LETTER **Communicated by Jiannan Lu**

Block-Regularized $m \times 2$ Cross-Validated Estimator of the Generalization Error

Ruibo Wang
wangruibo@sxu.edu.cn
Yu Wang
wangyu@sxu.edu.cn
Jihong Li*
Li_ML@sxu.edu.cn
School of Software, Shanxi University, Taiyuan 030006, P.R.C.

Xingli Yang
yangxingli@sxu.edu.cn
Jing Yang
yangjing@email.edu.cn
School of Mathematical Sciences, Shanxi University, Taiyuan 030006, P.R.C.

A cross-validation method based on m replications of two-fold cross validation is called an $m \times 2$ cross validation. An $m \times 2$ cross validation is used in estimating the generalization error and comparing of algorithms' performance in machine learning. However, the variance of the estimator of the generalization error in $m \times 2$ cross validation is easily affected by random partitions. Poor data partitioning may cause a large fluctuation in the number of overlapping samples between any two training (test) sets in $m \times 2$ cross validation. This fluctuation results in a large variance in the $m \times 2$ cross-validated estimator. The influence of the random partitions on variance becomes serious as m increases. Thus, in this study, the partitions with a restricted number of overlapping samples between any two training (test) sets are defined as a block-regularized partition set. The corresponding cross validation is called block-regularized $m \times 2$ cross validation ($m \times 2$ BCV). It can effectively reduce the influence of random partitions. We prove that the variance of the $m \times 2$ BCV estimator of the generalization error is smaller than the variance of $m \times 2$ cross-validated estimator and reaches the minimum in a special situation. An analytical expression of the variance can also be derived in this special situation. This conclusion is validated through simulation experiments. Furthermore, a practical construction method of $m \times 2$ BCV by a two-level orthogonal array is provided. Finally, a conservative estimator is proposed for the variance of estimator of the generalization error.

*Corresponding author

Neural Computation **29, 519–554** (2017)
doi:10.1162/NECO_a_00923
© 2017 Massachusetts Institute of Technology

1 Introduction

In machine learning research, a cross-validation method is commonly used in model selection, estimation of the generalization error, and comparison of algorithm performances. Several versions of cross validation have been developed: repeated learning-testing (RLT), standard K-fold cross validation, Monte Carlo cross validation, 5×2 cross validation and blocked 3×2 cross validation (Dietterich, 1998; Alpaydin, 1999; Friedman, Hastie, & Tibshirani, 2001; Nadeau & Bengio, 2003; Arlot & Celisse, 2010; Yildiz, 2013; Wang, Wang, Jia, & Li, 2014). Among them, the standard two-fold cross validation has received considerable attention because of its simplicity and ease of use. For example, Nason (1996) employed two-fold cross validation and its variants to choose a threshold for wavelet shrinkage. Fan, Guo, and Hao (2012) used two-fold cross validation in variance estimation in an ultra-high-dimensional linear regression model. Stanišić and Tomović (2012) used two-fold cross validation in a frequent item set mining task. In practice, to improve the accuracy of estimation, data partitioning is conducted a number of times (i.e., two-fold cross validation is implemented in multiple replications). The generalization error is often estimated based on the average of the replicated two-fold cross validations.

Cross validation based on m replications of two-fold cross validation is called $m \times 2$ cross validation; it is achieved by randomly splitting the data into two equal-sized blocks m times. The $m \times 2$ cross validation is widely used in machine learning. Dietterich (1998) provided a t-test for use in the comparison of algorithms based on 5×2 cross validation. Alpaydin (1999) proposed a combined 5×2 cross-validated F-test along the line of the 5×2 cross-validated t-test and demonstrated its superiority through simulated comparisons. Yildiz (2013) adjusted the 5×2 cross-validated t-test and conducted comparison experiments on multiple real-life data sets in the UC Irvine Machine Learning Repository of databases widely used by the machine learning community (Lichman, 2013).

However, the performance of the $m \times 2$ cross-validation method often relies on the quality of data partitioning and the accuracy (variance) of the $m \times 2$ cross-validated estimator of the generalization error. Traditionally, a data set is randomly split into multiple different training and test data sets of equal size; training sets (test sets) from any two independent partitions contain common samples regardless of how the data set is split. The number of common samples is defined as the number of overlapping samples, which are defined in section 2. Markatou, Tian, Biswas, and Hripcsak (2005) theoretically proved that the number of overlapping samples follows a hypergeometric distribution with the mathematical expectation of $n/4$ (where n is the size of a data set). Example 2 and plot 2 in Wang et al. (2014) showed that the variance of estimation of generalization error increases when the number of overlapping samples deviate from $n/4$ in the classification situation with the support vector machine classifier. Example 1 further validates

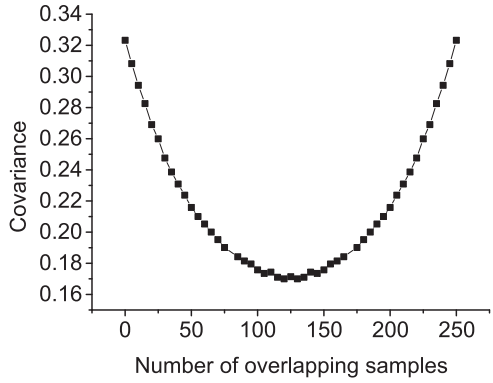

Figure 1: An example of covariance on a simulated regression data set.

the impact of the number of overlapping samples on variance in a simple linear regression situation.

Example 1. Let $D_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$ be the data set in which the predictor vector $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$ is drawn from a multivariate normal distribution $\mathcal{N}_n(0, \Sigma_{p\times p})$. For the covariance matrix $\Sigma_{p\times p}$, all diagonal elements are equal to 1, off-diagonal elements of the fourth column and fourth row of the matrix are equal to $\sqrt{\varphi}$, and the other elements are equal to φ. The response variable is

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{1.1}$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ with the first four coordinates $(b, b, b, -3b\sqrt{\varphi})$ and 0 elsewhere. The lasso method is used as a learning algorithm. The squared loss function is used as the loss function (Fan & Lv, 2008). To simulate the covariance with regard to the number of overlapping samples, we set $n = 500$, $p = 500$, $b = 5$, and $\varphi = 0.5$. The simulation result is depicted in Figure 1 .

Figure 1 shows that the initial decrease and subsequent increase in the covariance of any two two-fold cross-validated estimators corresponds to an increase in the number of overlapping samples. The covariance reaches the minimum when the number of overlapping samples is $n/4$. This condition implies that poor data partitioning may cause a large fluctuation in the number of overlapping samples between any two training (test) sets and thus result in a large variance of the cross-validated estimator.

Wang, Li, and Li (2015) showed that the quantiles of the maximum numbers of the overlapping samples deviated from the $n/4$ increase as m

increases; thus, their influence on the variance of estimation of the generalization error intensifies as m increases.

For this reason, Wang et al. (2014) proposed a new blocked 3×2 cross-validation method with an equal number of overlapping samples in any two training (test) sets of $n/4$ and provided an accurate theoretical expression of the variance of the blocked 3×2 cross-validated estimator of the generalization error. However, they did not investigate deeply the optimal property of the estimation of the generalization error based on blocked 3×2 cross validation. In a more general case with $m \geq 3$, $m \times 2$ cross validation with a restricted number of overlapping samples is called block-regularized $m \times 2$ cross validation (abbreviated as $m \times 2$ BCV). In this letter, we study the property of the $m \times 2$ BCV estimator of generalization error theoretically and provide a novel construction algorithm of data partitioning for an $m \times 2$ BCV. Furthermore, we provide an empirical guide for selection of the replication count of m and propose a conservative estimator of variance of the block-regularized $m \times 2$ cross-validated estimator.

This letter is organized as follows. Section 2 introduces several basic notations and definitions. Section 3 presents a theoretical analysis of the variance of the $m \times 2$ BCV estimator of the generalization error. A construction method of $m \times 2$ BCV based on a two-level orthogonal array is provided in section 4. Section 5 discusses the choice of m in $m \times 2$ BCV. The developed variance estimators are described in section 6. Section 7 presents the simulation experiments, and section 8 concludes.

2 Notations and Definitions

We assume that data set D_n consists of n samples (i.e., $D_n = \{z_i : z_i = (x_i, y_i), i = 1, \ldots, n\}$), where z_is are independently sampled from unknown distribution $\mathcal{P}$, x_i is a predictor variable vector, and y_i is a response variable. $\mathcal{A}(D_n)$ denotes the prediction model trained on data set D_n by learning algorithm $\mathcal{A}$. We let $L(.,.)$ be the loss function. In this letter, zero-one loss is used for classification problems and squared loss is used for regression problems. Then the generalization error of algorithm $\mathcal{A}$ is defined as

$$\mu(n) \triangleq E_{D_n,z}[L(\mathcal{A}(D_n), z)]. \tag{2.1}$$

Generally the generalization error is estimated by some kind of cross validation in practice. In this study, we consider $m \times 2$ cross validation. In $m \times 2$ cross validation, each standard two-fold cross validation is conducted by randomly splitting the entire data set into two equal-sized blocks. Several notations and definitions follow:

Definition 1. *The $\mathcal{S} \triangleq (I^{(t)}, I^{(v)})$ is called a partition of index set $\mathcal{I}$, where $I^{(t)}$ and $I^{(v)}$ are random index sets from $\mathcal{I} = \{1, 2, \ldots, n\}$ of data set D_n. $I^{(t)}$ and $I^{(v)}$ satisfy*

$I^{(t)} \cup I^{(v)} = \mathcal{I}$, $I^{(t)} \cap I^{(v)} = \varnothing$, and $|I^{(t)}| = |I^{(v)}| = \frac{n}{2}$. Then $\mathbb{S} = \{< \mathcal{S}_i, \mathcal{S}_i^\top >: \mathcal{S}_i = (I_i^{(t)}, I_i^{(v)}), \mathcal{S}_i^\top = (I_i^{(v)}, I_i^{(t)}), i = 1, 2, \ldots, m\}$ is the set of $m \times 2$ partitions for $\mathcal{I}$.

Let $D^{(t)} = \{z_i : i \in I^{(t)}\}$ and $D^{(v)} = \{z_i : i \in I^{(v)}\}$ denote the training and test sets, respectively. Then $D_n = D^{(t)} \cup D^{(v)}$. $D^{(t)}$ and $D^{(v)}$ serve as training or test sets in a two-fold cross validation.

Remark 1. For the two-fold cross validation, the training set $D^{(t)}$ and the test set $D^{(v)}$ are both generally called data blocks.

Definition 2. *For any two partitions $\mathcal{S}_i = (I_i^{(t)}, I_i^{(v)})$ and $\mathcal{S}_j = (I_j^{(t)}, I_j^{(v)})$ in $\mathbb{S}$, $\phi_{ij} = |I_i^{(t)} \cap I_j^{(t)}|$ is defined as number of overlapping samples between $\mathcal{S}_i$ and $\mathcal{S}_j$, where $\phi_{ij} = x$, $0 \le x \le n/2$ and $i, j = 1, 2, \ldots, m$. Matrix $\Phi = (\phi_{ij})$ can be regarded as a measure of partition set $\mathbb{S}$.*

Remark 2. In fact, the ith and jth elements of $\mathbb{S}$ are $\langle \mathcal{S}_i, \mathcal{S}_i^\top \rangle$ and $\langle \mathcal{S}_j, \mathcal{S}_j^\top \rangle$, respectively. These two pairs of partitions can result in four numbers of overlapping samples through a comparison of $\mathcal{S}_i$ and $\mathcal{S}_j$, $\mathcal{S}_i^\top$ and $\mathcal{S}_j$, $\mathcal{S}_i$ and $\mathcal{S}_j^\top$, and $\mathcal{S}_i^\top$ and $\mathcal{S}_j^\top$. However, if we let the number of overlapping samples between $\mathcal{S}_i$ and $\mathcal{S}_j$ be ϕ_{ij}, the other three numbers of overlapping samples are equal to $n/2 - \phi_{ij}$, $n/2 - \phi_{ij}$ and ϕ_{ij}. Moreover, these four numbers have the same distribution. Therefore, we simply consider ϕ_{ij}, the number of overlapping samples between $\mathcal{S}_i$ and $\mathcal{S}_j$, in definition 2.

Generally ϕ_{ij} is an integer-valued random variable in $[0, n/2]$. Markatou et al. (2005) proved that ϕ_{ij} is drawn from hypergeometric distribution, and its expectation is $n/4$. If there are more than two partitions, multiple numbers of overlapping samples should be considered. Furthermore, all differences between the multiple numbers of overlapping samples and $n/4$ should be regularized to reduce the variance of the cross-validated estimator of the generalization error. On the basis of these intuitions, we propose a new partitioning method, block-regularized cross-validation partitions, to control the differences. Our method aims to control the difference between each number of overlapping samples and $n/4$ into smaller than its expectation in a random situation. This expectation is provided by Wang et al. (2015) and is expressed as

$$E\left|\phi_{ij} - \frac{n}{4}\right| = \frac{n^2}{4(n-1)} \frac{\binom{2n'-1}{n'}\binom{2n'-1}{n'-1}}{\binom{n-2}{2n'-1}} + \frac{n(n-2)}{8(n-1)} \frac{\binom{2n'-1}{n'-1^2}}{\binom{n-2}{2n'-2}} - \frac{n}{4} \frac{\binom{2n'}{n'^2}}{\binom{n}{2n'}}, \quad (2.2)$$

where n is the data set size and $n' = n/4$.

