. Natural variability may be introduced into the manufacturing process, causing fluctuations between different batches [11,12]. Under these circumstances, conventional analytical techniques can identify the variations, but they cannot maintain quality consistence across herbal products. For this reason, there is an urgent need to address the problem of variability from the production point of view, because the quality of herbal products is affected by the manufacturing processes to a large extent [13].

Nowadays, many technological systems have been adapted from the chemical and pharmaceutical industries, and used to modernize the ways in which herbal products are processed. These techniques include solvent extraction, macroporous resin column chromatography, high-speed counter-current chromatography and various dosage preparation methods [14–18]. The overall manufacturing process of herbal medicine often consists of multiple processing units, which could also be called multistage batch process [19]. Through the serial processing stages, the desired quality is transformed from the starting materials to the final products step by step. Downstream units are influenced by upstream units. All parameters of the manufacturing process more or

* Corresponding authors at: School of Chinese Materia Medica, Beijing University of Chinese Medicine, South of Wangjing Middle Ring Road, Chaoyang District, Beijing City, 100102, PR China. Tel.: +86 10 84738621; fax: +86 10 84738661.
E-mail addresses: shixinyuan01@163.com (X. Shi), yjqiao@263.com (Y. Qiao).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.08.008

less affect the final product [20]. Although the manufacture of herbal products is now subjected to the Good Manufacturing Practice (GMP) standards, there is a lack of strategies to coordinate the relationships between different production units and to optimize the process parameters, in order to assure consistent product quality. Nevertheless, developing such strategies is challenging, because of the complex kinetics and thermodynamics of these processes, and of the unclear mechanisms by which active ingredients are transferred.

In order to tackle the abovementioned difficulties, systematic methodologies including multistage batch process modeling, monitoring, control and optimization, are needed to effectively control and improve the product quality during the multistage manufacturing process [21–23]. And some new concepts, e.g. plant wide optimization (PWO), could also be applied to facilitate optimal operation conditions consistent with the quality objectives of the large scale production system [24–26]. Based on these thoughts, a new strategy named target-oriented overall process optimization (TOPO) is brought forward to address the problem of variability in the concentrations of active ingredients in herbal products. The rest part of this article is organized as follows. First, a brief introduction is made about the mathematical foundations of TOPO. Then, technical details of the TOPO are illustrated. The method of expanding PLS modeling proposed by A. Pomerantsev et al. [27] will be employed in the TOPO strategy together with the Bayesian optimization technique. After that, effects of the proposed TOPO strategy were tested in a seven-unit manufacturing process used to produce the Lonicerae Japonicae extract. Finally, a summary of this paper is provided.

2. Overview of TOPO

The target-oriented overall process optimization (TOPO) strategy proposed in this study integrates the expanding PLS regression method and the Bayesian approach together. The word "expanding" here means a series of PLS models are built at the end of each stage, where the quality variables can also be predicted [27,28]. For a multistage manufacturing system with many process variables, it is time consuming and even impossible to optimize all combinations of variables. Therefore, with the help of established series of PLS models, optimization operation is designed to start from the second stage to the last stage. For a certain stage to be optimized, all the historical data, as well as the measurements from its previous stages are utilized. The goal of TOPO is to consistently provide optimal assurance for the product quality meeting defined specifications.

2.1. Mathematical fundamentals of TOPO

2.1.1. Partial least square regression

Partial least square (PLS) regression is a popular chemometric tool and is widely applied in industrial research, development and production. In the presence of historical production data which are often none-designed and not orthogonal, or contain noisy and collinear process variables, PLS method deserves the property to grasp hidden relationships between process variables and quality variables.

The main purpose of PLS regression is to build a linear model relating the independent data $\mathbf{X}$ (size $m \times n$, m is the number of observations and n is number of variables) with the response data $\mathbf{Y}$ (size $m \times q$, q is the number of responses):

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

where $\mathbf{B}$ (size $n \times q$) is the matrix of regression coefficients; $\mathbf{E}$ is a noise term and has the same dimension with $\mathbf{Y}$. The basic assumption of PLS method is that there is a small number of latent variables (LVs) [29], which are linear combination of the original $\mathbf{X}$ variables, and can capture most of information in the calibration data for predicting the responses. These latent variables are also known as $\mathbf{X}$-scores, by which the PLS model can be written as follows:

$$\mathbf{Y} = \mathbf{TV} + \mathbf{E} \tag{2}$$

where $\mathbf{T}$ (size $m \times p$, and p corresponds to the number of latent variables) is the matrix of $\mathbf{X}$-scores; $\mathbf{V}$ (size $p \times q$) is the matrix of inner regression coefficients for $\mathbf{T}$. PLS method produces the $\mathbf{T}$ through a weighting matrix $\mathbf{W}$ and loading matrix $\mathbf{P}$:

$$\mathbf{T} = \mathbf{XW}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{3}$$

where $\mathbf{W}$ is a $n \times p$ matrix and is computed to maximize the covariance between the scores and responses. $\mathbf{P}$ is a $n \times p$ matrix. Two popular algorithms can be employed to compute the scores matrix, i.e. the NIPALS algorithm and SIMPLS algorithm [30,31]. Once $\mathbf{T}$ is obtained, the inner regression coefficients $\mathbf{V}$ in Eq. (2) is estimated by regressing $\mathbf{Y}$ on $\mathbf{T}$ via ordinary least square regression (OLS) procedures:

$$\widehat{V} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}. \tag{4}$$

Given a new sample vector $\mathbf{x}$ (size $n \times 1$), the $\mathbf{x}$ is firstly projected onto the latent space, generating a score vector $\mathbf{t}$ (size $p \times 1$):

$$\mathbf{t} = \left(\mathbf{W}^{\mathrm{T}}\mathbf{P}\right)^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{x} \tag{5}$$

Then, the corresponding response $\widehat{y}$ could be predicted according to Eq. (2):

$$\widehat{y} = \mathbf{t}^{\mathrm{T}}\widehat{V}. \tag{6}$$

Prediction can also be made directly from original variables of the sample $\mathbf{x}$ according to Eq. (1), where $\mathbf{B}$ is estimated as follows:

$$\widehat{B} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{V}. \tag{7}$$

The number of latent variables determines the complexity of the model. Therefore, the test set validation method and cross validation method (e.g. leave one out, LOO) are usually introduced to help select the optimal number of LVs and to test the predictive ability during the model construction [32]. Some chemometric indicators, such as root mean square error of calibration (RMSEC), root mean square error of cross validation (RMSECV), root mean square error of prediction (RMSEP), ratio of performance to deviation (RPD), predicted residual error sum square (PRESS) and bias, are often used to assess the performance of the established model [33,34]. For example, the PRESS index is calculated as:

$$\mathrm{PRESS} = \sum_{i=1}^{m}\left(\widehat{y}_i - y_i\right)^2. \tag{8}$$

The PRESS value shows the sum of squares of deviation between the predicted and the true property of the validation sample i during cross-validation, and it decreases as the LVs increase. When the PRESS value tends to be constant, the optimum number of LVs is obtained.

