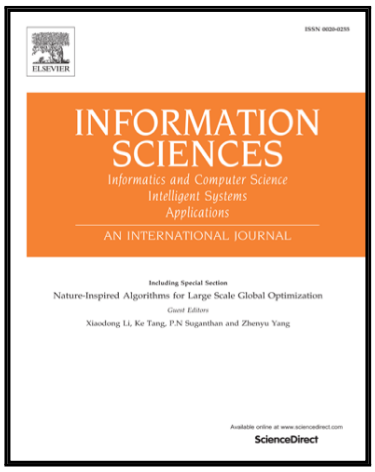

Markov Cross-validation for Time Series Model Evaluations

Gaoxia Jiang, Wenjian Wang

PII: S0020-0255(16)31100-8
DOI: 10.1016/j.ins.2016.09.061
Reference: INS 12557

To appear in: *Information Sciences*

Received date: 21 July 2016
Revised date: 5 September 2016
Accepted date: 28 September 2016

Please cite this article as: Gaoxia Jiang, Wenjian Wang, Markov Cross-validation for Time Series Model Evaluations, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.09.061

Markov Cross-validation for Time Series Model Evaluations

Gaoxia Jiang, Wenjian Wang*

School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China

Abstract

Cross-validation (CV) is a simple and universal tool to estimate generalization ability, however, existing CVs do not work well for periodicity, overlapping or correlation of series. The corresponding three criteria aimed at describing these properties are presented. Based on them, we put forward a novel Markov cross-validation (M-CV), whose data partition can be seen as a Markov process. The partition ensures that samples in each subset are neither too close nor too far. In doing so, overfitting model or information loss of series, which may result in underestimation or overestimation of the error, can be avoided. Furthermore, subsets from M-CV partition could well represent the original series, and it may be extended to time series or stream data sampling. Theoretical analysis shows that M-CV is the unique one which meets all of above criteria among current CVs. In addition, the error estimation on subsets is proved to have less variance than that on original series, therefore it ensures the stability of M-CV. Experimental results demonstrate that the proposed M-CV has lower bias, variance and time consumption than other CVs.

Keywords: Model evaluation; Markov cross-validation; time series

*Corresponding author. Tel. :+86 0351 7017566.
Email addresses: jianggaoxia@sxu.edu.cn (Gaoxia Jiang), wjwang@sxu.edu.cn (Wenjian Wang)

1. Introduction

Time series appears in many fields, e.g. economics, meteorology, finance, medicine and many others. Time series processing techniques mainly include prediction, smoothing, regression and others. Time series prediction (auto-regression) aims to forecast future values by the past series. Time series smoothing is to build an approximating function that attempts to capture important patterns of series. And time series regression is to create a functional relation between the response series and exogenous variables. Evaluating the performance of models of time series is an important problem when choosing the better one among various available models or parameters. Many aspects of models, e.g. generalization ability or error, complexity, interpretability, should be considered. For time series models, generalization error may be the most important factor, so most literatures about comparing time series models focus on it. The key problem for comparison of generalization ability is how to estimate generalization error.

There are some traditional approaches to estimate generalization error at present. Hold-out is an estimator with low computational complexity. Its downside is that the results are highly dependent on the choice for data split [20]. The bootstrap estimator is known to have better performance on small samples. However, in all situations of severe overfit, the estimator is downwardly biased [6]. Cross-validation is an estimator widely used to estimate generalization error for its practicability and flexibility. The above estimators have been compared in related researches [10,11]. Kohavi [11] studied above methods, and the results indicated that the best method for model selection is 10-fold stratified cross-validation. Kim [10] performed an empirical study to compare the 0.632C bootstrap estimator with the repeated 10-fold cross-validation and the repeated one-third hold-out estimator, and the results showed that the repeated CV estimator is recommended for general use. Currently, cross-validation is widely accepted in data analysis and machine learning, and serves as a standard procedure for performance estimation and model selection.

There are some new CVs for time series in recent years. Bergmeir [4] proposed blocked cross-validation (BCV) in evaluating prediction accuracy. Opsomer [17] found that cross-validation will fail when the correlation between errors of time series exists. To solve the correlation, three new CVs called modified cross-validation (MCV), partitioned cross-validation (PCV) and hv-blocked cross-validation (hvBCV) were presented [7,19].

How to measure the generalization error is crucial for comparing time series models because different measurements may provide opposite results, e.g., models with low mean absolute error could have large mean relative error. Salzberg [23] proposed using k-fold CV followed by appropriate hypothesis test to compare models rather than the average accuracy. Many subsequent studies about comparing algorithms are in the schema of cross-validation and hypothesis test (CV & HT)[8]. The variance of error estimation is needed in most hypothesis tests. In addition, Rodriguez [21,22] compared the estimator for different folds of CV and concluded that if the aim is to compare classifiers with similar bias, 2-fold CV is advocated because it has the lowest variance. Therefore, the variance of estimator is very important for comparing models.

The variance of errors is usually estimated before hypothesis tests. On the one hand, the classical variance estimator would be grossly underestimated due to the overlap between training and testing sets [2,3,23]. On the other hand, if series autocorrelation is present, the test error will also be underestimated, but CV is not able to detect this [19]. Existing CVs do not solve above problems at the same time.

This paper aims to design an effective error estimation method for time series models. Considering the periodicity, overlapping or correlation of series, M-CV with Markov property is proposed. Its randomness and independence could overcome the above problems, and the equiprobability and representativeness could balance CV subsets. Furthermore, its low variance could promote the error estimation. These characters ensure that M-CV could provide an effective and accurate estimation of generalization error.

The paper is organized as follows. In Section 2 three criteria are summarized for model evaluation of time series. Based on them, M-CV methodology is proposed. In Section 3 and section 4, some sound properties of M-CV are subsequently illustrated and it is compared with other CVs in theory and experiments. Section 5 concludes.

2. M-CV methodology

2.1. Time series model

This paper focuses on time series smoothing model. Time series smoothing or fitting is a basic representation technique which can be used for distance measures, time series compression, clustering and so on [24].

For a common time series $S = \{y_{t_i}\}, (i = 1, 2, \cdots, n)$, the conventional time series smoothing aims to estimate a function $f(\cdot)$ which could reflect the real series to some extent. It is essentially single-input regression. Time series could be expressed as: $y_{t_i} = f(t_i) + \epsilon_i$, where ϵ_i denotes noise component.

2.2. CV criteria

2.2.1. Randomness of partition

Seasonal and cyclical components usually exist in time series. If series is partitioned periodically in CV procedure, models are likely to learn biased information and may produce inaccurate error estimation. This can be illustrated by the following example.

Fig. 1 shows monthly series of carbon dioxide content in Mauna Loa within 16 years (1965.1∼1980.12) [9]. Two sub-series (series in April and October) and smoothed curves are plotted in Fig. 1. It can be observed that the original series has an obvious seasonal component. The values in April and October are peaks and valleys of series, respectively. Obviously, the two smoothed curves are biased for the whole series. Moreover, if a model is trained on peak points and tested on valley points, the prediction error will be overestimated. Thus periodic partition should be avoided. This can be achieved by the partition with randomness.

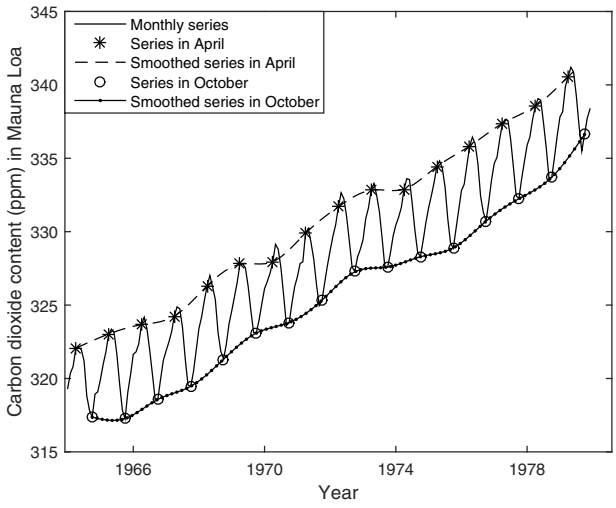

Fig. 1. Smoothed curves on periodical sub-series

2.2.2. Independence of test errors

The variance of test errors is usually estimated by sample variance. However, if we do not take into account the error correlations due to the overlap between training or test sets, naive variance estimator will seriously underestimate the variance [2,20]. There is no overlap between test sets because each example of the original data set is used once and only once as a test example [2]. For most k-fold CVs, there are additional dependencies between training sets. An exception is 2-fold CV whose test errors are independent since the training sets do not overlap [1,2,8].

