Journal of Nonparametric Statistics
Vol. 00, No. 00, Month 200x, 1–16

RESEARCH ARTICLE

A K-fold Averaging Cross-validation Procedure

Yoonsuh Jung[a*] and Jianhua Hu[b]

[a]*Department of Statistics, University of Waikato, Hamilton, 3240, New Zealand*;
[b]*University of Texas MD Anderson Cancer Center,Houston, TX 77030, USA*

(*February 2014*)

Cross-validation type of methods have been widely used to facilitate model estimation and variable selection. In this work, we suggest a new K-fold cross validation procedure to select a candidate 'optimal' model from each hold-out fold and average the K candidate 'optimal' models to obtain the ultimate model. Due to the averaging effect, the variance of the proposed estimates can be significantly reduced. This new procedure results in more stable and efficient parameter estimation than the classical K-fold cross validation procedure. In addition, we show the asymptotic equivalence between the proposed and classical cross validation procedures in the linear regression setting. We also demonstrate the broad applicability of the proposed procedure via two examples of parameter sparsity regularization and quantile smoothing splines modeling. We illustrate the promise of the proposed method through simulations and a real data example.

Key words: Cross-validation; Model Averaging; Model Selection

1. Introduction

Cross-validation (CV) has been widely used for model selection in regression problems. Its theoretical properties and empirical performance have been extensively discussed in the literature including the pioneering work of Stone (1974, 1977) and Geisser (1975). Burman (1989) studied the properties of leave-one-out CV and K-fold CV procedures. Shao (1993) and Rao and Wu (2005) provided theoretical insights and asymptotic theories of model selection with a fixed number of variables for linear models. Wong (1983) examined consistency of CV in kernel nonparametric regression. Zhang (1993) showed superiority of leave-K-out CV over leave-one-out CV in linear regression. Kohavi (1995)

*Corresponding author. Email: yoonsuh@waikato.ac.nz

ISSN: 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

assessed the performance in terms of model estimation and variable selection at various values of K_1 and K_2 respectively in K_1-fold CV and leave-K_2-out CV. For robust model selection, Ronchetti et al. (1997) suggested a robust loss function to measure the prediction error. Substantial practical and theoretical results of traditional and newer CV approaches were summarized in a review paper by Arlot and Celisse (2010).

In a typical K-fold CV procedure for a linear model, the data set is randomly and evenly split into K parts (if possible). A candidate model is built based on $K-1$ parts of the data set, called a training set. Prediction accuracy of this candidate model is then evaluated on a test set containing the data in the hold-out part. By respectively using each of the K parts as the test set and repeating the model building and evaluation procedure, we choose the model with the smallest cross-validation score (typically, the mean squared prediction error $MSPE$) as the 'optimal' model. Given p independent variables, there are in total of 2^p-1 possible models. In the K-fold CV procedure, each model is in fact evaluated K times. Therefore, a single 'optimal' model is selected via $K(2^p-1)$ times of model evaluation.

In this work, we propose a new cross-validation procedure with the core idea of first choosing K candidate 'optimal' models to build the ultimate model. On each training set, the 2^p-1 models are fitted and the candidate 'optimal' model is selected with the smallest cross-validation score obtained from the corresponding test set. This procedure is repeated on K training and test sets to obtain K candidate 'optimal' models. At last, the ultimate model is obtained with its parameter estimates as the average values across K candidate 'optimal' models. Therefore, we call the proposed method K-fold averaging cross-validation (ACV). Due to the averaging effect, efficiency of the final parameter estimates obtained by ACV improves over that of the traditional K-fold CV. We note that parameter estimates of CV and ACV are identical when all the candidate 'optimal' models from ACV are identical to the model selected by the traditional CV.

The proposed ACV procedure is also applicable to high dimensional data to determine the amount of penalty imposed by a regularization method, e.g., LASSO (Tibshirani 1996). In fact, the ACV procedure has broad applications as CV to problems wherein regularization parameters associated with a penalization method need to be determined. This will be demonstrated via an application to smoothing spline based nonparametric modeling (Nychka et al. 1995).

The paper is organized as follows. We describe the explicit expression of the ACV estimator and its asymptotic property in linear regression models in Section 2. We discuss the connection and distinction between ACV and CV in the applications of LASSO and smoothing spline based modeling in Section 3. We illustrate the proposed method through simulation studies in Section 4 and through application to a real data set in Section 5. And we provide some concluding remarks in Section 6.

2. *K*-fold averaging cross-validation in a linear model

2.1. *Basic setting*

Consider a linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{Y} = (y_1, \ldots, y_n)'$ is the response vector, $\mathbf{X} = \{x_{ij}\}$ is an $n \times p$ design matrix for the full model, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is a vector of *iid* random variables with mean 0 and variance σ^2. We assume that p is fixed. All the observations are divided into K parts such that every part is mutually exclusive. Throughout this paper, we assume that K is finite and fixed. In this paper, we consider deterministic predictors, but when $\mathbf{X}$ is random, the results are valid almost surely if the two assumptions in Section 2.3 and the assumption on ϵ are satisfied almost surely for given $\mathbf{X}$.

We use n_k to denote the sample size in the k^{th} fold, where $\sum_{k=1}^{K} n_k = n$. We define M_k as the selected model based on the k^{th} hold-out fold (test set), and let its corresponding design matrix be $\mathbf{X}_k$ with all observations. We further assume that $\mathbf{X}_k$ is full rank for any k.

When the k^{th} fold is a test set, a design matrix in the k^{th} fold is defined as $\mathbf{X}_{n_k}$, and the design matrix of the corresponding training set is denoted as $\mathbf{X}_{-n_k}$. $\mathbf{Y}_k$ and $\mathbf{Y}_{-k}$ are the response observations corresponding to $\mathbf{X}_{n_k}$ and $\mathbf{X}_{-n_k}$, respectively.

The choice of a candidate 'optimal' model is made based on the mean squared prediction error in the hold-out fold. The M_k can be represented as a subset of $\{1, ..., p\}$. Let $|M_k|$ be the number of elements (or, equivalently the number of independent variables) in M_k. Further, let $X_{M_k,-n_k}$ be the design matrix composed of variables of M_k in the training set which has a size of $n - n_k$ by $|M_k|$. And we denote $\tilde{\beta}_k$ as an estimator of β with M_k. For example, the $\tilde{\beta}_k$ obtained from the least squares method can be expressed as $(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}$, which is a length-$|M_k|$ vector.

