

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321222744>

Multiple predicting K -fold cross-validation for model selection

Article in *Journal of Nonparametric Statistics* · November 2017

DOI: 10.1080/10485252.2017.1404598

CITATIONS

124

READS

3,257

1 author:



Yoonsuh Jung

Korea University

31 PUBLICATIONS 438 CITATIONS

SEE PROFILE

Multiple Predicting K -fold Cross-validation for model selection

Yoonsuh Jung

Department of Statistics, Korea University

Seoul, South Korea

yoons77@korea.ac.kr

Abstract

K -fold cross-validation is widely adopted as a model selection criterion. In K -fold cross-validation, $K - 1$ folds are used for model construction and the hold-out fold is allocated to model validation. This implies model construction is more emphasized than the model validation procedure. However, some studies show that more emphasis on the validation procedure can result in improved model selection. Specifically, leave- m -out cross-validation with n samples can achieve variable-selection consistency when m/n approaches to 1. In this work, a new cross-validation method is proposed under the framework of K -fold cross-validation. The proposed method uses $K - 1$ folds of the data for model validation, while the other fold is for model construction. This provides $K - 1$ predicted values for each observation. These values are averaged to produce a final predicted value. Then, the model selection based on the averaged predicted values can reduce variation in the assessment due to the averaging. The variable-selection consistency of the suggested method is established. Its advantage over K -fold cross-validation with finite samples are examined under linear, nonlinear, and high dimensional models.

KEYWORDS: Cross-validation; K -fold Cross-validation; Model Selection; Tuning Parameter Selection

1 Introduction

Since the introduction of leave-one-out cross-validation (LOOCV) by Stone (1974), cross-validation has been one of the most popular methods for model selection. Cross-validation (CV) has been considered under various situations such as standard least squares (Shao; 1993; Zhang; 1993), correlated data (Burman et al.; 1994; Carmack et al.; 2009, 2012), and for bandwidth selection in density estimation Chow et al. (1987). An exhaustive literature review is provided by Arlot and Celisse (2010).

To reduce computational cost of LOOCV, Geisser (1975) proposed K -fold cross-validation. It partitions a data set into K nearly equal size, then $(K - 1)$ fold is used to construct a model, whereas the left-out sample is used to validate. During the iteration of this procedure for K times, each of the K folds is successively assigned as validation data. In practice, typical choice of K is between 5 and 10. Burman (1989) argued that $K < 5$ might cause an issue.

Shao (1993) revealed that LOOCV is inconsistent in model selection under a linear model, and proved that leave- m -out cross-validation is variable-selection consistent under certain conditions. One of the key conditions in Shao (1993) is that m should be of equal magnitude as sample size n asymptotically. That is, $m/n \rightarrow 1$. This implies that variable selection consistency in the linear model is maintained when most of the observations are used for validation. However, the traditional K -fold CV method is deviated from this concept since only one fold of the data (hold-out fold) is usually used for model validation. This motivates us to modify K -fold CV to improve the model selection by allocating $(K - 1)$ folds of the sample for model validation, while the other fold is used for construction. By repeating the procedure for K times, we will have $(K - 1)$ predicted values for

each observation. Notice that the usual K -fold CV yields one predicted value for each observation. As the proposed method provides $(K - 1)$ predicted values for each observation, we may perform more accurate model assessment. We call the proposed method multiple predicting cross-validation (*MPCV*).

One should consider the suggested method when one fold of the data provides reasonable fit. As economic data often contains considerable number of samples compared to the number of variables, the performance of *MPCV* can be satisfactory in these types of data. Our simulation studies under the linear model suggest that we have at least three times more samples than the number of significant variables in one-fold of data. For example, if we have five significant variables (out of a total of 10 variables) in the data set, then approximately 15 observations need to be allocated for model construction. As we do not know the truly significant variables in real data, a pilot fitting can provide rough clue for choosing K . Alternatively, conservative allocation of 30 samples (which is three times more than the total of 10 variables) in each fold would work. If this requires more samples than we have, we may use smaller K . Some heuristic rules supported by the theoretical properties are given. However, heuristic rules may not be applied beyond the linear model.

Zhang (1993) proved that K -fold CV is inconsistent in variable-selection, although computational cost is lower than LOOCV. Zhang (1993)'s finding coincides with the conclusion of Shao (1993) as the traditional K -fold CV keeps the proportion of samples in validation fold as $m/n = 1/K$. However, the proposed method maintains the ratio to be $(K - 1)/K$, and thus, can be shown to be consistent in variable-selection under mild conditions such as $K = O(\log(n))$, that is one of the distinctions compared to the K -fold CV.

Another advantage arises from the multiple predictions. As we use $(K - 1)$ folds of the data for model validation or assessment, the proposed procedure yields $(K - 1)$ predicted values. These values are averaged to produce a final predicted value which may be more accurate compared to the single predicted value by K -fold CV.

The details of the proposed methods under the linear model, and its consistent properties are described in Section 2.1 and Section 2.2, respectively.

The proposed methods can be adopted in modeling high dimensional data ($n < p$). Since high dimensional data usually requires the use of penalization methods, variable selection is often conducted by choosing one or two penalty parameter(s). In such a case, the above rule for choosing K is not needed, and more flexible choice of K is possible. We examine the various choices of K with $n = 200$ and $p = 500$. The details and the consistency of *MPCV* under the high dimensional model are discussed in Section 2.3. However, the results we illustrate here should be confined to parametric models. Yang (2007) revealed that the sample size in the evaluation data does not have to dominate under some nonparametric models.

To examine the performance of the proposed methods under various models, we primarily compare K -fold CV to *MPCV* in Section 3 with simulated data sets under linear, non-linear, and high dimensional models. The other model selection criteria such as AIC (Akaike; 1973), BIC (Schwarz; 1978), and their variants are also compared. An analysis of Philippine income data in Section 4 shows parsimonious variable selection of *MPCV*, while maintaining similar estimation accuracy to K -fold CV.

2 Multiple Predicting Cross-validation

2.1 MPCV in Linear Model

First, we consider the linear model of

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad (1)$$

where $\mathbf{Y} = (y_1, \dots, y_n)'$ is a response vector, $\mathbf{X} = (x_{ij}), i = 1, \dots, n, j = 1, \dots, p$ be the deterministic design matrix, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ is a vector of *iid* random errors with mean 0. As we split

the data set into K folds, we introduce more notations. Let \mathbf{x}_j be the j th column vector of \mathbf{X} , and $\mathbf{X}_{k,D}$ be the sub-matrix of the design matrix in the k th fold with variables in set $D \subset \{1, \dots, p\}$. Thus, there are $2^p - 1$ possible different choices of set D excluding the intercept model in model (1). Let \mathbf{Y}_k be the response vector of size n_k in the k th fold. Throughout this paper, we use subscript k to notate the sub-matrix with k th fold, and $(-k)$ for the sub-matrix without k th fold.

We assume that the data set is equally divided into K folds with $n_K = n/K$ samples in each fold. Then, the mean squared prediction error (MSPE) by K -fold CV is given as

$$\sum_{k=1}^K \|\mathbf{Y}_k - \mathbf{X}_{k,D} \hat{\beta}_{(-k),D}\|^2 / n, \quad (2)$$

where $\hat{\beta}_{(-k),D}$ is the least squares estimate of β using variables in set D excluding samples in the k th fold. After calculating (2) for all possible models (or, possible sets of D), K -fold CV chooses one model which produces minimum MSPE.