2.2.3. Independence between training set and test set

If a series is autocorrelated, the model is easily overfitted and the test error will be underestimated [19]. Independence can be assured by leaving a certain distance between training and test samples. In other words, if a sample appears in test set, all other correlated samples have to be removed from the training set to avoid overfitting on the sample. Thus CV partition on time series has to leave a certain distance to keep the independence between training set and test set [4,5,12].

2.3. M-CV

To meet the above criteria, we proposed a novel Markov cross-validation. The main idea is to partition dataset into a few subsets which could well represent the original series, and complete 2-fold CV on each subset. Here come three questions. (1) How many subsets would be suitable? (2) How to partition dataset to obtain representative subsets? (3) How to complete 2-fold CV?

To leave a certain distance between training set and test set, the original series $S = \{y_t\}, (t = 1, 2, \cdots, n)$ is split into positive subset S^+ and negative subset S^- firstly. Just a random split is far away from meeting above criteria.

On one hand, if the distance of adjacent samples in the same subset (S^+ or S^-) is too large, the representativeness of the subset will be damaged. Thus we set a rule for the split that three adjacent samples of S cannot belong to the same subset. The rule can be achieved by the following design.

$$y_t \in \begin{cases} S^+ & \text{if } y_{t-1}, y_{t-2} \in S^- \\ S^- & \text{if } y_{t-1}, y_{t-2} \in S^+ \\ S^+ \text{ or } S^- & \text{otherwise} \end{cases} \tag{1}$$

On the other hand, the partition of positive and negative subsets couldn't meet the third criterion because samples in the same subset may be adjacent. Thus further split is needed for the two subsets. To keep the balance, each subset is split into m sub-subsets. Here m is a parameter related to series autocorrelation order p, i.e. large p needs large distance between training set and test set which can be produced by large m. More precisely, M-CV with too little m couldn't meet the third criterion, and too large m will damage the representativeness of each sub-subset. The calculation of the minimal m which meets the third criterion is given by

$$m = \begin{cases} 2p/3 & \text{if } p\%3 = 0 \\ 2\lfloor p/3 \rfloor + 2 & \text{otherwise} \end{cases} \tag{2}$$

where % is modulo operation, and $\lfloor \cdot \rfloor$ denotes round down function.

After partitioning S to S^+/S^- and finding m, the following steps become easy. As the first partition (S^+/S^-) is partly random, the second partition can be periodic. Every m-th sample in S^+ is selected as an element of a sub-subset, then S^+ is split into m sub-subsets (Nos. $1\sim m$) just like the systematic sampling. S^- can be split into m sub-subset (Nos. $(m+1) \sim 2m$) in the same way. So we have $\bigcup_{u=1}^{m} S_u = S^+$, $\bigcup_{u=m+1}^{2m} S_u = S^-$ and $S^+ \bigcup S^- = S$. After the two steps, the original series S is partitioned into $2m$ sub-subsets, and thus it is named $2m$ partition.

For the third question, the training set and test set of 2-fold CV are selected from a sub-subset according to whether the ordinal number of each sample in the sub-subset is odd or even. 2-fold CV is completed on all $2m$ sub-subsets in this way.

The above steps show the way of partitioning the original series into $2m$ sub-subsets and the way of completing 2-fold CV on each set. As the key first partition has the property that the attribution of any sample depends only upon those of the former two, the new CV is named Markov CV. The detailed implement is shown in Algorithm 1.

To show M-CV intuitively, an example is given below. Fig. 2 shows that 100 equal spaced samples in a series are partitioned into four subsets as $p = 2$. It can be seen that samples of subsets are not circularly selected from the original series. Each sample is trained and tested only once and there is no overlap between training set and test set. Generally, the time distance of y_{t_i} and y_{t_j} is $|t_i - t_j|$. And the time distance (hereafter called distance simply) between solid line and dash line in each subset, i.e., the distance between

Algorithm 1 The procedure of M-CV.

Require:

Time series $S = \{y_t\}, (t = 1, 2, \cdots, n)$ with autocorrelation order p;

Ensure:

Prediction errors $\hat{e}_u$;

if $p\%3 = 0$ **then**
  $m = 2p/3 + 1$;
  $m = 2\lfloor p/3 \rfloor + 2$.
end if // **end the calculation of partition number** $2m$
i=1, j=-1;
The state (d) is randomly initialized by one of the following four cases:
  $d_1 = i{+}{+}$, $d_2 = i{+}{+}$;
  $d_1 = i{+}{+}$, $d_2 = j{-}{-}$;
  $d_1 = j{-}{-}$, $d_2 = i{+}{+}$;
  $d_1 = j{-}{-}$, $d_2 = j{-}{-}$.
for $t = 3, \cdots, n$ **do**
  if $(d_{t-1} > 0)\&(d_{t-2} > 0)$ **then**
    $d_t = j{-}{-}$;
  else $\{(d_{t-1} < 0)\&(d_{t-2} < 0)\}$
    $d_t = i{+}{+}$;
  else $\{rd > 0.5\}$
    $//rd$ is a random number from a uniform distribution on [0,1]
    $d_t = j{-}{-}$;
  else
    $d_t = i{+}{+}$;
  end if
end for // **end Markov iteration**
$Id = d\%m + 1 + I(d > 0) \cdot m$; $//I(\cdot)$ is indicator function
$S_u = \{y_t | Id_t = u\}, u = 1, 2, \cdots, 2m.$ // **end** $2m$ **subsets partition**
for $u = 1, 2, \cdots, 2m$ **do**
  Divide S_u into two groups or folds S_{uo} and S_{ue} according to whether the ordinal number of a sample in S_u is odd or even.
  Obtain prediction errors $\hat{e}_u$ by completing 2-fold CV with S_{uo} and S_{ue}.
end for // **end 2-fold CV on each subset**

training set and test set, is beyond the autocorrelation order. In a word, M-CV satisfies the above three criteria in this case.

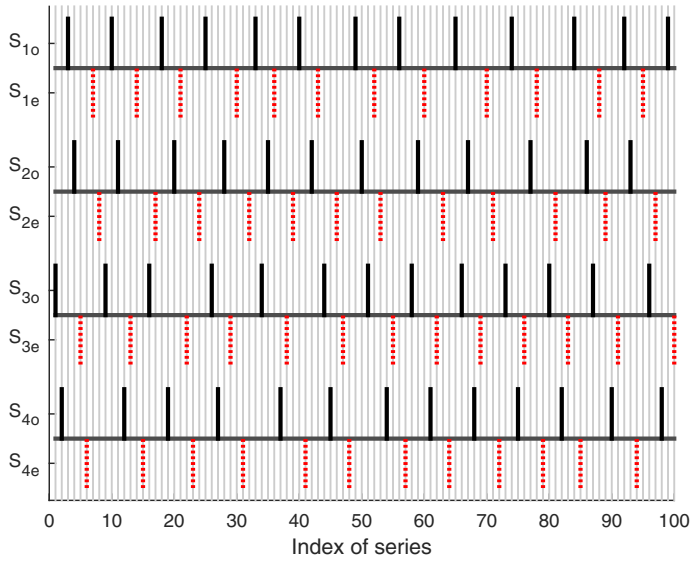

Fig. 2. An example of M-CV (p=2, $2m$=4)

3. Properties discussion and comparison

3.1. Properties of M-CV

3.1.1. Randomness and equiprobability of partition

When the former two samples do not belong to the same subset (S^+ or S^-), the next one is partitioned into any subset randomly (see Eq. (1)). So M-CV has randomness and meets the first criterion. The following proposition shows that the M-CV partition has not only randomness but also equiprobability, so the subsets are relatively balanced.