Since the $\tilde{\beta}_k$s are averaged over k to obtain the ultimate parameter estimates, we need to ensure that each of them is of length p. For this purpose, we introduce the transformation matrix $\mathbf{T}_k$, which is of size $p \times |M_k|$. Each column vector of $\mathbf{T}_k$ is composed of a 1 and $(p-1)$ 0s. For example, if $\mathbf{x}_2$ and $\mathbf{x}_4$ are selected and $p = 6$, then $M_k = \{2, 4\}$, and $\mathbf{T}_k$ contains two column vectors of $(0, 1, 0, 0, 0, 0)'$ and $(0, 0, 0, 1, 0, 0)'$, where the position of 1 in the i^{th} column is specified in the i^{th} element of M_k, and the variables not contained in M_k are indexed as 0. As a result, the size of $\tilde{\beta}_k$ is changed to p via multiplying by $\mathbf{T}_k$. Going back to the example, $\tilde{\beta}_k = (1, 2)'$ yields $\mathbf{T}_k\tilde{\beta}_k = (0, 1, 0, 2, 0, 0)'$. And $\mathbf{T}_k$ also indicates the selected model with $\mathbf{X}\mathbf{T}_k = \mathbf{X}_k$.

To measure the discrepancy between the observations and the predicted values, we use a squared error loss. The $MSPE$ evaluated with the data $\mathbf{Y}_k$ can be written as

$$MSPE_k = \frac{1}{n_k}(\mathbf{Y}_k - \mathbf{X}_{n_k}\hat{\beta}_{-n_k})'(\mathbf{Y}_k - \mathbf{X}_{n_k}\hat{\beta}_{-n_k}), \tag{1}$$

where the least squares estimate $\hat{\beta}_{-n_k} = \mathbf{T}_k(\mathbf{X}'_{M_k,-n_k}\mathbf{X}_{M_k,-n_k})^{-1}\mathbf{X}'_{M_k,-n_k}\mathbf{Y}_{-k}$. Note that $\hat{\beta}_{-n_k}$ depends on the variables contained in M_k. Because $\hat{\beta}_{-n_k}$ is not our final estimator, convergence of (1) to 0 does not necessarily deduce consistency of the estimator defined in (2); whereas it does in the method of leave-n_k-out cross-validation described in Zhang (1993) and Shao (1993). Rather, we choose M_k or equivalently $\mathbf{X}_k$ by minimizing (1), and use it to construct the ultimate estimator. The explicit form of the proposed parameter estimator based on the least squares method is provided below.

2.2. *A new estimator*

In the traditional K-fold CV, a selected model is what produces smallest prediction error where the number of predicted values is the same as sample size. That is, K-fold CV

selects a model which minimizes $\sum_{k=1}^{K} MSPE_k/K$.

The fundamental distinction of our method is to choose a candidate 'optimal' model $\mathbf{X}_k$ based on the hold-out fold, and then iterate this procedure for $k = 1, \ldots, K$. Note that the model selection at each iteration provides an opportunity to choose different (up to K) plausible models, which is not the case in K-fold CV. At last, the K candidate 'optimal' models are averaged to yield an ultimate model.

The regression coefficient estimator of the model optimally chosen by the k^{th} fold is

$$\hat{\beta}_k = \mathbf{T}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y},$$

where $\mathbf{X}_k$ is an $n \times |M_k|$ design matrix containing all the n values of the explanatory variables in M_k selected by the k^{th} fold according to (1). Note again that $\mathbf{T}_k$ is the transformation matrix to ensure the length of $\hat{\beta}_k$ to be p. Then, the ultimate estimator of the regression coefficients in the K-fold ACV procedure is defined as

$$\hat{\beta}_{ACV} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k = \frac{1}{K}\sum_{k=1}^{K}\mathbf{T}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y}, \tag{2}$$

Accordingly, the fitted value of $\mathbf{Y}$ by ACV is

$$\hat{\mathbf{Y}}_{ACV} = \mathbf{X}\hat{\beta}_{ACV} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y} = \mathbf{H}_{ACV}\mathbf{Y}.$$

Let $\mathbf{P}_k = \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'$, then $\mathbf{H}_{ACV} = \sum_{k=1}^{K}\mathbf{P}_k/K$. Since $\mathbf{X}$ is deterministic predictors, we have,

$$\begin{aligned} &E(\mathbf{X}\hat{\beta}_{ACV}) = \mathbf{H}_{ACV}E(\mathbf{Y}) = \mathbf{H}_{ACV}\mathbf{X}\beta, \\ &Var(\mathbf{X}\hat{\beta}_{ACV}) = Var(\mathbf{H}_{ACV}\mathbf{Y}) = \mathbf{H}_{ACV}'\mathbf{H}_{ACV}\sigma^2. \end{aligned} \tag{3}$$

Although each $\mathbf{P}_k$ is a projection matrix, $\mathbf{H}_{ACV}$ is not a projection matrix in general. This implies that $\hat{\beta}_{ACV}$ is likely a biased estimator (Hoaglin and Welsch 1978). However, efficiency of the estimator can be substantially gained from averaging. To account for variability induced by randomly splitting the data, we repeat the splitting step several times in both CV and ACV procedures. Alternatively, we can use a balanced split. Note that there are $n!/(n_1!n_2!\cdots n_K!)$ different ways of splitting data. Then, all observations have equal probability of being included in the training set and in the test set, thus there is no random variation arises from the splitting. In practice, the random variation from the split is ignorable, and only a few times of splitting observed to be enough in the simulation studies.

Connection with *Bagging*: The bootstrap aggregating approach (*Bagging*) suggested by Breiman (1996) is similar to our proposed method in terms of the repeated parameter estimation and variable selection steps. The *Bagging* procedure first generates a sample of size n. Based on the generated sample, it performs model selection according to a criterion such as K-fold CV and obtains the least squares estimator, $\hat{\beta}_1^{Bagging}$. After repeating this process B times, $(\hat{\beta}_1^{Bagging}, \ldots, \hat{\beta}_B^{Bagging})$ are obtained and averaged to form a final estimate. The distinction is that Bagging procedure requires generating bootstrap samples from the original data set, while the proposed ACV always use original data

set and average only 'optimal' models obtained from each hold-out fold to form a final model.

