ELSEVIER

Chemometrics and Intelligent Laboratory Systems 56 (2001) 1–11

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Monte Carlo cross validation

Qing-Song Xu [a], Yi-Zeng Liang [b,*]

[a] College of Mathematics and Econometrics, Hunan University, Changsha, People's Republic of China
[b] College of Chemistry and Chemical Engineering, Central South University, Changsha, Hunan 410083 People's Republic of China

Received 6 July 2000; accepted 27 November 2000

Abstract

In order to choose correctly the dimension of calibration model in chemistry, a new simple and effective method named Monte Carlo cross validation (MCCV) is introduced in the present work. Unlike leave-one-out procedure commonly used in chemometrics for cross validation (CV), the Monte Carlo cross validation developed in this paper is an asymptotically consistent method in determining the number of components in calibration model. It can avoid an unnecessary large model and therefore decreases the risk of over-fitting for the calibration model. The results obtained from simulation study showed that MCCV has an obviously larger probability than leave-one-out CV in choosing the correct number of components that the model should contain. The results from real data sets demonstrated that MCCV could successfully choose the appropriate model, but leave-one-out CV could not.

Keywords: Model; Number of components; Leave-one-out; Cross validation; Monte Carlo

1. Introduction

Multivariate partial least squares regression (PLSR) and principle components regression (PCR) modeling are very useful tools for the analysis of high-dimensional data. In multivariate calibration, chemical samples of a compound are depicted by spectra (such as ultraviolet and near infrared spectra) that are recorded to give more than several hundred variables. PLSR (or PCR) provides adapted access to model these kinds of high dimensional data. However, it is difficult to determine the adapted number of PLS components (or PCs) that should be used in the model and, at the same time, to make the determined model to have the best predictive ability for future samples.

Many methods, such as Akaike information criterion [1], the C_p statistics [2], the jackknife and the bootstrap [3,4] and cross validation (CV) [5–8], can be used to ascertain the number of components included into model. Among these methods, CV is of most applications in chemometrics. It is a method of evaluating given models according to the predictive ability and to determine appreciate components included into the model. In the literature, CV is generally referred to as the simplest leave-one-out cross validation unless announced specially. However, there is a compelling problem for leave-one-out CV. As pointed in Ref. [9], leave-one-out CV often cause over-fitting, and on average, it gave an under-estima-

* Corresponding author. Tel.: +86-731-882-2841; fax: +86-731-882-5637.
E-mail address: yzliang@public.cs.hn.cn (Y.-Z. Liang).

0169-7439/01/$ - see front matter
PII: S0169-7439(00)00122-2

tion of the true predictive error. Many other chemometricians also perceived this shortage of leave-one-out CV [10,11]. They were very careful when using CV and making some improvements over the leave-one-out CV criterion [12,13]. The reason for leave-one-out CV having such a deficiency is that it is an asymptotically inconsistent method [4,14,15] (the other methods mentioned above share the same deficiency). The consequence of this method is that it tends to include unnecessary components into the model and make the model larger than it should be. Therefore, the model with the number of components determined by leave-one-out CV often performs good in calibration, but poor in prediction.

On the other hand, much attention was paid to CV with more than one sample left out at a time in validation. Multifold CV has appeared in Ref. [16]. The leave-two-out CV performed better than leave-one-out CV [17]. The theoretic results about multifold CV can be found in Refs. [18–20].

Monte Carlo cross validation (MCCV) was first considered in Ref. [21]. This method has been shown asymptotically consistent [14], but it is rarely used in chemometrics. In the presented paper, we introduce MCCV method in a multivariate calibration problem. Although MCCV is a consistent method for linear model in the large data cases, it is a mystery when used in a small data set for calibration. We try to gain some insight as to see how the scale of the samples in the calibration data set ought to be left out at a time in validation and whether MCCV is an effective method for determining the number of components in the calibration model. To accomplish this goal, simulated experiments are performed. Not only the level of random error, but also the degree of collinearity in the spectra of compounds is taken into consideration. Finally, two real examples are discussed in detail.

2. Theory and method

The following linear calibration model is considered:

$$\begin{cases} \boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}; \\ E(\boldsymbol{e}) = 0, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I} \end{cases} \tag{1}$$

where $\boldsymbol{\beta}$ is regression coefficient vector, and $\mathbf{X}$ observation matrix, $\boldsymbol{e}$ random error vector, $\boldsymbol{y}$ response vector, respectively. $\mathbf{I}$ is the identity matrix. Suppose that $\mathbf{X}$ is an $n \times m$ matrix; $\boldsymbol{y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors and $\boldsymbol{\beta}$ is an $m \times 1$ vector. $E(\bullet)$ and $\mathrm{Cov}(\bullet)$ denote the expectation and covariance, respectively. Estimator of regression coefficient vector is obtained by least squares (LS):

$$\hat{\beta}_{\mathrm{L}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{y} \tag{2}$$

Typical chemical data, such as the spectroscopic data, tend to be characterized by many independent variables on relatively fewer observations ($m > n$), or more generally, $\mathrm{rank}(\mathbf{X}) < \min(n, m)$. There is high collinearity among the independent variables. It is well known that under this situation, the estimator of regression coefficient vector by LS is unstable, leading to poor prediction accuracy [22]. In this situation, because of the unavoidable deficiencies of the data, it is better to choose the PLS (or PCR) model [23]. In this paper, we only consider the PLS regression model.