Based on the above analysis, we propose a definition of the block-regularized cross-validation partitions as follows.

Definition 3. *Given a set of partitions $\mathbb{S}$ of $m \times 2$ for all ϕ_{ij}s $(i \neq j)$ in Φ, if regularized condition $|\phi_{ij} - n/4| \leq c$ is satisfied, in which $c \geq 0$ is called a regularization parameter, then the partition set $\mathbb{S}$ is called a block-regularized partition set and denoted as $\mathbb{S}^b$ and the measure Φ^b, accordingly. The $m \times 2$ cross validation on $\mathbb{S}^b$ is called block-regularized $m \times 2$ cross validation(abbreviated as $m \times 2$ BCV). When $n = 4l, l \in N^+$ and $c = 0$, that is, $\phi_{ij} \equiv n/4$, the corresponding $m \times 2$ BCV is called a balanced $m \times 2$ BCV and denoted as $\mathbb{S}^*$. In this case, the measure Φ for $\mathbb{S}$ degenerates into a constant matrix, denoted as Φ^*.*

Remark 3. The regularization parameter c should not exceed the expectation of $|\phi_{ij} - n/4|$ of the $m \times 2$ cross validation, which is defined in equation 2.2.

In the following sections, the $m \times 2$ cross validation is abbreviated as $m \times 2$ CV. Definitions of some estimators of the generalization error (Friedman et al., 2001) are provided in the following paragraphs.

Definition 4. *For a given partition $\mathcal{S} = (I^{(t)}, I^{(v)})$, the hold-out estimator (HO estimator) of $\mu(n)$ is defined as*

$$\hat{\mu}_{HO}(\mathcal{S}) \triangleq \frac{1}{|I^{(v)}|} \sum_{j \in I^{(v)}} L(\mathcal{A}(D^{(t)}); z_j) = \frac{2}{n} \sum_{j \in I^{(v)}} L(\mathcal{A}(D^{(t)}); z_j). \tag{2.3}$$

The standard two-fold cross-validated estimator (S2CV estimator) of $\mu(n)$ can be written as

$$\hat{\mu}(\mathcal{S}) \triangleq \frac{1}{2}\hat{\mu}_{HO}(\mathcal{S}) + \frac{1}{2}\hat{\mu}_{HO}(\mathcal{S}^\top), \tag{2.4}$$

where $\mathcal{S}^\top = (I^{(v)}, I^{(t)})$.

The $m \times 2$ cross-validated estimator ($m \times 2$ CV estimator)of $\mu(n)$ can be expressed as

$$\hat{\mu}_{m\times 2}(\mathbb{S}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \hat{\mu}(\mathcal{S}_i), \tag{2.5}$$

where, $\hat{\mu}(\mathcal{S}_i)$ is the S2CV estimator for partition $\mathcal{S}_i$. Accordingly, the estimator of $\mu(n)$ based on $\mathbb{S}^b$ is denoted as $\hat{\mu}_{m\times 2}(\mathbb{S}^b)$ which is a block-regularized $m \times 2$ cross-validated estimator ($m \times 2$ BCV estimator) of $\mu(n)$.

3 Theoretical Analysis of $\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}))$

The variance of the $m \times 2$ CV estimator can be decomposed as

$$\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})) = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(\hat{\mu}(\mathcal{S}_i)) + \frac{1}{m^2}\sum_{i\neq j, i, j=1,2,\ldots,m}\mathrm{Cov}(\hat{\mu}(\mathcal{S}_i), \hat{\mu}(\mathcal{S}_j)). \tag{3.1}$$

The $\mathrm{Var}(\cdot)$ should be expressed as $\mathrm{Var}_{D,\mathbb{S}}(\cdot)$ exactly. But this letter considers the original sample size n fixed and considers only the measure Φ for $\mathbb{S}$. From the perspective of measure Φ for $\mathbb{S}$, all $\mathrm{Var}(\hat{\mu}(\mathcal{S}_i))$s should be the same for each $S_i \in \mathbb{S}$. Due to all samples z_is in the data set D_n are i.i.d, $\mathrm{Cov}(\hat{\mu}(\mathcal{S}_i), \hat{\mu}(\mathcal{S}_j))$ depends on only the number of overlapping samples $\phi_{ij} = |I_i^{(t)} \cap I_j^{(t)}|$. Thus,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi) &= \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m}\hat{\mu}(\mathcal{S}_i)|\Phi\right)\\ &= \frac{1}{m}\mathrm{Var}(\hat{\mu}(\mathcal{S}_1)) + \frac{1}{m^2}\sum_{i\neq j, i, j=1}^{m}\mathrm{Cov}(\hat{\mu}(\mathcal{S}_i), \hat{\mu}(\mathcal{S}_j)|\phi_{ij}),\end{aligned} \tag{3.2}$$

where $\mathrm{Var}(.|\Phi)$ is with regard to random samples D_n.

Motivated by experimental design in statistics (Wu & Hamada, 2011), we attempt to design a set of partitions $\mathbb{S}$ to reduce the effects of random variable ϕ_{ij}. We will prove that when $\phi_{ij} = n/4$, for all $i \neq j$, $\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi)$ reaches the minimum, that is, $\mathbb{S}^*$ (balanced $m \times 2$ BCV) satisfies

$$\mathbb{S}^* = \mathrm{argmin}_{\mathbb{S}}\,\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi). \tag{3.3}$$

In the expression of $\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi)$ in equation 3.2, the key issue is to comprehensively analyze the properties of $\mathrm{Cov}(\hat{\mu}(\mathcal{S}_i), \hat{\mu}(\mathcal{S}_j)|\phi_{ij})$. Lemma 1 characterizes the lower convex property of $\mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_i), \hat{\mu}_{HO}(\mathcal{S}_j)|\phi_{ij})$. Lemma 2 characterizes the minimum property of $\mathrm{Cov}(\hat{\mu}(\mathcal{S}_i), \hat{\mu}(\mathcal{S}_j)|\phi_{ij})$ at $\phi_{ij} = n/4$.

Lemma 1. *We let $e_j(\mathcal{S}) \triangleq L(\mathcal{A}(D^{(t)}); z_j)$ be the loss function on z_j for $j = 1, 2, \ldots, n/2$ and $\mathcal{S}_1 = (I_1^{(t)}, I_1^{(v)})$ and $\mathcal{S}_2 = (I_2^{(t)}, I_2^{(v)})$ be two random partitions*

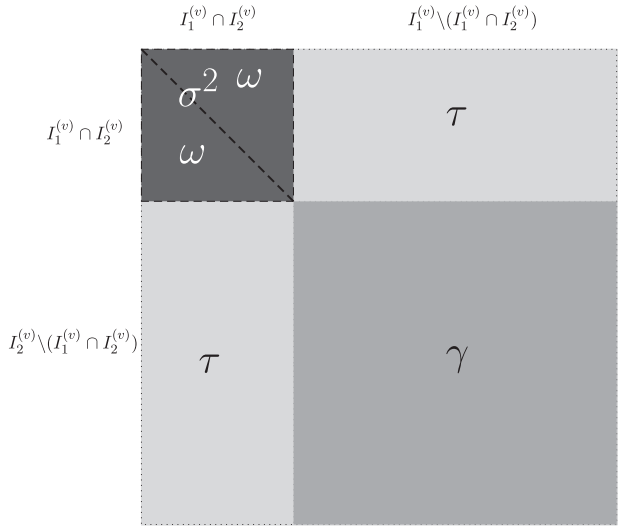

Figure 2: Demo of values of parameters σ^2, ω, τ, and γ.

of $\mathcal{I} = \{1, \ldots, n\}$, and ϕ is the number of overlapping samples between $\mathcal{S}_1$ and $\mathcal{S}_2$. We have:

i. For $i, j \in \{1, 2, \ldots, n/2\}$, when $\phi = n/4$, $Cov(e_i(\mathcal{S}_1), e_j(\mathcal{S}_2)|\phi)$ has the following form:

$$Cov(e_i(\mathcal{S}_1), e_j(\mathcal{S}_2)|\phi = n/4) = \begin{cases} \sigma^2 & i = j,\ i, j \in (I_1^{(v)} \cap I_2^{(v)}) \\ \omega & i \neq j,\ i, j \in (I_1^{(v)} \cap I_2^{(v)}) \\ \gamma & i \in I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}) \text{ and} \\ & j \in I_2^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}) \\ \tau & \text{others} \end{cases}$$

where $\sigma^2, \omega, \gamma, \tau$ are constants, as shown in Figure 2.

ii. If denoting $f(x) \triangleq Cov(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2)|\phi = x)$, $f(x)$ can be expressed as a quadratic polynomial function of x:

$$f(x) \triangleq \frac{4}{n^2}\left[(\omega + \gamma - 2\tau) \cdot x^2 + (\sigma^2 - \omega - n\gamma + n\tau) \cdot x + \frac{n^2}{4}\gamma\right]. \tag{3.4}$$

Therefore, $f(x)$ is a lower convex function with regard to x when $\omega + \gamma \geq 2\tau$.

Proof. From the definition of a partition, we know that four index subsets $I_1^{(t)} \cap I_2^{(t)}$, $I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$, $I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)})$, and $I_1^{(v)} \cap I_2^{(v)}$ can be obtained from $\mathcal{S}_1 = (I_1^{(t)}, I_1^{(v)})$ and $\mathcal{S}_2 = (I_2^{(t)}, I_2^{(v)})$. For $\mathcal{S}_1$, we have

$$\begin{aligned}\mathcal{I} &= I_1^{(t)} \bigcup I_1^{(v)} \\ &= [I_1^{(t)} \cap I_2^{(t)}] \cup [I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})] \bigcup [I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)})] \cup [I_1^{(v)} \cap I_2^{(v)}].\end{aligned}$$

For $\mathcal{S}_2$, the following equation holds:

$$\begin{aligned}\mathcal{I} &= I_2^{(t)} \bigcup I_2^{(v)} \\ &= [I_1^{(t)} \cap I_2^{(t)}] \cup [I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})] \bigcup [I_2^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)})] \cup [I_2^{(v)} \cap I_2^{(v)}].\end{aligned}$$

Obviously we can get

$$I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)}) = I_2^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}), I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)}) = I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}) \tag{3.5}$$

$$\begin{aligned}&|I_1^{(t)} \cap I_2^{(t)}| = |I_1^{(v)} \cap I_2^{(v)}| = x, \\ &|I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})| = |I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})| = \frac{n}{2} - x.\end{aligned}$$

Then,

$$\begin{aligned}f(x) &\triangleq \mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2) | \phi = x) \\ &= \frac{4}{n^2} \mathrm{Cov}\left(\sum_{i \in I_1^{(v)}} e_i(\mathcal{S}_1), \sum_{j \in I_2^{(v)}} e_j(\mathcal{S}_2) | \phi = x \right) \\ &= \frac{4}{n^2} \left[\mathrm{Cov}\left(\sum_{i \in (I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}))} e_i(\mathcal{S}_1), \sum_{k \in (I_2^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}))} e_k(\mathcal{S}_2 | \phi = x) \right) \right. \\ &\quad + \mathrm{Cov}\left(\sum_{i \in (I_1^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}))} e_i(\mathcal{S}_1), \sum_{j \in (I_1^{(v)} \cap I_2^{(v)})} e_j(\mathcal{S}_2) | \phi = x \right) \\ &\quad \left. + \mathrm{Cov}\left(\sum_{j \in (I_1^{(v)} \cap I_2^{(v)})} e_j(\mathcal{S}_1), \sum_{k \in (I_2^{(v)} \backslash (I_1^{(v)} \cap I_2^{(v)}))} e_k(\mathcal{S}_2) | \phi = x \right) \right.\end{aligned}$$

$$+ \operatorname{Cov}\left(\sum_{j \in (I_1^{(v)} \cap I_2^{(v)})} e_j(\mathcal{S}_1), \sum_{j \in (I_1^{(v)} \cap I_2^{(v)})} e_j(\mathcal{S}_2) | \phi = x \right) \Bigg] \tag{3.6}$$

Finally, we obtain

$$\begin{aligned} f(x) &= \frac{4}{n^2}\left[\left(\frac{n}{2} - x^2\right)\gamma + 2x\left(\frac{n}{2} - x\right)\tau + x\sigma^2 + (x^2 - x)\omega\right] \\ &= \frac{4}{n^2}\left[(\omega + \gamma - 2\tau)\cdot x^2 + (\sigma^2 - \omega - n\gamma + n\tau)\cdot x + \frac{n^2}{4}\gamma\right], \end{aligned}$$

in which $f(x)$ is a lower convex function with regard to x when $\omega + \gamma > 2\tau$.

Remark 4. In actuality, the parameters of σ^2, ω, γ, and τ have relationships with x. Nevertheless, we mainly focus on the values of these parameters at the point of $x = n/4$ because the expectation of the number of overlapping samples is $n/4$.