Connection with *BMA*: The *ACV* method also shares some similarity with Bayesian model averaging (BMA) (Raftery et al. 1997) where the posterior distribution of β given data D is

$$Pr(\beta|D) = \sum_{v=1}^{V} Pr(\beta|M_v, D)Pr(M_v|D),$$

where $V = 2^p - 1$. This is a weighted average of the posterior distribution of β, wherein the weight corresponds to the posterior model probabilities. Based on this formulation, we can view ACV as a special case of BMA by setting $Pr(M_v|D) = 1/K$ if $M_v \in \{M_1, \ldots, M_K\}$ and $Pr(M_v|D) = 0$ otherwise. Thus, we average those 'optimally' selected models based on the $MSPE_k$ defined in (1) with the equal weight, instead of averaging across all $2^p - 1$ possible models as in BMA. In fact, ACV can be modified to obtain the similar form of BMA. We define $\hat{\beta}_k^{FMA} = \sum_{v=1}^{V} w_{k,v}\hat{\beta}_{k,v}$, where $\hat{\beta}_{k,v}$ is the estimate of the v^{th} model and its corresponding weight is $w_{k,v}$. Because we concern finding the smallest $MSPE_k$ among $2^p - 1$ possible ones using data in the k^{th} hold-out fold, we use the $MSPE_k$ values from the v^{th} model, $MSPE_{k,v}$, to estimate $w_{k,v}$. For example, we can use $w_{k,v} = 1/\sqrt{MSPE_{k,v}}$ that is normalized to sum up to 1. Hence, we can build a new version of model averaging as $1/K \sum_{k=1}^{K} \hat{\beta}_k^{FMA}$. The properties of this model averaging method will not be pursued in this work.

Remark: The ACV procedure can be integrated with robust model estimation. For example, instead of the mean one may consider using a component-wise median estimation among the candidate models. M-estimator (Huber and Ronchetti 2009) or other types of robust estimators can also be used.

2.3. *Theoretical Properties*

We denote the true model as M_*. A model M can be classified into two categories (Shao 1993):

- Category I: $M \supseteq M_*$
- Category II: At least one component of M_* is not contained in M.

Now, we introduce the following assumptions :
Assumption A. $n/n_k \to K$ for all $k = 1, \ldots, K$.
Assumption B. For M_k in Category II,

$$\liminf_{n\to\infty} n^{-1}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}\beta = b_k > 0, \text{ and } n^{-1}\beta'\mathbf{X}'\mathbf{P}_k\mathbf{X}\beta \to 0.$$

Assumption B is adopted from Zhang (1993). Notice that b_k is non-zero when M_k is in Category II, but becomes zero when M_k is in category I.

We define the mean squared error (MSE) of ACV as

$$MSE_{ACV} = \frac{1}{n}E((\mathbf{X}\beta - \mathbf{X}\hat{\beta}_{ACV})'(\mathbf{X}\beta - \mathbf{X}\hat{\beta}_{ACV})). \quad (4)$$

Note that MSE_{ACV} is defined based on the ultimate parameter estimates and uses all the explanatory variables, which is different from $MSPE_k$ used to select the candidate 'optimal' model in the hold-out fold k. Using (3), (4) is expressed as

$$\begin{aligned} MSE_{ACV} &= \frac{1}{n} tr(Var(\mathbf{X}\hat{\beta}_{ACV})) + \frac{1}{n}\|E(\mathbf{X}\hat{\beta}_{ACV} - \mathbf{X}\beta)\|^2 \\ &= \frac{1}{n} tr(\mathbf{H}'_{ACV}\mathbf{H}_{ACV})\sigma^2 + \frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{H}_{ACV})'(\mathbf{I}_n - \mathbf{H}_{ACV})\mathbf{X}\beta, \end{aligned} \tag{5}$$

where $\mathbf{I}_n$ is an identity matrix of size n, and $tr(\cdot)$ is a trace.

Now, let the model chosen by the traditional K-fold CV as $\mathbf{X}_o$. Then, the MSE of the traditional K-fold CV (MSE_{CV}) is obtained by replacing $\mathbf{H}_{ACV}$ with the projection matrix of $\mathbf{X}_o$.

$$MSE_{CV} = \frac{d}{n}\sigma^2 + \frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_{CV})\mathbf{X}\beta, \tag{6}$$

where $d = tr(\mathbf{P}_{CV}) \leq p$ and $\mathbf{P}_{CV} = \mathbf{X}_o(\mathbf{X}'_o\mathbf{X}_o)^{-1}\mathbf{X}'_o$. Then, it is readily seen that $MSE_{CV} = \frac{d}{n}\sigma^2 = O(n^{-1})$ when $\mathbf{X}_o$ is in Category I, since the second term on the right-hand side (RHS) of equation (6) is 0. When $\mathbf{X}_o$ is in Category II, MSE_{CV} is dominated by $\frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_{CV})\mathbf{X}\beta$, which is greater than 0 by Assumption B.

In Theorem 1, we show that the two estimates chosen by CV and ACV are asymptotically equivalent as the sample size increases.

Theorem 2.1: *Let $M_f = \cup_{k=1}^K M_k$ be a union of M_k for $k = 1, \ldots, K$. Under Assumptions A and B,*
(a) $MSE_{ACV} = MSE_{CV} + O(n^{-1})$ if M_f and $\mathbf{X_o}$ are in Category I.
(b) $MSE_{ACV} = MSE_{CV} + o(1)$ if M_f and $\mathbf{X_o}$ are in Category II.

Note that the rate of decrease in both MSE_{ACV} and MSE_{CV} is subject to inclusion of the true model. This implies that inclusion of the true model will determine the asymptotic behavior of ACV and CV, even though the two sets of parameter estimates are asymptotically equivalent.