2.1. PLS regression

The linear model (1) is considered. Taking the notation of Refs. [24,25], the observation matrix can be decomposed as follows:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^t + \mathbf{t}_2\mathbf{p}_2^t + \ldots + \mathbf{t}_k\mathbf{p}_k^t + \mathbf{R} \tag{3}$$

where $\mathbf{t}_i$, $\mathbf{p}_i$ are the PLS scores and loadings; $\mathbf{R}$ is the residual matrix. The number k denotes the number of PLS components that are introduced into model (1). If there is no measurement errors in the data matrix, then

$$\mathbf{X} = \sum_{i=1}^{q} \mathbf{t}_i\mathbf{p}_i^t = \mathbf{T}\mathbf{P}^t \tag{4}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_q]$, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_q]$. The subscript q denotes the rank of matrix $\mathbf{X}$. Each PLS score $\mathbf{t}_i$ is a combination of the column vector of observation matrix $\mathbf{X}$, that is

$$\mathbf{T} = \mathbf{X}\mathbf{H} \tag{5}$$

The model (1) can be rewritten as follows:

$$\boldsymbol{y} = \mathbf{T}\mathbf{P}^t\boldsymbol{\beta} + \boldsymbol{e} = \mathbf{T}\alpha + \boldsymbol{e} \tag{6}$$

The matrices **T**, **P** and **H** can be partitioned as follows:

$$\mathbf{T} = \left[\mathbf{t}_1, \mathbf{t}_2 \ldots, \mathbf{t}_k, \vdots \mathbf{t}_{k+1}, \mathbf{t}_{k+2}, \ldots, \mathbf{t}_q\right] = \left[\mathbf{T}_k \vdots \mathbf{T}_{(k+1)}\right]$$

$$\mathbf{P} = \left[\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k, \vdots \mathbf{p}_{k+1}, \mathbf{p}_{k+2}, \ldots, \mathbf{p}_q\right] = \left[\mathbf{P}_k \vdots \mathbf{P}_{(k+1)}\right]$$

$$\mathbf{H} = \left[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k, \vdots \mathbf{h}_{k+1}, \mathbf{h}_{k+2}, \ldots \mathbf{h}_q\right] = \left[\mathbf{H}_k \vdots \mathbf{H}_{(k+1)}\right] \quad (7)$$

Inserting Eq. (7) into Eq. (6), we get

$$\boldsymbol{y} = \mathbf{T}_k \mathbf{P}_k^t \boldsymbol{\beta} + \mathbf{T}_{(k+1)} \mathbf{P}_{(k+1)}^t \boldsymbol{\beta} + \boldsymbol{e}$$

Let $\alpha^t = [\alpha_k^t, \alpha_{(k+1)}^t] = [(\mathbf{P}_k^t \boldsymbol{\beta})^t, (\mathbf{P}_{(k+1)}^t \boldsymbol{\beta})^t]$, then

$$\boldsymbol{y} = \mathbf{T}_k \alpha_k + \mathbf{T}_{(k+1)} \alpha_{(k+1)} + \boldsymbol{e} \quad (8)$$

When the number of PLS components that are introduced into model is k, the latter q–k PLS components have very small variance and are considered as representatives of noises or the cause of collinearity in the data set. Therefore, the latter q–k elements of α are regarded as zeros, and only the former k components remain in the model.

$$\boldsymbol{y} = \mathbf{T}_k \alpha_k + \boldsymbol{e} \quad (9)$$

The number of components k is also called the dimension of the model. The least square solution of Eq. (9) is

$$\hat{\alpha}_k = \left(\mathbf{T}_k^t \mathbf{T}_k\right)^{-1} \mathbf{T}_k^t \boldsymbol{y} \quad (10)$$

The fitted value of $\boldsymbol{y}$ is

$$\hat{\boldsymbol{y}} = \mathbf{T}_k \hat{\alpha}_k = \mathbf{X}\mathbf{H}_k \left(\mathbf{T}_k^t \mathbf{T}_k\right)^{-1} \mathbf{H}_k^t \mathbf{X}^t \boldsymbol{y} \quad (11)$$

The PLS estimator $\hat{\beta}_k$ of $\boldsymbol{\beta}$ with the former k components remaining in the model can be acquainted from Eq. (11).

$$\hat{\beta}_k = \mathbf{H}_k \left(\mathbf{T}_k^t \mathbf{T}_k\right)^{-1} \mathbf{H}_k^t \mathbf{X}^t \boldsymbol{y} \quad (12)$$

2.2. Cross validation and Monte Carlo cross validation

The fundamental step after the data are available is to determine the number of components (dimension) for the derived model (9). There are total q possible different models taking the pattern of Eq. (9) corresponding to $k = 1, 2, \ldots, q$. How to determine k is the problem. For general CV, the n samples (the rows of X) are split into two parts. The first part (calibration set), denoted as S_c, contains n_c samples for fitting the models. The second part (validation set), denoted as S_v, contains $n_v = n - n_c$ samples for validating the model. There are total $\binom{n}{n_v}$ different forms of sample splits. For each sample split, the model is fitted by the n_c samples of the first part S_c.

$$\hat{\beta}_{S_c k} = \mathbf{H}_{S_c k} \left(\mathbf{T}_{S_c k}^t \mathbf{T}_{S_c k}\right)^{-1} \mathbf{H}_{S_c k}^t \mathbf{X}_{S_c}^t \boldsymbol{y}_{S_c} \quad (k = 1, 2, \ldots, q) \quad (13)$$

where rows of $\mathbf{X}_{S_c}$ and $\boldsymbol{y}_{S_c}$ are corresponding to the samples in the calibration set; $\mathbf{H}_{S_c k}$ and $\mathbf{T}_{S_c k}$ are determined by Eqs. (4) and (5) based on $\mathbf{X}_{S_c}$. The samples in the validation set $\mathbf{X}_{S_v}$ are treated as if they were future ones. The fitted model then predicts response vector $\boldsymbol{y}_{S_v}$.