Proposition 1. *Each sample has the same probability of belonging to any subset in M-CV, i.e.* $P\{y_t \in S_u\} = \frac{1}{2m}, (u = 1, 2, \cdots, 2m)$ *for any* t $(t = 1, 2, \cdots, n)$.

Proof. Suppose that subset pair of any two adjacent samples y_t, y_{t+1} has four cases or states: $St_1(S^+S^+), St_2(S^-S^-), St_3(S^+S^-)$ and $St_4(S^-S^+)$. Let α_t be the probability vector of four states for y_t, y_{t+1}.

Considering the partition of positive and negative subsets (see Eq. (1)), the data partition can be seen as a Markov process. The initial state probabilities of being one of above four states can be denoted as a vector $\alpha_1 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$, i.e., the subsets including y_1, y_2 have the same probability of being any one of the four states. From the iteration step, the state transition probability matrix is:

$$M = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{3}$$

We will prove the proposition by mathematical induction. It is obvious $\alpha_1(1) = \alpha_1(2) = \alpha_1(3) = \alpha_1(4)$. Assume $\alpha_t(1) = \alpha_t(2)$ and $\alpha_t(3) = \alpha_t(4)$, α_t can then be denoted as $\begin{pmatrix} c & c & \frac{1}{2} - c & \frac{1}{2} - c \end{pmatrix}$, where c is a constant between 0 and 0.5.

$\alpha_{t+1} = \alpha_t \cdot M = \begin{pmatrix} \frac{1-2c}{4} & \frac{1-2c}{4} & \frac{1+2c}{4} & \frac{1+2c}{4} \end{pmatrix}$. It can be observed that $\alpha_{t+1}(1) = \alpha_{t+1}(2)$ and $\alpha_{t+1}(3) = \alpha_{t+1}(4)$ still hold. We can conclude that $\alpha_t(1) = \alpha_t(2)$ and $\alpha_t(3) = \alpha_t(4)$ hold for any t. Thus $P\{y_t \in S^+\} = \alpha_t(1) + \alpha_t(3) = \alpha_t(2) + \alpha_t(4) = P\{y_t \in S^-\} = \frac{1}{2}$.

Both S^+ and S^- are split into m balanced sub-subsets, so $P\{y_t \in S_u\} = \frac{1}{2m}$ holds for any t. □

The above property indicates that each sample randomly belongs to a subset under Markov constraint. The randomness of $2m$ partition in M-CV solves the drawbacks of periodic partition.

3.1.2. Boundedness of sample distance

Without loss of generality, assume the time distance between any two adjacent samples in original series to be 1. The bound of distance between samples of a subset is discussed in Proposition 2.

Proposition 2. *The distance between any two adjacent samples $y_{u_k}, y_{u_{k+1}}$ of a subset $S_u = \{y_{u_k}\}, k = 1, 2, \cdots, n/2m, u \in \{1, 2, \cdots, 2m\}$, is bounded. That is $\frac{3m-1}{2} \leq |u_k - u_{k+1}| \leq 3m$.*

Proof. Assume that $S_u \subset S^+$. In S^+ , there must be $m-1$ samples between any two adjacent samples of S_u. For any m, there are at most two samples of S^- between any two adjacent samples of S^+. It means that the maximal distance between any two adjacent samples of S^+ is 3. When any two adjacent

samples of S^+ are separated by two samples of S^-, the distance between any two adjacent samples of S_u reaches the maximum, i.e. $max|u_k - u_{k-1}| = 3m$.

When any two adjacent samples of S^- are separated by two samples of S^+, S^+ is the most compact, and the distance between any two adjacent samples of S_u reaches the minimum. If $m\%2 = 0$, the minimal distance between any two adjacent samples of S_u is $\frac{3m}{2}$; otherwise it is $\frac{3(m-1)}{2} + 1$. Therefore, $|u_k - u_{k+1}| \geq \frac{3m-1}{2}$.

It gives the same result when $S_u \subset S^-$. □

To illustrate this proposition, two examples are given. Let Id be a vector about subsets, and $Id(t) = u$ means $y_t \in S_u$.

For $m = 4$, $S_1 \bigcup S_2 \bigcup S_3 \bigcup S_4 = S^+$ and $S_5 \bigcup S_6 \bigcup S_7 \bigcup S_8 = S^-$. When $Id = (\cdots, 1, 2, 5, 3, 4, 6, 1, \cdots)$, any two adjacent samples of S^- are separated by two samples of S^+. Now two examples in S_1 have the minimal distance $(3m/2 = 6)$; When $Id = (\cdots, 1, 5, 6, 2, 7, 8, 3, 5, 6, 4, 7, 8, 1, \cdots)$, any two adjacent samples of S^+ are separated by two samples of S^-, and two examples in S_1 have the maximal distance $(3m = 12)$.

For $m = 5$, $S_1 \bigcup S_2 \bigcup S_3 \bigcup S_4 \bigcup S_5 = S^+$ and $S_6 \bigcup S_7 \bigcup S_8 \bigcup S_9 \bigcup S_{10} = S^-$. When $Id = (\cdots, 1, 2, 6, 3, 4, 7, 5, 1, \cdots)$, any two adjacent samples of S^- are separated by two samples of S^+ and two examples in S_1 have the minimal distance $((3m - 1)/2 = 7)$; When $Id = (\cdots, 1, 6, 7, 2, 8, 9, 3, 10, 6, 4, 7, 8, 5, 9, 10, 1, \cdots)$, any two adjacent samples of S^+ are separated by two samples of S^- and two examples in S_1 have the maximal distance $(3m = 15)$.

The boundedness of sample distance shows that any two adjacent samples in any M-CV sub-subset will be neither too close nor too far. As 2-fold CV is applied on each sub-subset, the distance between training sample and test sample is also bounded.

If samples in training and test sets have a short distance, the model will be easily overfitted. If they have a long distance, the test errors are usually overestimated. The boundedness of M-CV prevents the above two situations. Furthermore, the lower bound makes M-CV satisfying criterion 3, and the upper bound helps each subset to represent the original series well.

3.1.3. Independences

The independences described in criteria 2 and 3 are discussed in the following propositions.

Proposition 3. *There is no overlap between training or test sets of M-CV.*

The result obviously holds as it is a special case of 2-fold CV on each subset. Therefore, M-CV meets criterion 2.

Proposition 4. *The distance between any two samples in a subset is beyond the autocorrelation order, i.e.,* $|u_k - u_{k'}| > p$.

Proof. According to Eq. (2), for $y_{u_k}, y_{u_{k'}}$ in subset S_u, when $m\%2 = 1$ (or $p\%3 = 0$), $|u_k - u_{k'}| \geq min\{|u_{k-1} - u_k|, |u_k - u_{k+1}|\} \geq \frac{3m-1}{2} = p + 1 > p$; When $m\%2 = 0$ (or $p\%3 \neq 0$), $|u_k - u_{k'}| \geq min\{|u_{k-1} - u_k|, |u_k - u_{k+1}|\} \geq \frac{3m}{2} = 3 \cdot \lfloor \frac{p}{3} \rfloor + 3 > p$.

And thus $|u_k - u_{k'}| > p$ holds for any p. □

It means that the distance between any training sample and test sample in a subset is larger than the autocorrelation order p. So M-CV satisfies criterion 3.

From the above analysis, M-CV matches all three criteria summarized in subsection 2.2. Two more good properties will be described in the following subsections.

3.1.4. Representativeness

Proposition 2 shows that the distance between any two samples in a subset is not larger than 1.5 times of the partition number ($2m$), and each sample in a subset is related to samples within the distance p. So any subset contains information of more than $n/2m$ samples as series autocorrelation. The upper bound of distance ensures that M-CV could well represent the original series.

A definition which reflects the representativeness of a subset in relation to original series is given here.

Definition 1. *Coverage rata (CR) of a subset* S_u *in relation to* $S = \{y_t\}, (t = 1, 2, \cdots, n)$:

$$CR(S_u) = \frac{\sum_{t=1}^{n} I\left(\min_{y_{u_k} \in S_u} |u_k - t| \leq p\right)}{n}.$$

CR denotes how many samples in the original series are related to samples of the subset (within a distance p). It measures the capability of retaining the information of the original series. Fig. 3 shows an example. The original series is partitioned into four subsets as p=2. Fig. 3 plots the first subset S_1

and its coverage information. The solid lines denote the indexes of samples of S_1. The dash lines denote the indexes of samples uncovered by S_1. And dash dot lines denote the indexes of samples covered by S_1. There are three uncovered samples whose minimal distance from any samples of S_1 is larger than p. In this case, $CR(S_1) = 97\%$.