Remark: Some existing work use a slightly different but more realistic definition of MSE, since β is unknown in practice, which can be written as $MSE_{ACV} = \frac{1}{n}E((\mathbf{Y} - \mathbf{X}\hat{\beta}_{ACV})'(\mathbf{Y} - \mathbf{X}\hat{\beta}_{ACV}))$, and $MSE_{CV} = \frac{1}{n}E((\mathbf{Y} - \mathbf{X}\hat{\beta}_{CV})'(\mathbf{Y} - \mathbf{X}\hat{\beta}_{CV}))$. Using this definition, we can derive the same asymptotic equivalence of $\hat{\beta}_{ACV}$ and $\hat{\beta}_{CV}$.

Now, we show the reduced variance of $\hat{\beta}_{ACV}$ under the finite sample.

Lemma 2.2: *Under the setting in Section 2.1, we have*

$$\det(var(\hat{\beta}_{ACV})) = \det(var(\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k)) \leq \det(var(\hat{\beta}_{CV})).$$

The inequality holds when $\hat{\beta}_k$s are all identical.

This implies that the variance of $\hat{\beta}_{ACV}$ is smaller than $\hat{\beta}_{CV}$ (as $\sqrt{\det(A_k)}$ measures the *volume* of A_k) due to the averaging effect except for the case when $\hat{\beta}_k$s are all identical. Under our empirical experiments in Section 4, the magnitude of the reduced variance

tends to be larger than the allowed bias, thus results in more accurate model selection by ACV. This indicates the proposed method could take advantage of bias-variance tradeoff.

3. Integration of *ACV* with other methods

Lasso penalization: we introduce some applications of ACV beyond the traditional linear regression problem with a fairly small p as discussed in Section 2. When the number of variables p is large, it is impractical to fit all $2^p - 1$ models. Rather, we use a penalization method such as LASSO (Tibshirani 1996), which is widely used for handling high-dimensional data. Such penalization methods convert a model selection problem to controlling the amount of regularization applied to the parameter estimates. Specifically, LASSO is defined as

$$\hat{\beta}^{LASSO}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} (||\mathbf{Y} - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|).$$

Because λ takes a real number in $[0, \infty)$, we cannot consider all the possible models. We examine some candidate values of λ in a reasonable range. With K-fold CV, a $\hat{\lambda}_{CV}$ that produces the smallest cross validated score is claimed as 'optimal.'

In this case, ACV intends to select K candidate 'optimal' values of $\hat{\lambda}_k$ based on the cross-validated score from hold-out fold $k = 1, \ldots, K$. Let $\hat{\beta}^{LASSO}_{-n_k}(\lambda)$ be a LASSO estimate that is obtained based on all the samples excluding the ones in the k^{th} fold. The ultimate estimate of λ follows as $\hat{\lambda}_{ACV} = \frac{1}{K} \sum_{k=1}^{K} \hat{\lambda}_k$, where

$$\hat{\lambda}_k = \arg\min_{\lambda \in [0,\infty)} (\frac{1}{n_k} ||\mathbf{Y}_k - \mathbf{X}_{n_k} \hat{\beta}^{LASSO}_{-n_k}(\lambda)||^2).$$

Such way of averaging enables obtaining a more stable estimate of the penalization parameter. Because the possible number of λ values is infinite, a manageable solution is to use several values of λ on a discrete grid for assessment. Note that $\hat{\lambda}_{ACV}$ is likely to be more accurate than the estimate of CV using the same grid, due to the averaging effect of ACV.

Moreover, the computational efficiency of ACV is similar to CV. In both the methods, the major computation comes from estimating model with the $(K-1)$ folds of training data and from evaluating the performance with the hold-out fold. For ACV, the additional steps of choosing K candidate 'optimal' values and averaging could significantly improve the performance at trivial additional computation cost. We make comparisons between $\hat{\beta}^{LASSO}(\hat{\lambda}_{CV})$ and $\hat{\beta}^{LASSO}(\hat{\lambda}_{ACV})$ via simulation studies that will be described later.

Quantile smoothing splines: The cross-validation method is also frequently used in many different versions of smoothing splines to control the amount of roughness of the fitted function. Quantile smoothing splines are developed for targeting conditional quantiles in a nonparametric fashion. There are several different definitions of such smoothing splines; we use the method proposed by Bloomfield and Steiger (1983) and Nychka et al.

(1995). Basically, a quantile smoothing spline is taken as the minimizer of

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

where $\rho_q(u) = u\{q - I(u < 0)\}$ is the check function of Koenker and Bassett (1978) and $0 < q < 1$. With a conditional probability density of $f_{Y|X}(Y|X)$, the q^{th} conditional quantile function $g_q(x)$ is a function of x such that $q = \int_{-\infty}^{g_q(x)} f_{Y|X}(y|x)dy$. Here, the roughness of the fitted curves is determined by the value of λ. We define $\hat{\lambda}^q_{CV}$ as the minimizer of the cross-validated score where the check loss function is used for validation. Similar to the procedure of conducting ACV under LASSO, we can obtain the ACV version of $\hat{\lambda}^q_{ACV}$ for quantile smoothing splines. We use simulations to investigate the empirical properties of $\hat{\lambda}^q_{CV}$ and $\hat{\lambda}^q_{ACV}$.

4. Simulations

In empirical studies, it is desirable for K, the number of folds, to be reasonably large in order to receive the benefit of averaging. We consider $K = 5$ and 10 throughout the simulation studies. We investigate the performance of CV and ACV in three scenarios: the traditional linear regression model, LASSO regularization method, and quantile smoothing splines. For these scenarios, the corresponding ACV methods described in Section 2 and Section 3 are used. The classical CV procedure is described as follows.

The CV Procedure

(1) Randomly and evenly split the data set into K folds.
(2) Use $K-1$ folds of data as a training data set to fit the linear model (or, LASSO, and quantile smoothing splines).
(3) With the fitted models in (2), predict the value of the response variable in the hold-out fold.
(4) From the response variable in hold-out fold (say, k^{th} fold), calculate mean squared prediction error by $MSPE_k = \sum_{i\in\aleph_k}(y_i - \hat{y}_i^{pred})^2/n_k$, where $\hat{y}_i^{pred}$ is the predicted value for y_i and $\aleph_k$ is the data set in k^{th} fold.
(5) Repeat (2) through (4) for K times so that each of K fold is used as a hold-out fold from which we obtain $MSPE_1, \ldots, MSPE_K$.
(6) Each candidate model obtains a prediction performance measure $\sum_{k=1}^{K} MSPE_k/K$. The 'optimal' model which minimizes $\sum_{k=1}^{K} MSPE_k/K$ will be selected.