$$\hat{\boldsymbol{y}}_{S_c k} = \mathbf{X}_{S_v} \hat{\beta}_{S_c k} \qquad (k = 1, 2, \ldots, q) \quad (14)$$

The average squared prediction error (ASPE) over all samples in validation set is

$$\mathrm{ASPE}(S_v, k) = \frac{1}{n_v} \| \boldsymbol{y}_{S_v} - \hat{\boldsymbol{y}}_{S_v k} \|^2 \quad (15)$$

where $\|\bullet\|$ stands for Euclidean norm of a vector. Let S be the set whose elements are all from the validation sets corresponding to $\binom{n}{n_v}$ different forms of sample splits. The cross validation criterion with n_v left out of the model is defined as

$$\mathrm{CV}_{n_v}(k) = \frac{1}{\binom{n}{n_v}} \sum_{S_v \in S} \mathrm{ASPE}(S_v, k) \quad (16)$$

$\mathrm{CV}_{n_v}(k)$ is calculated for every kth component as it is added to the model. The optimal k^*, which gives a minimum value of CV_{n_v} ($k = 1, 2, \ldots, q$), is the number of components that should be contained into the model.

The simplest CV, with $n_v = 1$ (leave-one-out), is largely used in chemometrics. However, it was

proven that the model chosen by CV_1 ($n_v = 1$) is asymptotically incorrect [14,15]. It tends to include unnecessary excessive components into the model and consequently bring about over-fitting. The reason CV_1 inclines to choose a larger model is that it emphasizes calibration but not validation. For every split of n samples, $n-1$ samples are used for calibration, whereas only one sample is used for validation. The larger the n_c, the lesser the influence of validation on $CV_{n_v}(k)$. As for calibration, the more components are included in the model, the better the model is fitted. Therefore, it is not difficult to understand that leave-one-out CV has such a deficiency.

Under the conditions that $n_c \to \infty$ and $n_v/n \to 1$, it has been proven by Shao [14] that the probability for cross validation with n_v left out for validation to choose the correct model tends to 1. In this sense, the $CV_{n_v}(k)$ criterion (Eq. 16) is asymptotically consistent. For the data sets of finite size, with the increasing of samples that are left out for validation, the probability of selecting the model with the correct number of components also increases [20]. However, the computation of CV_{n_v} with large n_v is not applicable (the computation complexity of CV_{n_v} is exponential). Monte Carlo cross validation (MCCV) [14] is a simple and effective method: randomly split the samples into two parts $S_c(i)$ (of size n_c) and $S_v(i)$ (of size n_v); repeat the procedure N times ($i = 1, 2, \ldots, N$). The repeated MCCV criterion is defined:

$$\mathrm{MCCV}_{n_v}(k) = \frac{1}{Nn_v}\sum_{i=1}^{N} \| y_{S_v(i)} - \hat{y}_{S_v(i)} \|^2 \qquad (17)$$

By means of the Monte Carlo method, the amount of computation complexity can be reduced substantially. Theoretically, the fewer samples used in model calibration, the more repeat times are needed and $N = n^2$, in general, is enough in order to make MCCV_{n_v} perform as well as CV_{n_v} [20].

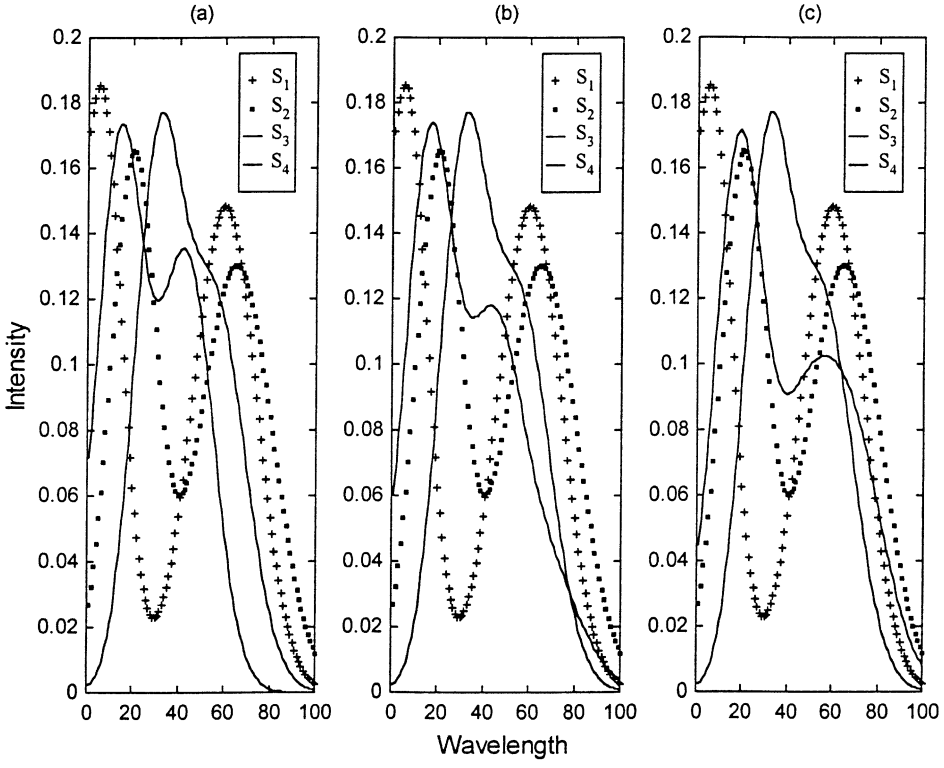

Fig. 1. The spectra for the four pure compounds. (a) Low collinearity; (b) middle collinearity; (c) high collinearity.