In contrast, *MPCV* constructs a model using one fold of the data, and validate it with $(K - 1)$ folds. Thus, the prediction error from *MPCV* when the k th fold is utilized for the model construction with the variables in set D is

$$PE(k, D) = \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \hat{\beta}_{k,D} \quad (3)$$

where $\mathbf{Y}_{(-k)}$ is a response vector without samples in the k th fold, $\mathbf{X}_{(-k),D}$ is a design matrix with variables in set D excluding samples in the k th fold, and $\hat{\beta}_{k,D}$ is the least squares estimate with variables in the set D using samples in the k th fold. When we calculate (3) for all k , each observation is used only once for model construction $\hat{\beta}_{k,D}$ in *MPCV*. But, we will have $(K - 1)$ predicted values of $(\hat{y}_{i,1}, \dots, \hat{y}_{i,K-1})$ except for one case when y_i is used for model construction. Note that we use each observation $(K - 1)$ times for constructing $\hat{\beta}_{(-k),D}$ in K -fold CV given in (2). Then, we define $\hat{y}_{i,D}$ to be a mean of $(\hat{y}_{i,1}, \dots, \hat{y}_{i,K-1})$, and utilize it as a final prediction value for y_i with variables in set D for $i = 1, \dots, n$. As $\hat{y}_{i,D}$ is the mean of the $(K - 1)$ predicted values of $\hat{y}_{i,k}$ s, it is the minimizer of the mean of the $(K - 1)$ squared predicted errors, $\sum_{k=1}^{K-1} (y_i - \hat{y}_{i,k})^2$. Now, the mean squared prediction error with variables D from *MPCV* is

$$\sum_{i=1}^n (y_i - \hat{y}_{i,D})^2 / n. \quad (4)$$

Then, *MPCV* chooses the model which minimizes (4). The general *MPCV* procedure is provided in Algorithm 1.

ALGORITHM 1. (Multiple Predicting Cross-validation)

1. Randomly divide the data set into K folds into (nearly) equal size.
2. Construct a model from the k th fold, then obtain prediction errors from the other samples not in the k th fold.
3. Repeat Step 2 for $k = 1, \dots, K$ until obtain $(K - 1)$ predicted errors for all observations.
4. Set the average of $(K - 1)$ predicted errors, $\hat{y}_{i,D}$, as a final predicted value of y_i for $i = 1, \dots, n$.
5. Select a model D which minimizes the mean squared prediction error given in equation (4).

When compared to K -fold CV, we note that the additional computation for *MPCV* is the averaging at Step (4) in Algorithm 1. With this minor extra computing, we may potentially gain significant

accuracy in validation. An important point is that *MPCV* has a significantly lower computational cost relative to *KCV* due to the training sets being smaller. The above algorithm can be applied not only to a model selection from linear models, but also to a broad range of model selections from more general models.

Efron and Tibshirani (1997) argued that cross-validation is nearly unbiased for the future error rate, but often highly variable. In this view, the Step 4 in Algorithm 1 can significantly lower the undesirable variation in validation step by taking the mean of the multiple predicted values. This is analogous to the bias-variance tradeoff. That is, giving up some accuracy in model construction may result in better model selection by the highly improved validation, in principle.

2.2 Theoretical Properties under Linear Model

Let $\mathbf{H}_{(-k),D} = \mathbf{X}_{(-k),D}(\mathbf{X}'_D \mathbf{X}_D)^{-1} \mathbf{X}'_{(-k),D}$, where \mathbf{X}_D is a design matrix with all samples composed of variables in D . Then, $PE(k, D)$ in (3) can be rewritten as

$$\begin{aligned} PE(k, D) &= \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D}(\mathbf{X}'_{k,D} \mathbf{X}_{k,D})^{-1} \mathbf{X}'_{k,D} \mathbf{Y}_k \\ &= \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D}(\mathbf{X}'_D \mathbf{X}_D - \mathbf{X}'_{(-k),D} \mathbf{X}_{(-k),D})^{-1} \mathbf{X}'_{k,D} \mathbf{Y}_k \end{aligned} \quad (5)$$

Note that we require any fold of the data $X_{k,D}$ to be full column rank. Thus, the value of K should be chosen considering the sample size. Utilizing Sherman-Morrison formula Sherman and Morrison (1950), which states $(A - B'B)^{-1} = A^{-1} + A^{-1}B'(I - BA^{-1}B')^{-1}BA^{-1}$, we set $A = \mathbf{X}'_D \mathbf{X}_D$ and $B = \mathbf{X}_{(-k),D}$. Then (5) becomes

$$\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \{ (\mathbf{X}'_D \mathbf{X}_D)^{-1} + (\mathbf{X}'_D \mathbf{X}_D)^{-1} \mathbf{X}'_{(-k),D} (\mathbf{I} - \mathbf{H}_{(-k),D})^{-1} \mathbf{X}_{(-k),D} (\mathbf{X}'_D \mathbf{X}_D)^{-1} \} \mathbf{X}'_{k,D} \mathbf{Y}_k,$$

where \mathbf{I} is an identity matrix of size $n - n_K$. As $\mathbf{X}'_{k,D} \mathbf{Y}_k = \mathbf{X}'_D \mathbf{Y} - (\mathbf{X}'_{(-k),D} \mathbf{Y}_{(-k)})$, we can further modify (5) to

$$\begin{aligned} PE(k, D) &= \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \hat{\beta}_D + \mathbf{H}_{(-k),D} \mathbf{Y}_{(-k)} + \mathbf{H}_{(-k),D} (\mathbf{I} - \mathbf{H}_{(-k),D})^{-1} \mathbf{H}_{(-k),D} \mathbf{Y}_{(-k)} \\ &\quad - \mathbf{H}_{(-k),D} (\mathbf{I} - \mathbf{H}_{(-k),D})^{-1} \mathbf{X}_{(-k),D} \hat{\beta}_D \\ &= \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \hat{\beta}_D + \mathbf{H}_{(-k),D} (\mathbf{I} - \mathbf{H}_{(-k),D})^{-1} (\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \hat{\beta}_D) \\ &= (\mathbf{I} - \mathbf{H}_{(-k),D})^{-1} (\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),D} \hat{\beta}_D), \end{aligned} \quad (6)$$

where $\hat{\beta}_D = (\mathbf{X}'_D \mathbf{X}_D)^{-1} \mathbf{X}'_D \mathbf{Y}$. (6) shows $PE(k, D)$ can be obtained from the estimate using full data. Since we are considering a linear model, mean squared prediction error with variables D from *MPCV* excluding the k th fold will be less than or equal to maximum of $\|PE(k, D)\|^2$. That is,

$$\sum_{i=1}^{n-n_K} (y_i - \hat{y}_{i,D})^2 / (n - n_K) \leq \max_{k=1, \dots, K} \|PE(k, D)\|^2 / (n - n_K).$$

Mean squared prediction error (of the k th fold) from K -fold CV is rewritten as

$$\frac{1}{n} \sum_{k=1}^K \|(\mathbf{I} - \mathbf{H}_{k,D})^{-1} (\mathbf{Y}_k - \mathbf{X}_{k,D} \hat{\beta}_D)\|^2. \quad (7)$$

Zhang (1993) revealed (7) converges to 0 when the selected model D contains all variables in the true model, but D need not to be equal to the true model D^* ; That is, $D \supset D^*$. Now, we notate the selected model by K -fold CV and *MPCV* as D_{KCV} and D_{MPCV} , respectively. Zhang (1993) revealed that $\lim_{n \rightarrow \infty} P(D_{KCV} \neq D^*) > 0$. This result corresponds to the findings of Shao (1993) since K -fold CV does not meet requirements of $m/n \rightarrow 1$ where m is the sample size in the validation. In *MPCV*, the proportion of samples for validation is $(K - 1)/K = 1 - 1/K$, which

approaches to 1 if K increases as n increases. We show that $MPCV$ is variable-selection consistent under the mild conditions given below.

First, let $\mathbf{H}_D = \mathbf{X}_D(\mathbf{X}'_D\mathbf{X}_D)^{-1}\mathbf{X}'_D$ be the projection matrix consist of variables in set D , and \mathcal{A} be the collection of all subsets of $\{1, \dots, n\}$ with size $n_K = n/K$. We consider the following conditions.

(C1). $\mathbf{X}'\mathbf{X} = O(n)$ and $(\mathbf{X}'\mathbf{X})^{-1} = O(1/n)$,

(C2). $\lim_{n \rightarrow \infty} \max_{i \leq n} h_{i,D} = 0$ for any D , where $h_{i,D}$ is the i th diagonal element of \mathbf{H}_D ,

(C3). $\liminf_{n \rightarrow \infty} \beta' \mathbf{X}'(\mathbf{I}_n - \mathbf{H}_D)\mathbf{X}\beta/n > 0$ when $D \not\supset D^*$, but is 0 when $D \supset D^*$.