4.1. *Linear model*

Assume a linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ as described in Section 2.1 with ϵ_is following an iid standard normal distribution. We consider the number of samples $n = 200$. The X is generated from a multivariate normal distribution with mean 0 and the correlation between the i^{th} and j^{th} covariates $\rho = 0.5^{|i-j|}$. We consider three cases of the true values of β: (i) $\beta = (3, 0, 0, 0, 0, 0, 0, 0)'$; (ii) $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$; and (iii) $\beta = (2, 2, 2, 2, 2, 2, 2, 2)'$ that represent sparse, intermediate, and dense parameter space, respectively.

Table 1. Estimate of MSE (multiplied by 10^2) and its standard error (in parentheses, multiplied by 10^2) based on 1000 MC samples for CV, ACV and *Bagging*.

	β=(3,0,0,0,0,0,0,0)'	β=(3,1.5,0,0,2,0,0,0)'	β=(2,2,2,2,2,2,2,2)'
CV, $K=5$	516.4 (13.6)	593.1 (13.3)	806.3 (13.0)
ACV, $K=5$	405.8 (8.7)	521.1 (5.8)	806.3 (13.0)
$reduction(\%)$	21.43	12.15	0
CV, $K=10$	509.3 (13.7)	590.9 (13.3)	806.3 (13.0)
ACV, $K=10$	370.1 (7.6)	553.4 (5.8)	806.3 (13.0)
$reduction(\%)$	27.34	6.35	0
Bagging	811.4(13.1)	808.3(13.1)	810.6(13.0)

To measure accuracy of the estimates, we use the mean squared error (MSE) defined as

$$\begin{aligned} MSE &= E_{\hat{\beta}}||\mathbf{X}\hat{\beta} - \mathbf{X}\beta||^2 \\ &= E_{\hat{\beta}}\{(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)\}. \end{aligned} \tag{7}$$

With L Monte Carlo (MC) samples ($L=1000$), the estimate $\widehat{MSE}$ can be computed as $1/L\sum_{l=1}^{L}\{(\hat{\beta}^l-\beta)'\mathbf{X}_l'\mathbf{X}_l(\hat{\beta}^l-\beta)\}$, where $\hat{\beta}^l$ is the estimate of β for the l^{th} MC sample.

In addition, *Bagging* procedure is applied to the same MC samples. Since CV and ACV choose a model from $2^8-1=255$ possible models, we let *Bagging* procedure to select from 255 re-sampled (with replacement) data sets from one MC sample. The size of re-sampled data set is 200. The results of both CV and ACV are reported in Table 1. In Table 1, "$reduction(\%)$" denotes the reduction of MSE achieved by ACV in comparison to CV. We see that the largest reduction is attained in case (i) which is the sparse scenario, while no reduction is made in the dense case (iii). In the dense case (iii), ACV is able to select the true model K times which leads to the accurate ultimate model via averaging, and thus results in the identical values of MSE as CV. Performance of Bagging is consistent over the three cases, and thus relatively works fine in the dense case, but is outperformed by CV and ACV.

4.2. *Lasso penalization*

We also investigate the performance of CV and ACV in the case of large p with the usage of sparsity penalization. We use the similar simulation set-up with $n=200$ in the previous section, but increase p to 1000 and set $\rho=0$. The β is set to have 50 non-zero values and 950 zeros. Two different sets of non-zero values are considered as follows: (i) 25 values are randomly generated from $Unif(1,2)$, and the other 25 from $Unif(-2,-1)$; and (ii) 25 values are taken to be 1 and the others are with values of -1. We interrogate 200 equally spaced values of λ in (0.001, 0.4) with $K=10$.

We compare CV and ACV in terms of parameter estimation and identification of true non-zero parameters. We again use MSE in (7) to measure the estimation accuracy. The mean and its standard error of 1000 MSE values, respectively, obtained from 1000 MC samples in each case are reported in Table 2. We see that ACV reduces MSE by about 30% comparing to CV in each case. To investigate accuracy of detecting the 50 true

Table 2. Estimates of MSE and its standard error (in parentheses) based on 1000 MC samples for CV and ACV.

	CV	ACV	$reduction(\%)$
case (i)	573.2 (17.8)	381.3 (6.3)	33.5
case (ii)	487.4 (13.3)	342.8 (5.6)	29.7

Table 3. The distribution of the number of detected true non-zero βs of CV subtracted from that of ACV, based on 1000 MC samples.

	-6 to -4	-3	-2	-1	0	1	2	3	4	5 to 7	8 to 15
case (i)	14	17	51	140	352	170	108	62	31	39	16
case (ii)	14	15	61	138	306	181	112	58	46	48	21

non-zero parameters, we focus on the 50 variables with the largest absolute values of the parameter estimates, and record the number of detected true non-zero ones among them. In Table 3, we report the distribution of the number of detected true non-zero βs of CV subtracted from that of ACV. In case (i), ACV and CV show the identical detection accuracy for 352 times out of 1000 runs. Note that ACV performs better than CV for 426 times, while CV outperforms ACV for 222 times. In case (ii), ACV detects more true non-zero βs than CV for 466 times, but fewer for 228 times. It is also noticeable that ACV outperforms CV with the number of accurate detection as high as 15, while CV outperforms ACV with no more than 6 accurate detection.