2.3. Confirmation of the correct model

For calibration problems, prediction, not calibration, is the main purpose. Thus, in this paper, the correct number of components in the model is confirmed based on the performance in prediction. After the number of components that should be included into the model is settled down by CV. The model with the determined number of components is then constructed by the full data set. Whether this model is correct depends on its performance on a new prediction set. Let the rows of matrix $\mathbf{X}_p$ be the samples in the prediction set and $\boldsymbol{y}_p$ be their response vector. The mean squared error of prediction (MSEP) is as follows.

$$\mathrm{MSEP}(k) = \frac{1}{n_p} \sum_{i=1}^{n_p} \|\boldsymbol{y}_p - \hat{\boldsymbol{y}}_{pk}\|^2 \quad (18)$$

where $\hat{\boldsymbol{y}}_{pk}$ is the predicted response vector by the model including k PLS components in it, and n_p is the size of prediction set. If k^*, determined by CV, makes MSEP(k^*) the minimum among MSEP(k) ($k = 1, 2, \ldots, q$), then it is said that the determined model is the correct one.

3. Data

Although MCCV can asymptotically select the model with the correct number of components, its performance on the data set with finite number of samples needs further investigation. In this section, the simulated data, near infrared data and ultraviolet data are used for exploring the method. All the data sets are centered before computation.

3.1. Simulated data

In order to investigate the possessions of MCCV under the different situations, a set of simulated data has been created for the four chemical component calibration models.

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e} \quad (19)$$

where every row of $\mathbf{X}$ is a combination of spectra of four pure compounds; $\boldsymbol{y}$ is a concentration vector of one of the compounds. The normalized spectra of pure compounds $\mathbf{s}_i^t$ ($\|\mathbf{s}_i\| = 1$, $i = 1, 2, 3, 4$) are shown in Fig. 1. In favor of exploring the influences of collinearity in data on MCCV method, we only

Table 1
The concentration combination of four compounds

No.	x_1	x_2	x_3	x_4	No.	x_1	x_2	x_3	x_4
1	0.1000	0.8000	0.4000	0.0667	21	0.1000	1.0000	0.1333	0.7333
2	0.3000	0.5000	0.8000	0.0667	22	0.2000	0.7000	0.5333	0.7333
3	0.5000	0.2000	0.2000	0.1333	23	0.3000	0.4000	0.9333	0.8000
4	0.7000	0.9000	0.6000	0.1333	24	0.4000	0.1000	0.3333	0.8000
5	0.9000	0.6000	1.0000	0.2000	25	0.5000	0.8000	0.7333	0.8667
6	0.2000	0.3000	0.3333	0.2000	26	0.6000	0.5000	0.0667	0.8667
7	0.4000	1.0000	0.7333	0.2667	27	0.7000	0.2000	0.4667	0.9333
8	0.6000	0.7000	0.1333	0.2667	28	0.8000	0.9000	0.8667	0.9333
9	0.8000	0.4000	0.5333	0.3333	29	0.9000	0.6000	0.2667	1.0000
10	1.0000	0.1000	0.9333	0.3333	30	1.0000	0.3000	0.6667	1.0000
11	1.0000	0.9000	0.2667	0.4000	31	0.1000	0.3000	0.2000	0.3000
12	0.8000	0.6000	0.6667	0.4000	32	0.2000	0.6000	0.4000	0.5000
13	0.6000	0.3000	0.0667	0.4667	33	0.3000	0.9000	0.1000	0.2000
14	0.4000	1.0000	0.4667	0.4667	34	0.4000	0.1000	0.3000	0.5000
15	0.2000	0.7000	0.8667	0.5333	35	0.5000	0.4000	0.5000	0.2000
16	0.1000	0.4000	0.2000	0.5333	36	0.6000	0.7000	0.1000	0.4000
17	0.3000	0.1000	0.6000	0.6000	37	0.7000	1.0000	0.1000	0.3000
18	0.5000	0.8000	1.0000	0.6000	38	0.8000	0.2000	0.5000	0.4000
19	0.7000	0.5000	0.4000	0.6667	39	0.9000	0.5000	0.2000	0.1000
20	0.9000	0.2000	0.8000	0.6667	40	1.0000	0.8000	0.4000	0.3000