2.1.2. Bayesian approach

Under the framework of Bayes' theorem, the Bayesian inference combines the prior knowledge about the model parameters with information from measured data [35]. In process optimization, the Bayesian approach provides a natural way of making inference on future response $\widetilde{y}$ from its posterior predictive distribution. Using the classical

(i.e. Jeffreys) non-informative prior on **B** and **E** of the linear model in Eq. (1), namely:

$$p(\mathbf{B},\mathbf{E}) \propto p(\mathbf{B})p(\mathbf{E}) \tag{9}$$

where $p(\mathbf{B}) \propto$ constant, and $p(\mathbf{E}) \propto |\mathbf{E}|^{-(q+1)/2}$, the prior density for **B** and **E** must be of the form:

$$p(\mathbf{B},\mathbf{E}) \propto |\mathbf{E}|^{-(q+1)/2}. \tag{10}$$

The posterior predictive density for $\tilde{y}$ can be obtained by the Student t distribution with v degree of freedom (df) [36]:

$$p(\tilde{y}|\mathbf{X},\mathbf{Y},x) \propto \left(1+\frac{1}{v}\left(\tilde{y}-x^{\mathrm{T}}\widehat{B}\right)^{\mathrm{T}}\mathbf{H}\left(\tilde{y}-x^{\mathrm{T}}\widehat{B}\right)\right)^{\frac{v+q}{2}} \tag{11}$$

where $v = m-p-q+1$ (12)

$$\mathbf{H} = \frac{vS^{-1}}{1+x^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}x}; \tag{13}$$

$$S = \left(\mathbf{Y}-\mathbf{X}\widehat{B}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}\widehat{B}\right); \tag{14}$$

$\widehat{B}$ is estimated by Eq. (7). It can be seen that the posterior predictive distribution of response accounts for the uncertainty from both the unknown parameters **B** and the unknown error **E** of the model. However, the prerequisite for estimating the posterior density of $\tilde{y}$ is that there is a small amount of collinearity among different variables in the **X** data. Otherwise, the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ may not exist [37]. If this prerequisite is not satisfied, the collinear variables should be carefully excluded from analysis or combined into a single index beforehand. Some more sophisticated technique (e.g. the Bayesian latent factor regression [38]) may be considered, but this is beyond the scope of this article.

Monte Carlo simulation method could be used to easily approximate the posterior prediction distribution of $\tilde{y}$. Firstly, a vector **u** (size $1 \times q$) is sampled from a normal distribution $\mathbf{N}(0, \mathbf{H}^{-1})$, and an independent random number c is simulated from the chi-square distribution with df v. Then, the estimation for $\tilde{y}$ is calculated from:

$$\tilde{y} = \mathbf{u}\sqrt{\frac{v}{c}} + x^{\mathrm{T}}\widehat{B}. \tag{15}$$

After a large number of **u** and c are simulated, the posterior prediction distribution of $\tilde{y}$ can be acquired and used for inference.

2.2. Implementation procedures of TOPO

With the basic mathematical methods described earlier, 4 major phases and 10 steps are designed and included in the proposed approach, as shown in Fig. 1. The technical details for each phase and step are illustrated as follows:

Phase I Target definition

Step 1 The optimization target is defined as quality specifications, and is denoted as **O**. The specifications for each response should include the lower and upper limits, or at least one of them. For the jth stage (or unit) to be optimized, the optimization problem can be expressed as:

$$\mathbf{x_{jopt}} = \arg\max \mathbf{P}\left(\tilde{y}_j \in \mathbf{O}|\mathrm{Data}, \mathbf{x_{jcon}}\right)$$

$$\mathbf{x_{jcon}} \in \mathbf{L_j} \tag{17}$$

where Data represents all the available information. $\mathbf{L_j}$ is the acceptable range of controlled process variables in the jth stage. $P(\cdot)$ is the Bayesian posterior predictive probability of $\tilde{y}_j$ meeting the target. The objective of optimization is to find $\mathbf{x_{jopt}}$ with maximum $P(\cdot)$.

Phase II Data pretreatment

Step 2 Assume that there is a collection of historical batch production data, and the number of process parameters does not change over time. Then, the process variables are organized into a data matrix **X**, while the quality variables are organized to form a data matrix **Y**.

Step 3 All data are scaled properly in order to eliminate dimension differences.

Step 4 The data arrangement is an important step before process modeling. According to the expanding PLS modeling technique, the whole process is supposed to involve k production stages. Each stage can be represented by a data block $\mathbf{X_j}$ ($1 \le j \le k$). Then, the whole data matrix **X** is composed by a series of data blocks from $\mathbf{X_1}$ to $\mathbf{X_k}$. At the jth stage, all the previous data blocks $\mathbf{X_1}$ to $\mathbf{X_{j-1}}$, together with $\mathbf{X_j}$, are arranged to form the joint data block $\mathbf{X_{(j)}}$.

$$\mathbf{X_{(j)}} = \left[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X_j}\right]. \tag{18}$$

The k joint data blocks will be used in the modeling step.

Phase III Process modeling

Step 5 Select the best number of latent factors by LOO cross validation (or K-fold cross validation), as well as chemometric indicators.

Step 6 By relating the joint data block $\mathbf{X_{(j)}}$ and the responses **Y**, the process PLS model $\mathrm{XY_j}$ at jth stage can be established as:

$$\mathbf{XY_j}: \mathbf{Y} = \mathbf{X_{(j)}}\mathbf{B_j} + \mathbf{E_j}. \tag{19}$$

Along with enlargement of the joint data blocks, k PLS models are built (i.e. expanding PLS modeling).

Step 7 The parameters used in the jth process PLS model, i.e. $\mathbf{B_j}$, are stored for the following computation.