In order to clearly interpret the condition $\omega + \gamma > 2\tau$, we provide some intuitive clarifications of ω, γ, and τ:

- ω is the covariance of two loss functions with test samples of each pair in block $I_1^{(v)} \cap I_2^{(v)}$. Specifically, the first loss function uses training set of $I_1^{(t)}$ and the test sample of $z_i, \forall i \in I_1^{(v)} \cap I_2^{(v)}$. The second loss function uses $I_2^{(t)}$ as the training set and tests on $z_j, \forall j \in I_1^{(v)} \cap I_2^{(v)}$ and $i \neq j$. Given that z_i and z_j do not appear in the two training sets and are independent, ω merely measures the correlations caused by the two training sets, assuming that the correlation is affected by nothing else except the training and test sets. Moreover, $i = j$ corresponds to σ^2.
- τ is the covariance of two loss functions with two training sets of $I_1^{(t)}$ and $I_2^{(t)}$ and two test samples of $z_i, \forall i \in I_1^{(v)} \cap I_2^{(v)}$ and $z_j, \forall j \in I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$ (or $z_j, \forall j \in I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$). Given that z_j occurs in the training set of $I_1^{(t)}$ (or $I_2^{(t)}$), τ measures not only the correlation caused by the two training sets but also the correlation caused by the appearance of test sample z_j in training set $I_1^{(t)}$ (or $I_2^{(t)}$). Therefore, τ is greater than ω.
- γ is the covariance of two loss functions with two test samples of z_i, $\forall i \in I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$, and $z_j, \forall j \in I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$. The first loss function uses the training set $I_1^{(t)}$ and $z_i, \forall i \in I_2^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$ as a test sample, and the second loss function uses the training set $I_2^{(t)}$ and $z_j, \forall j \in I_1^{(t)} \backslash (I_1^{(t)} \cap I_2^{(t)})$ as a test sample. Test samples z_i and z_j both occur in the other's training set. Therefore, γ measures the correlations caused

by the two training sets and the appearance of both test samples in the other's training sets. Intuitively, γ is greater than τ—$\gamma > \tau > \omega$.

Furthermore, $\omega + \gamma - 2\tau$ actually measures the differences between $\gamma - \tau$ and $\tau - \omega$. Specifically, it measures how the increment of covariance of two loss functions changes with the occurrence of one partition's test sample in the other partition's training set. $\tau - \omega$ indicates the increment of covariances caused by only one test sample occurring in the training set, and $\gamma - \tau$ indicates the increment of covariances caused by both test samples appearing in the training sets. Therefore, the intuitive interpretation of $\gamma + \omega > 2\tau$ is $\gamma - \tau > \tau - \omega$, that is, the increase from ω to γ is nonlinear and the increment between γ and τ is larger than that between τ and ω.

Remark 5. Proving that the condition $\omega + \gamma > 2\tau$ holds in broad families of loss functions, algorithms, and data populations is difficult. However, with squared loss function, proving that the condition $\omega + \gamma > 2\tau$ holds for mean regression and multivariate regression is possible. The detailed proofs are in the appendix. Moreover, some simulation results are presented in section 7.2 to illustrate that this condition is true.

Lemma 2. *We let $\hat{\mu}(\mathcal{S}_1)$ and $\hat{\mu}(\mathcal{S}_2)$ be two S2CV estimators of $\mu(n)$ on partitions $\mathcal{S}_1$ and $\mathcal{S}_2$, and $g(x) \triangleq Cov(\hat{\mu}(\mathcal{S}_1), \hat{\mu}(\mathcal{S}_2)|\phi = x)$. Then, for any $x \in [0, \frac{n}{2}]$, $g(x) = \frac{1}{2}(f(x) + f(\frac{n}{2} - x))$. Function $g(x)$ has the following two properties:*

i. Symmetry: $g(x) = g(\frac{n}{2} - x)$.
ii. Boundedness: $g(\frac{n}{4}) \leqslant g(x) \leqslant g(0) = Var(\hat{\mu}(\mathcal{S}_i))$, $i = 1, 2, \ldots, m$.

Proof. If $|I_1^{(t)} \cap I_2^{(t)}| = x$, then

$$|I_1^{(v)} \cap I_2^{(v)}| = x, \quad |I_1^{(t)} \cap I_2^{(v)}| = |I_1^{(v)} \cap I_2^{(t)}| = \frac{n}{2} - x.$$

From the definition of S2CV estimator, we have

$$\begin{aligned}
g(x) &\triangleq \mathrm{Cov}(\hat{\mu}(\mathcal{S}_1), \hat{\mu}(\mathcal{S}_2)|\phi = x) \\
&= \mathrm{Cov}\left(\frac{1}{2}(\hat{\mu}_{HO}(\mathcal{S}_1) + \hat{\mu}_{HO}(\mathcal{S}_1^\top)), \frac{1}{2}(\hat{\mu}_{HO}(\mathcal{S}_2) + \hat{\mu}_{HO}(\mathcal{S}_2^\top))|\phi = x\right) \\
&= \frac{1}{4}[\mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2)|\phi = x) \\
&\quad + \mathrm{Cov}\left(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2^\top)|\phi = \frac{n}{2} - x\right) \\
&\quad + \mathrm{Cov}\left(\hat{\mu}_{HO}(\mathcal{S}_1^\top), \hat{\mu}_{HO}(\mathcal{S}_2)|\phi = \frac{n}{2} - x\right) \\
&\quad + \mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_1^\top), \hat{\mu}_{HO}(\mathcal{S}_2^\top)|\phi = x)] \qquad (3.7)
\end{aligned}$$

According to the definition of $f(x) \triangleq \mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2)|\phi = x)$, we have

$$\mathrm{Cov}(\hat{\mu}_{HO}(\mathcal{S}_1^\top), \hat{\mu}_{HO}(\mathcal{S}_2^\top)|\phi = x) = f(x)$$

and

$$\begin{aligned}
&\mathrm{Cov}\left(\hat{\mu}_{HO}(\mathcal{S}_1^\top), \hat{\mu}_{HO}(\mathcal{S}_2)|\phi = \frac{n}{2} - x\right) \\
=&\mathrm{Cov}\left(\hat{\mu}_{HO}(\mathcal{S}_1), \hat{\mu}_{HO}(\mathcal{S}_2^\top)|\phi = \frac{n}{2} - x\right) \\
=&f\left(\frac{n}{2} - x\right).
\end{aligned}$$

Thus,

$$g(x) = \frac{1}{2}\left(f(x) + f\left(\frac{n}{2} - x\right)\right). \tag{3.8}$$

Obviously, $g(x) = g(\frac{n}{2} - x)$, that is, $g(x)$ is a symmetric function, and its symmetry axis is $x = \frac{n}{4}$. In particular, $g(\frac{n}{4}) = f(\frac{n}{4})$.

According to the property of covariance, we have

$$\begin{aligned}
g(x) =& \mathrm{Cov}(\hat{\mu}(\mathcal{S}_1), \hat{\mu}(\mathcal{S}_2)|\phi = x) \\
&\leqslant \sqrt{\mathrm{Var}(\hat{\mu}(\mathcal{S}_1)|\phi = x) \cdot \mathrm{Var}(\hat{\mu}(\mathcal{S}_2)|\phi = x)}.
\end{aligned} \tag{3.9}$$

Together with the fact that $g(0) = \mathrm{Var}(\hat{\mu}(\mathcal{S}_i)|\phi = 0) = \mathrm{Var}(\hat{\mu}(\mathcal{S}_i))$, $i = 1, 2, \ldots, m$ and the symmetric property of $g(x)$, we have

$$g(x) \leqslant g(0) = g\left(\frac{n}{2}\right).$$

According to the lower convex property of $f(x)$ (lemma 1), based on Jensen's inequality, we can easily obtain $\frac{1}{2}f(x) + \frac{1}{2}f(\frac{n}{2} - x) \geqslant f(\frac{n}{4}) = g(\frac{n}{4})$, that is,

$$g(x) \geqslant g\left(\frac{n}{4}\right)$$

where, $x \in [0, \frac{n}{2}]$.

In the simulated experiments in section 7.3, we provide some simulated curves of function $g(x)$ and their approximations in parameters σ^2, ω, γ, and τ at a neighborhood of $x = \frac{n}{4}$.

Theorem 1. *Given a set of partitions $\mathbb{S}$ of n samples D_n, from the perspective of measure Φ for $\mathbb{S}$, the variance of an $m \times 2$ CV estimator of generalization error satisfies*

$$E_{\Phi} Var(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi) \geqslant E_{\Phi^b} Var(\hat{\mu}_{m\times 2}(\mathbb{S}^b)|\Phi^b) \geqslant E_{\Phi^*} Var(\hat{\mu}_{m\times 2}(\mathbb{S}^*)|\Phi^*), \tag{3.10}$$

where

$$E_{\Phi^*} Var(\hat{\mu}_{m\times 2}(\mathbb{S}^*)|\Phi^*) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\rho_2 \tag{3.11}$$

- $\sigma^2 = f(\frac{n}{2}) = Var(\hat{\mu}_{HO}(\mathcal{S}_i))$ *is the variance of HO estimators with the size of a training set of* $\frac{n}{2}$.
- $\rho_1 = \frac{f(0)}{f(\frac{n}{2})} = \frac{Cov(\hat{\mu}_{HO}(\mathcal{S}_i), \hat{\mu}_{HO}(\mathcal{S}_i^T))}{Var(\hat{\mu}_{HO}(\mathcal{S}_i))}$ *is the correlation coefficient between two HO estimators within an S2CV estimator.*
- $\rho_2 = \frac{f(\frac{n}{4})}{f(\frac{n}{2})} = \frac{Cov(\hat{\mu}_{HO}(\mathcal{S}_i), \hat{\mu}_{HO}(\mathcal{S}_j))}{Var(\hat{\mu}_{HO}(\mathcal{S}_i))}$ *is the correlation coefficient of any two S2CV estimators in an $m \times 2$ BCV estimator.*

Proof. We can easily equation 3.10 from lemma 2. Specifically, to prove the first inequality in equation 3.10, we introduce a random variable $\varphi = n/4 - |\phi - n/4|$ in which $\phi \in \Phi$. Due to $0 \leq \phi \leq n/2$, we can obtain $0 \leq \varphi \leq n/4$. Then, by employing Jensen's inequality on $g(\varphi)$, we can obtain

$$E_{\phi} g(\phi) = E_{\varphi} g(\varphi) \geq g(E\varphi) = g\left(\frac{n}{4} - E|\phi - n/4|\right) \geq g\left(\frac{n}{4} - c\right). \tag{3.12}$$

Using the symmetric property of $g(x)$ clarifies that

$$E_{\phi} g(\phi) \geq g\left(\frac{n}{4} \pm E|\phi - n/4|\right) \geq g\left(\frac{n}{4} \pm c\right). \tag{3.13}$$

Thus, the first inequality holds. The second inequality can be derived directly because $g(\phi)$ reaches its minimum at $\phi = n/4$.

Furthermore, the variance of a balanced $m \times 2$ BCV estimator can be decomposed into combinations of hold-out estimators as follows:

$$E_{\Phi^*} \mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}^*)|\Phi^*) = E_{\Phi^*} \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} \hat{\mu}(\mathcal{S}_i)|\Phi^*\right)$$

$$
\begin{aligned}
&= \frac{1}{m} g(0) + \frac{m-1}{m} g\left(\frac{n}{4}\right) \\
&= \frac{1}{m}\left(\frac{1}{2}\left(f(0) + f\left(\frac{n}{2}\right)\right)\right) + \frac{m-1}{m} f\left(\frac{n}{4}\right) \\
&= \frac{1}{2m} f\left(\frac{n}{2}\right)\left(1 + \frac{f(0)}{f(\frac{n}{2})}\right) + \frac{m-1}{m} f\left(\frac{n}{2}\right)\frac{f(\frac{n}{4})}{f(\frac{n}{2})} \\
&= \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\rho_2.
\end{aligned}
$$

Corollary 1. *For any two partition sets $\mathbb{S}_1^*$ and $\mathbb{S}_2^*$ of balanced $m \times 2$ BCV,*

$$Var(\hat{\mu}_{m\times 2}(\mathbb{S}_1^*)|\Phi) = Var(\hat{\mu}_{m\times 2}(\mathbb{S}_2^*)|\Phi).$$

Corollary 2. *The variance of a balanced $m \times 2$ BCV estimator obviously decreases with the increment of m. As m increases, the proportion of the second part $\frac{m-1}{m}\sigma^2\rho_2$ in variance $Var(\hat{\mu}_{m\times 2}(\mathbb{S}^*)|\Phi)$ becomes large.*

4 Nested Construction Algorithm of $\mathbb{S}^b$ for $m \times 2$ BCV

Although $m \times 2$ BCV has good properties, it has no comprehensive use in practical applications if it cannot be easily constructed. A classical construction method for $\mathbb{S}^b$ for $m \times 2$ BCV is provided in McCarthy (1976). The construction method employs rows of an orthogonal array. Specifically, the data set D_n is divided into blocks based on columns of an orthogonal array. Then, a partition in $\mathbb{S}^b$ can be derived by combining the blocks according to the levels of each row of the orthogonal array. A weakness of the construction method is that it is not nested; that is, the new $\mathbb{S}^b$ for larger m does not include the previous $\mathbb{S}^b$ for small m; thus, it should be reconstructed from the beginning to create $m \times 2$ BCV with a larger m. Accordingly, training or testing models should be retaken for any different m of $m \times 2$ BCV.

In this section, we propose a nested construction algorithm of $\mathbb{S}^b$ for $m \times 2$ BCV. The algorithm can construct $\mathbb{S}$ for $m \times 2$ BCV, along with an increment of m. The nested construction algorithm and its theoretical guarantee is presented in theorem 2:

Theorem 2. *Assuming that data set D_n of size n can be split into $4k$ (k is a given value from $\{1, 2, \ldots, n/4\}$) disjoint and almost equal-sized blocks (such that the maximum difference of sizes of any two blocks is one), a partition set $\mathbb{S} = \{\mathcal{S}_i = (I_i^{(t)}, I_i^{(v)}), i = 1, 2, \ldots, 4k-1\}$ can be constructed by using an orthogonal array $OA(4k, 2^{4k-1})$ according to the following two steps:*

Table 1: Orthogonal Array $OA(8, 2^7)$.