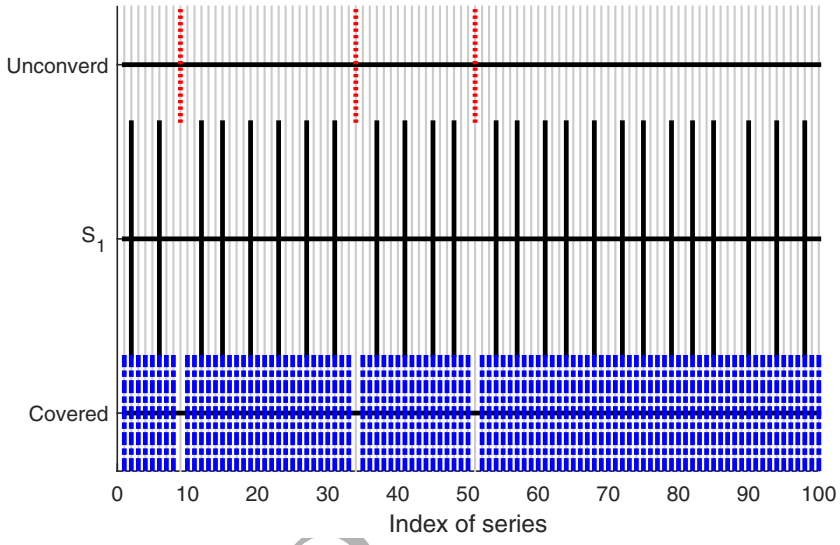

Fig. 3. Coverage of S_1 in relation to the original series

Why the partition number is $2m$, but not $2(m-1)$ or $2(m+1)$? It will be answered by the third criterion and representativeness.

Table 1 lists the autocorrelation order and three kinds of partition number. By the lower bound in Proposition 2, it is not hard to find that $2m$ is the minimal partition number in M-CV which could make the minimal distance between samples of subsets be larger than p. It means that M-CV with partition number less than $2m$ may not meet criterion 3.

Fig. 4 shows the minimal gaps between samples of subsets with $2(m-1)$, $2m$, and $2(m+1)$ partitions. In Fig. 4, the minimal gaps have the following sequence: $2(m+1) > 2m > 2(m-1)$. In other words, the minimal gap increases with the partition number. And the minimal gaps with $2(m+1)$ and $2m$ partitions are beyond p, but it is not for $2(m-1)$ partition. So M-CV with $2(m-1)$ partition doesn't meet criterion 3.

Table 1. Different partition numbers for p

p	2(m-1)	2m	2(m+1)
1	2	4	6
2	2	4	6
3	4	6	8
4	6	8	10
5	6	8	10
6	8	10	12
7	10	12	14
8	10	12	14
9	12	14	16
10	14	16	18

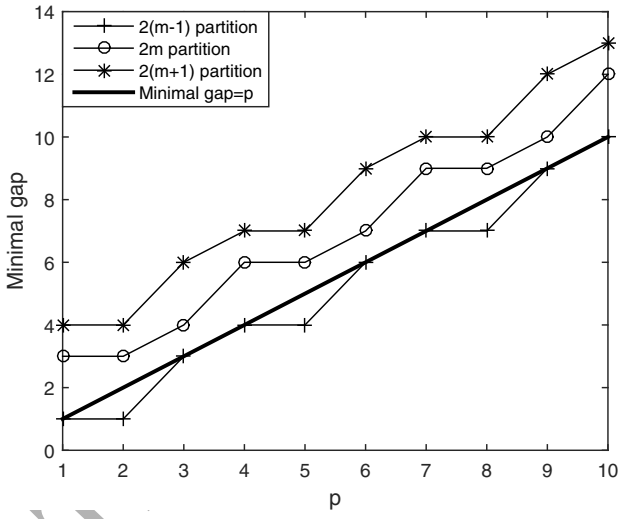

Fig. 4. Minimal gaps of different partitions

The representativeness helps M-CV find the best partition number. The average CRs of M-CV with $2(m-1), 2m$, and $2(m+1)$ partitions are displayed in Fig. 5. It can be observed that CR increases with p for each partition and is very close or equal to 100% when $p > 4$. The average CRs have the following sequence: $2(m-1) > 2m > 2(m+1)$. So M-CV subsets with less partition number have better representativeness. It also shows that subsets from $2m$ partition still retain almost all of the information in the original series except for p=1 in which situation CR is not bad (75%). In general, considering criterion 3 and the representativeness, $2m$ partition is

the best balance for M-CV. The test errors are usually overestimated if the model is trained on part of original dataset. While it prevents M-CV from overestimating error that CRs of M-CV subsets approach to 100%.

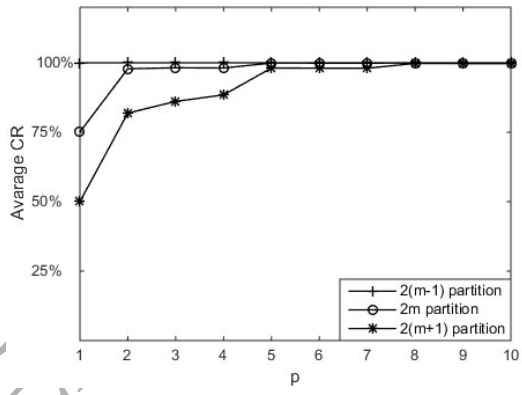

Fig. 5. Average CRs of different partitions (simulated on 10000 samples, repeated 10 times)

3.1.5. Low variance

As most subsets could well represent the original data set, it can be assumed that the performance of model trained on all data set is similar to that on any subset. Or particularly, test errors from all data and any subset have the same distribution, i.e., $Var(\hat{e}) = Var(\hat{e}_u)$, where $Var(\cdot)$ is variance

function, e_i and e_{u_k} are test errors of models trained on S and subset S_u, and $\hat{e} = \frac{1}{2}\sum_{i=1}^{n} e_i$, $\hat{e}_u = \frac{2m}{n}\sum_{k=1}^{n/2m} e_{u_k}$, $\hat{e}_{M-CV} = \frac{1}{2m}\sum_{u=1}^{2m}\hat{e}_u$.

Proposition 5. *The variance of M-CV:* $\hat{e}_{M-CV} = \frac{Var(\hat{e})}{2m}$.

Proof. As $S_u \bigcap S_{u'} = \varnothing$ for $u \neq u'$, $\hat{e}_u$ is independent of $\hat{e}_{u'}$.
$Var(\hat{e}_{M-CV}) = Var(\frac{1}{2m}\sum_{u=1}^{2m}\hat{e}_u) = \frac{1}{4m^2}\sum_{u=1}^{2m} Var(\hat{e}_u) = \frac{Var(\hat{e}_u)}{2m} = \frac{Var(\hat{e})}{2m}$. □

It means that the variance of M-CV is less than that of CV on the original data set. In other words, 2-fold CV after $2m$ partition is better than a direct 2-fold CV on original data set from the view of variance. The variance difference between them will be more notable for larger m or p because the variance of M-CV is inversely proportional to m. Note that m and p are fixed values for a given series. M-CV variance cannot be regulated by changing m or p.

3.2. Comparisons of time series CVs

M-CV is compared with other four time series CVs (BCV, MCV, PCV, hvBCV). Their main ideas are listed in Table 2.

Table 2. CVs for time series

CVs	Parameter	Full name	Main idea
BCV	b	Blocked CV	Samples are partitioned into b continues subset or block.
MCV	l	Modified CV	Leaving out the $2l$+1 samples surrounding the test observation.
PCV	g	Partitioned CV	Samples are partitioned into g subsets by taking every g-th samples and CV is performed for each subset.
hvBCV	h,v	h-v Blocked CV	Remove v samples around $2h$+1 test samples.
M-CV	p/m	Markov CV	The partition of each sample is independent of others except for the former two.