Phase IV Process optimization

Step 8 For a new batch process, the process variables are denoted as vectors **x**. Suppose that the jth stage of a new batch is going to be optimized. The process variables $\mathbf{x_j}$ are firstly divided into the observed ones $\mathbf{x_{jobs}}$ and controlled ones $\mathbf{x_{jcon}}$. The observed variables could be read, but not be adjusted. Only the controlled variables can be manipulated and optimized. Therefore, the observed variables together with variables $\mathbf{x_{(j-1)}}$ from previous ($j-1$) stages are combined into the fixed variables $\mathbf{x_{jfixed}}$. According to the optimization range and control precision of the controlled variables of the jth stage, an exhaustive method is conducted to form a grid of controlled variables (i.e. $\mathbf{L_j}$ in Eq. (17)), the row number of which is denoted as N. Then, each row in $\mathbf{L_j}$ is combined with $\mathbf{x_{jfixed}}$, forming the matrix $\mathbf{D_j}$ for exploration:

$$\mathbf{D_j} = \left[\mathbf{x_{jfixed}} \otimes \mathbf{g}, \mathbf{L_j}\right] = \begin{pmatrix} \mathbf{x_{jfixed}} & \mathbf{x_{jcon-1}} \\ \mathbf{x_{jfixed}} & \mathbf{x_{jcon-2}} \\ \vdots & \vdots \\ \mathbf{x_{jfixed}} & \mathbf{x_{jcon-N}} \end{pmatrix} \tag{20}$$

where $\otimes$ is Kronecker product operator; **g** is $N \times 1$ column matrix with all elements that equal one.

Step 9 By applying the approach introduced in Section 2.1.2, the Bayesian posterior predictive distribution for the

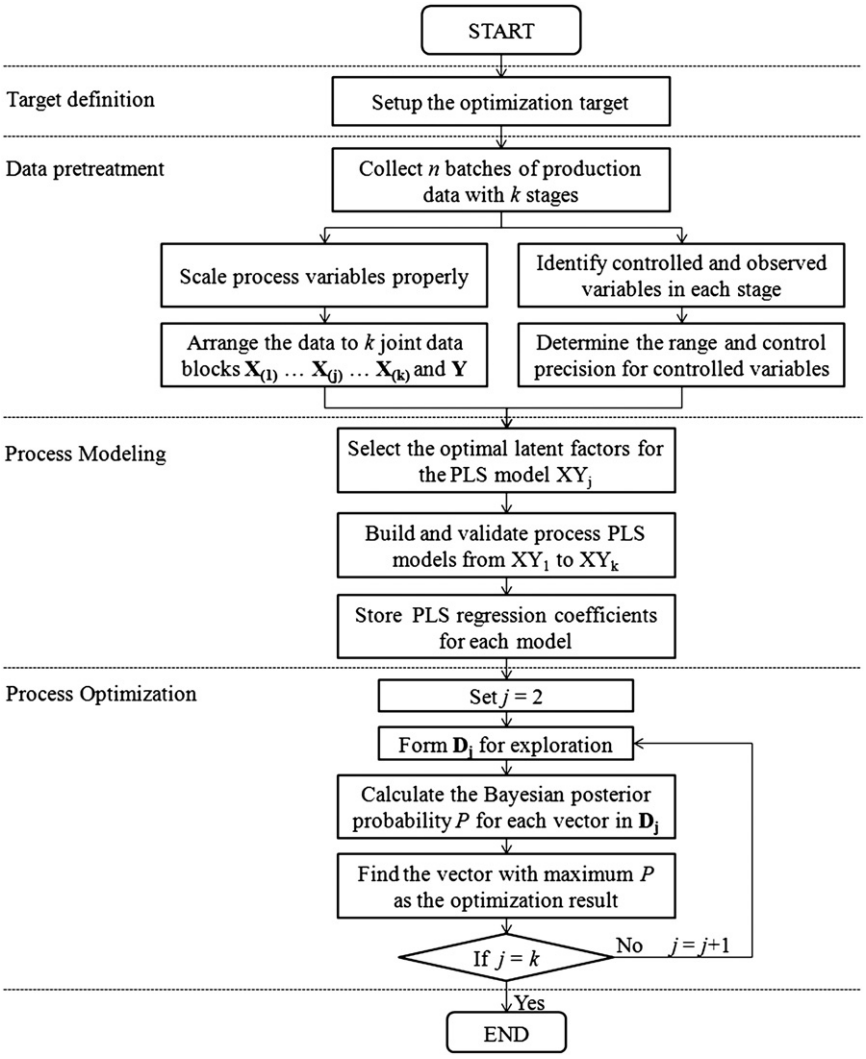

Fig. 1. Implementation procedures of TOPO.

row vector in $\mathbf{D_j}$ can be approximated. In this paper, the number of Monte Carlo simulation was set to 10,000. Then, $P(\cdot)$ for each row in $\mathbf{D_j}$ is calculated from proportion. The best setup $\mathbf{x_{jopt}}$ corresponds to $\mathbf{x_{jcon}}$ with the maximum probability.

Step 10 Based on the series of expanding PLS models, steps 8 and 9 are repeated from the second stage to the end stage.

3. Real world data

3.1. Process description

The overall production process of Lonicerae Japonicae extract used as an example in this paper belongs to the pharmaceutical manufacturing system of Qingkailing Injection, as specified in *Chinese Pharmacopoeia* (2010 Edition, Volume I). The whole process consists of seven key unit operations ($\mathbf{X_1}$, $\mathbf{X_2}$, …, $\mathbf{X_7}$), such as water extraction, concentration, alcohol precipitation, and alcohol recovery. It is also a typical herbal medicine production process. During the seven stages of production, the Chlorogenic acid in the starting material (Lonicerae Japonicae Flos) is gradually separated and transferred into the Lonicerae Japonicae liquid extract which is stored and used in the subsequent dosage preparation process.

Generally, 28 process parameters ($x_1, x_2, \ldots, x_{28}$) and one quality attribute (y) which represents the concentration of Chlorogenic acid determined by HPLC, are taken into consideration in the overall production process. Every variable has a specific meaning, for example, x_1 represents the quality of input material. Each unit has a different number of process parameters, as shown in Table 1. Some units have two types of process variables, the controlled and the observed.

Fig. 2 shows an abstract illustration of the overall process. Each stage is denoted in the form of a rectangle. The length of each rectangle is proportional to the number of process variables, which is given in the middle of it. The process variable data block can be expanded as the process continues, forming a series of joint data blocks ($\mathbf{X_{(1)}}$, $\mathbf{X_{(2)}}$, …, $\mathbf{X_{(7)}}$),

Table 1
Properties of parameters at every stage of the production process.