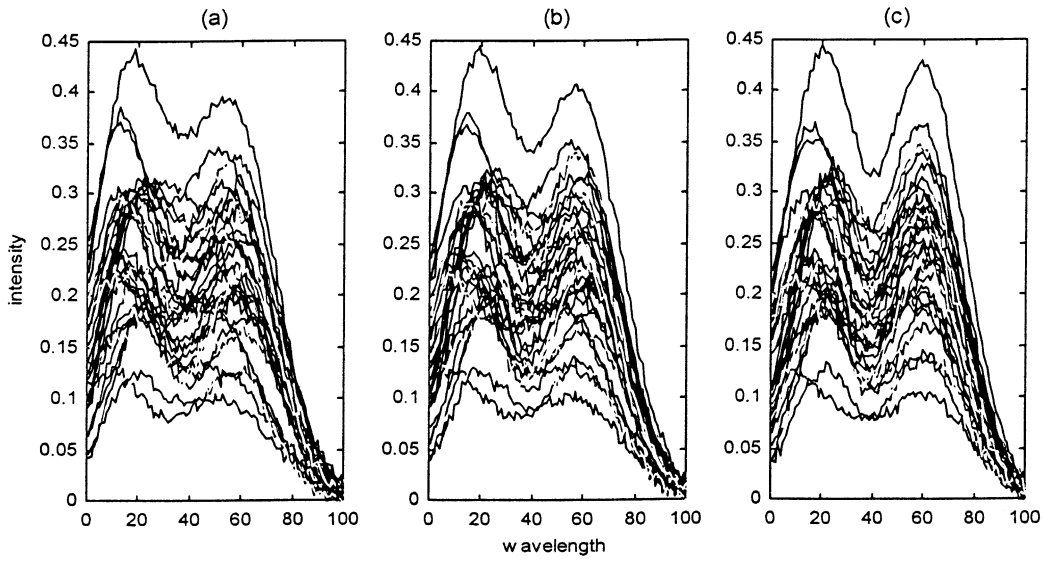

Fig. 2. The spectra of mixtures with $\sigma = 0.004$. (a) Low collinearity; (b) middle collinearity; (c) high collinearity.

change the fourth spectra $\mathbf{s}_4$ in (a), (b) and (c). The correlation coefficients of $\mathbf{s}_2$ and $\mathbf{s}_4$ are 0.8124, 0.9050 and 0.9760, respectively, corresponding to low, middle and high degree of collinearity in data. The levels of random errors are also taken into consideration. The random errors obey normal distribution. The standard deviations of them, say σ, are 0.002 (low level) and 0.004 (high level), respectively. A total of 40 samples are collected. The concentration compositions of these samples are listed in Table 1. The spectra of these mixtures (samples) are shown in Fig. 2. These samples serve as the data set for MCCV to determine the number of PLS components and then to construct the model. The concentration vector of third compound is used as response vector $\boldsymbol{y}$.

3.2. Near infrared data

This near infrared data came from Næs published in Ref. [10]. It was also used as an example in Ref. [11]. It consists of measurements on 28 samples. The NIR spectra are collected at 19 different NIR wavelengths and the percentage of protein is used as response. There are three outlier samples among 28 samples. The other remaining 25 samples are used for the data set in this paper. More explicit explanation for the data is referred to Ref. [10].

3.3. Ultraviolet data

This data consists of ultraviolet measurements of mixtures of naphthalene, anthracene, fluorene and phenanthrene. The ultraviolet spectra are collected at intervals of 1 nm between the wavelengths 200 and 280 nm. The response $\boldsymbol{y}$ is the concentration (0.1 mg

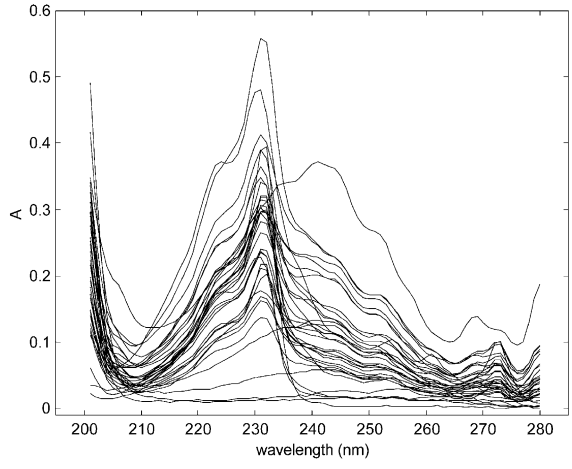

Fig. 3. Ultraviolet spectra.

Table 2
The values of AMSPE(k) for prediction set

σ		Component number									
		1	2	3	4	5	6	7	8	9	10
0.002	(a)	0.2395	0.1083	0.0164	0.0044	0.0048	0.0051	0.0053	0.0054	0.0055	0.0055
	(b)	0.2443	0.1074	0.0171	0.0044	0.0048	0.0052	0.0053	0.0054	0.0055	0.0055
	(c)	0.2443	0.0953	0.0289	0.0043	0.0047	0.0051	0.0052	0.0053	0.0054	0.0054
0.004	(a)	0.2419	0.1103	0.0178	0.0087	0.0096	0.0103	0.0107	0.0109	0.0109	0.0110
	(b)	0.2424	0.1062	0.0185	0.0088	0.0096	0.0103	0.0107	0.0109	0.0110	0.0110
	(c)	0.2399	0.0942	0.0294	0.0088	0.0096	0.0103	0.0107	0.0110	0.0110	0.0111

ml^{-1}) of phenanthrene. Thirty five samples are collected. Their spectra are shown in Fig. 3.

4. Results and discussion

4.1. Simulated data

The mixture spectra for 40 samples with high lever of random errors are shown in Fig. 2. For the MCCV, each simulation is calculated for 80 times ($N = 80$) and 500 simulations are computed for every situation. In order to confirm the correct model, another 100 samples for each simulation, used as a prediction set, are generated with the concentrations of four compounds chosen randomly from the uniform distribution in [0,1]. The average of mean squared error (AMSEP) of prediction for 500 simulations is defined as follows.

$$\mathrm{AMSEP}(k) = \frac{1}{500}\sum_{i=1}^{500}\mathrm{MSEP}_i(k) \qquad (20)$$

where i denotes the ith simulation. The number, say k^*, which makes AMSEP (k^*) the minimum among all AMSEP(k) ($k = 1, 2, \ldots, q$) is taken as the correct number of components that should be contained into the model.