(C4). $\lim_{n \rightarrow \infty} \max_{A \in \mathcal{A}} \|\sum_{i \in A} \mathbf{x}_i \mathbf{x}'_i / n_K - \sum_{i \in A^c} \mathbf{x}_i \mathbf{x}'_i / (n - n_K)\| = 0$, where \mathbf{x}_i is the i th row of \mathbf{X} .

(C5). $K = O(\log(n))$.

(C2) assumes that there are no extreme high-leverage points as a model with d variables yields $\sum_{i=1}^n h_{i,D} = d$, and thus an average size of $h_{i,D}$ is d/n . In this subsection, we assume d is fixed. (C3) implies the estimation error does not disappear when the selected model D misses at least one variable in the D^* . The estimation error approaches to zero when all the variables in D^* are contained in the selected model D . For understanding (C4), notice that $\mathbf{x}_i \mathbf{x}'_i$ is Fisher information matrix regarding β with the i th observation. Thus, we assume by (C4) that the average information contained in the construction samples is asymptotically equal to that in the validation samples. More details of (C4) is explained in Section 4.4 of Shao (1993). Finally, we allow the number of folds, K , increase slowly as n increases in (C5). For the practical choice of K , the nearest integer value of $\log(n)$ works well, which we suggest as a rule for choosing K . Its finite performance is illustrated by simulation studies in Section 3.

Now, we focus attention on the fact that Step 2 in Algorithm 1 is equivalent to leave- $(n - n_K)$ -out cross-validation with single split as $(n - n_K)$ samples are used for validation. It is because n_K samples are used for model construction and $(n - n_K)$ samples for validation. Then, with similar arguments in Shao (1993), we can show that the mean square of (5) is

$$\frac{1}{n - n_K} \|PE(k, D)\|^2 = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}_D \hat{\beta}_D\|^2 + \frac{n + n_K}{n_K(n - 1)} \sum_{i=1}^n h_{i,D} (y_i - \mathbf{x}'_{i,D} \hat{\beta}_D)^2 + o_p(1/n_K). \quad (8)$$

Then, the following Lemma shows specific form of (8) for the case of $D \supset D^*$.

Lemma 1. *Suppose (C1) through (C5) hold. When $D \supset D^*$, we have,*

$$\frac{1}{n - n_K} \|PE(k, D)\|^2 = \epsilon' \epsilon / n + d \sigma^2 / n_K + o_p(1/n_K), \quad (9)$$

where d is the number of variables in model D .

As we assume the fixed p , d is also fixed. However, the arguments given here can be extended to slowly increasing p such as $p = O(\log(n))$. In this case, the increasing rate of d , and the relationship between K , p , and d should be further examined. But, considering $d = O(\log(n))$, $p = O(\log(n))$, and $d < p$, we can arrive at the same variable-selection consistency in Theorem 1. Because the result in Lemma 1 does not depend on the specific fold, Lemma 1 holds for all $k = 1, \dots, K$. Thus, the result of Lemma 1 also represents the MSPE for the $MPCV$ for the case of $D \supset D^*$. Therefore, the candidate model $D(\supset D^*)$ is distinguished by the second term $d \sigma^2 / n$ in RHS of (9). On the contrary, the sample size for model construction in K -fold CV is $n - n_K$. This makes the order of the second and third term in (19) equal, thus $d \sigma^2 / n$ cannot discriminate the models in $D \supset D^*$. Therefore K -fold CV does not achieve variable selection consistency, but the selected model is a random walk. The details are in Corollary 1 of Zhang (1993). Note that $d \sigma^2 / n$ is the smallest if and only if the selected model $D = D^*$ among $D \supset D^*$.

Now, we consider the models that at least one variable in D^* is not included in the candidate model, i.e., $D \not\supset D^*$. In this under-fitted case, a similar argument to equation (3.23) in Shao (1993)

shows that

$$\frac{1}{n - n_K} \|PE(k, D)\|^2 = \epsilon' \epsilon / n + \beta' \mathbf{X}' (\mathbf{I}_n - \mathbf{H}_D) \mathbf{X} \beta / n + o_p(1), \quad (10)$$

where the second term in RHS does not converge to zero by (C3). Intuitively, this assumption is reasonable since we do not expect to attain the true regression surface without some important variables. Again, (10) is asymptotically independent of k . Therefore, it is the same for all $k = 1, \dots, K$. Thus, the MSPE value of the suggested method when $D \not\supseteq D^*$ is asymptotically equivalent to (10). Then, when combined with the conclusion from Lemma 1, we can have the following variable-selection consistency of *MPCV*.

Theorem 1. *Suppose the conditions (C1) through (C5) holds, then,*

$$\lim_{n \rightarrow \infty} \text{Prob}(D_{MPCV} = D^*) = 1.$$

2.3 MPCV in High Dimensional Modeling

High dimensional regression models often select the variables via the selection of penalty parameter. Under the model in (1), the penalized regression coefficient can be defined as a minimizer of

$$\frac{1}{n} \sum_{i=1}^n \rho(y_i - \mathbf{x}'_i \beta) + \lambda P(\beta), \quad (11)$$

where $\rho(\cdot)$ is a general loss function, and $P(\beta)$ is a penalty function. With the squared error loss, L1-type penalty yields LASSO (Tibshirani; 1996), and L2-type penalty corresponds to ridge regression (Hoerl and Kennard; 1970). Although there has been no theoretic justification of K -fold CV under high dimensional model, it has been the most commonly embedded in R packages for selecting penalty parameter such as *glmnet* for implementing Friedman et al. (2010); Simon et al. (2011), *glmplot* for Park and Hastie (2007), *genlasso* for Tibshirani and Taylor (2011), *lqa* for Zou (2006), *elasticnet* for Zou and Hastie (2005), *ncvreg* for Fan and Li (2001), and *quantreg* for Wu and Liu (2009) among many others. Thus, the proposed methods could be practically useful.

The Algorithm 1 in Section 2.1 can be applied in selecting the penalty parameter with high dimensional data, where we select an ‘optimal’ parameter value at the step (5) in Algorithm 1. Specifically, *MPCV* selects the penalty parameter λ such that

$$\hat{\lambda}_{MPCV} = \arg \min_{\lambda \in [0, \lambda_{max}]} \frac{1}{n} \sum_{k=1}^K \sum_{i \in (k)} \rho(y_i - \mathbf{x}'_{i,\lambda} \sum_{\kappa \in \{1, \dots, K\} \setminus \{k\}} \hat{\beta}_{\kappa, \lambda} / (K - 1)), \quad (12)$$

where (k) is the set of samples in the k th fold, $\hat{\beta}_{\kappa, \lambda}$ is the minimizer of (11) with samples in κ th fold, and $\mathbf{x}_{i,\lambda}$ is the i th observation containing variables corresponding to the λ given. The selected parameter by K -fold CV is defined as

$$\hat{\lambda}_{KCV} = \arg \min_{\lambda \in [0, \lambda_{max}]} \frac{1}{n} \sum_{k=1}^K \sum_{i \in (k)} \rho(y_i - \mathbf{x}'_{i,\lambda} \hat{\beta}_{(-k), \lambda}), \quad (13)$$

where (k) is the set of samples in the k th fold. Again, the distinctions between the proposed method and K -fold CV are the additional averaging of $(K - 1)$ predicted values obtained for each observation, and assigning $(K - 1)$ folds of the data for validation to improve the estimation of the prediction or validation errors.

As high dimensional data ($n < p$) becomes popular, many penalized regression models have been proposed. However, current literature on the cross-validation for selecting the penalty parameter under high dimension is somewhat under-developed. Another area of model selection methods with high dimensional data is the *BIC*-type of criteria. Some examples include the suggestions by Chen and Chen (2008) and Chen and Chen (2012) under linear and generalized linear model, respectively, and by Pan and Shen (2007) under model-based clustering. The method by Wang et al. (2009) allows p to increase, but requires $n > p$ for variable-selection consistency. In fact, Chen and Chen (2008) and Chen and Chen (2012) also require the selected model to be well-posed, i.e. $p < n$ although p increases with n .