Stage	Unit operation	Process parameters		
		Range	Controlled	Observed
1	Water extraction	x_1–x_5	None	x_1–x_5
2	Concentration	x_6	x_6	None
3	75% alcohol precipitation	x_7–x_{13}	x_7, x_8, x_{11}, x_{13}	x_9, x_{10}, x_{12}
4	Alcohol recovery	x_{14}–x_{16}	x_{14}, x_{15}	x_{16}
5	85% alcohol precipitation	x_{17}–x_{23}	$x_{17}, x_{18}, x_{21}, x_{23}$	x_{19}, x_{20}, x_{22}
6	Alcohol recovery	x_{24}–x_{26}	x_{24}, x_{25}	x_{26}
7	Purification	x_{27}–x_{28}	x_{27}, x_{28}	None

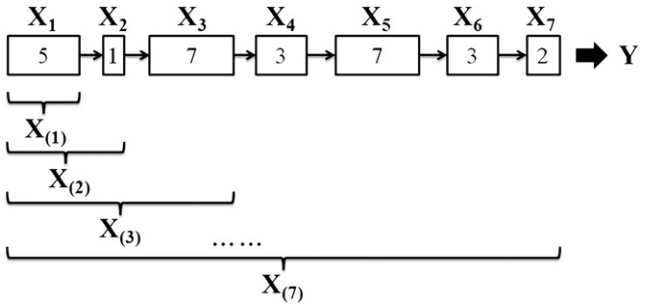

Fig. 2. Graphical description of the overall production process of Lonicerae Japonicae extract.

according to Step 4 of TOPO. At the end of any stage, the process engineer may assess about the finished process operations and decide whether to take any corrective action for the next set of operation. In this sense, TOPO could be used to facilitate reasonable solutions.

3.2. Target product quality profile

The U.S. Food and Drug Administration (FDA) published a set of guidelines describing the basic components of the target product profile (TPP) in 2007 [39]. In general, TPP covers the overall clinical safety and efficacy intent of drug development [40]. TPP is qualitative and implicit. Only when TPP is translated into target product quality profile (TPQP), a quantitative surrogate for TPP, can it be used to optimize a manufacturing process [41].

In our case, Chlorogenic acid has already been proved to be one of the most effective components in both the herb, Lonicerae Japonicae Flos, and the Qingkailing Injection product [42]. In practice, the range of concentrations of Chlorogenic acid has shown to be the most important TPQP for the Lonicerae Japonicae intermediate. A histogram showing Chlorogenic acid concentrations varied between 0.30 and 5.97 $mg \cdot mL^{-1}$ (properly scaled to the range between −1 and 1) is shown in Fig. 3. It is based on normal historical production data, including 173 batch records from 2008 to 2011. Although the concentration values are all within the predefined specifications, the increasingly stringent regulatory requirements set by quality control standards for Chinese medicine injections show that the variability of this concentration must be reduced still further. For this reason, a higher level of TPQP, i.e. a narrower range of Chlorogenic acid concentrations, is needed from the quality improvement point of view.

In order to set up the reasonable lower and upper limits (i.e. **O** in Step 1 of TOPO) for the control of Chlorogenic acid, a moving window method was applied. The window size represents the expected variation coverage for the concentration of Chlorogenic acid, and was set up to 1.0 $mg \cdot mL^{-1}$ in this study. Then, the window was moved from 0 to 6 $mg \cdot mL^{-1}$ at the step of 0.1 $mg \cdot mL^{-1}$. During the movement of the window, the number of historical batch records hit at the window was saved. Finally, the window located between 2.5 and 3.5 $mg \cdot mL^{-1}$ with the maximum number of batch records (i.e. 61) was selected as the target optimization interval (i.e. black dashed lines properly scaled to −0.2240 and 0.1287 in Fig. 3). In this way, the target response interval is narrow enough to control the variability, while representing enough normal operating conditions.

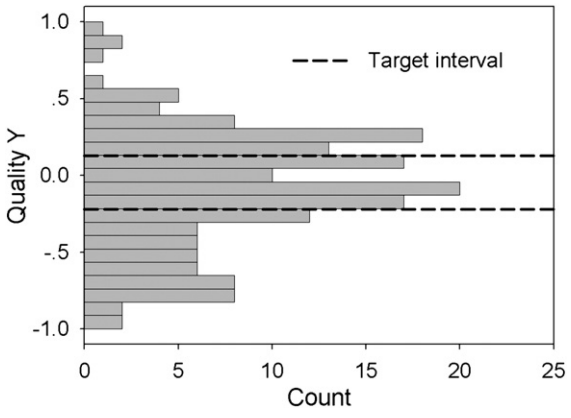

Fig. 3. Target product quality profile (TPQP).

4. Results and discussion

4.1. Computer implementation

All programs involved in this paper were implemented using MATLAB 7.0 platform (MathWorks Inc., U.S.). Process PLS regression models were developed using PLS_Toolbox 2.1 (Eigenvector Research Inc., U.S.). Multivariate normal random numbers and Chi-square random numbers used in the Monte Carlo simulation were generated using the corresponding distribution functions from the MATLAB Statistics Toolbox.

4.2. Pretreatment of data

According to Steps 2, 3 and 4 of TOPO, the process variables **X** and quality variable **y** of the collected 173 batch production data, were coded to the range between −1 and 1. Then, 115 batch data were selected as the calibration set using the Kennard and Stone algorithm. The remaining 58 batch data were treated as the validation set and the optimization object (i.e. the control set). The acceptable minimum and maximum bounds for the controlled process variables were established according to both the historical records and the process engineers' advice (Table 2).

To confirm whether the available data satisfied the prerequisites proposed in Section 2.1.2, the variance inflation factor (VIF) was used to measure the collinearity of variables in the joint data block $\mathbf{X}_{(j)}$ formed at different stages of the process (Table 3). As suggested in previous works, the collinearity was considered weak if the VIF was less than 10 [43]. In the present case, almost all the VIF values were within the suggested limit, except for one variable with VIF of 10.25 in $\mathbf{X}_{(7)}$. However, this did not influence the accuracy of the calculations that took place during the last stage optimization.

4.3. Calibration and validation of the process PLS models

During the Phase 3 of TOPO, seven PLS models were developed using the NIPALS algorithm, relating the process variables and the quality variable. Each calibration model has an optimum LV number. Not enough information would be obtained if LVs are too small. This phenomenon

Table 2
Meaning and optimization range of controlled parameters.

Parameter	Name	Range
x_6	Relative density (room temperature)	−1–1
x_7	Temperature of liquid drug extract	−0.3913–0.7681
x_8	Time of alcohol adding	−1–0.8433
x_{11}	Volume of alcohol added	−0.9022–0.9308
x_{13}	Time of refrigeration	−0.9914–0.2499
x_{14}	Vacuum degree	−1–1
x_{15}	Temperature of ethanol recovery	−0.6667–0.5833
x_{17}	Temperature of liquid drug extract	−0.8182–1
x_{18}	Time of alcohol adding	−0.7143–1
x_{21}	Volume of alcohol added	−0.8821–0.9047
x_{23}	Time of refrigeration	−0.9789–0.8619
x_{24}	Vacuum degree	−1–1
x_{25}	Temperature of ethanol recovery	−1–1
x_{27}	Time of agitation	−0.9322–0.7627
x_{28}	Time of refrigeration	−0.9816–0.9450

Table 3
Variance inflation factor (VIF) calculated for the variables in the joint data block at different stages.