Take the case of a series with autocorrelation order 2, the procedures of above CVs (in a round) are displayed in Fig. 6. It can be observed that test samples keep a certain distance for all CVs except BCV. Most samples are trained on each train-test round for BCV, MCV and hvBCV, and their time complexities are larger than those of the other two.

Some theoretical properties and time complexity are listed in Table 3. '$\surd$' and '$\times$' denote whether CVs meet three criteria in subsection 2.2. Because samples are partitioned or validated randomly in M-CV and PCV, both of them meet criterion 1. As samples in all CVs except for M-CV could be

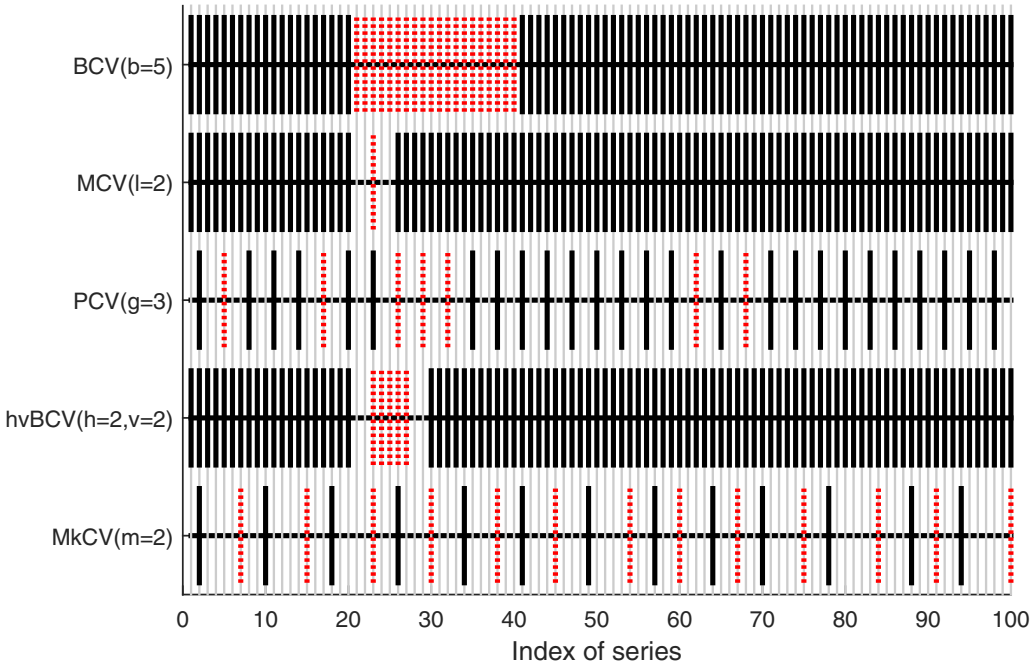

Fig. 6. Five CVs in one train-test round (solid line: training set; dash line: test set)

overlapped between training sets, only M-CV meets criterion 2. The minimal distance between training and test sets could be set by parameters of CVs except for BCV, thus only BCV doesn't meet criterion 3. Table 3 indicates that only M-CV meets all three criteria. Assume that the time complexity of a smoothing model is n^2, we could conclude that PCV and M-CV don't raise the complexity.

Table 3. Comparisons of five CVs

CVs	**Criterion 1**	**Criterion 2**	**Criterion 3**	**Complexity (for n^2 model)**
BCV	×	×	×	n^3
MCV	×	×	√	n^3
PCV	√	×	√	n^2
hvBCV	×	×	√	n^3
M-CV	√	√	√	n^2

4. Experiments and analysis

In this section, five CVs are compared in respect of accuracy, efficiency and stability. And M-CV errors of four time series models are compared by statistical test.

4.1. Data sets, models and evaluation indicators

Ten synthetic or real-world time series (see Table 4) are smoothed by four models with relatively good parameters (see Table 5) in our experiments.

Table 4. Datasets

	Dataset	Sampling interval	Size	Start	End	Source
1	Sinc data (1)	0.012	1000	-6	6	$sin(\pi t)/(\pi t) + \epsilon, \epsilon \sim N(0, 0.05^2)$
2	Sinc data (2)	0.012	1000	-6	6	$sin(\pi t)/(\pi t) + \epsilon, \epsilon \sim AR(2), Var(\epsilon) = 0.05^2$
3	Sinc data (3)	0.012	1000	-6	6	$sin(\pi t)/(\pi t) + \epsilon, \epsilon \sim N(0, 0.3^2)$
4	Sinc data (4)	0.012	1000	-6	6	$sin(\pi t)/(\pi t) + \epsilon, \epsilon \sim AR(1), Var(\epsilon) = 0.3^2$
5	Gold price	Month	426	1978/12	2014/5	World Gold Council: https://www.gold.org/
6	Gold price	Day	1000	2010/8/13	2014/6/12	World Gold Council: https://www.gold.org/
7	Count of total rental bikes	Day	730	2011/1/1	2012/12/30	UCI: http://archive.ics.uci.edu/ml/index.html
8	EER in China	Month	248	1994/1	2014/8	BIS: http://www.bis.org/
9	Unemployment	Month	728	1954/1	2014/8	https://datamarket.com/
10	Dow-Jones industrial index	Month	290	1968/8	1992/9	https://datamarket.com/

Table 5. Parameters of time series smoothing models

Model	Parameter	Values on series Nos.									
		1	2	3	4	5	6	7	8	9	10
BS	Smoothing parameter	0.1	0.5	1	5	10	20	50	100	200	500
LPR	Bandwidth	0.1	0.5	1	5	10	20	50	100	200	500
GS	Bandwidth	0.01	0.1	0.5	1	2	5	10	20	30	50
EP	Bandwidth	0.01	0.05	0.1	0.5	1	2	5	8	10	15

B-spline smoothing (BS) model is a combination of polynomial pieces and difference penalties. The smoothness is controlled by a penalty parameter [13]. For local polynomial regression (LPR), a low-degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is estimated [16]. In Nadaraya-Watson kernel regression, smoothing function is a weighted estimate with kernel functions like Gaussian kernel (GS) and Epanechnikov kernel (EP) [14,18].

A small part (20% in our experiments) of time series is randomly selected as *out-set*. The remaining part called *in-set* is partitioned according to each

CV procedure. Time series smoothing models are trained and tested on *in-set*, then the estimate of prediction error ($\hat{PE}$) could be calculated from test error. $\hat{PE}$ in the mean squared error (MSE) form is calculated in the following way:

$$\hat{PE} = \frac{1}{n_1}\sum_{i=1}^{n_1}[y_i - f(t_i)]^2 \tag{4}$$

where $f(\cdot)$ denotes a trained model, y_i is a sample in testing set, and n_1 is the size of testing set. The variance is calculated by estimations of all folds. The true prediction error (PE) is obtained by validating a well-trained model $f_{in-set}(\cdot)$ on *out-set* which consists of n_2 samples.

$$PE = \frac{1}{n_2}\sum_{i=1}^{n_2}[y_i - f_{in-set}(t_i)]^2 \tag{5}$$

$\hat{PE}$ of a good CV should be close to PE. To be consistent with Ref. [5], here we adopt mean absolute predictive accuracy error (MAPAE) which is defined as:

$$MAPAE = \frac{1}{r}\sum_{i=1}^{r}|\hat{PE}_i - PE_i| \tag{6}$$

where r is the number of *out-set*. In the following experiments, *out-set* is randomly selected 10 times for each series. Generally, the less MAPAE is, the better the CV estimator is.

The accuracy, efficiency and stability of M-CV are measured by MAPAE, time consumption and variance, respectively.

4.2. Performances of CVs

Five CVs, including BCV($b = 5$), MCV($l = p$), PCV($g = 3$), hvBCV($h = p, v = 5$) and M-CV, are compared in this subsection. Here p is estimated by autocorrelation function of series. Tables 6-8 show MAPAE, time consumption and variance of five CVs with four models on ten series, respectively. The least values are in bold font. They are sorted in ascending order among different CVs, then we can obtain their ranks. Figs. 7-9 show the ranks of MAPAE, time consumption and variance, respectively.