Row Index	a	b	ab	c	ac	bc	abc
$1 \leftrightarrow I_1^{(8)}$	+	+	+	+	+	+	+
$2 \leftrightarrow I_2^{(8)}$	-	+	-	+	-	+	-
$3 \leftrightarrow I_3^{(8)}$	+	-	-	+	+	-	-
$4 \leftrightarrow I_4^{(8)}$	-	-	+	+	-	-	+
$5 \leftrightarrow I_5^{(8)}$	+	+	+	-	-	-	-
$6 \leftrightarrow I_6^{(8)}$	-	+	-	-	+	-	+
$7 \leftrightarrow I_7^{(8)}$	+	-	-	-	-	+	+
$8 \leftrightarrow I_8^{(8)}$	-	-	+	-	+	+	-

i. *The i-th column of orthogonal array $OA(4k, 2^{4k-1})$ corresponds to partition $\mathcal{S}_i = (I_i^{(t)}, I_i^{(v)})$, where $I_i^{(t)}$ is the set of rows of the "+" level of the i-th column. Similarly, the rows of the "-" level of the i-th column form the set $I_i^{(v)}$.*
ii. *According to step i, by taking over all columns of $OA(4k, 2^{4k-1})$, we can obtain $4k - 1$ partitions $\mathcal{S}_i$s of I for $i = 1, 2, \ldots, 4k - 1$.*

Then, $\mathbb{S} = \{\mathcal{S}_i = (I_i^{\{(t)\}}, I_i^{\{(v)\}}), i = 1, 2, \ldots, 4k - 1\}$ is a block-regularized partition set $\mathbb{S}^b$ with regularized condition $|\phi_{ij} - n/4| \leqslant k$ for any $i \neq j$.

Proof. Orthogonal array $OA(4k, 2^{4k-1})$ corresponds to a matrix with $4k$ rows and $4k - 1$ columns. The elements of the matrix consist of "+" and "-", which are called levels in statistics. For any two columns in $OA(4k, 2^{4k-1})$, there are only four combinations of $(+, +)(+, -)(-, +)(-, -)$ and equal replicated times for each combination. This condition means that the replicated time for each combination is $n/4$, that is, the corresponding number of the same samples in any two columns is $n/4$. Thus, $4k - 1$ replications of two-fold cross validation constructed by the above operation form the $\mathbb{S}$ of $(4k - 1) \times 2$.

Since the maximum difference of size of any two of $4k$ disjoint and equal-sized blocks is one and testing sets from any two independent partitions contain common k blocks, $|\phi_{ij} - n/4| \leqslant k$,for any $i \neq j$.

Example 2. This example illustrates the construction process of $\mathbb{S}^b$ of 7×2. Index set $\mathcal{I}$ from data set D_n is split into $4k = 8$ blocks denoted as $I_i^{(8)}$, $i = 1, 2, 3, \ldots, 8$. Orthogonal array $OA(8, 2^7)$ is employed (see Table 1). Then $\mathbb{S}^b$ of 7×2 is constructed with Table 2. When $n = 400$, the expectation of $|\phi_{ij} - n/4|$ of $m \times 2$ CV is about 3.98. However, our construction algorithm can constrain k to 2.

Remark 6. For data set D_n with sample size n, according to the construction method of theorem 2, the maximum value of m in $\mathbb{S}^b$ of $m \times 2$ should be

Table 2: Mapping between S_i and $I_i^{(8)}$ for $\mathbb{S}^b$ of 7×2.

Partition	$I_i^{(t)}$	$I_i^{(v)}$
$\mathcal{S}_1$	$I_1^{(8)}, I_3^{(8)}, I_5^{(8)}, I_7^{(8)}$	$I_2^{(8)}, I_4^{(8)}, I_6^{(8)}, I_6^{(8)}$
$\mathcal{S}_2$	$I_1^{(8)}, I_2^{(8)}, I_5^{(8)}, I_6^{(8)}$	$I_3^{(8)}, I_4^{(8)}, I_7^{(8)}, I_8^{(8)}$
$\mathcal{S}_3$	$I_1^{(8)}, I_4^{(8)}, I_5^{(8)}, I_8^{(8)}$	$I_2^{(8)}, I_3^{(8)}, I_6^{(8)}, I_7^{(8)}$
$\mathcal{S}_4$	$I_1^{(8)}, I_2^{(8)}, I_3^{(8)}, I_4^{(8)}$	$I_5^{(8)}, I_6^{(8)}, I_7^{(8)}, I_8^{(8)}$
$\mathcal{S}_5$	$I_1^{(8)}, I_3^{(8)}, I_6^{(8)}, I_8^{(8)}$	$I_2^{(8)}, I_4^{(8)}, I_5^{(8)}, I_7^{(8)}$
$\mathcal{S}_6$	$I_1^{(8)}, I_2^{(8)}, I_7^{(8)}, I_8^{(8)}$	$I_3^{(8)}, I_4^{(8)}, I_5^{(8)}, I_6^{(8)}$
$\mathcal{S}_7$	$I_1^{(8)}, I_4^{(8)}, I_6^{(8)}, I_7^{(8)}$	$I_2^{(8)}, I_3^{(8)}, I_5^{(8)}, I_8^{(8)}$

$n-1$ because the $OA(4k, 2^{4k-1})$ employed is a saturated orthogonal array (Wu & Hamada, 2011).

Remark 7. The blocked 3×2 cross validation provided by Wang et al. (2014) is a special case of the proposed $m \times 2$ BCV with $m = 3$. In fact, the construction method of blocked 3×2 cross validation is in accordance with our method constructed based on $OA(4, 2^3)$.

The construction of $\mathbb{S}^b$ of 7×2 is intuitively related to $\mathbb{S}^b$ of 3×2. In data partitioning for $\mathbb{S}^b$ of 3×2, each of the four blocks from $\mathbb{S}^b$ of 3×2 is split further into two equal-sized subblocks. These eight blocks can also be used to construct $\mathbb{S}^b$ of 7×2. In essence, the partitions for $\mathbb{S}^b$ of 7×2 include the partitions for $\mathbb{S}^b$ of 3×2.

Generally when $4k = 2^p$, $\mathbb{S}^b$ of $(2^p - 1) \times 2$ can be constructed based on $\mathbb{S}^b$ of $(2^{p-1} - 1) \times 2$. Specifically, $\mathbb{S}^b$ of $(2^p - 1) \times 2$ is expanded from $\mathbb{S}^b$ of $(2^{p-1} - 1) \times 2$. In this letter, this construction method is called the nested construction algorithm. It is formulated as follows:

1. Construct an orthogonal array $OA(2^p, 2^{2^p-1})$ based on $OA(2^{p-1}, 2^{2^{p-1}-1})$, $p \geq 3$ (Wu & Hamada, 2011). Specifically, the $OA(2^{p-1}, 2^{2^{p-1}-1})$ corresponds to a Hardmard matrix H. Then matrix $\begin{bmatrix} H & H \\ H & -H \end{bmatrix}$ is still a Hardmard matrix, which corresponds to $OA(2^p, 2^{2^p-1})$.
2. Split all blocks used in $\mathbb{S}^b$ of $2^{p-1} \times 2$ into two nearly equal-sized subblocks. For any $j \in \{1, 2, \ldots, 2^{p-1}\}$, the original jth block should be split evenly and denoted as the jth subblock and the $(j + p)$th subblock.
3. Generate the $(j + p)$th partition in $\mathbb{S}^b$ of $(2^p - 1) \times 2$ by employing step i of theorem 2 on the $(j + p)$th column of the $OA(2^p, 2^{2^p-1})$ and the blocks of step ii.

Table 3: Orthogonal Array $OA(4, 2^3)$.

Row Index	a	b	ab
$1 \leftrightarrow I_1^{(4)}$	+	+	+
$2 \leftrightarrow I_2^{(4)}$	-	+	-
$3 \leftrightarrow I_3^{(4)}$	+	-	-
$4 \leftrightarrow I_4^{(4)}$	-	-	+

Table 4: Mapping of Blocks and Partitions between 3×2 BCV and 7×2 BCV.

	$I_i^{(t)}$		$I_i^{(v)}$	
Partition	3×2 BCV	7×2 BCV	3×2 BCV	7×2 BCV
$\mathcal{S}_1$	$I_1^{(4)}, I_3^{(4)}$	$I_1^{(8)}, I_3^{(8)}, I_5^{(8)}, I_7^{(8)}$	$I_2^{(4)}, I_4^{(4)}$	$I_2^{(8)}, I_4^{(8)}, I_6^{(8)}, I_8^{(8)}$
$\mathcal{S}_2$	$I_1^{(4)}, I_2^{(4)}$	$I_1^{(8)}, I_2^{(8)}, I_5^{(8)}, I_6^{(8)}$	$I_3^{(4)}, I_4^{(4)}$	$I_3^{(8)}, I_4^{(8)}, I_7^{(8)}, I_8^{(8)}$
$\mathcal{S}_3$	$I_1^{(4)}, I_4^{(4)}$	$I_1^{(8)}, I_4^{(8)}, I_5^{(8)}, I_8^{(8)}$	$I_2^{(4)}, I_3^{(4)}$	$I_2^{(8)}, I_3^{(8)}, I_6^{(8)}, I_7^{(8)}$
$\mathcal{S}_4$		$I_1^{(8)}, I_2^{(8)}, I_3^{(8)}, I_4^{(8)}$		$I_5^{(8)}, I_6^{(8)}, I_7^{(8)}, I_8^{(8)}$
$\mathcal{S}_5$		$I_1^{(8)}, I_3^{(8)}, I_6^{(8)}, I_8^{(8)}$		$I_2^{(8)}, I_4^{(8)}, I_5^{(8)}, I_7^{(8)}$
$\mathcal{S}_6$		$I_1^{(8)}, I_2^{(8)}, I_7^{(8)}, I_8^{(8)}$		$I_3^{(8)}, I_4^{(8)}, I_5^{(8)}, I_6^{(8)}$
$\mathcal{S}_7$		$I_1^{(8)}, I_4^{(8)}, I_6^{(8)}, I_7^{(8)}$		$I_2^{(8)}, I_3^{(8)}, I_5^{(8)}, I_8^{(8)}$

The following example illustrates the nested construction of $\mathbb{S}^b$ of 7×2 based on $\mathbb{S}^b$ of 3×2.

Example 3. $\mathbb{S}^b$ of 3×2 is based on $OA(4, 2^3)$(see Table 3) and the four blocks $(I_1^{(4)}, I_2^{(4)}, I_3^{(4)}, I_4^{(4)})$. The upper left-hand corner 4×3 subarray in Table 1 is identical to $OA(4, 2^3)$ and the fourth column of $OA(8, 2^7)$. Next, the four blocks $(I_1^{(4)}, I_2^{(4)}, I_3^{(4)}, I_4^{(4)})$ are split into eight subblocks $(I_1^{(8)}, I_2^{(4)}, I_3^{(4)}, \ldots, I_8^{(8)})$ using the following rules:

$$I_1^{(4)} \leftrightarrow I_1^{(8)}, I_5^{(8)} \quad I_2^{(4)} \leftrightarrow I_2^{(8)}, I_6^{(8)},$$
$$I_3^{(4)} \leftrightarrow I_3^{(8)}, I_7^{(8)} \quad I_4^{(4)} \leftrightarrow I_4^{(8)}, I_8^{(8)}.$$

Finally, the partitions of $\mathcal{S}_4, \mathcal{S}_5, \mathcal{S}_6, \mathcal{S}_7$ in $\mathbb{S}^b$ of 7×2 are derived using the last four columns of $OA(8, 2^7)$ and eight subblocks. All the partitions in $\mathbb{S}^b$ of 3×2 and $\mathbb{S}^b$ of 7×2 are compared in Table 4. Their first three partitions are illustrated as identical.

Corollary 2 indicates that the increase in m in $m \times 2$ BCV reduces the variance of the estimator of the generalization error. Thus, continually adding

the number of partitions on the basis of the previous cross-validated estimators to form the next $m \times 2$ BCV is very useful in practical experiments.

5 Selection of m

In practical applications, providing a selection method of m is necessary. equation 3.11 of theorem 1 shows that $E\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}|\Phi^*)) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\rho_2$. As m increases, the magnitude of variance reduction declines as well, although the variance gradually decreases. Considering the reduction rate of variance,

$$\frac{E\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}|\Phi^*)) - E\mathrm{Var}(\hat{\mu}_{(m+1)\times 2}(\mathbb{S}|\Phi^*))}{E\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}|\Phi^*))} = \frac{1+\rho_1-2\rho_2}{(m+1)(1+\rho_1)+2(m^2-1)\rho_2}. \tag{5.1}$$

If the value is very small, such as smaller than α (5% or 1%), additional replications are not required. Hence, the problem of determining m can be solved with this idea.

However, ρ_1 and ρ_2 in the reduction rate of variance are unknown. The values of ρ_1 and ρ_2 should be related to the sample size, the used model (algorithm), and so on. It should not be related to m. Based on the large number of simulation experiments conducted by Wang et al. (2014), we believe that the ranges of the values of ρ_1 and ρ_2 should be $0 < \rho_1, \rho_2 < 1/2$.