The value of AMSEP (k) on prediction set in all cases is shown in Table 2. It is seen from the table that the value of AMSEP (k) goes down quickly as the number of component increases. But after they reach the minima at $k = 4$, it increases slightly as the number of component increases. Therefore, the correct number of components of the model is always four, being equal to the number of chemical compounds.

The frequencies for the MCCV to choose the number of components are collected in Tables 3 and 4 for the model with low and high lever random errors, respectively. It is interesting to see from Table 3 that whether there is high collinearity or not in the data, the MCCV with $n_v = 1$ always gives the poorest performances. The very important thing that should be noticed is the frequencies for MCCV to choose the correct number of components have a maximum at $n_v = 20$ or $n_v = 25$. This implies that the probability for MCCV to choose the correct

Table 3
The frequencies for the MCCV to choose the number of components contained in the model when $\sigma = 0.002$

n_v	(a)				(b)				(c)			
	3	4	5	≥ 6	3	4	5	≥ 6	3	4	5	≥ 6
1	0	0.750	0.135	0.115	0	0.717	0.175	0.108	0	0.710	0.138	0.152
5	0	0.770	0.140	0.090	0	0.747	0.145	0.108	0	0.797	0.110	0.093
10	0	0.832	0.115	0.053	0	0.810	0.102	0.088	0	0.837	0.125	0.038
15	0	0.857	0.098	0.045	0	0.790	0.140	0.070	0	0.812	0.130	0.058
20	0	0.865	0.093	0.042	0	0.858	0.112	0.030	0	0.848	0.112	0.040
25	0	0.855	0.093	0.052	0	0.843	0.115	0.042	0	0.875	0.090	0.035
30	0	0.775	0.138	0.087	0	0.795	0.113	0.092	0	0.765	0.130	0.105

Table 4
The frequencies for the MCCV to choose the number of components contained in the model when $\sigma = 0.004$

n_v	(a)				(b)				(c)			
	3	4	5	≥ 6	3	4	5	≥ 6	3	4	5	≥ 6
1	0	0.700	0.168	0.132	0	0.712	0.158	0.130	0	0.732	0.150	0.118
5	0	0.765	0.135	0.100	0	0.807	0.125	0.068	0	0.777	0.135	0.088
10	0	0.810	0.110	0.080	0	0.797	0.130	0.073	0	0.800	0.138	0.062
15	0	0.820	0.132	0.048	0	0.840	0.095	0.065	0	0.785	0.140	0.075
20	0	0.795	0.133	0.072	0	0.840	0.120	0.040	0	0.820	0.105	0.075
25	0	0.793	0.130	0.077	0	0.813	0.120	0.067	0	0.783	0.150	0.067
30	0	0.550	0.255	0.195	0	0.573	0.222	0.205	0	0.695	0.160	0.145

number of components is the largest if 50% or 60% samples are left out for validation for this model. It is also worth noticing that the frequencies for MCCV to choose three components for the model are zero in all the cases. This indicates that it is almost impossible for the MCCV method to choose the underfitting model. In addition, one could see that collinearity seems to have no adverse influence upon choosing correct number of components for MCCV methods, when random errors are not very big.

Table 4 shows some differences. The frequencies go down quickly at $n_v = 30$, after they reach maximum at $n_v = 15$ or $n_v = 20$ and have become apparently smaller than they are at $n_v = 1$. This fact indicates that it needs more samples in the calibration set in order to obtain the correct model with the largest probability, when there are large random errors in the data. As shown in Table 2, the maxima of frequencies in all the cases in Table 4 for MCCV to choose correct model are slightly smaller than that in Table 3. These results indicate that the larger random errors decrease the possibility to choose the correct model. In general, in order to obtain good performances for MCCV method in the cases in Table 4, 40–50% of samples in the data should be left out for validation.

4.2. Near infrared data

As pointed out in Ref. [10], high degree of collinearity among the spectral variables makes the analysis of this data difficult. For cross validation with $n_v = 1$, the result is shown in Fig. 4. The value of CV reaches its minimum at $k = 7$. It was noticed [10] that there was clear over-fitting at $k = 7$ for the data. The number of components that ought to be included in the model, thus, is difficult to determine by CV_1. Based on the principal component analysis and much of his prolific experiences, Næs [10] suggested using the model with three principal components for prediction. Höskudsson [11] used many methods, such as Akaike information criterion, the C_p statistics and H-error criterion, as well as CV with $n_v = 1$ and $n_v = 0.1 \sim 0.2n$, to determine the number of components for this data. But it was hard to obtain a convincing conclusion.

The MCCV given by Eq. (16) in Section 2.2 is used in the different cases with $n_v = 1, 4, 7, 10, 13,$

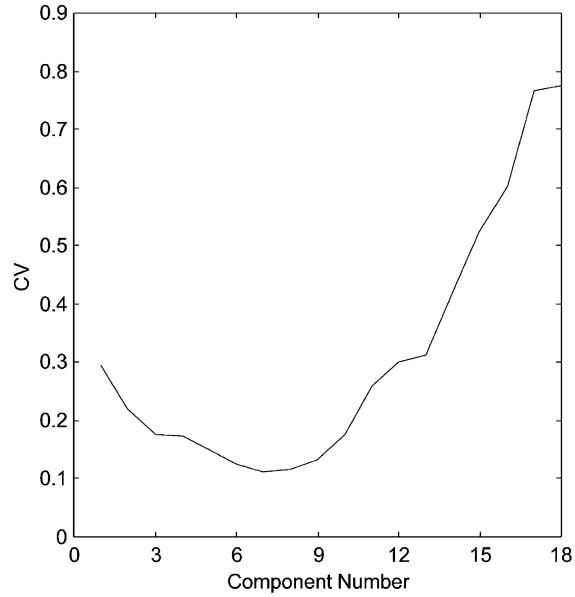

Fig. 4. Leave-one-out CV plot for NIR data.