From Table 6, the frequencies of having least MAPAE for five CVs are 0/40, 5/40, 4/40, 5/40 and 26/40, respectively. They are 6/10, 8/10, 7/10,

Table 6. MAPAE of five CVs

model	CVs	Series Nos.									
		1	2	3	4	5	6	7	8	9	10
BS	BCV	1.13E-02	1.12E-02	1.48E-02	1.43E-02	1.73E+09	2.38E+06	9.06E+12	2.72E+03	2.04E+01	8.10E+07
	MCV	1.15E-06	2.00E-08	6.07E-06	2.73E-06	7.55E+08	1.00E+04	1.02E+12	1.48E+02	1.38E+01	4.70E+06
	PCV	1.15E-03	1.59E-05	3.87E-05	**1.50E-06**	6.63E+10	2.78E+07	2.70E+12	1.52E+04	1.34E+01	6.53E+09
	hvBCV	**1.07E-06**	2.86E-08	6.15E-06	4.18E-06	**7.20E+08**	1.13E+04	1.02E+12	2.00E+02	**1.31E+01**	6.22E+06
	M-CV	1.10E-06	**1.01E-08**	**5.70E-06**	2.34E-06	7.41E+08	**6.46E+03**	**9.37E+11**	**1.22E+02**	1.41E+01	**5.11E+05**
LPR	BCV	2.56E-02	1.73E-03	4.14E-01	1.04E+00	3.58E+10	1.10E+09	8.74E+16	1.15E+06	6.79E+01	2.89E+11
	MCV	1.28E-04	1.71E-06	3.26E-04	1.39E-02	5.22E+08	2.61E+05	9.83E+14	**1.88E+03**	1.49E+00	8.91E+07
	PCV	1.75E-05	5.43E-05	5.87E-07	6.17E-05	1.64E+09	6.10E+07	2.36E+12	6.23E+03	**1.47E+00**	4.67E+11
	hvBCV	4.10E-03	3.31E-04	2.08E-02	2.90E-01	2.88E+09	1.06E+06	2.89E+16	9.56E+03	1.10E+01	2.64E+09
	M-CV	**1.06E-08**	**9.84E-07**	**2.97E-07**	**1.22E-05**	**5.78E+07**	**2.35E+02**	**4.92E+11**	2.50E+04	1.69E-02	**1.35E+07**
GS	BCV	1.40E-02	1.41E-02	1.56E-02	1.58E-02	1.02E+09	5.07E+08	4.47E+12	5.66E+02	4.02E+00	2.06E+07
	MCV	1.18E-07	3.51E-07	**1.12E-07**	**2.97E-06**	3.60E+07	1.54E+05	**9.24E+10**	2.86E+02	2.08E-01	5.37E+06
	PCV	1.69E-07	1.00E-07	4.09E-07	4.14E-06	2.50E+07	3.15E+05	1.61E+11	1.22E+02	7.33E-02	5.84E+06
	hvBCV	1.07E-06	4.54E-06	5.96E-07	1.75E-05	6.88E+07	2.34E+05	1.07E+11	4.21E+02	6.25E-01	7.78E+06
	M-CV	**8.40E-09**	**9.77E-09**	1.61E-07	3.88E-06	**6.51E+06**	**7.86E+04**	1.60E+11	**4.76E+01**	**1.94E-02**	**1.75E+05**
EP	BCV	1.68E-06	1.30E-07	6.43E-06	5.34E-05	1.73E+07	6.10E+04	7.47E+11	6.73E+01	4.41E-02	5.43E+05
	MCV	1.97E-08	3.23E-07	1.19E-06	**1.79E-06**	6.71E+06	6.47E+04	1.01E+11	3.47E+01	1.95E-01	1.80E+06
	PCV	2.80E-07	**7.98E-08**	**3.96E-07**	2.12E-06	8.06E+06	2.31E+04	1.35E+11	2.66E+01	2.25E-02	9.36E+05
	hvBCV	**8.34E-09**	2.73E-07	4.70E-07	6.26E-06	9.20E+06	3.62E+04	**9.33E+10**	3.66E+01	2.50E-01	8.14E+05
	M-CV	1.49E-08	2.45E-07	4.05E-07	1.93E-06	**1.72E+06**	**3.51E+03**	1.73E+11	**4.35E+00**	**1.79E-03**	**5.74E+04**

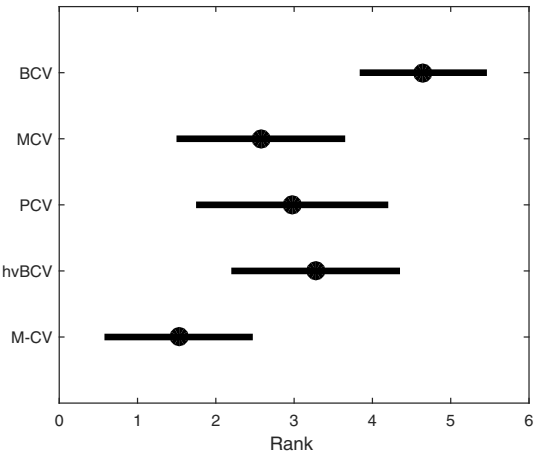

Fig. 7. Rank of five CVs' MAPAEs

5/10 for M-CV on four models, respectively. For ten series, there are 4/4 on two series, 3/4 on three series, 2/4 on four series and 1/4 on one series for M-CV. When M-CV does not have the least MAPAE, its result is comparable with the least (less than 110% of the least MAPAE) in 6 cases. It means that the real frequency of being accurate CV for M-CV reaches 80% (32/40).

In Fig. 7, mean ranks have the following sequence: M-CV<MCV<PCV <hvBCV<BCV. BCV has the largest rank in most cases. MCV, PCV and hvBCV usually have median ranks. M-CV is more likely to have the least rank (minimal MAPAE). Thus it is more accurate than other CVs.

In Table 7, the frequencies of having least time consumptions for five CVs are 6/40, 0/40, 3/40, 1/40 and 30/40, respectively. They are 9/10, 10/10, 6/10, 5/10 for M-CV on four models, respectively. So M-CV with BS or LPR tends to have the least time consumption. There are 4/4 on three series, 3/4 on four series and 2/4 on three series for M-CV. It means M-CV has the least time consumption on most series. When M-CV does not have the least time consumption, its result is comparable with the least (less than 110% of the least time consumptions) in 8 cases. So the real frequency of being efficient CV for M-CV reaches 95% (38/40).

In Fig. 8, mean ranks have the following sequence: M-CV<PCV<BCV <hvBCV <MCV. The least rank appears more frequently in M-CV and PCV because they have less complexity (see Table 3). MCV and hvBCV usually

Table 7. Time consumptions of five CVs (Seconds)