We let ARRV denote the averaged reduction rate of variance over the range of $0 < \rho_1, \rho_2 < 1/2$ regardless of the model used. Hence, we recommend determining m by limiting ARRV to smaller than α:

$$ARRV \triangleq 4\int_0^{0.5}\int_0^{0.5} \frac{1+\rho_1-2\rho_2}{(m+1)(1+\rho_1)+2(m^2-1)\rho_2} d\rho_1 d\rho_2 < \alpha. \tag{5.2}$$

Table 5 shows that the 3×2 BCV provided by Wang et al. (2014) has an ARRV of less than 10%. If one wishes the averaged reduction rate of variance to be smaller than $\alpha = 5\%$, one should make m at least larger than 5. This may provide an explanation as to why 5×2 cross validation is empirically recommended by several researchers in the comparison of algorithm performance (Dietterich, 1998; Alpaydin, 1999; Yildiz, 2013). Furthermore, if one wishes the averaged reduction rate of variance to be smaller than $\alpha = 1\%$, one must make $m \geq 16$.

Table 5: Averaged Reduction Rate of Variance with Regard to m.

m	ARRV	α	Scheme of Cross Validation
2	0.1552		
3	0.0984	<10%	3×2 BCV
4	0.0688		
5	0.0516		
6	0.0404	<5%	6×2 BCV
7	0.0324		
11	0.0168		
15	0.0105		
16	0.0095	<1%	16×2 BCV

Table 6: Comparison of New and Old Notations.

New Notations	Old Notations	Meaning
$\hat{\mu}_{m\times 2}$	$\hat{\mu}_{m\times 2}(\mathbb{S}^*)$	Balanced $m \times 2$ BCV estimator
$\hat{\mu}^{(i)}$	$\hat{\mu}(\mathcal{S}_i)$	The ith S2CV estimator in $\hat{\mu}_{m\times 2}(\mathbb{S}^b)$
$\hat{\mu}_1^{(i)}$	$\hat{\mu}_{HO}(\mathcal{S}_i)$	One HO estimator in $\hat{\mu}(\mathcal{S}_i)$
$\hat{\mu}_2^{(i)}$	$\hat{\mu}_{HO}(\mathcal{S}_i^\top)$	Another HO estimator in $\hat{\mu}(\mathcal{S}_i)$

6 Estimation of $\mathbf{Var}(\hat{\boldsymbol{\mu}}_{m\times 2}(\mathbb{S}^*))$

Before introducing the variance estimator, we provide this theorem:

Theorem 3. *Universal unbiased estimator of* $Var(\hat{\mu}_{m\times 2}(\mathbb{S}^*))$ *does not exist.*

Here, universal refers to the estimation statistic that it is valid under all distributions of samples. The proof of theorem 3 is similar to that provided by Bengio and Grandvalet (2004) and Yang, Wang, Wang, and Li (2014); thus, it is omitted in this letter.

For a simple and clear expression of the idea, the notations in the above sections are simplified in Table 6.

6.1 Estimators of $\mathbf{Var}(\hat{\boldsymbol{\mu}}_{m\times 2})$. In this section, we consider a generic estimator of $\mathrm{Var}(\hat{\mu}_{m\times 2})$ that depends on the within-block and between-block sample variances. Similar to Wang et al. (2014), the compromise of the within-block and between-blocks sample variances can be expressed,

$$\widehat{\mathrm{Var}}(\hat{\mu}_{m\times 2}) = \lambda_1 \frac{1}{m^2}\sum_{i=1}^{m}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)})^2 + \lambda_2 \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\mu}^{(i)} - \hat{\mu}_{m\times 2})^2, \tag{6.1}$$

where λ_1 and λ_2 are two hyperparameters to tune the relative importance of the within-block and between-block sample variances in the variance estimator because the within-block sample variance is almost a liberal estimator of variance.

The expectation of $\widehat{\mathrm{Var}}(\hat{\mu}_{m\times 2})$ in equation 6.1 is

$$E(\widehat{\mathrm{Var}}(\hat{\mu}_{m\times 2})) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{1}{m}\sigma^2\left[\left(\lambda_1 + \frac{m}{2}\lambda_2 - \frac{1}{2}\right) - \left(\lambda_1 - \frac{m}{2}\lambda_2 + \frac{1}{2}\right)\rho_1 - m\lambda_2\rho_2\right]. \tag{6.2}$$

Whether this type of estimator is liberal or conservative depends on the selection of the values of λ_1 and λ_2. However, finding universal good values of λ_1 and λ_2 is difficult. We provide several specific pairs of values to construct three variance estimators:

1. $\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$: when $\lambda_1 = \frac{m}{2}$ and $\lambda_2 = 0$:

$$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2}) \triangleq \frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)})^2, \tag{6.3}$$

$$E(\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\left[\frac{1}{2} - \frac{m+1}{2(m-1)}\rho_1\right],$$

$$E(\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})) - \mathrm{Var}(\hat{\mu}_{m\times 2}) = \frac{m-1}{m}\sigma^2\left[\frac{1}{2} - \frac{m+1}{2(m-1)}\rho_1 - \rho_2\right]. \tag{6.4}$$

   As long as ρ_1 and ρ_2 satisfy $\frac{1}{2} > \rho_2 + \frac{m+1}{2(m-1)}\rho_1$, $E(\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})) > \mathrm{Var}(\hat{\mu}_{m\times 2})$.

2. $\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$: when $\lambda_1 = \frac{m}{2}$ and $\lambda_2 = \frac{m-1}{m}$:

$$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2}) \triangleq \frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}_{m\times 2})^2, \tag{6.5}$$

$$E(\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\left[1 - \frac{1}{m-1}\rho_1 - \rho_2\right], \tag{6.6}$$

$$E(\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})) - \mathrm{Var}(\hat{\mu}_{m\times 2}) = \frac{m-1}{m}\sigma^2\left[1 - \frac{1}{(m-1)}\rho_1 - 2\rho_2\right]. \tag{6.7}$$

   When ρ_1 and ρ_2 satisfy $\frac{1}{2} > \rho_2 + \frac{1}{2(m-1)}\rho_1$, $E(\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})) > \mathrm{Var}(\hat{\mu}_{m\times 2})$.

3. $\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})$: when $\lambda_1 = \frac{m}{2}$ and $\lambda_2 = \frac{m+1}{m}$:

$$\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2}) \triangleq \frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{2}(\hat{\mu}_k^{(i)} - \hat{\mu}^{(i)})^2 + \frac{m+1}{m(m-1)}\sum_{i=1}^{m} (\hat{\mu}^{(i)} - \hat{\mu}_{m\times 2})^2, \tag{6.8}$$

$$E(\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})) = \frac{1}{2m}\sigma^2(1+\rho_1) + \frac{m-1}{m}\sigma^2\left[\frac{m}{m-1} - \frac{m+1}{m-1}\rho_2\right],$$

$$E(\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})) - \text{Var}(\hat{\mu}_{m\times 2}) = \frac{m-1}{m}\sigma^2\left[\frac{m}{m-1} - \frac{2m}{m-1}\rho_2\right]. \tag{6.9}$$

As long as $\frac{1}{2} > \rho_2$, $E(\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})) > \text{Var}(\hat{\mu}_{m\times 2})$.

Remark 8. According to proposition 2 in Wang et al. (2014), if the loss function does not depend on actual examples and the underlying algorithm, then $\rho_2 = 0.5$. However, this is a balanced result. In fact, the loss function must depend on the data and the algorithm. Thus, ρ_2 should be smaller than 0.5. The simulation experiments in the work of Wang et al. (2014) validated that $\frac{1}{2} > \rho_2, \rho_1 > 0$.

6.2 Comparison of the Estimators of $\text{Var}(\hat{\mu}_{m\times 2})$. To provide a convincing statistical inference in the comparison of algorithm performance and interval estimation, conservative variance estimation should be provided. The admissible (ρ_1, ρ_2) regions where the above three variances are conservative estimators are shown in Figure 3 for $m = 3, 5, 7, 9$.

Figure 3 shows that $\widehat{\text{Var}}_1(\hat{\mu}_{m\times 2})$ being conservative in a large region of $0 < \rho_1, \rho_2 < \frac{1}{2}$ cannot be guaranteed. Although the coverage region increases as m increases, it can at most cover three-quarters of the entire region (see Table 7). For $\widehat{\text{Var}}_2(\hat{\mu}_{m\times 2})$, its expression is simple and clear. The admissible (ρ_1, ρ_2) region satisfying the condition $\frac{1}{2} > \rho_2 + \frac{1}{2(m-1)}\rho_1$ covers most of the region of $0 < \rho_1, \rho_2 < 1/2$. As m increases, its coverage region becomes close to the entire region. Furthermore, the admissible region of the third variance estimator, $\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})$, covers all area of the region. Therefore, the $\widehat{\text{Var}}_3(\hat{\mu}_{m\times 2})$ is the most conservative estimator among the three estimators.

Based on this analysis, we recommend $\widehat{\text{Var}}_2(\hat{\mu}_{m\times 2})$ as an estimator of $\text{Var}(\hat{\mu}_{m\times 2})$. The simulation experiments in section 7.5 show that $\widehat{\text{Var}}_2(\hat{\mu}_{m\times 2})$ is a suitable estimator of $\text{Var}(\hat{\mu}_{m\times 2})$.

7 Simulation Study

In this section, we demonstrate the following through simulated experiments:

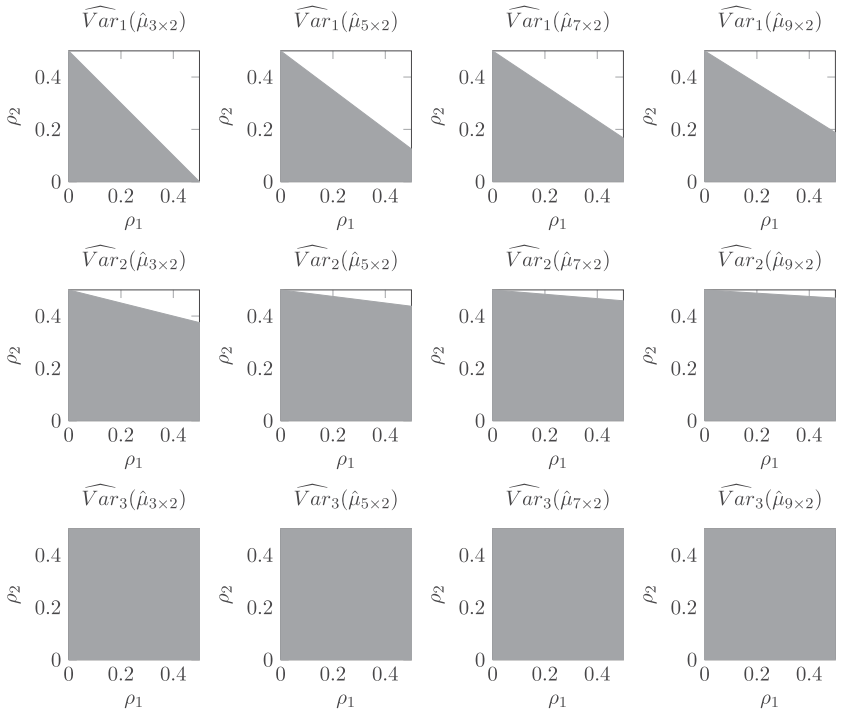

Figure 3: Admissible (ρ_1, ρ_2) regions where the three variance estimators are conservative.

Table 7: Comparison of the Three Variance Estimators.

Case	Estimator	Condition That $E(\widehat{\text{Var}}(\hat{\mu}_{m\times2})) > \text{Var}(\hat{\mu}_{m\times2})$	Ratio of the Admissible (ρ_1, ρ_2) Region and Entire Region
1	$\widehat{\text{Var}}_1(\hat{\mu}_{m\times2})$	$\frac{1}{2} > \rho_2 + \frac{m+1}{2(m-1)}\rho_1$	$\frac{3m-5}{4(m-1)} \times 100\%$
2	$\widehat{\text{Var}}_2(\hat{\mu}_{m\times2})$	$\frac{1}{2} > \rho_2 + \frac{1}{2(m-1)}\rho_1$	$\frac{4m-5}{4(m-1)} \times 100\%$
3	$\widehat{\text{Var}}_3(\hat{\mu}_{m\times2})$	$\frac{1}{2} > \rho_2$	100%

1. Is coefficient $\omega + \gamma - 2\tau$ in the function $f(x)$ larger than 0?
2. How well can $f(x)$ be approximated in parameters σ^2, ω, γ, and τ, and how well can $g(x) \triangleq \text{Cov}(\hat{\mu}(S_1), \hat{\mu}(S_2)|\phi = x)$ be approximated in these parameters in the neighborhood of $x = \frac{n}{4}$?
3. How large is the difference of the variances between the $m \times 2$ CV and $m \times 2$ BCV estimators of the generalization error?
4. Which of the following is a suitable estimator: $\widehat{\text{Var}}_1(\hat{\mu}_{m\times2})$, $\widehat{\text{Var}}_2(\hat{\mu}_{m\times2})$ and $\widehat{\text{Var}}_3(\hat{\mu}_{m\times2})$?

Table 8: Coefficients of $f(x)$ on Simulation Data Set.