The model given in (11) is quite general including nonlinear model and M-estimator. To provide theoretic justification of *MPCV* under a specific model, we consider the linear LASSO model. Then, $\tilde{\beta}_\lambda$ minimizes

$$\frac{1}{2n} \sum_{i=1}^n (y_i - x_i' \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (14)$$

The consistency of LASSO estimates has been established under the low dimension with fixed p by Knight and Fu (2000) and with increasing p by Zhao and Yu (2006). Zhang and Huang (2008) showed the consistency of LASSO under $n < p$, but requires the number of important variables is less than n . This is not a strong restriction, rather a reasonable assumption because relatively small number of variables show large signal and most of the others are ignorable in high dimensional data such as microarray and mass spectrum data. Similarly, we confine the selected number of variables (d) and the true number of variables (d^*) are smaller than n , although we consider $n < p$. In this paper, we focus on establishing the consistency of the proposed *MPCV* when LASSO is used as a modeling procedure.

First, let $\mathbf{H}_\lambda = \mathbf{X}_\lambda (\mathbf{X}_\lambda' \mathbf{X}_\lambda)^{-1} \mathbf{X}_\lambda'$ be the projection matrix where \mathbf{X}_λ contains d variables under given λ . It should be noted that \mathbf{X}_λ is defined as long as $d < n$, thus it is limited to certain range of λ . Then, the conditions of (C1) through (C5) in Section 2.2 are adapted where \mathbf{X} , \mathbf{H}_D , and \mathbf{x}_i are replaced by \mathbf{X}_λ , \mathbf{H}_λ , and $\mathbf{x}_{i,\lambda}$, respectively. We call them (D1), (D2), (D3), (D4), and (D5), respectively. We consider additional condition on λ for the LASSO estimate to be a consistent estimator.

(D6). $\lambda = o(1/(K\sqrt{n}))$.

(D7). $\text{rank}(\mathbf{X}_{k,\lambda}) < n_K$, where $\mathbf{X}_{k,\lambda}$ is a design matrix from the k th fold.

Then, under the LASSO estimate which minimizes (14), *MPCV* is variable-selection consistent.

Theorem 2. Suppose the conditions (D1) through (D7) hold, then, under the LASSO estimate,

$$\lim_{n \rightarrow \infty} \text{Prob}(D_{\text{MPCV}} = D^*) = 1$$

To remove the bias caused by the shrinkage in LASSO estimate, Yu and Feng (2013) suggested modified cross-validation criterion defined as

$$\frac{1}{m} \|\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),\lambda} \tilde{\beta}_{k,\lambda}\|^2 - \frac{\lambda^2(n-m)^2}{(m)} \mathbf{M}_{k,\lambda}' \mathbf{M}_{k,\lambda}, \quad (15)$$

where $\mathbf{M}_{k,\lambda} = \mathbf{X}_{k,\lambda} (\mathbf{X}_{(-k),\lambda}' \mathbf{X}_{(-k),\lambda})^{-1} (\text{sgn}(\tilde{\beta}_{k,\lambda}))$ with LASSO estimate of $\tilde{\beta}_{k,\lambda}$. The second term in (15) removes the systematic bias by the shrinkage. Although we use the first term only as a loss function, under the assumed conditions here, the second term tends to zero as the sample size increases. Yu and Feng (2013) argued the modified cross-validation procedure is consistent when $(n-m)/n \rightarrow 0$ and $(n-m) \rightarrow \infty$. This holds for our proposed algorithms, while $(n-m)/n = (K-1)/K > 0$ in K -fold CV. Zhang and Yang (2015) revealed the sample size in the validation set should be dominating under the high dimensional model for the best procedure to be selected. This is in line with the result in Theorem 2, although our result implies variable-selection consistency.

3 Simulations

In this section, we mainly compare K -fold CV with $MPCV$ under the linear model, high-dimensional model and non-linear model with simulated data sets and discuss advantages and/or disadvantages of the proposed methods.

3.1 Linear Model

Under the model (1), \mathbf{X} is generated from multivariate normal distribution with mean 0 and correlation matrix Σ whose element in $\{i, j\}$ position is $\rho^{|i-j|}$. The errors are *iid* and follow standard normal distribution. We set $p=8$, $n=200$, and $\rho=0.2$, and 0.5. Thus, there are $2^8-1=255$ candidate models. Now, for the true regression coefficients β , we investigate eight scenarios from the sparsest case to the densest case. That is, $\beta_1 = (1, 0, 0, 0, 0, 0, 0, 0)$, $\beta_2 = (1, 1, 0, 0, 0, 0, 0, 0)$, $\beta_3 = (1, 1, 1, 0, 0, 0, 0, 0)$, $\beta_4 = (1, 1, 1, 1, 0, 0, 0, 0)$, $\beta_5 = (1, 1, 1, 1, 1, 0, 0, 0)$, $\beta_6 = (1, 1, 1, 1, 1, 1, 0, 0)$, $\beta_7 = (1, 1, 1, 1, 1, 1, 1, 0)$, and $\beta_8 = (1, 1, 1, 1, 1, 1, 1, 1)$. Following the above description, we generate 300 data sets for each scenario. For the proper choice of K , as mentioned in Section 1, it is desirable to have at least $n_K > 3d^*$ where d^* is the number of non-zero regression coefficients. Although we do not know d^* in practice, a rough estimate of d^* will suffice. As our theoretic condition requires $K = c \log(n)$ with $1 \leq c \leq 2$, we try $K=5$ or 10 since $\log(200) \approx 5.3$. The likelihood based selection methods such as BIC (Schwarz; 1978), AIC (Akaike; 1973), and corrected AIC ($AICc$) by Cavanaugh (1997) are compared, too.

To gauge the performance of the methods, we use two criteria of estimation accuracy and identification accuracy. For identification accuracy, we check the number of truly positive and the number of false positive variables as we know the truth in the simulations. For the estimation accuracy, mean squared error (MSE) from the r th simulated data set is defined as

$$(\hat{\beta}^r - \beta)' \Sigma (\hat{\beta}^r - \beta) + (\hat{\beta}_0^r)^2, \quad (16)$$

where $\hat{\beta}_0^r$ is the estimated intercept where the true intercept is zero. To get $\hat{\beta}^r$ and $\hat{\beta}_0^r$, we first choose the optimal model by the described criteria such as KCV and $MPCV$, where prediction error is gauged. Once the optimal model with minimum prediction error is selected, then, we fit the optimal model using all the samples to get $\hat{\beta}^r$ and $\hat{\beta}_0^r$, and calculate (16). The mean of MSE values from 300 simulated data sets are summarized in Table 1. In terms of estimation, $MPCV$ outperforms all the other methods in most of the scenarios. The MSE reduction achieved by $MPCV$ is the largest in β_1 and the magnitude of reduction gradually decreases as we move to dense scenario. In β_8 , there is no MSE reduction. All the considered methods choose the correct model in all 300 data sets with β_8 .

The huge reduction in MSE with β_1 (about 59% reduction) stems from the improved identification of $MPCV$ over K -fold CV. Interestingly, the MSE from BIC is consistently lower than that by AIC and KCV , where AIC and KCV show similar performance.

The number of correct identification in eight cases are given in Table 2. K -fold CV detects the true model only 90 times out of 300 data sets, whereas $MPCV$ detects 281 times. BIC detects true model more than K -fold CV in most of the cases. As the scenario changes from sparse to dense, K -fold CV identifies the correct model more accurately.

To measure the identification accuracy of both the methods, we compare the number of false positive (FP) and false negative (FN) from each method. Specifically, whenever the correct model is not selected, we count the number of FP variables in the incorrectly selected models. The results are summarized in Table 2. Regardless of the scenarios, $MPCV$ selects about 1 FP variable on average, whereas FP from K -fold CV is considerably higher than 1 in most cases. When FN is examined in a similar way, none are detected from both the K -fold CV and $MPCV$, thus omitted. As the results from β_8 are all the same, it is not recorded in Table 2.

Table 1: Mean of MSE and its standard error (in parenthesis) for AIC , AIC_c , BIC , K -fold CV (KCV), and $MPCV$ from 300 simulated data set. All values are multiplied by 10^3 . $K=5$ is used with the mark *.