model	CVs	Series Nos.									
		1	2	3	4	5	6	7	8	9	10
BS	BCV	1.41E+00	1.27E+00	1.54E+00	1.51E+00	5.81E-01	5.46E+00	1.02E+00	3.44E-01	1.00E+00	3.76E-01
	MCV	2.84E+02	2.52E+02	3.06E+02	2.99E+02	4.52E+01	1.07E+03	1.44E+02	1.45E+01	1.40E+02	1.89E+01
	PCV	2.96E+00	2.80E+00	3.37E+00	3.22E+00	1.53E+00	1.19E+01	2.28E+00	1.01E+00	2.29E+00	1.05E+00
	hvBCV	2.72E+01	**3.12E+01**	5.17E+01	2.96E+01	5.55E+00	1.77E+02	1.75E+01	1.79E+00	1.37E+01	3.13E+00
	M-CV	**3.43E-01**	3.34E-01	**3.82E-01**	**3.87E-01**	**2.04E-01**	**1.37E+00**	**2.94E-01**	**1.48E-01**	**2.89E-01**	**1.18E-01**
LPR	BCV	4.23E+00	4.20E+00	4.18E+00	4.07E+00	2.59E-01	1.64E+01	1.74E+00	9.51E-02	1.61E+00	7.99E-02
	MCV	1.65E+00	1.63E+00	1.64E+00	1.63E+00	1.14E-01	7.10E+00	6.24E-01	4.54E-02	6.36E-01	5.65E-02
	PCV	3.27E-01	3.27E-01	1.73E-01	2.82E-01	8.09E-02	1.36E+00	1.97E-01	5.05E-02	1.67E-01	4.27E-02
	hvBCV	3.21E+00	3.36E+00	3.38E+00	3.40E+00	1.25E-01	1.43E+01	1.10E+00	4.48E-02	1.09E+00	5.58E-02
	M-CV	**1.38E-01**	**1.27E-01**	**1.37E-01**	**1.27E-01**	**4.53E-02**	**5.70E-01**	**7.42E-02**	**2.91E-02**	**7.01E-02**	**1.85E-02**
GS	BCV	3.17E-02	3.22E-02	3.21E-02	3.25E-02	1.37E-02	2.55E-01	2.06E-02	**8.45E-03**	2.20E-02	**9.22E-03**
	MCV	7.04E-02	6.93E-02	6.89E-02	6.97E-02	2.56E-02	6.06E-01	4.60E-02	1.39E-02	4.37E-02	1.63E-02
	PCV	**2.68E-02**	2.59E-02	2.58E-02	2.59E-02	1.45E-02	**1.17E-01**	1.89E-02	9.92E-03	1.84E-02	1.06E-02
	hvBCV	4.69E-02	4.58E-02	4.62E-02	4.22E-02	1.71E-02	4.44E-01	2.73E-02	9.45E-03	2.63E-02	1.14E-02
	M-CV	2.71E-02	**2.50E-02**	**2.60E-02**	**2.41E-02**	**1.36E-02**	1.36E-01	**1.75E-02**	8.81E-03	**1.71E-02**	9.51E-03
EP	BCV	2.07E-02	2.21E-02	2.18E-02	2.07E-02	**9.65E-03**	1.50E-01	**1.62E-02**	**7.64E-03**	1.60E-02	**8.15E-03**
	MCV	5.59E-02	5.85E-02	5.71E-02	5.43E-02	2.22E-02	4.58E-01	4.01E-02	1.40E-02	4.08E-02	1.67E-02
	PCV	1.93E-02	2.02E-02	1.77E-02	1.63E-02	1.14E-02	**6.67E-02**	1.46E-02	9.59E-03	1.67E-02	9.44E-03
	hvBCV	4.72E-02	3.18E-02	3.47E-02	2.60E-02	1.39E-02	2.79E-01	2.10E-02	9.84E-03	2.23E-02	1.44E-02
	M-CV	**1.58E-02**	**1.78E-02**	**1.66E-02**	**1.41E-02**	9.70E-03	6.80E-02	1.30E-02	8.50E-03	**1.45E-02**	8.91E-03

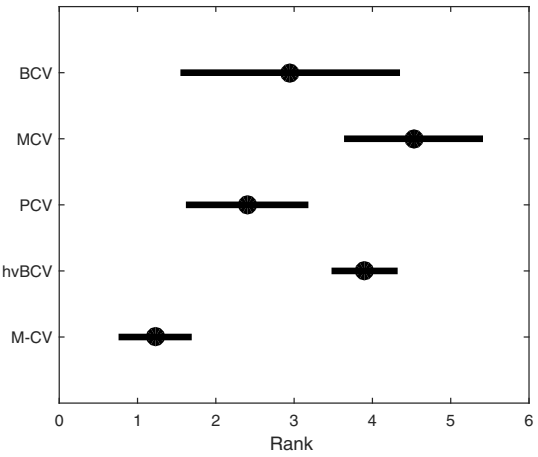

Fig. 8. Rank of five CVs' time consumptions

spend more time as models are trained almost n times on most samples. Generally speaking, M-CV has less time complexity in most cases and it is an efficient CV.

In Table 8, the frequencies of having the least variance for five CVs are 0/40, 12/40, 6/40, 1/40 and 21/40, respectively. They are 2/10, 7/10, 7/10, 5/10 for M-CV on four models, respectively. So M-CV with LPR or GS tends to have the least variance. There are 4/4 on one series, 3/4 on two series, 2/4 on four series and 1/4 on three series for M-CV. Generally, M-CV is superior to others in respect of variance.

In Fig. 9, mean ranks have the following sequence: M-CV<MCV<hvBCV <PCV<BCV. BCV has the largest variance in most cases. MCV, PCV and hvBCV usually have median variances. M-CV is more likely to have minimal variance. Thus it is a stable error estimator.

Above experimental results indicate that M-CV is advocated for estimating the error of time series model.

4.3. Model comparisons and top models

As M-CV errors of any two models are paired results, their comparison is completed by paired tests. The detailed procedures are as follows.

1) Normality test. Test whether the two sets of errors are normal distributed by Kolmogorov-Smirnov (K-S) test [15];

Table 8. Variances of five CVs

model	CVs	Series Nos.									
		1	2	3	4	5	6	7	8	9	10
BS	BCV	2.69E-04	3.15E-04	1.05E-03	1.49E-03	1.45E+08	8.96E+05	1.42E+12	2.85E+02	1.79E+00	1.76E+07
	MCV	**3.20E-11**	**8.11E-11**	**2.76E-09**	7.37E-08	**4.16E+05**	5.80E+01	**4.20E+09**	**1.82E+00**	**3.21E-04**	**4.70E+04**
	PCV	1.13E-03	1.47E-05	2.93E-05	4.77E-07	4.04E+10	2.60E+07	2.06E+12	8.17E+03	4.45E+00	6.19E+09
	hvBCV	1.52E-09	1.88E-10	8.89E-09	1.05E-07	1.16E+06	1.75E+02	7.19E+09	1.46E+01	8.53E-03	3.03E+05
	M-CV	6.69E-10	1.25E-10	1.54E-08	**1.29E-08**	2.07E+06	**7.29E+00**	7.99E+09	7.51E+00	3.78E-03	1.25E+05
LPR	BCV	5.22E-03	3.78E-04	8.86E-02	2.97E-01	6.19E+09	1.31E+07	2.54E+16	7.59E+04	4.60E+00	2.65E+10
	MCV	1.84E-06	**1.59E-08**	9.34E-06	7.37E-05	4.19E+07	1.11E+04	3.28E+13	**1.05E+02**	1.07E-02	**4.72E+06**
	PCV	4.01E-06	1.22E-05	1.05E-07	6.24E-06	2.60E+08	3.39E+06	1.43E+11	1.77E+03	3.27E-02	8.99E+10
	hvBCV	4.12E-05	4.00E-06	1.62E-04	5.79E-03	2.10E+08	4.19E+04	1.92E+15	5.03E+02	3.68E-01	9.47E+08
	M-CV	**8.37E-10**	6.82E-07	**5.95E-08**	**3.99E-08**	**1.75E+07**	**2.79E+01**	**1.14E+10**	5.88E+03	**1.83E-03**	1.25E+07
GS	BCV	1.37E-04	4.96E-05	2.77E-05	8.14E-05	6.50E+06	9.48E+05	3.65E+10	5.60E+00	2.09E-02	5.79E+05
	MCV	2.24E-09	1.59E-08	**9.49E-09**	3.33E-08	1.29E+06	4.27E+02	3.83E+09	8.99E+00	1.14E-03	1.04E+05
	PCV	3.21E-09	3.06E-09	1.38E-08	**2.13E-08**	9.56E+05	2.90E+02	4.61E+09	**2.47E+00**	5.32E-04	9.32E+04
	hvBCV	3.75E-09	3.27E-09	3.52E-08	6.94E-08	1.11E+06	5.30E+02	4.99E+09	2.66E+00	2.92E-03	4.58E+04
	M-CV	**4.82E-10**	**1.13E-10**	2.35E-08	4.13E-08	**2.12E+05**	**1.19E+02**	**3.68E+09**	4.15E-01	**6.35E-05**	**1.30E+04**
EP	BCV	4.14E-07	2.28E-08	7.03E-07	9.34E-06	8.19E+06	1.45E+04	9.10E+10	1.67E+01	6.69E-03	1.97E+05
	MCV	1.74E-09	6.30E-09	6.86E-08	5.38E-08	4.10E+05	3.43E+02	8.23E+09	3.07E-01	5.94E-03	1.15E+05
	PCV	1.01E-08	**7.50E-10**	**5.54E-09**	**1.81E-08**	3.52E+05	4.08E+02	**4.19E+09**	3.98E-01	3.69E-04	4.57E+04
	hvBCV	**4.55E-10**	8.59E-09	2.90E-08	4.08E-08	2.88E+05	2.38E+02	6.09E+09	1.79E-01	2.97E-03	3.14E+04
	M-CV	6.54E-10	3.00E-09	1.77E-08	3.73E-08	**1.26E+05**	**4.83E+01**	8.03E+09	**3.52E-02**	**5.94E-05**	**5.21E+03**