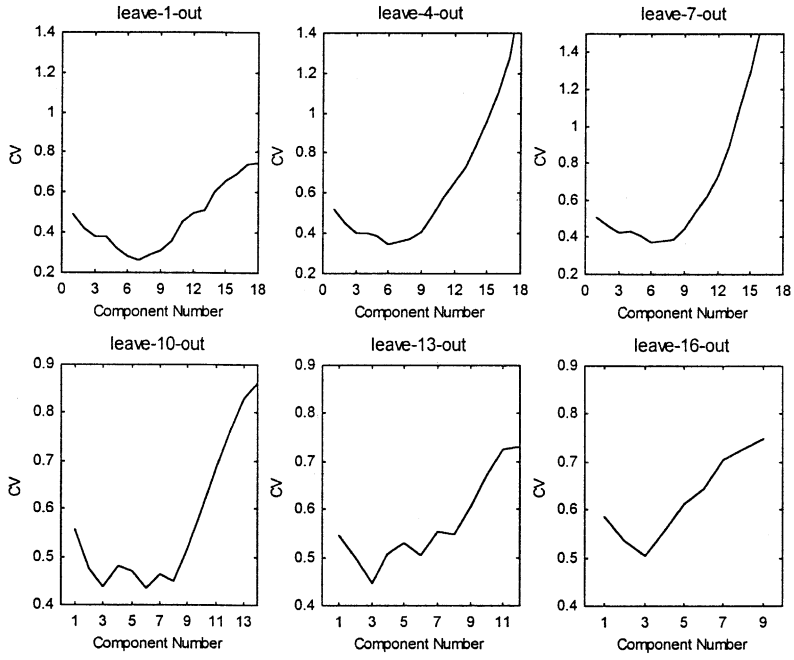

Fig. 5. The CV plot based on MCCV method in various cases for NIR data.

16 and $N = 2.5n$. The results are shown in Fig. 5. One can see from the figure that for $n_v = 1, 4, 7, 10$, CV(k) reach their local minima at $k = 3$, and the global minima, for $n_v = 13, 16$ at $k = 3$. If the first local minimum of CV(k) is used as criterion for determining the number of the components contained into the model, then the CV with any number of samples left out for validation could choose the expected model. But this kind of criterion seems to be a little arbitrary. For MCCV with no other than 50%

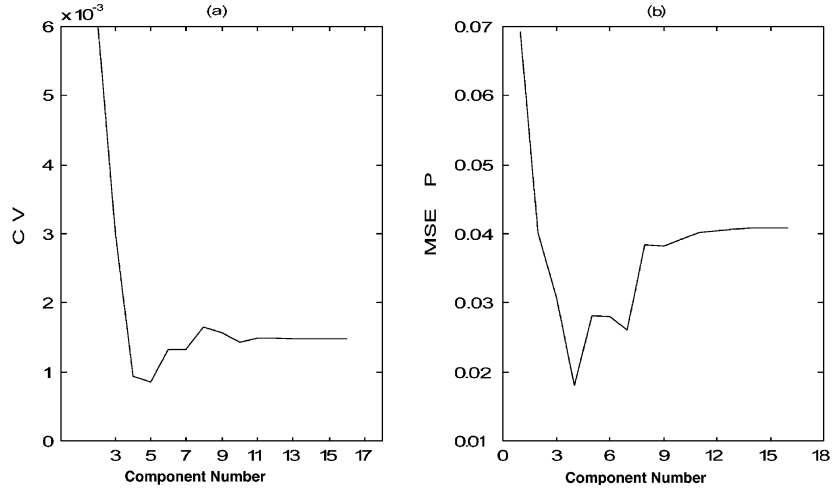

Fig. 6. (a) Leave-one-out CV plot for ultraviolet data. (b) The mean squared error of prediction for ultraviolet data.