ϱ		β_1	β_2	β_3	β_4
0.2	AIC	32.48 (1.43)	34.71 (1.43)	37.09 (1.42)	38.93 (1.41)
	AIC_c	32.10 (1.41)	34.01 (1.41)	36.59 (1.41)	38.43 (1.39)
	BIC	16.42 (1.05)	20.79 (1.11)	25.99 (1.16)	30.28 (1.19)
	KCV	31.76 (1.46)	34.30 (1.43)	36.75 (1.42)	39.18 (1.40)
	$MPCV$	11.28 (0.67)	17.28 (0.78)	23.16 (0.92)	27.64 (0.99)
0.5	AIC	32.15 (1.46)	33.76 (1.46)	36.08 (1.45)	38.18 (1.42)
	AIC_c	32.57 (1.44)	33.07 (1.45)	35.45 (1.44)	37.54 (1.42)
	BIC	16.79 (1.16)	20.43 (1.19)	24.31 (1.21)	28.46 (1.23)
	KCV	31.76 (1.46)	33.31 (1.43)	36.27 (1.45)	38.02 (1.43)
	$MPCV$	13.00 (0.84)	17.02 (0.85)	21.12 (0.91)	26.66 (1.08)
ϱ		β_5	β_6	β_7	β_8
0.2	AIC	41.68 (1.42)	43.73 (1.41)	45.44 (1.40)	47.51 (1.39)
	AIC_c	41.22 (1.42)	43.51 (1.41)	45.20 (1.39)	47.51 (1.39)
	BIC	35.03 (1.25)	38.51 (1.26)	43.32 (1.32)	47.51 (1.39)
	KCV	41.78 (1.40)	43.96 (1.39)	45.78 (1.39)	47.51 (1.39)
	$MPCV$	32.94 (1.11)	37.19 (1.18)	42.71 (1.32)	47.51 (1.39)
0.5	AIC	40.60 (1.43)	42.57 (1.42)	45.12 (1.40)	47.51 (1.39)
	AIC_c	40.34 (1.43)	42.45 (1.42)	44.98 (1.40)	47.51 (1.39)
	BIC	33.07 (1.25)	38.35 (1.33)	43.21 (1.37)	47.51 (1.39)
	KCV	40.47 (1.45)	42.72 (1.44)	45.25 (1.42)	47.51 (1.39)
	$MPCV$	31.70 (1.19)	37.69 (1.31)	43.60*(0.91)	47.51* (1.39)

Table 2: Number of correctly selected model (O) among 300 simulated data sets. Mean number of false positive variables (+) and their standard error (in the parentheses) among incorrectly selected models are given. $K=5$ is used with the mark *.

$\varrho = 0.2$	β_1	β_2	β_3	β_4	β_5	β_6	β_7
<i>AIC</i> (O)	97	112	128	159	178	204	255
<i>AICc</i> (O)	98	118	137	166	188	210	261
<i>BIC</i> (O)	251	259	266	270	275	285	292
<i>KCV</i> (O)	77	104	123	139	159	189	240
<i>MPCV</i> (O)	285	281	283	288	287	294	281*
<i>AIC</i> (+)	1.75 (.06)	1.60 (.06)	1.42 (.05)	1.36 (.05)	1.24 (.04)	1.14 (.04)	1.00 (.00)
<i>AICc</i> (+)	1.75 (.06)	1.60 (.06)	1.44 (.05)	1.37 (.05)	1.26 (.04)	1.14 (.04)	1.00 (.00)
<i>BIC</i> (+)	2.24 (.14)	2.12 (.14)	1.85 (.15)	1.67 (.13)	1.44 (.10)	1.27 (.12)	1.00 (.00)
<i>KCV</i> (+)	1.78 (.06)	1.67 (.06)	1.47 (.05)	1.38 (.05)	1.25 (.04)	1.14 (.03)	1.00 (.00)
<i>MPCV</i> (+)	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00* (.00)
$\varrho = 0.5$	β_1	β_2	β_3	β_4	β_5	β_6	β_7
<i>AIC</i> (O)	100	117	136	154	178	220	259
<i>AICc</i> (O)	106	127	145	167	184	224	263
<i>BIC</i> (O)	258	264	270	280	285	288	295
<i>KCV</i> (O)	90	108	117	141	165	208	249
<i>MPCV</i> (O)	281	277	283	290	286	290	277*
<i>AIC</i> (+)	1.68 (.06)	1.60 (.06)	1.51 (.05)	1.36 (.05)	1.25 (.04)	1.06 (.03)	1.00 (.00)
<i>AICc</i> (+)	1.69 (.06)	1.63 (.06)	1.53 (.06)	1.38 (.05)	1.27 (.04)	1.05 (.03)	1.00 (.00)
<i>BIC</i> (+)	1.90 (.15)	1.78 (.15)	1.70 (.16)	1.65 (.18)	1.20 (.11)	1.00 (.00)	1.00 (.00)
<i>KCV</i> (+)	1.81 (.06)	1.65 (.05)	1.56 (.05)	1.38 (.05)	1.25 (.04)	1.10 (.03)	1.00 (.00)
<i>MPCV</i> (+)	1.05 (.05)	1.04 (.04)	1.00 (.00)	1.00 (.00)	1.07 (.07)	1.00 (.00)	1.00* (.00)

From Table 1 and Table 2, it is clearly shown that the variable selection by *MPCV* is superior under the linear model partly due to the increased size of validation sample and the averaging effect from the multiple predicted values.

3.2 High Dimensional Model

To gauge the performance of *MPCV* under high dimensional models, we utilize the simulation settings in Section 3.1. We increased the number of variables from 8 to 500, while the other conditions remain unchanged. We follow the general concept of sparseness in Zhang and Huang (2008), which states exact selection for all $\beta_j \neq 0$ is unattainable or undesirable when many of β_j s are close to zero, but not exactly equal to zero. This is what is typically found in high-dimensional data, and it is difficult to reason that, say, 50 variables (out of 500) are truly large coefficients, while all the others are exactly zero. And it is more realistic to consider the scenario in which many of the coefficients are non-zero, but only small portion of them are moderate to large. Thus, we set half of the regression coefficients β_0 to zero, and generate the other 250 values from $N(0, 0.5^2)$. By the considered setting, about 50 coefficients (out of 500) show signal to noise ratio (β/σ) more than 0.5. Thus, we aim to capture the top 50 estimates, and discard the minor coefficients.

We fit LASSO (Tibshirani; 1996) and ridge regression (Hoerl and Kennard; 1970) to the simulated data sets, and measure the estimation accuracy by MSE in (16). For gauging the identification accuracy, first, we selected two sets of 50 variables with the largest absolute estimates of β_0 by K -fold CV and *MPCV*. From the each set of 50 estimates, we count the number of truly significant variables which overlap with the top 50 largest values of $|\beta_0|$. In addition, false positive detection was examined by the number of variables among the selected top 50 variables, which are truly zero. For the penalty parameter λ in (11), 200 equidistant values in $[0.001, 0.25]$ and $[0.05, 2.5]$ are used

Table 3: Mean of MSE , number of truly significant variables (TS), and number of false positive (FP) for K -fold CV (KCV) and our method ($MPCV$) from 500 simulated data set. Standard errors are in the parentheses. $EBIC(0)$, $EBIC(.5)$, and $EBIC(1)$ are $EBIC$ with $\gamma=0, 0.5$, and 1, respectively.