4.3. *Quantile smoothing splines*

In this section, we investigate applicability of the proposed method ACV in the problem of quantile smoothing splines. We consider the mean of the observations from a sinusoid curve with period 1, which is expressed as

$$y_i = sin(2\pi x_i) + \epsilon_i, \quad i = 1, ..., n,$$

where x_is are iid from the standard uniform distribution with $n = 200$. We consider two distribution settings for the iid ϵ_i's: (i) $N(0, 0.2^2)$; and (ii) shifted exponential distribution with the median 0 and the standard deviation 0.2. A fine grid search of the smoothing parameter is conducted to select the 'optimal' value by CV and ACV with L MC samples ($L = 1000$) and $K = 5$. We adopt the empirical estimate of MSE by MC samples as,

$$MSE = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \hat{g}^l(x_i))^2,$$

where $g(x_i)$ is the true underlying function, and $\hat{g}^l(x_i)$ indicates the fitted value at x_i from the l^{th} MC sample. The results at several quantiles are shown in Table 4, where $reduction(\%)$ stands for the percentage of reduction in the average value of MSE achieved by ACV comparing to CV. Under the normal error distribution in case (i),

Table 4. Estimate of MSE and its standard error (in parentheses) based on 1000 MC samples for CV and ACV. All numbers are multiplied by 10^4.

case (i)	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5
CV	84.25(1.8)	63.39(1.2)	55.72(1.1)	51.63(1.0)	50.92(1.0)
ACV	83.13(1.8)	58.83(1.2)	50.55(1.0)	46.78(0.9)	45.10(0.8)
$reduction(\%)$	1.32	7.21	9.28	9.4	11.42
case (ii)	q=0.1	q=0.25	q=0.5	q=0.75	q=0.9
CV	2.21(0.06)	5.02(0.12)	13.14(0.28)	35.84(0.08)	89.67(2.0)
ACV	1.81(0.04)	4.28(0.09)	11.81(0.23)	33.68(0.07)	90.41(2.3)
$reduction(\%)$	18.28	14.75	10.08	6.03	-0.82

the result for $q > 0.5$ is not reported because of the similarity to $q < 0.5$. The considerable reductions achieved by ACV indicate that its averaging strategy results in better selection of the penalization parameter. For both the cases, quantiles in the region of high conditional density of the errors lead to greater reduction (%) in MSE, comparing to the low density regions. The rationale is that fewer observations are available to estimate the targeted quantile in the low density regions, and thus the advantage of ACV is not as prominent. This can be seen at $q = 0.9$ in case (ii) that the performances of these two methods are very similar with $n = 200$. To verify this finding, we increase the sample size to 400 and observe 5% reduction in MSE by using ACV. Thus, it is important to keep a sufficient number of observations in the hold-out fold to take advantage of ACV.

Another phenomenon often observed is data over-fitting by CV. In Figure 1, we demonstrate this point by two plots of the fitted curves that produce the maximum (left panel) and minimum (right panel) values of MSE_{ACV}/MSE_{CV} among 1000 MC samples, respectively. In the left panel, ACV in fact performs similarly to CV. In contrast, the right panel shows the obvious over-fitting by CV, manifested by its jagged curve.

5. A real example

Isaacs et al. (1983) established the reference percentiles of children for the serum concentrations of immunoglobulin-G (IgG) in grams per liter, based on 298 children of ages 6 months to 6 years. They considered various polynomial regressions treating IgG as the response and age (in month) as an explanatory variable. The best-fit model provided by Isaacs et al. (1983) is $IgG^{\frac{1}{2}} = 1.16 + 0.2715 * age^{\frac{1}{2}} - 0.01195 * age$ for the 50^{th} percentile, which is drawn as a dashed line in Figure 2. Royston and Altman (1994) revisited this data and used additive polynomial regressions with smoother fitted lines to eliminate the jaggedness near the boundary ages of 6 months and 6 years. Herein, we target selecting an appropriate smoothing penalty parameter for quantile smoothing splines to obtain well-behaved reference percentiles.

To find the 'optimal' $\hat{\lambda}$, we examined 2000 $\log(\lambda)$ values within (-35,-0.2). First, the data is randomly divided into 5 (nearly) equal parts, among which one contains 58 observations and each other part contains 60 observations. To account for the variability due to random split, we iterate the procedure 200 times for both ACV and CV. The penalization parameter values, denoted by $\hat{\lambda}_{ACV}$ and $\hat{\lambda}_{CV}$, are obtained by averaging their respective 200 estimates of λ. The fitted conditional quantile lines of the IgG data

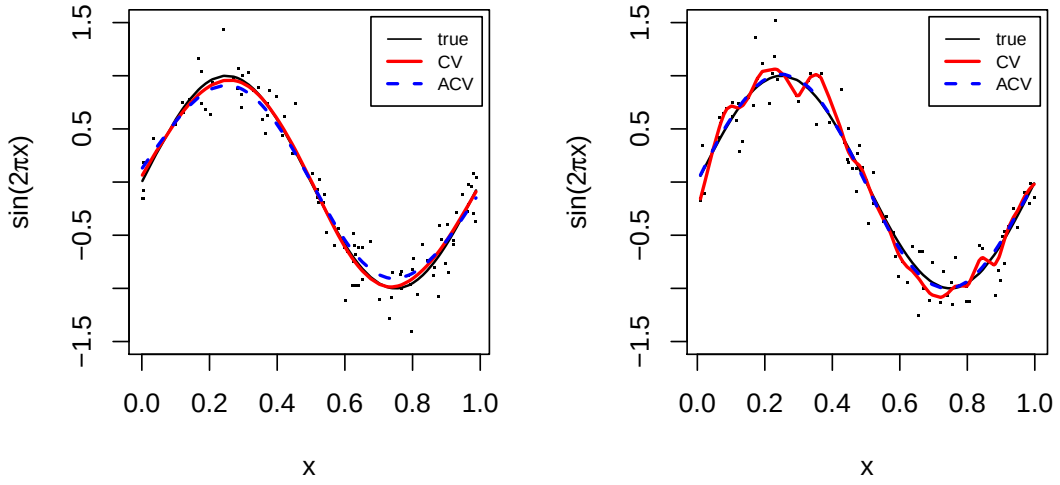

Figure 1. Under a normal error distribution, the fitted model selected by ACV performs relatively 'worst' compared to that selected by CV (left), and ACV performs relatively 'best' compared to CV (right) among 1000 simulated data sets.