Stage	Data block	Variance inflation factor		
		Minimum	Maximum	Mean
1	$\mathbf{X}_{(1)}$	1.05	1.67	1.37
2	$\mathbf{X}_{(2)}$	1.07	1.71	1.35
3	$\mathbf{X}_{(3)}$	1.10	4.98	2.22
4	$\mathbf{X}_{(4)}$	1.13	5.67	2.24
5	$\mathbf{X}_{(5)}$	1.18	9.86	3.45
6	$\mathbf{X}_{(6)}$	1.24	10.00	3.47
7	$\mathbf{X}_{(7)}$	1.26	10.25	3.40

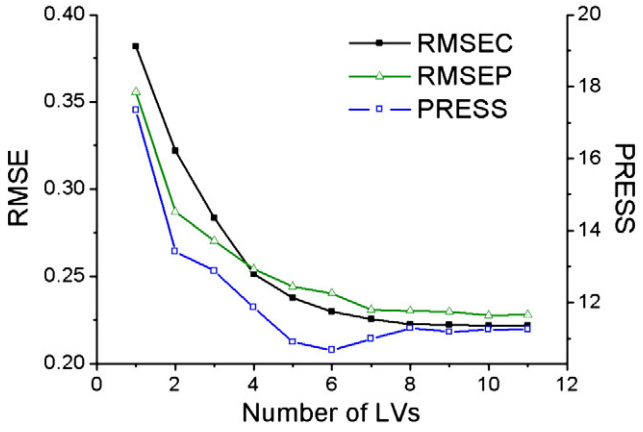

Fig. 4. Selection of the optimal number of the latent variables for PLS model XY_7.

is called "under-fitting" of a model. And an excessive inclusion of LVs may increase the accuracy of calibration but decrease the model predictivity, which indicates that the model is misrepresented or "over-fitted". In this paper, the number of latent factors was optimized using both the LOO cross-validation and an external validation set. Without loss of generality, taking model XY_7 established during the last stage for example, the PRESS, RMSEC, and RMSEP values stabilized at 8 latent factors, which were chosen as the optimum LVs (as shown in Fig. 4). The correlation diagrams under these 8 factors are shown in Fig. 5. The r values are 0.8393 and 0.7680 for the calibration set and validation set, respectively.

After the calibration models were established, their predictive ability was *verified* using the validation set. The results are listed in Table 4. It could be found that the performance of these PLS models was improved stage by stage. This was proved by the decreasing RMSEC, RMSEP, and RMSECV indexes, and by the increasing r values. One rational explanation for these phenomena is that the useful information carried in the process variables is gradually added into process PLS models. It should be noted, however, that all the r values are not very high, ranging between about 0.52 and 0.84. This could be due to several reasons. Firstly, some process variables (e.g. the refrigeration time represented by x_{13} and x_{23}) were recorded manually in our case. Hence, the accuracy of these variables could not be adequately assured. Secondly, some meaningful process parameters, like the time and speed of agitation in the alcohol precipitation stage, were unfortunately not preserved. Thirdly, high-dimensional interactions between different process variables, as well as interactions between different stages, were not fully taken into consideration when performing the PLS modeling. Despite these limitations, the *established* PLS models were able to capture the primary pattern of the multistage batch process, and could be used for optimization. Moreover, the Bayesian approach introduced could assure the reasonable inference, based on even low fidelity models.

Table 4
Performance of the process PLS models established during different stages.

Stage	LVs	Calibration				Validation			
		r_{cal}	RMSEC	RMSECV	$BIAS_{cal}$	r_{val}	RMSEP	RPD	$BIAS_{val}$
1	5	0.5258	0.3501	0.3669	0.2866	0.6333	0.2905	1.23	0.2293
2	6	0.5357	0.3475	0.3709	0.2826	0.6381	0.2893	1.23	0.2317
3	7	0.7477	0.2718	0.3179	0.2208	0.7865	0.2242	1.59	0.1699
4	7	0.7623	0.2649	0.3186	0.2098	0.7853	0.2272	1.57	0.1682
5	8	0.8003	0.2454	0.3177	0.1994	0.7718	0.2294	1.56	0.1799
6	8	0.8156	0.2368	0.3210	0.1852	0.7762	0.2265	1.58	0.1765
7	8	0.8393	0.2225	0.3132	0.1779	0.7680	0.2302	1.55	0.1785

The RPD value is also a sensitive index closely related to the model predictive performance. According to previous works, PLS model with a RPD value larger than 1.5 may be acceptable [44]. As shown in Table 4, RPD values for the first and second stage models were both less than 1.5, which may have been because of the limited number of process variables included in the two stages. However, after Stage 3, the RPD values all exceeded 1.5 and did not vary too much. For this reason, Stage 3 was considered an important operation unit with respect to the quality variable. This is closely consistent with results observed in real production situations, in which alcohol precipitation plays a major role in the process of preparing Lonicerae Japonicae extract [11].

4.4. Optimization for a single batch

Every batch in the control set was investigated and optimized according to procedures in Phase 4 of TOPO. For a single batch, the optimization was simulated to adjust the values of the controlled variables

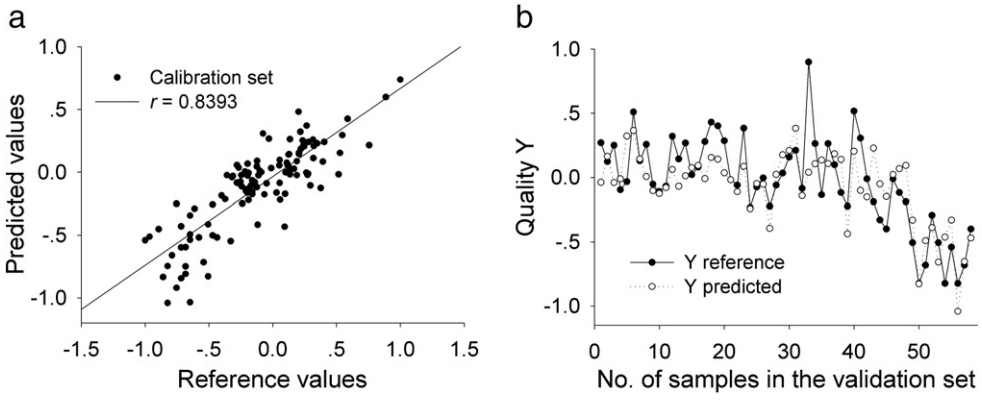

Fig. 5. Correlation between the reference and prediction values under PLS model XY_7. (a) Calibration set. (b) Validation set.

Table 5
Bayesian predictive distribution during the course of optimization for Batch 49.