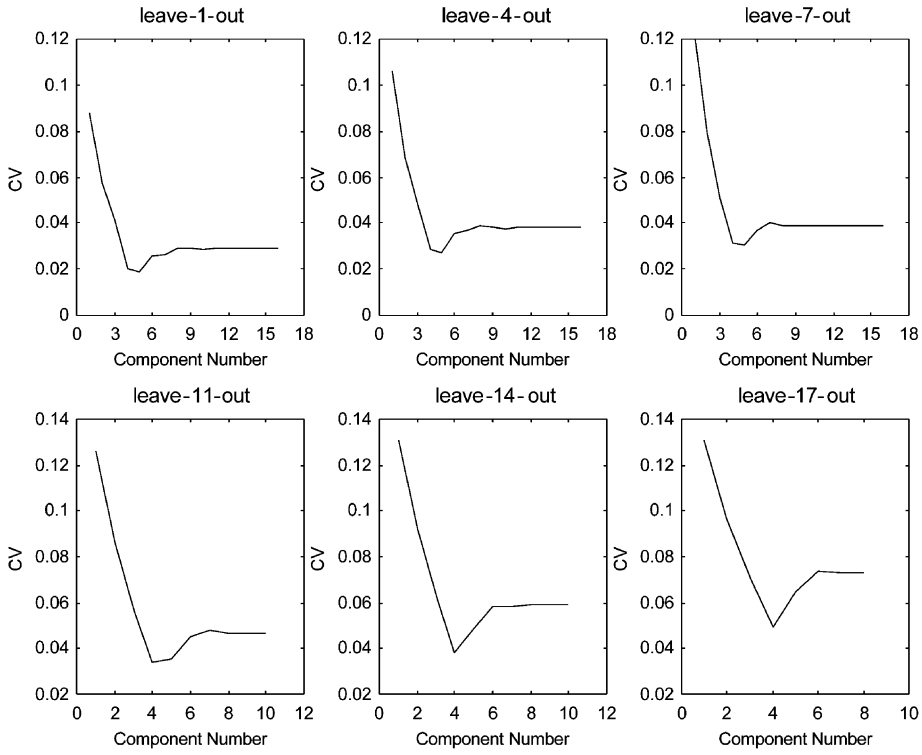

Fig. 7. The CV plot based on MCCV method in various cases for ultraviolet data.

(or more) of samples left out for validation, one can obtain the adapted model for this data set.

4.3. Ultraviolet data

In order to investigate the performance of MCCV for this data set, the samples are split into two parts. One consists of the first 25 samples and is used for the data set for MCCV. The other samples are used for confirming the correct number of components that should be contained into the model. Based on the leave-one-out CV, the model should contain five components. The results are shown in Fig. 6a. However, the values of MSEP(k) (Eq. (17)) shown in Fig. 6b, reach minimum at $k = 4$. Thus, the adapted model should be the one that contains only four components. The model determined by leave-one-out CV includes more components by one.

The MCCV are fulfilled with $N = 80$ and $n_v = 1$, 4, 7, 11, 14, 17. It is seen from Fig. 7 that the CV values for MCCV are very close to each other at $k =$ 4, 5 in the cases $n_v = 1$, 4, 7, 11. But they differ obviously in the cases $n_v = 14$, 17. Thus, the MCCV with more than 50% of samples left out can determine the number of components in the model without suspicion. These results manifest once again that the MCCV with only several or 10–30% of samples left out for validation also tends to involve more components into model in truth and are in concert with the simulation results nicely.

5. Conclusions

Since leave-one-out cross validation is an asymptotically inconsistent method in determining the number of components in multivariate calibration model, it prefers to choose an unnecessary large model and possibly cause over-fitting for prediction. There is a need for a consistent method by which less risk of over-fitting could be taken for small data sets. Monte Carlo cross validation is just the one to be

wanted. For the simulated data set and the two real data sets studied here, the following can be concluded.

(1) MCCV has an obviously larger probability than leave-one-out CV in choosing the correct number of components that the model should contain. The probability has a maximum for a small data set as the number of the samples left out for validation increases. For the examples in this paper, 40–60% of all samples is recommended to be left out for validation for MCCV. It should be pointed out, however, that the recommended percentage of the samples that is left out for validation may be even higher for larger data sets.

(2) It is hard for MCCV or CV to choose the model of under-fitting. The number of components determined by MCCV (or CV) is always not less than the model should contain.

(3) The high levels of random errors defy MCCV to determine the correct number of components for the model. The probabilities decrease as random errors in the data set increase. And lesser number of samples is needed for validation if one wants to obtain the largest probability.

(4) The collinearity in the data set has little influence on the probability for MCCV to choose the accurate model, when random errors are not big in data set.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of People's Republic of China (Grant No. 29735150). The authors are grateful to Prof. R. Manne for his valuable comments.

References

[1] H. Akaike, IEEE Trans. Autom. Control 19 (1974) 716–723.
[2] C.L. Mallows, Technometrics 15 (1973) 661–675.
[3] B. Efron, J. Am. Stat. Assoc. 78 (1983) 316–331.
[4] B. Efron, J. Am. Stat. Assoc. 81 (1986) 461–470.
[5] D.M. Allen, Technometrics 16 (1974) 125–127.
[6] M. Stone, J. R. Stat. Soc. B 36 (1974) 111–147.
[7] G. Wahba, S. Wold, Commun. Stat. 4 (1975) 1–17.
[8] S. Wold, Technometrics 20 (1978) 397–405.
[9] H.A. Martens, P. Dardenne, Chemom. Intell. Lab. Syst. 44 (1998) 91–121.
[10] T. Næs, Chemom. Intell. Lab. Syst. 5 (1989) 155–168.
[11] A. Höskudsson, Chemom. Intell. Lab. Syst. 32 (1996) 37–55.
[12] Unscrambler for Windows, User's Guide, CAMO (1996) Trondheim, Norway.
[13] U.G. Indahl, T. Næs, J. Chemometrics 12 (1998) 261–278.
[14] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.
[15] M. Stone, J. R. Stat. Soc., B 39 (1977) 44–47.
[16] S. Geisser, J. Am. Stat. Assoc. 70 (1975) 320–328.
[17] G. Herrzberg, S. Tsukanov, Utilitas Mathematics 29 (1986) 109–216.
[18] P. Burman, Biometrika 76 (1989) 503–514.
[19] L. Breiman, J.H. Friedman, R.A. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.
[20] P. Zhang, Ann. Stat. 21 (1993) 299–313.
[21] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.
[22] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1989.
[23] S. Wold, Technometrics 35 (1993) 137–139.
[24] A. Höskuldsson, J. Chemometrics 2 (1988) 211–220.
[25] Q.-S. Xu, Y.-Z. Liang, H.-L. Shen, J. Chemometrics 14 (2000) 1–15.