Table 5. Average values of $\log(\hat{\lambda}_{CV})$ and $\log(\hat{\lambda}_{ACV})$ at various quantiles over 200 different random splits.

	q=0.05	q=0.1	q=0.25	q=0.5	q=0.75	q=0.9	q=0.95
CV	-15.49	-15.59	-16.73	-19.86	-4.02	-3.85	-3.84
ACV	-1.58	-1.62	-1.7	-1.64	-1.42	-1.5	-1.52

at $q = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9$, and 0.95 are plotted in blue from the bottom to the top in Figure 2 .

In the data set, we find mild heteroscedasticity of increasing variance as Age increases. It is reflected by the fit from both CV and ACV. It is clear that the fitted line at $q = 0.5$ provided by CV appears to be less smooth than ACV. We notice that ACV shares more similarity with the model given by Isaacs et al. (1983) which is the dashed line. At $q = 0.25$, the non-monotonic fitting line of CV implies that the IgG level of 3-year-old children is higher than that of 4-year-old children, which seems insensible from the perspective of biology. Overall, ACV provides smoother fit to data than CV. We also summarize the logarithm-transformed $\hat{\lambda}_{CV}$ and $\hat{\lambda}_{ACV}$ at various quantiles in Table 5. We can see that $\hat{\lambda}_{CV}$ changes considerably across the quantiles, whereas $\hat{\lambda}_{ACV}$ is stable (around -1.5) due to the averaging effect.

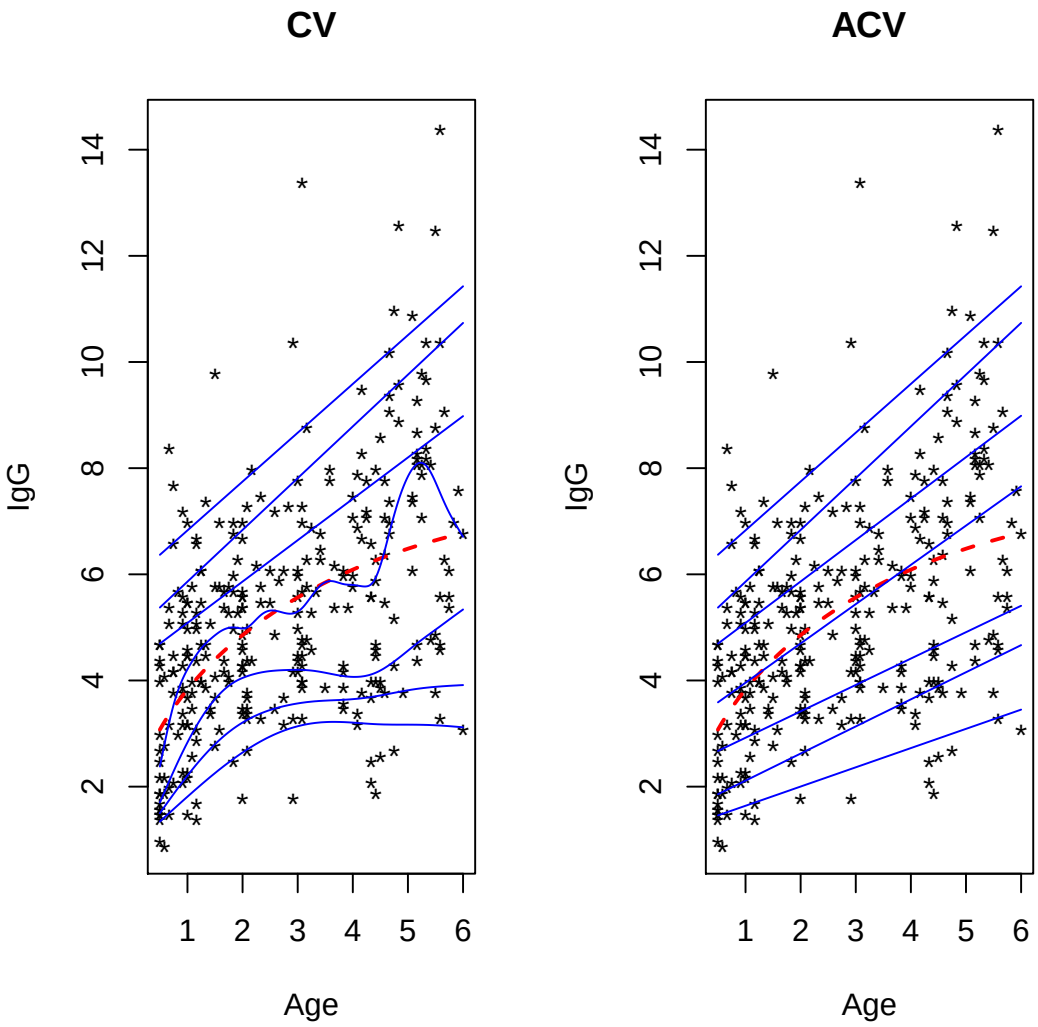

Figure 2. Fitted quantile smoothing splines at q=0.05, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.95, highlighted in blue from the bottom to the top, in the immunoglobulin-G data set. The dashed fitting line is obtained from the model given by Isaacs et al. (1983). The results of *CV* and *ACV* are shown in the left and right panels, respectively.

6. Conclusion

In this work, we propose a K-fold averaging cross-validation procedure for model selection and parameter estimation. We establish its theoretical property and show its promise via empirical investigation. Since cross-validation is actively employed in many areas of statistics, *ACV* can also be applied to a broad range of modeling procedures. For example, *ACV* can be easily used with penalized model selection method, for which selection of penalty parameters is of interest. Demonstrated through simulations with usage of LASSO, the *ACV* method outperforms the *CV* method in terms of mean squared error

and selection accuracy. We also investigate its applicability in quantile smoothing splines, and demonstrate its capability of providing more smooth data fit than the traditional CV method.

One shall pay attention to the size of the test set and selection of K when implementing ACV. A large value of K combined with a small sample size may cause insufficient fit of the data since K-fold ACV performs model selection based on $1/K$ of the data. We recommend including a sufficient number of observations in a test set. Note that the desirable size of the test set also depends on complexity of the underlying true model.

Acknowledgment

Hu's work was partially supported by the National Institute of Health Grants R01GM080503, R01CA158113, CCSG P30 CA016672, and 5U24CA086368-15.