Stage	Mean	Median	Std. deviation	IQR
1	−0.2426	−0.2427	0.3711	0.4998
2	−0.1196	−0.1195	0.3846	0.5182
3	−0.0066	−0.0044	0.3303	0.4459
4	−0.0831	−0.0844	0.3256	0.4356
5	−0.0554	−0.0526	0.3206	0.4245
6	−0.0401	−0.0418	0.3288	0.4327
7	−0.0152	−0.0108	0.3071	0.4100

from the second stage to the last stage, and the values of the observed variables were kept unchanged. All the possible combinations of the controlled variables during a certain stage were employed to form $\mathbf{L_j}$ in Eqs. (17) and (20). For example, in the fifth stage, $\mathbf{L_j}$ was generated from a grid of $9 \times 4 \times 6 \times 62$ data points.

Taking Batch 49 in the control set for example, the underlying optimization course was revealed by using the descriptive statistics of the Bayesian posterior prediction distribution at every stage, such as mean, median, standard deviation and interquartile range (IQR). IQR

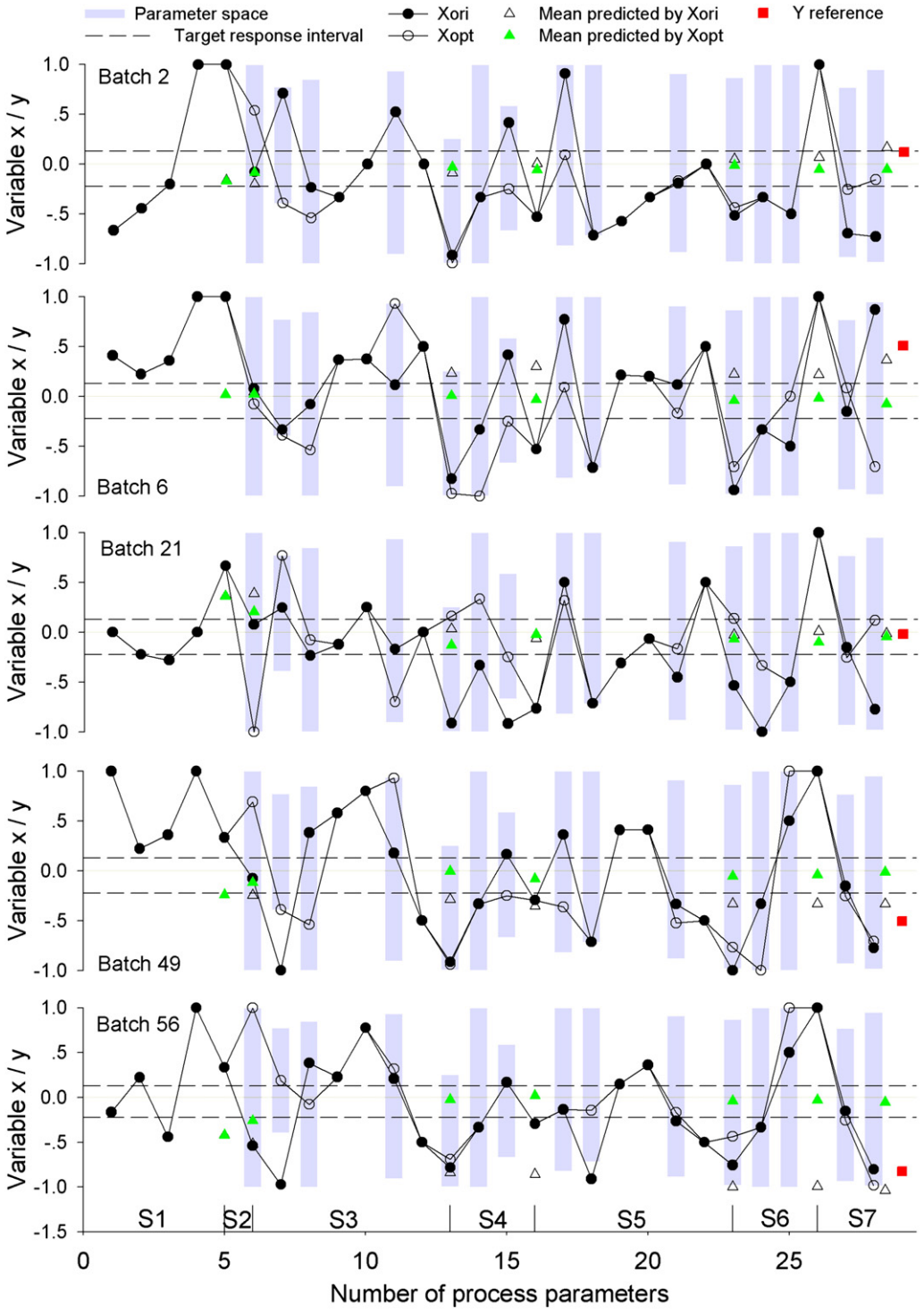

Fig. 6. Results of the overall process optimization for 5 batches selected from the validation set. $\mathbf{X_{ori}}$ indicates the original process variables and $\mathbf{X_{opt}}$ indicates the optimized process variables.

was considered a robust estimate for the spread of the data. Details of the results are shown in Table 5. As the optimization progressed, the standard deviation and IQR indexes tended to decrease. At the same time, the mean and the median values approached the center of the predefined target response interval (i.e. 0.0476) piece by piece, indicating that TOPO methods performed well.

4.5. Optimization for the control set

The efficiency of TOPO method was investigated on all 58 batches of the control set. Five distinctive batches (Nos. 2, 6, 21, 49 and 56), which had different starting materials and different final quality results were selected for analysis. As shown in Fig. 6, the values of the process variables before and after the optimization are compared. Within the acceptable parameter space, there was always a way of achieving the target response interval. Given the original and optimized process variables, the mean of Bayesian prediction distribution by original process variables gradually approached the **y** reference values, and the mean predicted by the optimized variable always lay around the center of the target interval. The variability of the input raw material could be mitigated during the multi-stage optimization.

The probability trajectory can be visualized by lining up the Bayesian posterior probability calculated at each stage. The probability trajectories of the above five selected batches under the original process variables are shown in Fig. 7a. These trajectories evolved without a definitive direction. As a result, Batches 6, 49, and 56 eventually failed to meet the target, and the reference value of Batch 2 stopped at the upper limit of the target interval.

The corresponding probability trajectories after applying TOPO are shown in Fig. 7b. The five trajectories all tended to increase, peaking at the final stage. After the overall process optimization, the optimized quality responses of the 58 batches in the control set were predicted using the optimized process variables based on the PLS model XY_7. The histograms for both the reference and optimized quality values were overlain to indicate the target achievement (as shown in Fig. 8). It became clear that the optimized quality values were all located within the target interval. These results demonstrated that the goal proposed in Section 2, that TOPO keeps the production process in line with the predefined target, was achieved.