$K = 5$	LASSO			ridge		
$\rho = 0.2$	MSE	TS	FP	MSE	TS	FP
$EBIC(0)$	47.97 (0.20)	20.76 (0.15)	10.87 (0.14)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
$EBIC(.5)$	46.78 (0.20)	22.02 (0.15)	9.92 (0.13)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
$EBIC(1)$	47.49 (0.18)	22.17 (0.15)	9.63 (0.13)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
KCV	46.97 (0.20)	21.94 (0.16)	9.98 (0.13)	42.96 (0.11)	22.96 (0.12)	8.94 (0.12)
$MPCV$	46.33 (0.20)	22.01 (0.15)	9.84 (0.13)	42.94 (0.11)	22.91 (0.12)	9.11 (0.12)
$\rho = 0.5$	MSE	TS	FP	MSE	TS	FP
$EBIC(0)$	39.37 (0.20)	21.01 (0.14)	10.46 (0.13)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
$EBIC(.5)$	40.34 (0.19)	21.82 (0.14)	10.05 (0.13)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
$EBIC(1)$	41.50 (0.17)	21.47 (0.14)	10.42 (0.12)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
KCV	39.34 (0.20)	21.73 (0.14)	10.09 (0.12)	35.33 (0.12)	23.56 (0.13)	9.14 (0.12)
$MPCV$	38.61 (0.19)	21.97 (0.14)	9.79 (0.13)	35.25 (0.12)	22.59 (0.13)	9.01 (0.12)
$K = 10$	LASSO			ridge		
$\rho = 0.2$	MSE	TS	FP	MSE	TS	FP
$EBIC(0)$	47.96 (0.20)	20.76 (0.15)	10.87 (0.14)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
$EBIC(.5)$	46.78 (0.20)	22.02 (0.15)	9.92 (0.13)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
$EBIC(1)$	47.49 (0.18)	22.17 (0.15)	9.93 (0.13)	47.53 (0.12)	19.66 (0.13)	11.35 (0.13)
KCV	47.11 (0.21)	21.71 (0.15)	10.02 (0.13)	42.97 (0.11)	22.92 (0.13)	8.97 (0.12)
$MPCV$	46.29 (0.20)	22.01 (0.15)	9.88 (0.13)	42.92 (0.12)	22.85 (0.13)	9.06 (0.12)
$\rho = 0.5$	MSE	TS	FP	MSE	TS	FP
$EBIC(0)$	39.37 (0.20)	21.00 (0.14)	10.46 (0.13)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
$EBIC(.5)$	40.34 (0.19)	21.82 (0.14)	10.05 (0.13)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
$EBIC(1)$	41.50 (0.17)	21.47 (0.14)	10.42 (0.12)	41.28 (0.14)	20.00 (0.13)	11.55 (0.13)
KCV	39.28 (0.20)	21.69 (0.14)	10.09 (0.13)	35.31 (0.12)	23.58 (0.12)	9.08 (0.12)
$MPCV$	38.72 (0.18)	21.98 (0.14)	9.82 (0.12)	35.25 (0.12)	23.49 (0.13)	9.15 (0.12)

for LASSO and ridge, respectively.

In addition, we consider the extended BIC ($EBIC$) by Chen and Chen (2012), which is designed for the case of $n < p$. For the tuning parameter γ in $EBIC$, we consider $\gamma=0, 0.5$, and 1 as suggested by the authors where $\gamma = 0$ is equivalent to traditional BIC .

The number of correctly identified variables among the selected 50 variables and MSE values from 500 simulated data sets are summarized in Table 3. The result from LASSO shows $MPCV$ produced lower MSE, higher number of truly significant (TS) variables among 50 selected variables, and lower false positive (FP) variables when compared to those from K -fold CV and $EBIC$. To further investigate the source of improvement, the estimated penalty parameter from both methods are plotted in Figure 1. For both LASSO and ridge, the variation in $\hat{\lambda}$ is considerably reduced by $MPCV$ compared to K -fold CV partly due to the averaging of multiple predicted values. However, this does not lead to considerable improvement in estimation and identification for the case of ridge regression.

3.3 Non-linear Model

Since cross-validation is commonly used in non-linear models, we compared K -fold CV and $MPCV$ under smoothing spline models where the selection of smoothing parameter is of interest. In

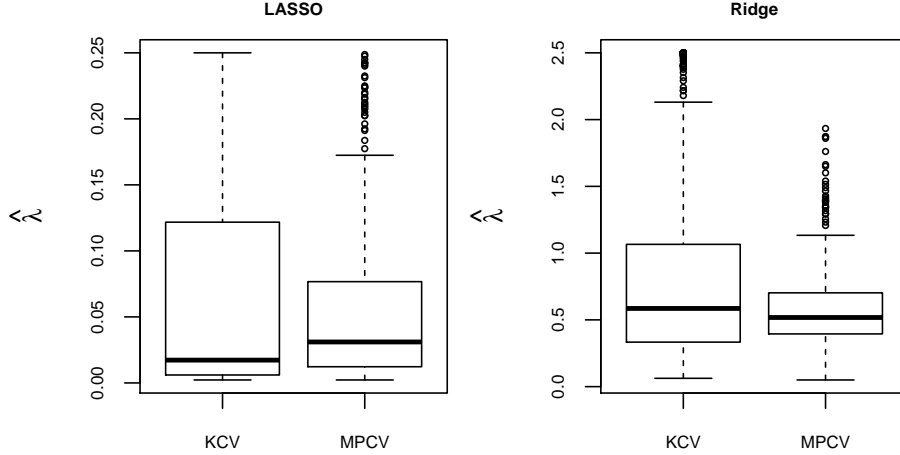


Figure 1: Estimated value of λ by K -fold CV and $MPCV$ from 300 simulated data sets.

Table 4: Mean of MSE for K -fold CV (KCV) and our method ($MPCV$) from 300 simulated data set. $N(0, 1)$ and t -distribution with 10 degrees of freedom $t(10)$ are used for the error distributions. Standard errors are in the parentheses. All numbers are multiplied by 10^3 .

n	K	$N(0, 1)$		$t(10)$	
		KCV	$MPCV$	KCV	$MPCV$
200	5	56.74 (2.70)	53.36 (1.56)	63.55 (3.33)	60.52 (1.69)
400	6	24.90 (1.11)	22.52 (0.69)	30.46 (1.57)	26.19 (0.87)
800	7	13.39 (0.67)	13.12 (0.37)	16.42 (0.75)	16.46 (0.51)

smoothing spline models, we seek a minimizer of the objective function,

$$\frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2 + \lambda \int (h''(x))^2 dx, \quad (17)$$

where $h''(x) = d^2h/dx^2$. The model considered is $y_i = f(x_i) + \epsilon_i = 1 + 3 \sin(2\pi x - \pi) + \epsilon_i$, $i = 1, \dots, n$. Standard normal distribution and t -distribution with 10 degrees of freedom are used for *iid* ϵ_i s with $n=200, 400$, and 800 . Using the suggested rule for K in condition (C5), the corresponding choice of K will be 5, 6, and 7, respectively. ($\log(200) \approx 5.3$, $\log(400) \approx 6.0$, and $\log(800) \approx 6.7$.) Mean squared error obtained by $MSE = \sum_{i=1}^n \{f(x_i) - \hat{y}_i\}^2 / n$ is reported in Table 4.

When $n = 200$, the MSE values from $MPCV$ is reduced by around 5%, and standard errors are much less compared to K -fold CV. The difference between the two methods disappears as we increase the sample size. As we conjecture from the smaller standard errors from $MPCV$, the variation in the 300 estimated λ s is smaller for $MPCV$ when compared to K -fold CV. The distribution of estimated λ is very similar to Figure 1, thus omitted.

Among the 300 simulated data sets, we further choose the relatively best performance of K -fold CV, which provides minimum value of MSE_{KCV}/MSE_{MPCV} , where the subscripts stand for the methods. Similarly, the data set with maximum value of MSE_{KCV}/MSE_{MPCV} was selected and fitted values from both methods were drawn to assess the relatively best performance of $MPCV$. The left panel of Figure 2 shows the former, while the right panel shows the latter. The black line is the true model, while blue line is a fitted model by KCV and dashed-and-dotted line is by