Configuration	ρ_2	$\omega + \gamma - 2\tau$	$\sigma^2 - \omega - n\gamma + n\tau$	γ
SREG + lm + 1000 + 100	0.4433	0.0023	3.7212	0.0029
SREG + rid + 1000 + 100	0.442	0.0022	3.7782	0.0028
SREG + lso + 1000 + 100	0.2822	0.0052	51.1965	−0.0005
SCLA + svm + 1000 + 20	0.4308	0.0001	−0.127	0.0002
SCLA + knn + 1000 + 20	0.3078	0.0001	−0.1592	0.0002

7.1 Experimental Setup of Simulations. For regression and classification situations, the experimental setups of regression and classification data sets are considered, respectively. Multiple classical machine learning algorithms are used in two types of data. The detailed settings of our simulated data sets and algorithms are as follows:

- *Simulated regression data set (SREG).* The predictor vector x_i contains p independent predictors, which are all extracted from a standard normal distribution. Response $y_i = \sqrt{3/p}\sum_{k=1}^{p} x_{ik} + \varepsilon_i$, in which $\varepsilon_i \sim N(0, 1)$. The setup of the data set comes from the work of Nadeau et al. (2003). We let $p < n$ and employ the usual linear regression (lm), ridge regression (rid), and lasso method (lso) to estimate generalization error. The loss function is a squared loss function. In this data set, we use $n = 1000$ and $p = 100$.
- *Simulated two-class classification data set (SCLA).* Data set $D_n = (x_i, y_i)_{i=1}^n$ is obtained with $Prob(Y = 1) = Prob(Y = 0) = \frac{1}{2}$ and $X|Y = 0 \sim N(0, I)$. For $Y = 1$, the first 10% of predictors are drawn from a normal distribution with a mean of 0.5 and a variance of 1; the other predictors are from a standard normal distribution. Here, we employ the support vector machine (svm) and the k-nearest neighborhood algorithm with $k = 5$ and triangular kernel (knn) as our classifiers. The loss function is the 0-1 loss function. This data set comes from the work of Tibshirani & Tibshirani (2009). The size n and dimensionality p of this data set are set to 1000 and 20.

In the following sections, we use data_name + algor_name + n + p to denote each simulation configuration. For example, SREG + lm + 1000 + 20 means that the experimental configuration consists of SREG data set with 1000 samples and 20 predictors and a linear regression algorithm.

7.2 Simulation Experiments for Question 1. The experimental purpose is to examine whether the coefficient $\omega + \gamma - 2\tau$ of the quadratic term of $f(x)$ is larger than zero. The results are shown in Table 8. The condition that $\omega + \gamma > 2\tau$ is satisfied in all cases.

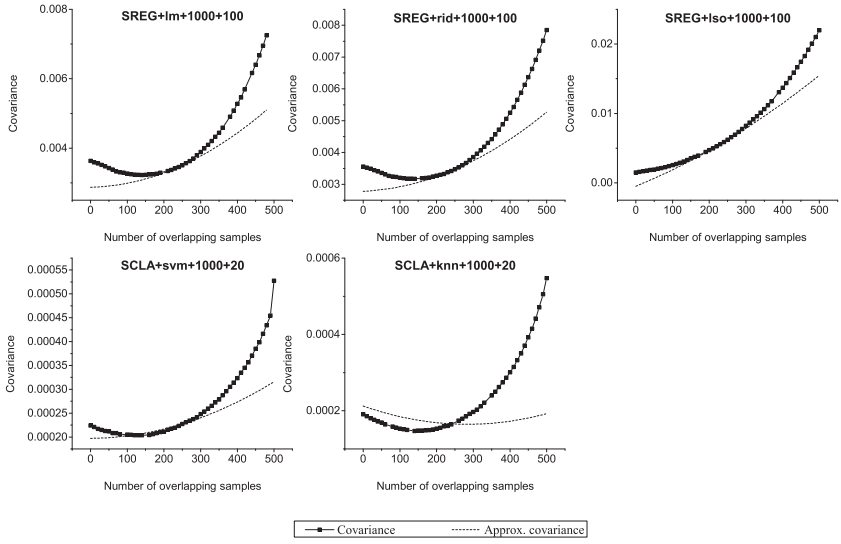

Figure 4: $f(x)$ and its approximation on the SREG and SCLA data set.

The values of ρ_2 are also provided in Table 8. From the table, we see that $0 < \rho_2 < 0.5$.

7.3 Simulation Experiments for Question 2. Some simulation results of two functions, $f(x)$ and $g(x)$, and their approximations in parameter σ^2, ω, τ, and γ in the neighborhood of $x = n/4$ are provided in Figures 4 and 5.

These results can well support that $f(x)$ and $g(x)$ are lower convex functions, as we proved in the two lemmas, and they can be well approximated in parameter σ^2, ω, τ, and γ in the neighborhood of $x = n/4$.

7.4 Simulation Experiments for Question 3. The purpose of this experiment is to examine whether the variance of the $m \times 2$ BCV estimator of the generalization error is not larger than that of the $m \times 2$ CV estimator. In this simulation, we randomly generated 10,000 data sets from a population and 1,000 sets of partitions for $m \times 2$ CV and $m \times 2$ BCV, respectively. Then, 10 million estimators of generalization error are derived and sample variances are computed as reported numeric variances. In these simulations, $m = 3, 5, 7, 9$ are employed. The experimental results are shown in Table 9. (The "scale" column provides orders of magnitude for each row.)

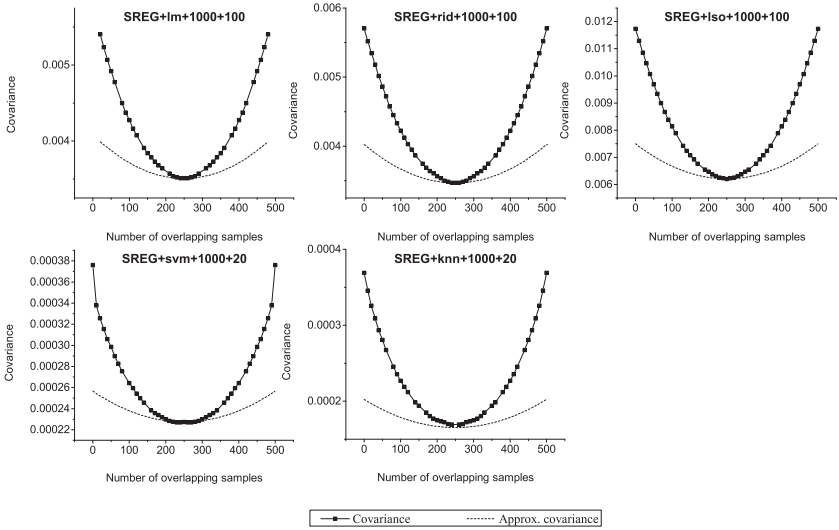

Figure 5: $g(x)$ and its approximation on SREG and SCLA data set.

Table 9: Variance Comparison on Simulation Data Sets.

Configuration	Scale	$m = 3$	$m = 5$	$m = 7$	$m = 9$
$m \times 2$ BCV					
SREG + lm + 1000 + 100	10^{-4}	7.5674	4.5481	3.2458	2.5202
SREG + rid + 1000 + 100	10^{-4}	7.4624	4.4860	3.2012	2.4850
SREG + lso + 1000 + 100	10^{-3}	1.8099	1.0841	0.7731	0.5996
SCLA + svm + 1000 + 20	10^{-5}	4.9337	2.9554	2.1120	1.6391
SCLA + knn + 1000 + 20	10^{-5}	6.6614	3.9830	2.8371	2.2051
$m \times 2$ CV					
SREG + lm + 1000 + 100	10^{-4}	7.5948	4.5602	3.2599	2.5355
SREG + rid + 1000 + 100	10^{-4}	7.4901	4.4958	3.2149	2.4998
SREG + lso + 1000 + 100	10^{-3}	1.8092	1.0852	0.7754	0.6036
SCLA + svm + 1000 + 20	10^{-5}	4.9454	2.9658	2.1196	1.6493
SCLA + knn + 1000 + 20	10^{-5}	6.6717	4.0049	2.8633	2.2228

For a clear comparison, the reduction percentage is defined as follows:

$$\text{Reduction percentage} = \frac{E_\Phi \text{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}|\Phi)) - E_{\Phi^b}\text{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}^b)|\Phi^b)}{E_\Phi \text{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi)} \cdot 100\%. \tag{7.1}$$

Table 10: Variance Reduction Percentages on Simulation Data Sets.

Configuration	Reduction Percentage (%)			
	$m = 3$	$m = 5$	$m = 7$	$m = 9$
SREG + lm + 1000 + 100	0.36	0.27	0.43	0.60
SREG + rid + 1000 + 100	0.37	0.22	0.43	0.59
SREG + lso + 1000 + 100	−0.04	0.10	0.30	0.66
SCLA + svm + 1000 + 20	0.23	0.35	0.36	0.62
SCLA + knn + 1000 + 20	0.15	0.55	0.91	0.80

Table 11: Comparison of Three Variance Estimators on Configurations SREG + rid + n + 20.

m	Variance	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
3	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.192125	0.030131	0.014093	0.008821	0.006376
3	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.151037	0.015067	0.007338	0.005044	0.003932
3	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.250211	0.027984	0.012346	0.007712	0.005587
3	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.349384	0.040901	0.017353	0.010380	0.007242
5	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.172204	0.027544	0.013121	0.008288	0.006042
5	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.153129	0.015071	0.007338	0.005040	0.003932
5	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.272423	0.030586	0.013356	0.008240	0.005922
5	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.332070	0.038343	0.016366	0.009840	0.006916
7	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.163635	0.026455	0.012683	0.008057	0.005902
7	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.152596	0.015064	0.007336	0.005040	0.003931
7	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.280363	0.031687	0.013779	0.008472	0.006063
7	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.322952	0.037228	0.015927	0.009616	0.006773

The reduction percentages for all simulation configurations are listed in Table 10.

Tables 9 and 10 reveal the following:

1. $E_{\Phi}\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S})|\Phi) > E_{\Phi^b}\mathrm{Var}(\hat{\mu}_{m\times 2}(\mathbb{S}^b)|\Phi^b)$ for all situations, that is, the variance of the $m \times 2$ CV estimator is larger than that of the $m \times 2$ BCV estimator.
2. As m increases from 3 to 9, the reduction percentage appears to increase as well. This condition means that as m increases, the effectiveness of $m \times 2$ BCV in the reduction of the variance of the estimator of the generalization error becomes increasingly evident.

7.5 Simulation Experiments for Question 4. Tables 11 and 12 compare three variance estimators on simulation data sets.

Table 12: Comparison of Three Variance Estimators on Configurations SCLA + svm + n + 20.

m	Variance	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$
3	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.002662	0.001442	0.000960	0.000721	0.000558
3	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.001270	0.000646	0.000436	0.000333	0.000270
3	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.002908	0.001384	0.000898	0.000657	0.000516
3	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.004545	0.002122	0.001359	0.000981	0.000761
5	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.002337	0.001293	0.000867	0.000657	0.000509
5	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.001270	0.000645	0.000437	0.000332	0.000271
5	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.003235	0.001532	0.000990	0.000721	0.000565
5	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.004217	0.001975	0.001266	0.000916	0.000713
7	$\mathrm{Var}(\hat{\mu}_{m\times 2})$	0.002195	0.001231	0.000828	0.000629	0.000488
7	$\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$	0.001271	0.000645	0.000437	0.000332	0.000271
7	$\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$	0.003375	0.001595	0.001029	0.000749	0.000586
7	$\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$	0.004077	0.001912	0.001227	0.000888	0.000692

From these results, we know that:

- $\widehat{\mathrm{Var}}_1(\hat{\mu}_{m\times 2})$ underestimates the variance in almost all cases. Meanwhile, $\widehat{\mathrm{Var}}_3(\hat{\mu}_{m\times 2})$ is a conservative estimator of true variance.
- $\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$ is conservative in most cases. When m increases from 3 to 7, it is a conservative estimator for all cases. Thus, in practical applications, it is suitable to use $\widehat{\mathrm{Var}}_2(\hat{\mu}_{m\times 2})$ as the estimator of variance.

7.6 Results on Real-Life Data Sets. In this section, we compare the variances between the $m \times 2$ BCV estimator and the $m \times 2$ CV estimator on multiple real-life data sets. All of these data sets were obtained from the UC Irvine machine learning repository (http://archive.ics.uci.edu.ml).

- The letter recognition data set (LETTER) is used to recognize a character based on an image.[1] The data set consists of 16 predictors and 26 classes. The data set setting is the same as that in the work of Bengio and Grandvalet (2004). Specifically, we combined the letters A to M as the first class and the remaining letters as the other class. We treat it as a binary-classification task, and then we employ the support vector machine (svm) and k-nearest neighborhood (knn).
- The wine quality data set(WQ) is used to predict wine quality.[2] Only part of the data set about white wine is used in our experiment. It has 4898 samples and 11 predictors. The response variable is the quality

[1]Letter recognition data set: http://archive.ics.uci.edu/ml/datasets/Letter+Recognition.

[2]Wine-quality data set: http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

grade of wine. We view it as a regression task and employ linear (lm), ridge (rid), and lasso regressions (lso).

- The air quality data set (AQ) is about the air quality in Italy. The response variable of the data set is absolute humidity.[3] The date and time tags of each record are removed, and the variable of the non-methane hydrocarbon concentration (NMHC) is removed because the majority of its values are missing. Furthermore, we remove all records having missing values. The final data set has 6941 records and 11 predictors. Linear (lm), ridge (rid), and lasso regressions (lso) are used.
- The connect-four data set (CON) is about the connect-four game.[4] This data set has 67,557 samples, 2 classes for the response variable, and 42 predictors. We regard it as a classification task and use support vector machine (svm) and k-nearest neighborhood (knn).
- The census income data set (ADULT) was extracted from a census database.[5] It has 32,561 samples and 14 predictors. After omitting rows with missing values, the remaining data set still has 30,162 samples. The task on this data set predicts whether a person makes over $50,000 annually. This task can be treated as a binary classification problem. Support vector machine (svm) and k-nearest neighborhood (knn) are employed as classifiers.