Besides, a simple univariate statistical process control (SPC) chart was created using the 58 TOPO probability values ranging from 0.3890 to 0.4917 at the last stage. The lower limit of the SPC chart, which equaled the mean minus three times the standard deviation, was set as the warning limit (i.e. 0.3803 represented as red lines in Fig. 7). Then the process engineer could monitor the production process. For example, the inferior performance of Batch 2, shown in Fig. 7a, may be attributed to improper operation during Stage 7. Assuming that TOPO

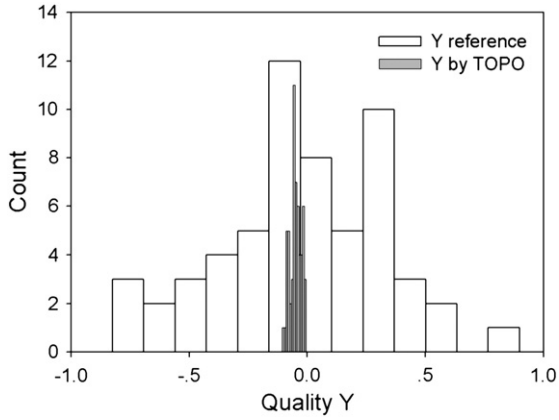

Fig. 8. Target achievement when TOPO was applied to the control set.

was introduced during the last stage, this unsatisfactory operation may have been prevented by maintaining the probability above the warning limit (shown as a green line segment in Fig. 7a). In this way, TOPO helps the process engineer understand and optimize the production process at a high level.

5. Conclusions

In this paper, a new strategy, target-oriented overall process optimization (TOPO) is brought forward. This system allows the user to continually monitor and optimize manufacturing processes to ensure consistent quality across herbal products. The feasibility of the proposed TOPO strategy was validated by applying it to the production of Lonicerae Japonicae extract. Results indicated that the problem of variability in the concentration of active ingredients in herbal products could be solved from an overall production process point of view using TOPO.

In general, TOPO explores the full potential of legacy batch production data with respect to understanding and optimizing the herbal production process. It can help process engineers monitor, control, and optimize the process of manufacturing of herbal medicine and ultimately reduce the variability in final products. For herbal prescriptions produced under the framework of ICH Q10 guidance (i.e. Pharmaceutical Quality System) [45], TOPO can serve for the process performance and product quality monitoring system, as well as the corrective action and preventative action (CAPA) system.

Batch process optimization is a long-standing problem [46], but overall process optimization is a new area in research of the field. With respect to reaching global optimization targets, it will be greatly

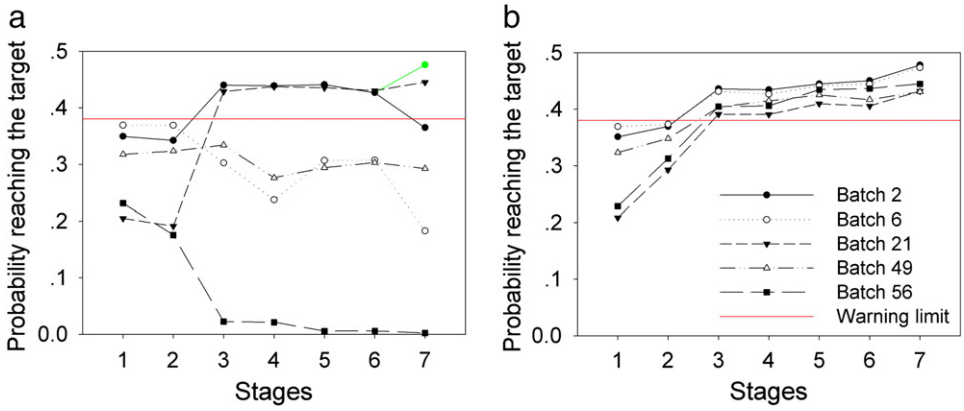

Fig. 7. Probability trajectories. (a) Without TOPO. (b) With TOPO. The green line segment in (a) shows that the inferior performance of Batch 2 during Stage 7 could be prevented via TOPO.

helpful to view the batch production process as a whole, rather than as discrete steps. TOPO is a new strategy for herbal medicine production and there is still room for improvement. Future studies may explore the multi-objective, high-dimensional, nonlinear and dynamic characteristics of the multistage batch process, and the economic ramifications of these variations.

Acknowledgments

The authors wish to thank the National Major Projects of Science and Technology named 'Creation of Major New Drugs' (No. 2010ZX09502-002, PR China). The authors would also like to thank the assistant manager Haiyan Zhou of Yabao Beizhongda (Beijing) Pharmaceutical Co., Ltd., for providing the historical production data.