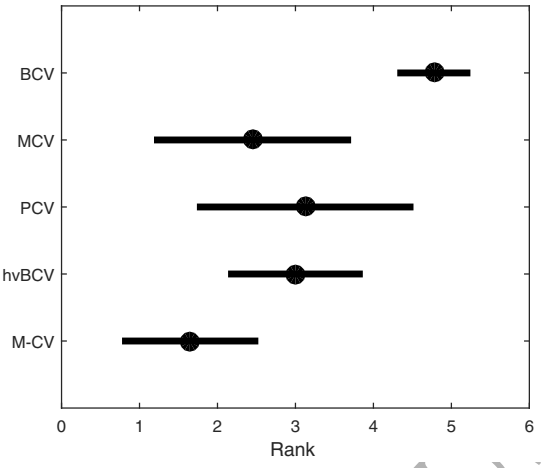

Fig. 9. Rank of five CVs' variances

2) Paired test. If they pass the normality test, compare them by paired t-test. If not, compare them by Wilcoxon signed-rank test [15].

Although there is only one smoothing model with minimal error for a series, good models may not be unique because the errors of two models may have no significant difference. Top models are some models which have no significant difference with minimal error model (MEM) by above tests. Usually we choose the model with least prediction error, such as model with minimal root of mean square error (RMSE), and ignore other top models. However, top models are worth exploration. For example, we could select a more efficient model among top models. These models also could be used for ensemble learning to improve the accuracy as their good performance and different modelling mechanisms.

Table 9 shows M-CV errors (in RMSE form) and the results of model comparison on each series. There is more than one top model on five series (Nos. 1, 2, 4, 7, 10).

The time consumptions of top models are listed in Table 10. It can be seen that the top models with minimal time consumption on series Nos. 1, 2, 4, 7 (EP, GS, EP, EP) could save a lot time expense than MEM on the premise of no loss in accuracy from statistical perspective.

Table 9. Comparison of M-CV errors

Series	BS	LPR	GS	EP	MEM	Top models
1	**0.022**	0.025	**0.023**	**0.023**	BS	BS, GS, EP
2	**0.014**	**0.014**	**0.014**	0.02	BS	BS, LPR, GS
3	0.055	0.064	**0.053**	0.056	GS	GS
4	**0.067**	0.072	0.07	**0.066**	BS	BS, EP
5	54.68	43.44	47.73	**42.73**	EP	EP
6	13.03	**10.84**	19.45	17.36	LPR	LPR
7	1138	1207	**1085**	**1099**	GS	GS, EP
8	2.392	10.031	2.298	**1.942**	EP	EP
9	0.327	0.273	0.424	**0.25**	EP	EP
10	25.33	**22.75**	**22.85**	**22.69**	EP	LPR, GS, EP

a. RMSEs of top models are shown in bold font.

Table 10. Time consumptions (seconds) of top models

Series	BS	LPR	GS	EP	MEM	Recommended
1	0.343	-	0.027	0.016	0.343	0.016
2	0.334	0.127	0.025	-	0.334	0.025
4	0.387	-	-	0.014	0.387	0.014
7	-	-	0.018	0.013	0.018	0.013
10	-	0.019	0.01	0.009	0.009	0.009

5. Conclusions

Traditional cross-validation may be inaccurate in estimating the error of time series model when facing periodicity, overlapping or correlation. M-CV is proposed to tackle these problems. In M-CV, a series is divided into $2m$ subsets. The distances between samples in each subset are within the specific limits, and thus it can avoid overfitting model or information loss of original series. In view of the representativeness of M-CV subset, it may provide a novel way for data stream sampling. Moreover, the partition number ($2m$) is determined automatically by the autocorrelation order of series, so it does not suffer from subjective interference in parameter tuning. Numerical experiments support that M-CV is an accurate, efficient and stable error estimation for time series smoothing.

The good properties of M-CV are based on accurate estimation of autocorrelation order. If not, its performance will be discounted and M-CV needs to be repaired or improved. In addition, time series regression faces problems similar to time series smoothing on dependence series. And M-CV provides a useful reference for the error estimation of time series regression models. These will be part of our future works.

Acknowledgements

The work described in this paper was partially supported by the National Natural Science Foundation of China (Nos. 61273291, 61503229), Research Project Supported by Shanxi Scholarship Council of China (No. 2016-004), and Graduate Innovation Project of Shanxi Province (No. 215548901016).

References

1. E. Alpaydin, Combined 5 × 2 cv F test for comparing supervised classification learning algorithms, Neural Computation 11(8) (1999) 1885-1892.

2. Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, The Journal of Machine Learning Research 5 (2004) 1089-1105.

3. C. Bergmeir, J. M. Bentez, On the use of cross-validation for time series predictor evaluation, Information Sciences 191 (2012) 192-213.

4. C. Bergmeir, M. Costantini, J. M. Bentez, On the usefulness of cross-validation for directional forecast evaluation, Computational Statistics and Data Analysis 76 (2014) 132-143.

5. C. Bergmeir, R. J. Hyndman, B. Koo, 2015. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction (No. 10/15). Monash University, Department of Econometrics and Business Statistics.

6. S. Borra, A. D. Ciaccio, Measuring the prediction error, A comparison of cross-validation, bootstrap and covariance penalty methods, Computational statistics and data analysis 54(12) (2010) 2976-2989.

7. C. K. Chu, J. S. Marron, Comparison of two bandwidth selectors with dependent errors, The Annals of Statistics (1991) 1906-1918.

8. T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10(7) (1998) 1895-1923.

9. R. J. Hyndman, 2012. <http://robjhyndman.com/tsdldata/>.

10. J. Kim, Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap, Computational Statistics and Data Analysis 53(11) (2009) 3735-3745.

11. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. 14th International Joint Conference on Artificial Intelligence, Morgan-Kaufmann, 1995, pp. 1137-1145.

12. R. M. Kunst, Cross validation of prediction models for seasonal time series by parametric bootstrapping, Austrian Journal of Statistics 37 (2008) 271-284.

13. Z. Lai, Z. Wang, Z. Zuo, et al, B-spline-based shape coding with accurate distortion measurement using analytical model, Neurocomputing 149 (2015) 1631-1646.

14. P. Malec, M. Schienle, Nonparametric kernel density estimation near the boundary, Computational Statistics and Data Analysis 72 (2014) 57-76.

15. D. C. Montgomery, G. C. Runger, Applied statistics and probability for engineers. John Wiley and Sons, 2010, pp. 581-584.

16. J. Opsomer, D. Ruppert, Fitting a bivariate additive model by local polynomial regression, The Annals of Statistics (1997) 186-211.

17. J. Opsomer, Y. Wang, Y. Yang, Nonparametric regression with correlated errors, Statistical Science (2001) 134-153.

18. F. M. Pouzols, A. Lendasse, Adaptive kernel smoothing regression for spatio-temporal environmental datasets, Neurocomputing 90 (2012) 59-65.

19. J. Racine, Consistent cross-validatory model-selection for dependent data: hv-block cross-validation, Journal of Econometrics 99(1) (2000) 39-61.

20. P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, in: Encyclopedia of Database Systems, Springer US, 2009, pp. 532-538.

21. J. D. Rodriguez, A. Perez, J. A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3) (2010) 569-575.

22. J. D. Rodriguez, A. Perez, J. A. Lozano, A general framework for the statistical analysis of the sources of variance for classification error estimators, Pattern Recognition 46(3) (2013) 855-864.

23. S. L. Salzberg, On comparing classifiers: Pitfalls to avoid and a recommended approach, Data Mining and Knowledge Discovery 1(3) (1997) 317-328.

24. X. Wang, A. Mueen, H. Ding, et al, Experimental comparison of representation methods and distance measures for time series data, Data Mining and Knowledge Discovery 26(2) (2013) 275-309.