The other settings of variance simulation are the same as those of the simulations on the SREG and SCLA data sets. Ten million replications are used for each experiment, including replications of 10,000 data sets and 1,000 partitions. All variances of the $m \times 2$ BCV and $m \times 2$ CV estimators are listed in Table 13. The corresponding reduction rates are given in Table 14.

From these results, we can obtain that the variance of the $m \times 2$ BCV estimator is smaller than that of the $m \times 2$ CV estimator for all configurations. Furthermore, as m increases, the variances of the $m \times 2$ BCV and $m \times 2$ CV estimators decrease and the reduction percentages increase for almost all configurations.

8 Conclusion

We developed a new data partitioning scheme for cross validation called block-regularized $m \times 2$ cross validation ($m \times 2$ BCV). In $m \times 2$ BCV, the differences between the number of overlapping samples and $n/4$ are controlled into smaller than its expectation of random situation. We discussed variance optimal property in all $m \times 2$ BCV estimators. Furthermore, we

[3] Air quality data set:http://archive.ics.uci.edu/ml/datasets/Air+Quality.
[4] Connect-4 data set: http://archive.ics.uci.edu/ml/datasets/Connect-4.
[5] Census income data set:http://archive.ics.uci.edu/ml/datasets/Census+Income.

Table 13: Variance Comparison on Real-Life Data Sets.

Configuration	Scale	$m = 3$	$m = 5$	$m = 7$	$m = 9$
$m \times 2$ BCV					
LETTER + svm + 500 + 256	10^{-5}	8.1657	4.8959	3.5017	2.7123
LETTER + knn + 500 + 256	10^{-4}	1.0508	0.6271	0.4469	0.3465
WQ + lm + 100 + 11	10^{-2}	1.1509	0.68899	0.48625	0.37031
WQ + rid + 100 + 11	10^{-2}	1.4007	0.82894	0.55728	0.42205
WQ + lso + 100 + 11	10^{-3}	3.6329	2.1599	1.5188	1.1654
AQ + lm + 200 + 11	10^{-7}	7.4631	4.4451	3.1432	2.4231
AQ + rid + 200 + 11	10^{-8}	3.7909	2.2534	1.5905	1.2153
AQ + lso + 200 + 11	10^{-7}	8.2932	4.9378	3.4873	2.6908
CON + svm + 500 + 42	10^{-4}	3.0035	1.6996	1.2739	0.99936
CON + knn + 500 + 42	10^{-4}	1.4920	0.89203	0.63534	0.49354
ADULT + svm + 500 + 14	10^{-5}	5.8201	3.5065	2.5294	1.9538
ADULT + knn + 500 + 14	10^{-5}	7.7733	4.6526	3.3153	2.5719
$m \times 2$ CV					
LETTER + svm + 500 + 256	10^{-5}	8.2011	4.9502	3.5278	2.7458
LETTER + knn + 500 + 256	10^{-4}	1.0540	0.6339	0.4521	0.3518
WQ + lm + 100 + 11	10^{-2}	1.1747	0.70407	0.49876	0.38629
WQ + rid + 100 + 11	10^{-2}	1.4065	0.84562	0.60355	0.47136
WQ + lso + 100 + 11	10^{-3}	3.6832	2.2152	1.5798	1.2307
AQ + lm + 200 + 11	10^{-7}	7.5392	4.5222	3.2349	2.5162
AQ + rid + 200 + 11	10^{-8}	3.8234	2.2979	1.6411	1.2772
AQ + lso + 200 + 11	10^{-7}	8.3728	5.0326	3.5938	2.7991
CON + svm + 500 + 42	10^{-4}	3.0443	1.8274	1.2922	1.0224
CON + knn + 500 + 42	10^{-4}	1.5024	0.90046	0.64282	0.50065
ADULT + svm + 500 + 14	10^{-5}	5.8495	3.6059	2.5599	1.9999
ADULT + knn + 500 + 14	10^{-5}	7.8270	4.7058	3.3588	2.6101

Table 14: Variance Reduction Percentages (%) on Real-Life Data Sets.

	Reduction Percentage			
Configuration	$m = 3$	$m = 5$	$m = 7$	$m = 9$
LETTER + svm + 500 + 256	0.43	1.10	0.74	1.22
LETTER + knn + 500 + 256	0.31	1.07	1.14	1.50
WQ + lm + 100 + 11	2.03	2.14	2.51	4.14
WQ + rid + 100 + 11	0.41	1.97	7.67	10.46
WQ + lso + 100 + 11	1.36	2.49	3.86	5.30
AQ + lm + 200 + 11	1.01	1.70	2.83	3.70
AQ + rid + 200 + 11	0.85	1.94	3.08	4.85
AQ + lso + 200 + 11	0.95	1.88	2.96	3.87
CON + svm + 500 + 42	1.34	6.99	1.41	2.25
CON + knn + 500 + 42	0.69	0.93	1.16	1.42
ADULT + svm + 500 + 14	0.50	2.76	1.19	2.30
ADULT + knn + 500 + 14	0.69	1.13	1.29	1.46

provided a nested construction algorithm of $m \times 2$ BCV based on a two-level orthogonal array. Finally, a conservative estimator of the variance of estimator of the generalization error is recommended.

In the case of non–indentical and independently distributed (i.i.d.) samples such as text data, the idea of the BCV framework proposed in this letter remains applicable, but the method should be extended. A good partitioning not only controls the percentage of overlapping parts between any two text training (test) sets in a partition scheme but also guarantees a small difference in the training and test sets. For text data in natural language processing areas, data partitioning could be implemented only in sentences, paragraphs, or documents in most specific tasks. This limitation may easily result in bad partitioning of text data, which can seriously affect the variance of the cross-validated estimator of the generalization error. Thus, to pursue good partitioning, additional measures of the differences of text data blocks should be introduced, such as measures of differences of token distribution, word frequency distribution, tag distribution, and sentence length distribution. Furthermore, the measures should be converted to the corresponding regularized conditions, and the conditions should be properly and carefully considered in the extension of the current BCV framework to further minimize the variance of the cross-validated estimator of the generalization error. Further in-depth studies can focus on how these measures can be converted to regularized conditions and introduced into the objective function of variance minimization and how the corresponding partitions can be constructed. These topics are included in our future research plan.

Appendix: Proof of $\omega + \gamma > 2\tau$

In this appendix, we provide our proofs of $\omega + \gamma > 2\tau$ for mean regression and multivariate regression with squared loss. Let $\mathcal{S}_1 = (I_1^{(t)}, I_1^{(v)})$ and $\mathcal{S}_2 = (I_2^{(t)}, I_2^{(v)})$ be two partitions of $\mathcal{I} = \{1, \ldots, n\}$, and the corresponding pairs of training and test sets are $(D_1^{(t)}, D_1^{(v)})$ and $(D_2^{(t)}, D_2^{(v)})$, respectively. Moreover, we simplify our loss function $L(\mathcal{A}(D^{(t)}), z_j)$ as $L(\hat{y}_{I^{(v)},j}, y_j)$, where y_j is the response value of test sample z_j and $\hat{y}_{I^{(v)},j}$ is the prediction value of y_j based on machine learning algorithm $\mathcal{A}$ and training set $D^{(t)} = \{z_i | i \in I^{(t)}\}$. A data set can be divided into the following parts:

- $A = \{a | a \in I_1^{(t)} \cap I_2^{(t)}\}$ is the index set of the common samples of training sets $D_1^{(t)}$ and $D_2^{(t)}$.
- $B = \{b | b \in I_1^{(t)} \backslash A\}$ is the index set of the samples occurring only in training set $D_1^{(t)}$, not in training set $D_2^{(t)}$.
- $C = \{c | c \in I_2^{(t)} \backslash A\}$ is the index set of the samples contained only in training set $D_2^{(t)}$, not in training set $D_1^{(t)}$.

- $D = \{d | d \in I_1^{(v)} \cap I_2^{(v)}\}$ is the index set of the samples that are not contained in training sets $D_1^{(t)}$ and $D_2^{(t)}$.

Under squared loss, ω, γ, and τ are expressed as follows:

- $\omega = \mathrm{Cov}((\hat{y}_{A\cup B,d} - y_d)^2, (\hat{y}_{A\cup C,d'} - y_{d'})^2), \forall d, d' \in D$ such that $d \neq d'$.
- $\gamma = \mathrm{Cov}((\hat{y}_{A\cup B,c} - y_c)^2, (\hat{y}_{A\cup C,b} - y_b)^2), \forall b \in B, c \in C$.
- $\tau = \mathrm{Cov}((\hat{y}_{A\cup B,c} - y_c)^2, (\hat{y}_{A\cup C,d} - y_d)^2), \quad \forall c \in C, d \in D, \quad$ or $\quad \tau = \mathrm{Cov}((\hat{y}_{A\cup B,d'} - y_{d'})^2, (\hat{y}_{A\cup C,b} - y_b)^2), \forall b \in B, d' \in D$.

The following sections are our proofs for the two regression situations.

A.1 Proof of $\omega + \gamma > 2\tau$ for Mean Regression. For mean regression, the algorithm uses only the n of response values $y_1, y_2, \ldots, y_n$ in a data set. These response values are identically and independently drawn from an unknown population. We assume the population mean and variance are μ and ψ^2, respectively. Moreover, mean regression uses the sample mean of the response values in the training set as a prediction value of a test sample: $\hat{y}_{A\cup B,d} = \bar{y}_{A\cup B}$, in which $\bar{y}_{A\cup B} = 2\sum_{i\in A\cup B} y_i/n$.

The covariances ω, γ, and τ can be decomposed directly as follows:

$$\begin{aligned}\omega = {} & \mathrm{Cov}(\bar{y}^2_{A\cup B}, \bar{y}^2_{A\cup C}) - 2\mathrm{Cov}(\bar{y}^2_{A\cup B}, 2\bar{y}_{A\cup C} y_{d'}) + 2\mathrm{Cov}(\bar{y}^2_{A\cup B}, y^2_{d'}) \\ & + \mathrm{Cov}(2\bar{y}_{A\cup B} y_d, 2\bar{y}_{A\cup C} y_{d'}) - 2\mathrm{Cov}(2\bar{y}_{A\cup B} y_d, y^2_{d'}) + \mathrm{Cov}(y^2_d, y^2_{d'}),\end{aligned} \tag{A.1}$$

$$\begin{aligned}\gamma = {} & \mathrm{Cov}(\bar{y}^2_{A\cup B}, \bar{y}^2_{A\cup C}) - 2\mathrm{Cov}(\bar{y}^2_{A\cup B}, \bar{y}_{A\cup C} y_b) - 2\mathrm{Cov}(\bar{y}^2_{A\cup C}, \bar{y}_{A\cup B} y_c) \\ & + \mathrm{Cov}(\bar{y}^2_{A\cup B}, y^2_b) + \mathrm{Cov}(\bar{y}^2_{A\cup C}, y^2_c) + 4\mathrm{Cov}(\bar{y}_{A\cup B} y_c, \bar{y}_{A\cup C} y_b) \\ & - 2\mathrm{Cov}(\bar{y}_{A\cup B} y_c, y^2_b) - 2\mathrm{Cov}(\bar{y}_{A\cup C} y_b, y^2_c) + \mathrm{Cov}(y^2_b, y^2_c),\end{aligned} \tag{A.2}$$

$$\begin{aligned}\tau = {} & \mathrm{Cov}(\bar{y}^2_{A\cup B}, \bar{y}^2_{A\cup C}) - 2\mathrm{Cov}(\bar{y}^2_{A\cup B}, \bar{y}_{A\cup C} y_b) - \mathrm{Cov}(\bar{y}^2_{A\cup C}, 2\bar{y}_{A\cup B} y_d) \\ & + \mathrm{Cov}(\bar{y}^2_{A\cup B}, y^2_b) + \mathrm{Cov}(\bar{y}^2_{A\cup C}, y^2_d) + 4\mathrm{Cov}(\bar{y}_{A\cup B} y_d, \bar{y}_{A\cup C} y_b) \\ & - 2\mathrm{Cov}(\bar{y}_{A\cup B} y_d, y^2_b) - 2\mathrm{Cov}(\bar{y}_{A\cup C} y_b, y^2_d) + \mathrm{Cov}(y^2_b, y^2_d).\end{aligned} \tag{A.3}$$

Therefore, we obtain the following:

$$\begin{aligned}& \omega + \gamma - 2\tau \\ & = 4\mathrm{Cov}(\bar{y}_{A\cup B} y_d, \bar{y}_{A\cup C} y_{d'}) + 4\mathrm{Cov}(\bar{y}_{A\cup B} y_c, \bar{y}_{A\cup C} y_b) - 8\mathrm{Cov}(\bar{y}_{A\cup B} y_d, \bar{y}_{A\cup C} y_b) \\ & = \frac{16}{n^2}\left[\mathrm{Cov}\left(y_c \sum_{b\in B} y_b, y_b \sum_{c\in C} y_c\right) - 2\mathrm{Cov}\left(y_d \sum_{b\in B} y_b, y_b \sum_{c\in C} y_c\right)\right]\end{aligned}$$

$$= \frac{16}{n^2}\left[\sum_{b'\in B}\sum_{c'\in C}\mathrm{Cov}(y_c y_{b'}, y_b y_{c'}) - 2\sum_{b'\in B}\sum_{c'\in C}\mathrm{Cov}(y_d y_{b'}, y_b y_c)\right].$$

If $b' \neq b$ and $c' \neq c$, equations $\mathrm{Cov}(y_c y_{b'}, y_b y_{c'}) = 0$ and $\mathrm{Cov}(y_d y_{b'}, y_b y_{c'}) = 0$ hold. Therefore, we obtain the following:

$$\begin{aligned}
&\omega + \gamma - 2\tau \\
&= \frac{16}{n^2}\Bigg[\sum_{c'\in C}\mathrm{Cov}(y_c y_b, y_b y_{c'}) + \sum_{b'\in B}\mathrm{Cov}(y_c y_{b'}, y_b y_c) - \mathrm{Var}(y_b y_c) \\
&\quad - 2\sum_{c'\in C}\mathrm{Cov}(y_d y_b, y_b y_{c'})\Bigg] \\
&= \frac{16}{n^2}\left[2\sum_{c'\in C}\mathrm{Cov}(y_c y_b, y_b y_{c'}) - 2\sum_{c'\in C}\mathrm{Cov}(y_d y_b, y_b y_{c'}) - \mathrm{Var}(y_b y_c)\right] \\
&= \frac{16}{n^2}(var(y_b y_c) - 2\mathrm{Cov}(y_d y_b, y_b y_c)),
\end{aligned} \tag{A.4}$$

where $\sum_{c'\in C}\mathrm{Cov}(y_c y_b, y_b y_{c'}) = \sum_{b'\in B}\mathrm{Cov}(y_c y_{b'}, y_b y_c)$ holds.