References

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

Bloomfield, P., and Steiger, W., *Least Absolute Deviations: Theory, Applications and Algorithms*, 1 ed., Boston: Birkhauser Boston (1983).

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123 – 140.

Burman, P. (1989), "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, 76, 503 – 514.

Geisser, S. (1975), "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, 70, 320–328.

Hoaglin, D.C., and Welsch, R.E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17 – 22.

Huber, P., and Ronchetti, E.M., *Robust Statistics*, 2 ed., Probability and Statistics, Wiley (2009).

Isaacs, D., Altman, D., Tidmarsh, C., Valman, H., and Webster, A. (1983), "Serum Immunoglobulin Concentrations in Preschool Children Measured by Laser Nephelometry: Reference Ranges for IgG, IgA, IgM," *Journal of Clinical Pathology*, 36, 1193 – 1196.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Kohavi, R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Nychka, D., Gray, G., Haaland, P., Martin, D., and O'Connell, M. (1995), "A Nonparametric Regression Approach to Syringe Grading for Quality Improvement," *Journal of the American Statistical Association*, 90, 1171–1178.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179 – 191.

Rao, C.R., and Wu, Y.W. (2005), "Linear model selection by cross-validation," *Journal of Statistical Planning and Inference*, 128, 231 – 240.

Ronchetti, E., Field, C., and Blanchard, W. (1997), "Robust Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 92, 1017–1023.

Royston, P., and Altman, D.G. (1994), "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43, 429 – 467.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 111 – 147.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society, Series B*, 39, 44–47.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267 – 288.

Wong, W.H. (1983), "On the Consistency of Cross-Validation in Kernel Nonparametric Regression," *The Annals of Statistics*, 11, 1136 – 1141.

Zhang, P. (1993), "Model Selection Via Multifold Cross Validation," *The Annals of Statistics*, 21, 299 – 313.

Appendix

Proof of Theorem 2.1

Proof:

$$tr(\mathbf{H}'_{ACV}\mathbf{H}_{ACV}) = \frac{1}{K^2}tr((\sum_{k=1}^{K}\mathbf{P}_k)'(\sum_{k=1}^{K}\mathbf{P}_k)) \leq \frac{1}{K}tr(\sum_{k=1}^{K}\mathbf{P}'_k\mathbf{P}_k) \tag{1}$$

Since $\mathbf{P}'_k\mathbf{P}_k = \mathbf{P}_k$, and $tr(\mathbf{P}_k) = |M_k|$, we have,

$$\frac{\sigma^2}{n}tr(\mathbf{H}'_{ACV}\mathbf{H}_{ACV}) \leq \frac{\sigma^2}{n}\frac{1}{K}\sum_{k=1}^{K}|M_k|, \tag{2}$$

where $\frac{1}{K}\sum_{k=1}^{K}|M_k| \equiv d^*$ is the mean number of selected variables from ACV. Thus, $|d - d^*| \leq p$, which leads to $|d - d^*|\frac{\sigma^2}{n} = O(n^{-1})$. The second term on the RHS of (5) will disappear when M_f is in Category I, which completes the proof of (a).

When M_f is in Category II, the second term on the RHS of (5) will be

$$\frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{H}_{ACV})'(\mathbf{I}_n - \mathbf{H}_{ACV})\mathbf{X}\beta \leq \max_{1\leq k\leq K}\frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_k)X\beta. \tag{3}$$

When it is subtracted by the second term on the RHS of equation (6), we have

$$\frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n-\mathbf{H}_{ACV})'(\mathbf{I}_n-\mathbf{H}_{ACV})\mathbf{X}\beta-\frac{1}{n}\beta'\mathbf{X}'(\mathbf{I}_n-\mathbf{P}_{CV})\mathbf{X}\beta \leq \max_{1\leq k\leq K}\frac{1}{n}\beta'\mathbf{X}'(\mathbf{P}_{CV}-\mathbf{P}_k)\mathbf{X}\beta. \tag{4}$$

By Assumption B, the RHS of equation (4) is $o(1)$, which completes the proof of (b). □

Proof of Lemma 2.2

Proof: First, define $\max_{1\leq k\leq K} A_k$ to be A_k with maximum determinant for $k = 1, ..., K$, where A_k is a positive definite matrix. Note that $\det(A_k) > 0$ as a property of positive

definite matrix, and this is the case here since we assume $\mathbf{X}_k$ is full column rank. From the facts that the value of correlation is always between -1 and 1, and $\det(A_k A_{k'}) = \det(A_k)\det(A_{k'})$, we have

$$\begin{aligned}\det(cov(\hat{\beta}_k, \hat{\beta}_{k'})) &\leq \det\left(\sqrt{var(\hat{\beta}_k)}\sqrt{var(\hat{\beta}_{k'})}\right) = \det\left(\sqrt{var(\hat{\beta}_k)}\right)\det\left(\sqrt{var(\hat{\beta}_{k'})}\right) \\ &\leq \det(\max_{1\leq k\leq K} var(\hat{\beta}_k)),\end{aligned}$$

where the last inequality directly comes from the definition. Therefore, we have,

$$\begin{aligned}\det(var(\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k)) &= \det\left(\frac{1}{K^2}\sum_{k=1}^{K} var(\hat{\beta}_k) + \frac{2}{K^2}\sum_{1\leq k<k'\leq K} cov(\hat{\beta}_k, \hat{\beta}_{k'})\right) \\ &\leq \det\left(\frac{1}{K}\max_{1\leq k\leq K} var(\hat{\beta}_k) + \frac{K(K-1)}{K^2}\max_{1\leq k\leq K} var(\hat{\beta}_k)\right) \\ &= \det(\max_{1\leq k\leq K} var(\hat{\beta}_k)) = \det(var(\hat{\beta}_{CV}))\end{aligned}$$

When $\hat{\beta}_k$s are identical for $k = 1, \ldots, K$, the above inequality becomes equality. Further, when the selected models from all the folds $\{\hat{\beta}_k\}$ are identical, it is in fact the same as a selected model by the traditional K-fold CV. Thus, we have $\hat{\beta}_{ACV} = \hat{\beta}_{CV}$, which confirms the last equality. □