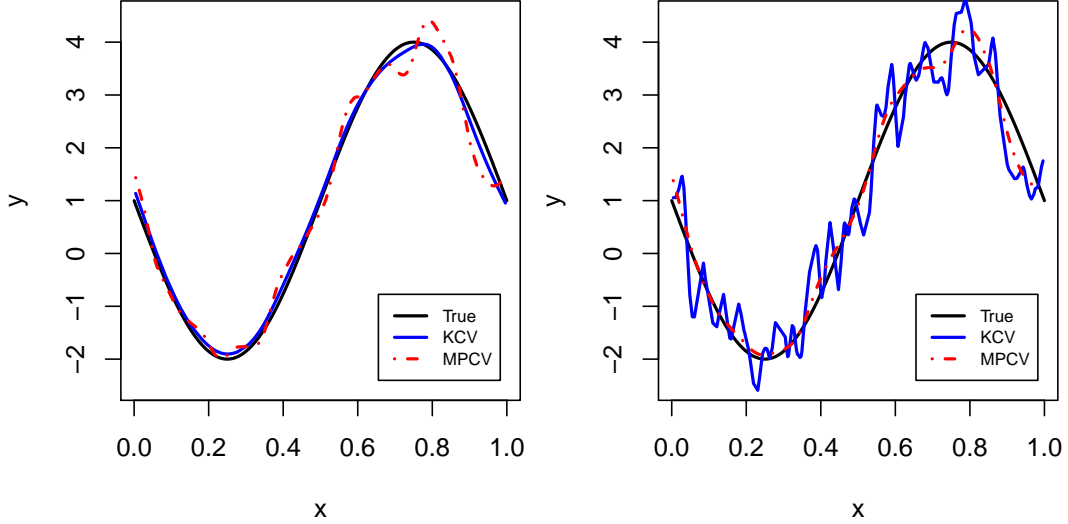


Figure 2: Relatively best performance of K -fold CV (left panel) and relatively best performance of $MPCV$ (right panel) among 300 simulated data sets with $K = 5$ and normal errors.

$MPCV$. We clearly see that KCV can select severely over-fitted model, while the worst case of $MPCV$ shows slight over-fits.

4 Real Data Analyses

4.1 Philippines' Income Data

Survey data from Philippines' National Statistics Office from 1997 comprises 632 responses with five variables of total income of households in Philippine peso (*income*), gender of household head (*gender*), family size (*size*), a factor with levels of "rural" and "urban" (*urbanity*), and a factor with four provinces of "Ilocos Norte" (*IN*), "Ilocos Sur" (*IS*), "La Union" (*LU*), and "Pangasinan". We call the last factor *province*. Taking *income* as a response variable, we fit multiple regression models to find significant variables. As there are four explanatory variables, there are 16 candidate models from the intercept model to a full model. Since *province* is a factor with four levels, we make three indicator variables for *IN*, *IS*, and *LU* to compare with "Pangasinan", which is a baseline province. When a full model is fitted, we have,

$$E(\text{income}) = 33223 + 3990\text{gender} + 8260\text{size} + 50387\text{urbanity} + 33177I(\text{IN}) + 29481I(\text{IS}) + 13513I(\text{LU}).$$

As the p-value for *gender* is 0.71, we regard it as not significant variable, while the other variables are significant. In detail, the p-values for *size* and *urbanity* are all less than 10^{-4} . For the indicator variables, *IN* and *IS* show significantly higher income than "Pangasinan" with p-value of 0.0188 and 0.0331. Since the true model is unknown, treating (or, assuming) both *size* and *urbanity* as significant variables and *province* as marginally significant, we compare the performance of K -fold CV and $MPCV$. In detail, after randomly splitting the data into K roughly equal part, we apply K -fold CV to select the 'best' model, which produces minimum MSPE. Then, we also apply $MPCV$ to the same data set to choose another 'best' model. To reduce the variation from random split, we repeat this procedure for 500 times. $K=5$ and 10 are used.

Again, we compared K -fold CV and $MPCV$ under two criteria of estimation and identification accuracy as we did in Section 3.1. However, the estimation accuracy for K -fold CV and $MPCV$ is similar to each other (less than 0.1% difference in MSE) at both K values. Although there is no difference in terms of estimation accuracy, we observe a significant difference in the selected models

Table 5: Number of selected model among 500 different splits of data.

selected model	$K=5$		$K=10$	
	KCV	$MPCV$	KCV	$MPCV$
<i>size, urbanity</i>	90	203	93	211
<i>gender, size, urbanity</i>	10	3	0	0
<i>size, urbanity, province</i>	378	293	407	289
<i>gender, size, urbanity, province</i>	22	1	0	0

Table 6: Mean of prediction error rates (and its standard error in the parentheses) from 100 different splits of test data.

Method	
KCV	0.055 (0.006)
$MPCV$	0 (0)

as illustrated in Table 5.

First, we focus on the variable *gender* which is not an important variable from the full model. When $K=5$, $MPCV$ chooses *gender* only three times or once, but K -fold CV selects 10 and 22 times, respectively. When $K=10$, both methods do not select the models with *gender*. Second, the model with two highly significant variables (*size* and *urbanity*) are picked for 203 and 211 times by $MPCV$ whereas K -fold CV picks only 90 and 93 times for $K=5$ and 10, respectively. As the MSE of K -fold CV and $MPCV$ are very close, $MPCV$ selects more parsimonious models without losing estimation accuracy.

4.2 Leukemia Microarray Data

Leukemia data is a gene expression data from Affymetrix oligonucleotide arrays analyzed by ?. The data set is available in R package `spikeslab`, which has $p = 3571$ and $n = 72$. The response variable is composed of two types of cancer; acute myeloid (AML) and acute lymphoblastic leukemias (ALL). Since the response variable is binary and $n < p$, we fit the data via the penalized logistic regression model with LASSO penalty. First, we randomly split the data into two parts. The first part contains 54 training samples, and the other 18 samples are used for the test data. We choose the penalty parameter λ from the training data by employing KCV and $MPCV$. We use $K = 3$ number of folds for KCV and $MPCV$. After selecting two λ values ($\hat{\lambda}_{KCV}$, and $\hat{\lambda}_{MPCV}$) by KCV and $MPCV$, respectively, we fit the LASSO penalized logistic regression model to the training data. Finally, we predict the 18 values of response variable in the test data and calculate the misclassification rate. To reduce the variation arisen from the random split, we repeat the described procedure for 100 times, and the mean prediction rate is presented in Table 6. $MPCV$ certainly shows better performance over that of KCV . Due to the limitation of sample size, we cannot produce the results for $K = 4$, which is, in fact, suggested by the proposed rule for choosing $K = \log(n)$.

5 Discussion

The traditional K -fold cross-validation can be improved by the suggested multiple predicting cross-validation by reducing the variation in the validation error. In turn, this elevates the variation in the model construction. However, as observed in this paper, considerable reduction in the validation

error often overcomes the increased uncertainty in the model construction. It is a consistent result with the findings of Shao (1993). The balance between two errors, construction and validation, can be tuned by the number of folds, K . We provide some guidance to select an appropriate value of K . However, it is not easy to provide a universal rule for the choice of K , since it depends on the sample size, number of parameters, structure of data, and so on. A naive rule is to choose K such that $K \approx \log(n)$ and $n/K > 3d$. The second condition means that we need certain amount of data for the model construction to capture the complexity of the data structure reasonably well. Otherwise, lack of precision in model construction will not be overcome by more precise model validation. From the similar perspective, we are likely to get the most benefit from *MPCV* when the sample size is relatively large compared to the number of parameters or to the complexity of data structure.

Acknowledgement

Jung's work was partially supported by National Research Foundation of Korea Grant NRF-2017R1C1B5017431 and by Korea University Grant (K1705711).