Given that $Ey_i = \mu$ and $\mathrm{Var}(y_i) = \psi^2$, $\mathrm{Var}(y_b y_c) = \psi^4 + 2\mu^2\psi^2$ and $\mathrm{Cov}(y_d y_b, y_b y_c) = \mu^2\psi^2$ can be known easily. Finally, we can obtain

$$\omega + \gamma - 2\tau = \frac{16}{n^2}\psi^4 > 0. \tag{A.5}$$

This proof supports that in the mean regression situation, the condition $\omega + \gamma > 2\tau$ holds.

A.2 Proof of $\omega + \gamma > 2\tau$ for Multivariate Linear Regression. For multivariate linear regression, we introduce an important expansion of the covariance of squared loss functions for the regression model. This expansion was developed by Markatou et al. (2005) and is expressed as follows.

Assume that our data set is $(\boldsymbol{x}_i, y_i)_{i=1}^n$ with n i.i.d. samples. y_i is the response variable, and $\boldsymbol{x}_i$ is our predictor vector consisting of p predictors, that is, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ be a coefficient vector. A multivariate regression model has the following form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{A.6}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ is the response vector, $\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$ is the design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is the noise vector, in which $\forall i, E\epsilon_i = 0, \mathrm{Var}(\epsilon_i) = \psi^2$; each pair of ϵ_i and ϵ_j in $\boldsymbol{\epsilon}$ is i.i.d. If $\boldsymbol{x}_i$ in $\mathbf{X}$ is random, the multivariate regression model (see equation A.6) is usually analyzed by conditioning on $\mathbf{X}$. Let $\mathcal{S} = (I^{(t)}, I^{(v)})$ be a partition, and $\boldsymbol{X}_{I^{(t)}}$ is a design matrix composed by $\{\boldsymbol{x}_i | i \in I^{(t)}\}$. Let $\hat{y}_{I^{(t)},i}$ be the prediction value of y_i using the training set on $\mathcal{S} = (I^{(t)}, I^{(v)})$. Thereafter, the covariance of the two squared losses with regard to the two partitions $\mathcal{S}_1$ and $\mathcal{S}_2$, as well as the two test samples of $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_{i'}, y_{i'})$, can be rewritten as

$$\begin{aligned} &\mathrm{Cov}((\hat{y}_{I_1^{(t)},i} - y_i)^2, (\hat{y}_{I_2^{(t)},i'} - y_{i'})^2 | \boldsymbol{X}) \\ &= 2\psi^4 tr\left\{ (\boldsymbol{x}_i \boldsymbol{x}_i^T) H^{-1}_{I_1^{(t)}} H_{I_1^{(t)} \cap I_2^{(t)}} H^{-1}_{I_2^{(t)}} (\boldsymbol{x}_{i'} \boldsymbol{x}_{i'}^T) H^{-1}_{I_1^{(t)}} H_{I_1^{(t)} \cap I_2^{(t)}} H^{-1}_{I_2^{(t)}} \right\}, \end{aligned} \tag{A.7}$$

where $H_{I^{(t)}} = \boldsymbol{X}^\top_{I^{(t)}} \boldsymbol{X}_{I^{(t)}}$, and $i \neq i'$.

Concerning the multivariate regression model given in equation A.6, we further assume that each sample in design matrix $\mathbf{X}$ is independently drawn from a population with mean $\boldsymbol{\nu}$ and covariance $\boldsymbol{\Sigma}$, in which $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)$ is the population mean vector and $\boldsymbol{\Sigma}$ is a $p \times p$ of the population covariance matrix.

According to equation A.7, we obtain

- $\omega = 2\psi^4 E[\boldsymbol{x}_d^\top H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} (\boldsymbol{x}_{d'} \boldsymbol{x}_{d'}^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} \boldsymbol{x}_d]$,
- $\gamma = 2\psi^4 E[\boldsymbol{x}_b^\top H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} (\boldsymbol{x}_c \boldsymbol{x}_c^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} \boldsymbol{x}_b]$,
- $\tau = 2\psi^4 E[\boldsymbol{x}_b^\top H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} (\boldsymbol{x}_{d'} \boldsymbol{x}_{d'}^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} \boldsymbol{x}_b]$
  or $\tau = 2\psi^4 E[\boldsymbol{x}_d^\top H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} (\boldsymbol{x}_c \boldsymbol{x}_c^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} \boldsymbol{x}_d]$,

in which the expectations are taken on the entire design matrix $\boldsymbol{X}$.

Therefore,

$$\begin{aligned} \omega + \gamma - 2\tau = 2\psi^4 tr\{ &E[(\boldsymbol{x}_c \boldsymbol{x}_c^\top - \boldsymbol{x}_{d'} \boldsymbol{x}_{d'}^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C} \\ &\cdot (\boldsymbol{x}_b \boldsymbol{x}_b^\top - \boldsymbol{x}_a \boldsymbol{x}_d^\top) H^{-1}_{A\cup B} H_A H^{-1}_{A\cup C}]\}. \end{aligned} \tag{A.8}$$

According to $H_A = \boldsymbol{X}_A^\top \boldsymbol{X}_A$, we obtain

$$\begin{aligned} \omega + \gamma - 2\tau = 2\psi^4 tr\{ &E[\boldsymbol{X}_A H^{-1}_{A\cup C} (\boldsymbol{x}_c \boldsymbol{x}_c^\top - \boldsymbol{x}_{d'} \boldsymbol{x}_{d'}^\top) H^{-1}_{A\cup B} \boldsymbol{X}_A^\top \\ &\cdot \boldsymbol{X}_A H^{-1}_{A\cup C} (\boldsymbol{x}_b \boldsymbol{x}_b^\top - \boldsymbol{x}_d \boldsymbol{x}_d^\top) H^{-1}_{A\cup B} \boldsymbol{X}_A^\top]\}. \end{aligned} \tag{A.9}$$

The design matrix $\boldsymbol{X}_A$ is composed of $\{\boldsymbol{x}_a | a \in A\}$ and $\boldsymbol{X}_A^\top \boldsymbol{X}_A = \sum_{a' \in A} \boldsymbol{x}_{a'} \boldsymbol{x}_{a'}^\top$. Given that the trace function of a matrix is the sum of all diagonal elements,

we obtain

$$tr\{E[\boldsymbol{X}_A H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)H_{A\cup B}^{-1}\boldsymbol{X}_A^\top \boldsymbol{X}_A H_{A\cup C}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)H_{A\cup B}^{-1}\boldsymbol{X}_A^\top]\}$$
$$= \sum_{a\in A}\sum_{a'\in A} E[\boldsymbol{x}_a^\top H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)H_{A\cup B}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_{a'}^\top H_{A\cup C}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)H_{A\cup B}^{-1}\boldsymbol{x}_a]. \tag{A.10}$$

Given that $\boldsymbol{x}_{a'}^\top H_{A\cup C}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)H_{A\cup B}^{-1}\boldsymbol{x}_a$ is a random number, we obtain

$$\boldsymbol{x}_{a'}^\top H_{A\cup C}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_a\boldsymbol{x}_d^\top)H_{A\cup B}^{-1}\boldsymbol{x}_a = \boldsymbol{x}_a^\top H_{A\cup B}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)H_{A\cup C}^{-1}\boldsymbol{x}_{a'}. \tag{A.11}$$

Thus, we obtain the following:

$$\begin{aligned}
\omega+\gamma-2\tau &= \sum_{a\in A}\sum_{a'\in A} E[\boldsymbol{x}_a^\top H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)H_{A\cup B}^{-1}\boldsymbol{x}_{a'} \\
&\quad \cdot \boldsymbol{x}_a^\top H_{A\cup B}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)H_{A\cup C}^{-1}\boldsymbol{x}_{a'}] \\
&= \sum_{a\in A}\sum_{a'\in A} E[tr\{H_{A\cup C}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)H_{A\cup B}^{-1}\boldsymbol{x}_{a'} \\
&\quad \cdot \boldsymbol{x}_a^\top H_{A\cup B}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)\}] \\
&= \sum_{a\in A}\sum_{a'\in A} tr\{E_{\boldsymbol{x}_a,\boldsymbol{x}_{a'}}[E(H_{A\cup C}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)|\boldsymbol{x}_a,\boldsymbol{x}_{a'}) \\
&\quad \cdot E(H_{A\cup B}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup B}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)|\boldsymbol{x}_a,\boldsymbol{x}_{a'})]\}.
\end{aligned} \tag{A.12}$$

For any $a, a' \in A$ by conditioning on $\boldsymbol{x}_a, \boldsymbol{x}_{a'}$, we have

$$\begin{aligned}
&E(H_{A\cup C}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup C}^{-1}(\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)|\boldsymbol{x}_a,\boldsymbol{x}_{a'})] \\
&\quad = E(H_{A\cup B}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup B}^{-1}(\boldsymbol{x}_b\boldsymbol{x}_b^\top - \boldsymbol{x}_d\boldsymbol{x}_d^\top)|\boldsymbol{x}_a,\boldsymbol{x}_{a'})].
\end{aligned} \tag{A.13}$$

Thus, we obtain the following:

$$\begin{aligned}
\omega+\gamma-2\tau &= \sum_{a\in A}\sum_{a'\in A} tr\{E_{\boldsymbol{x}_a,\boldsymbol{x}_{a'}}[E^2(H_{A\cup C}^{-1}\boldsymbol{x}_{a'}\boldsymbol{x}_a^\top H_{A\cup C}^{-1} \\
&\quad \cdot (\boldsymbol{x}_c\boldsymbol{x}_c^\top - \boldsymbol{x}_{d'}\boldsymbol{x}_{d'}^\top)|\boldsymbol{x}_a,\boldsymbol{x}_{a'})]\} \\
&> 0.
\end{aligned} \tag{A.14}$$

Therefore, the condition $\omega+\gamma > 2\tau$ holds for the multivariate regression model.

Acknowledgments

We thank the anonymous referee for helpful comments on an earlier version of this letter. This work was supported by the National Social Science Fund of China (NSSFC-16BTJ034), the National Natural Science Fund of China (NNSFC-61503228), the Natural Science Fund of Shanxi province (201601D011046), and the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

References

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

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

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

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

Fan, J., Guo, S., & Hao, N. (2012). Variance estimation using refitted cross-validation in ultra–high dimensional regression. *Journal of the Royal Statistical Society: Series B*, *74*(1), 37–65.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, *70*(5), 849–911.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Berlin: Springer.

Lichman, M. (2013). UCI Machine Learning Repository. Irvine: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.

Markatou, M., Tian, H., Biswas, S., & Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, *6*, 1127–1168.

McCarthy, P. J. (1976). The use of balanced half-sample replication in cross-validation studies. *Journal of the American Statistical Association*, *71*(355), 596–604.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281.

Nason, G. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B*, *58*(6), 463–479.

Stanišić, P., & Tomović, S. (2012). Frequent itemset mining using two-fold cross-validation model. In *Mediterranean Conference on Embedded Computing* (pp. 229–232). Piscataway, NJ: IEEE.

Tibshirani, R. J., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, *3*(2), 822–829.

Wang, Y., Li, J., & Li, Y. (2015). Measure for data partitioning in $m \times 2$ cross-validation. *Pattern Recognition Letters*, *65*, 211–217.

Wang, Y., Wang, R., Jia, H., & Li, J. (2014). Blocked 3 × 2 cross-validated t-test for comparing supervised classification learning algorithms. *Neural Computation*, *26*(1), 208–235.

Wu, C. J., & Hamada, M. S. (2011). *Experiments: Planning, analysis, and optimization*. Hoboken, NJ: Wiley.

Yang, X., Wang, Y., Wang, R., & Li, J. (2014). Variance of estimator of the prediction error based on blocked 3 × 2 cross-validation. *Chinese Journal of Applied Probability and Statistics*, *30*(4), 372–380.

Yildiz, O. T. (2013). Omnivariate rule induction using a novel pairwise statistical test. *IEEE Transactions on Knowledge and Data Engineering*, *25*(9), 2105–2118.

Received March 25, 2016; accepted October 7, 2016.