References

[1] I.A. Khan, T. Smillie, Implementing a "Quality by Design" approach to assure the safety and integrity of botanical dietary supplements, J. Nat. Prod. 75 (2012) 1665–1673.
[2] M.M. Robinson, X. Zhang, The World Medicines Situation 2011, Traditional Medicines: Global Situation, Issues and Challenges, World Health Organization, Geneva, Switzerland, 2011.
[3] R. Walker, Criteria for risk assessment of botanical food supplements, Toxicol. Lett. 149 (2004) 187–195.
[4] N. Sahoo, P. Manchikanti, S. Dey, Herbal drugs: standards and regulation, Fitoterapia 81 (2010) 462–471.
[5] F. Chau, H. Chan, C. Cheung, C. Xu, Y. Liang, O. Kvalheim, Recipe for uncovering the bioactive components in herbal medicine, Anal. Chem. 81 (2009) 7217–7225.
[6] X. Zheng, X. Gong, Q. Li, H. Qu, Application of multivariate curve resolution method in the quantitative monitoring transformation of salvianolic acid A using online UV spectroscopy and mass spectroscopy, Ind. Eng. Chem. Res. 51 (2012) 3238–3245.
[7] P. Li, L. Qi, E. Liu, J. Zhou, X. Wen, Analysis of Chinese herbal medicines with holistic approaches and integrated evaluation models, TrAC Trends Anal. Chem. 27 (2008) 66–77.
[8] O. Pelkonen, M. Pasanen, J.C. Lindon, K. Chan, L. Zhao, G. Deal, Q. Xu, T. Fan, Omics and its potential impact on R&D and regulation of complex herbal products, J. Ethnopharmacol. 140 (2012) 587–593.
[9] Y. Xie, Z. Jiang, H. Zhou, X. Cai, Y. Wong, Z. Liu, Z. Bian, H. Xu, L. Liu, Combinative method using HPLC quantitative and qualitative analyses for quality consistency assessment of a herbal medicinal preparation, J. Pharm. Biomed. Anal. 43 (2007) 204–212.
[10] P. Xie, A.Y. Leung, Understanding the traditional aspect of Chinese medicine in order to achieve meaningful quality control of Chinese materia medica, J. Chromatogr. A 1216 (2009) 1933–1940.
[11] B. Xu, Z. Wu, Z. Lin, C. Sui, X. Shi, Y. Qiao, NIR analysis for batch process of ethanol precipitation coupled with a new calibration model updating strategy, Anal. Chim. Acta 720 (2012) 22–28.
[12] B. Palumbo, G.D. Chiara, F. Sansone, R. Marrone, Technological scenarios of variation transmission in multistage machining processes, Qual. Reliab. Eng. Int. 27 (2011) 651–658.
[13] Y.Z. Liang, P. Xie, K. Chan, Quality control of herbal medicines, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 812 (2004) 53–70.
[14] K.N. Chan, C.C. Lau, K.L. Chow, K.M. Ko, Karl W.K. Tsim, K.M. Ng, Effect of extraction solvent on the bioactivity of an herbal formulation, Ind. Eng. Chem. Res. 48 (2009) 4852–4857.
[15] L. Yin, Y. Xu, Y. Qi, X. Han, L. Xu, J. Peng, C. Sun, A green and efficient protocol for industrial-scale preparation of dioscin from *Dioscorea nipponica* Makino by two-step macroporous resin column chromatography, Chem. Eng. J. 165 (2010) 281–289.
[16] J. Penga, Y. Jiang, G. Fan, B. Chen, Q. Zhang, Y. Chai, Y. Wu, Optimization suitable conditions for preparative isolation and separation of curculigoside and curculigoside B from *Curculigo orchioides* by high-speed counter-current chromatography, Sep. Purif. Technol. 52 (2006) 22–28.
[17] F. Xiong, C. Xiong, J. Yao, X. Chen, N. Gu, Preparation, characterization and evaluation of breviscapine lipid emulsions coated with monooleate-PEG-COOH, Int. J. Pharm. 421 (2011) 275–282.
[18] C. Lu, Y. Lu, J. Chen, W. Zhang, W. Wu, Synchronized and sustained release of multiple components in silymarin from erodible glyceryl monostearate matrix system, Eur. J. Pharm. Biopharm. 66 (2007) 10–219.
[19] C. Zhao, J. Fu, Q. Xu, Real-time dynamic hoist scheduling for multistage material handling process under uncertainties, AICHE J. 59 (2013) 465–482.
[20] X. Kong, Y. Yang, X. Chen, Z. Shao, F. Gao, Quality control via model-free optimization for a type of batch process with a short cycle time and low operational cost, Ind. Eng. Chem. Res. 50 (2011) 2994–3003.
[21] J. Shi, S. Zhou, Quality control and improvement for multistage systems: a survey, IIE Trans. 41 (2009) 744–753.
[22] Y. Yao, F. Gao, A survey on multistage/multiphase statistical modeling methods for batch processes, Annu. Rev. Control. 33 (2009) 172–183.
[23] A.G. Shetwan, V.I. Vitanov, B. Tjahjono, Allocation of quality control stations in multistage manufacturing systems, Comput. Ind. Eng. 60 (2011) 473–484.
[24] T. Tosukhowong, J. Lee, J. Lee, J. Lu, An introduction to a dynamic plant-wide optimization strategy for an integrated plant, Comput. Chem. Eng. 29 (2004) 199–208.
[25] A. Peschel, A. Jörke, K. Sundmacher, H. Freund, Optimal reaction concept and plant wide optimization of the ethylene oxide process, Chem. Eng. J. 207–208 (2012) 656–674.
[26] H. Li, J. Huang, Process goose queue (PGQ) approaches toward plantwide process optimization with applications in supervision-driven real-time optimization, Ind. Eng. Chem. Res. 51 (2012) 10848–10859.
[27] A. Pomerantsev, O. Rodionova, A. Höskuldsson, Process control and optimization with simple interval calculation method, Chemom. Intell. Lab. Syst. 81 (2006) 165–179.
[28] A. Höskuldsson, O. Rodionova, A. Pomerantsev, Path modeling and process control, Chemom. Intell. Lab. Syst. 88 (2007) 84–99.
[29] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[30] H. Wold, Path models with latent variables: the NIPALS approach, in: H.M. Blalock (Ed.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building, Academic Press, New York, 1975, pp. 307–357.
[31] S. De Jong, SIMPLS: an alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.
[32] R.R. Picard, R.D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. 79 (1984) 575–583.
[33] A. İspir, İ.T. Toğrul, Osmotic dehydration of apricot: kinetics and the effect of process parameters, Chem. Eng. Res. Des. 87 (2009) 166–180.
[34] P.C. Williams, K. Norris, Near-infrared Technology in the Agricultural and Food Industries, American Association of Cereal Chemists, St Paul, MN, 2001.
[35] M.C. Coleman, D.E. Block, Bayesian parameter estimation with informative priors for nonlinear systems, AICHE J. 52 (2006) 651–667.
[36] E. Del Castillo, Process Optimization: A Statistical Approach, Springer, Berlin, 2007.
[37] T. Næs, B. Mevik, Understanding the collinearity problem in regression and discriminant analysis, J. Chemom. 15 (2001) 413–426.
[38] H. Chen, B.R. Bakshi, P.K. Goel, Bayesian latent variable regression via Gibbs sampling: methodology and practical aspects, J. Chemom. 21 (2007) 578–591.
[39] FDA, Draft Guidance for Industry and Review Staff, Target Product Profile — A Strategic Development Process Tool, FDA, 2007.
[40] N.A. Charoo, A. Shamsher, A.S. Zidan, Z. Rahmand, Quality by design approach for formulation development: a case study of dispersible tablets, Int. J. Pharm. 423 (2012) 167–178.
[41] R.A. Lionberger, S. Lee, L. Lee, A. Raw, L. Yu, Quality by design: concepts for ANDAs, AAPS J. 10 (2008) 268–276.
[42] H. Zhang, P. Hu, G. Luo, Q. Liang, Y. Wang, S. Yan, Y. Wang, Screening and identification of multi-component in Qingkailing injection using combination of liquid chromatography/time-of-flight mass spectrometry and liquid chromatography/ion trap mass spectrometry, Anal. Chim. Acta 577 (2006) 190–200.
[43] R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, Qual. Quant. 41 (2007) 673–690.
[44] F. Shen, X. Niu, D. Yang, Y. Ying, B. Li, G. Zhu, J. Wu, Determination of amino acids in Chinese rice wine by Fourier transform near-infrared spectroscopy, J. Agric. Food Chem. 58 (2010) 9809–9816.
[45] ICH, Q10 Pharmaceutical Quality System, ICH, 2008.
[46] G.M. Troup, C. Georgakis, Process systems engineering tools in the pharmaceutical industry, Comput. Chem. Eng. 51 (2013) 157–171.