Appendix

Proof of Lemma 1. When $D \supset D^*$, the first term in the right-hand side of (8) can be modified as

$$\begin{aligned} \frac{(\mathbf{Y} - \mathbf{X}_D \hat{\beta}_D)'(\mathbf{Y} - \mathbf{X}_D \hat{\beta}_D)}{n} &= \frac{(\mathbf{Y} - \mathbf{H}_D \mathbf{Y})'(\mathbf{Y} - \mathbf{H}_D \mathbf{Y})}{n} \\ &= \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_D)\mathbf{Y}}{n} = \frac{(\mathbf{X}\beta + \epsilon)'(\mathbf{I} - \mathbf{H}_D)(\mathbf{X}\beta + \epsilon)}{n} \\ &= \frac{(\mathbf{X}\beta)'(\mathbf{I} - \mathbf{H}_D)(\mathbf{X}\beta)}{n} + \frac{2\epsilon'(\mathbf{I} - \mathbf{H}_D)(\mathbf{X}\beta)}{n} + \frac{\epsilon'(\mathbf{I} - \mathbf{H}_D)\epsilon}{n} \end{aligned} \quad (18)$$

As the selected model includes the true model ($D \supset D^*$), the first term in (18) goes to 0 as $n \rightarrow \infty$ by condition (C3). From the fact that $\text{plim}_{n \rightarrow \infty} \frac{\epsilon' \mathbf{X}_D}{n} = 0$, the second term in (18) is $o_p(1)$. Thus, (18) is $\epsilon'\epsilon/n - \epsilon'\mathbf{H}_D\epsilon/n + o_p(1)$. Then, (8) is equivalent to

$$\epsilon'\epsilon/n - \epsilon'\mathbf{H}_D\epsilon/n + \frac{n + n_K}{n_K(n-1)} \sum_{i=1}^n h_{i,D} (y_i - \mathbf{x}'_{i,D} \hat{\beta}_D)^2 + o_p(1/n_K), \quad (19)$$

By (18), $(y_i - \mathbf{x}'_{i,D} \hat{\beta}_D)^2$ is asymptotically equivalent to $\epsilon'\epsilon - \epsilon'\mathbf{H}_D\epsilon$. Since we know $\sum_{i=1}^n h_{i,D} = d$, or $h_{i,D} = O(1/n)$, (19) is re-expressed as

$$\epsilon'\epsilon/n - \epsilon'\mathbf{H}_D\epsilon/n + \frac{n + n_K}{n_K(n-1)} (d\sigma^2 + o_p(1)) + o_p(1/n_K). \quad (20)$$

Now, $\epsilon'\mathbf{H}_D\epsilon/n$ is dominated by $\frac{n+n_K}{n_K(n-1)}(d\sigma^2)$ because $E(\epsilon'\mathbf{H}_D\epsilon) = d\sigma^2$ and $n \cdot \frac{n+n_K}{n_K(n-1)}(d\sigma^2)$ is $O(n/n_K) = O(K) = O(\log(n))$ by condition (C5), which completes the proof. \square

Proof of Theorem 1. Based on the previous arguments, proof is trivial. When $D \subset D^*$, $D = D^*$ yields the minimum MSPE by Lemma 1, whereas $D \supsetneq D^*$ will not be selected as MSPE is not ignorable. Therefore, we select the true model D^* by *MPCV* with probability tend to 1. \square

Proof of Theorem 2. First, we consider the prediction error when the k th fold of the data is used for model construction. $PE(k, \lambda) = \mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),\lambda} \hat{\beta}_{k,\lambda}$, where $\hat{\beta}_{k,\lambda}$ is the estimated coefficients

from the k th fold. By Theorem 3 of Huang (2003), $\tilde{\beta}_{k,\lambda}$ is equivalent to $\hat{\beta}_k - \lambda(\mathbf{X}'_{k,\lambda}\mathbf{X}_{k,\lambda})^{-1}\text{sgn}(\hat{\beta}_k)$, where $\hat{\beta}_k$ is the least squares estimate from the k th fold and $\text{sgn}(\cdot)$ denotes the sign function. It should be noted that the least squares estimate can be fitted because $\mathbf{X}_{k,\lambda}$ can have $d(< n/K)$ variables by Condition (D7). Then, we have

$$\begin{aligned}
\|PE(k, \lambda)\|^2 &= \|\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),\lambda}\{\hat{\beta}_k - \lambda(\mathbf{X}'_{k,\lambda}\mathbf{X}_{k,\lambda})^{-1}\text{sgn}(\hat{\beta}_k)\}\|^2 \\
&= \|\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),\lambda}\hat{\beta}_k\|^2 + \lambda^2\|\mathbf{X}_{(-k),\lambda}(\mathbf{X}'_{k,\lambda}\mathbf{X}_{k,\lambda})^{-1}\text{sgn}(\hat{\beta}_k)\|^2 \\
&\quad + 2\lambda(\mathbf{Y}_{(-k)} - \mathbf{X}_{(-k),\lambda}\hat{\beta}_k)'(\mathbf{X}_{(-k),\lambda}(\mathbf{X}'_{k,\lambda}\mathbf{X}_{k,\lambda})^{-1}\text{sgn}(\hat{\beta}_k)) \\
&= \|PE(k, D)\|^2 + O_p(\lambda\sqrt{n - n_K}/n_K) + O_p(\lambda^2(n - n_K)/n_K^2) \\
&= \|PE(k, D)\|^2 + o_p(n^{-1}).
\end{aligned} \tag{21}$$

The first term of the last inequality in (21) means D is the selected model given λ .

The last inequality is from the fact that $O_p(\lambda\sqrt{n - n_K}/n_K) = o_p(1/(K\sqrt{n})\sqrt{n - n_K}/(n/K)) = o_p(1/n)$ by (D6). As $\|PE(k, \lambda)\|^2$ is asymptotically equivalent to $\|PE(k, D)\|^2$, the variable-selection consistency of $MPCV$ in Theorem 1 holds under the high dimensional modeling procedure considered in (14).

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Selected Papers of Hirotugu Akaike*, Springer, pp. 199 – 213.
- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys* **4**: 40–79.
- Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods, *Biometrika* **76**(3): 503 – 514.
- Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data, *Biometrika* **81**(2): 351 – 358.
- Carmack, P. S., Schucany, W. R., Spence, J. S., Gunst, R. F., Lin, Q. and Haley, R. W. (2009). Far casting cross-validation, *Journal of Computational and Graphical Statistics* **18**(4): 879 – 893.
- Carmack, P. S., Spence, J. S. and Schucany, W. R. (2012). Generalised correlated cross-validation, *Journal of Nonparametric Statistics* **24**(2): 269 – 282.
- Cavanaugh, J. E. (1997). Unifying the derivations for the akaike and corrected akaike information criteria, *Statistics & Probability Letters* **33**(2): 201 – 208.
- Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces, *Biometrika* **95**(3): 759 – 771.
- Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM, *Statistica Sinica* **22**: 555 – 574.
- Chow, Y. S., Geman, S. and Wu, L. D. (1987). Consistent cross-validated density estimation, *The Annals of Statistics* **11**(1): 25 – 38.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method, *Journal of the American Statistical Association* **92**(438): 548– 560.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348 – 1360.

- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1 – 22.
URL: <http://www.jstatsoft.org/v33/i01/>
- Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70**(350): 320–328.
- Hoerl, E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1): 55 – 67.
- Huang, F. (2003). Prediction error property of the lasso estimator and its generalization, *Australian & New Zealand Journal of Statistics* **45**(2): 217 – 228.
- Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics* **28**(5): 1356 – 1378.
- Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection, *Journal of Machine Learning Research* **8**: 1145–1164.
- Park, M. Y. and Hastie, T. (2007). L1 regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society. Series B (Methodological)* **69**(4): 659–677.
- Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**(2): 461–464.
- Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association* **88**(422): 486–494.
- Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix, *Annals of Mathematical Statistics* **21**(1): 124 – 127.
- Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent, *Journal of Statistical Software* **39**(5): 1–13.
URL: <http://www.jstatsoft.org/v39/i05/>
- Stone, M. (1974). Cross-validatory choice and the assessment of statistical predictions (with discussion), *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): 111 – 147.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267 – 288.
- Tibshirani, R. and Taylor, J. (2011). The solution path of the generalized lasso, *Annals of Statistics* **39**(3): 1335– 1371.
- Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society. Series B (Methodological)* **71**(3): 671 – 683.
- Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression, *Statistica Sinica* **19**: 801–817.
- Yang, Y. (2007). Consistency of cross validation for comparing regression procedures, *The Annals of Statistics* **35**(6): 2450 – 2473.
- Yu, Y. and Feng, Y. (2013). Modified cross-validation for penalized high-dimensional linear regression models, *Journal of Computational and Graphical Statistics* **23**(4): 1009 – 1027.
- Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression, *The Annals of Statistics* **36**(4): 1567 – 1594.

- Zhang, P. (1993). Model selection via multifold cross validation, *The Annals of Statistics* **21**(1): 299 – 313.
- Zhang, Y. and Yang, Y. (2015). Cross-validation for selecting a model selection procedure, *Journal of Econometrics* **187**(1): 95 – 112.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso, *Journal of Machine Learning Research* **7**: 2541 – 2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**(476): 1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Methodological)* **67**: 301 – 320.