Journal of Machine Learning Research 6 (2005) 1127–1168 Submitted 1/05; Revised 6/05; Published 7/05

Analysis of Variance of Cross-Validation Estimators of the Generalization Error

Marianthi Markatou MM168@COLUMBIA.EDU
Hong Tian HT2031@COLUMBIA.EDU
Department of Biostatistics
Columbia University
New York, NY 10032, USA

Shameek Biswas SPB2003@COLUMBIA.EDU
George Hripcsak GH13@COLUMBIA.EDU
Department of Biomedical Informatics
Columbia University
New York, NY 10032, USA

Editor: David Madigan

Abstract

This paper brings together methods from two different disciplines: statistics and machine learning. We address the problem of estimating the variance of cross-validation (CV) estimators of the generalization error. In particular, we approach the problem of variance estimation of the CV estimators of generalization error as a problem in approximating the moments of a statistic. The approximation illustrates the role of training and test sets in the performance of the algorithm. It provides a unifying approach to evaluation of various methods used in obtaining training and test sets and it takes into account the variability due to different training and test sets. For the simple problem of predicting the sample mean and in the case of smooth loss functions, we show that the variance of the CV estimator of the generalization error is a function of the moments of the random variables $Y = Card(S_j \bigcap S_{j'})$ and $Y^* = Card(S_j^c \bigcap S_{j'}^c)$, where S_j, $S_{j'}$ are two training sets, and S_j^c, $S_{j'}^c$ are the corresponding test sets. We prove that the distribution of Y and Y* is hypergeometric and we compare our estimator with the one proposed by Nadeau and Bengio (2003). We extend these results in the regression case and the case of absolute error loss, and indicate how the methods can be extended to the classification case. We illustrate the results through simulation.

Keywords: cross-validation, generalization error, moment approximation, prediction, variance estimation

1. Introduction

Progress in digital data acquisition and storage technology has resulted in the growth of very large databases. At the same time, interest has grown in the possibility of tapping these data and of extracting information from the data that might be of value to the owner of the database. A variety of algorithms have been developed to mine through these databases with the purpose of uncovering interesting characteristics of the data and generalizing the findings to other data sets.

One important aspect of algorithmic performance is the generalization error. Informally, the generalization error is the error an algorithm makes on cases that has never seen before. Thus, the generalization performance of a learning method relates to its prediction capability on the indepen-

dent test data. The assessment of the performance of learning algorithms is extremely important in practice because it guides the choice of learning methods.

The generalization error of a learning method can be easily estimated via either cross-validation or bootstrap. However, providing a variance estimate of the estimator of this generalization error is a more difficult problem. This is because the generalization error depends on the loss function involved, and the mathematics needed to analyze the variance of the estimator are complicated. An estimator of variance of the cross-validation estimator of the generalization error is proposed by Nadeau and Bengio (2003). In a later section of this paper we will discuss this estimator and compare it with the newly proposed estimator.

In this paper we address estimation of the variance of the cross validation estimator of the generalization error, using the method of moment approximation. The idea is simple. The cross validation estimator of the generalization error is viewed as a statistic. As such, it has a distribution. We then approximate the needed moments of this distribution in order to obtain an estimate of the variance. We present a framework that allows computation of the variance estimator of the generalization error for k fold cross validation, as well as the usual random set selection in cross validation. We address the problem of loss function selection and we show that for a general class of loss functions, the class of differentiable loss functions with certain tail behavior, and for the simple problem of prediction of the sample mean, the variance of the cross validation estimator of the generalization error depends on the expectation of the random variables $Y = Card(S_j \bigcap S_{j'})$ and $Y^* = Card(S_j^c \bigcap S_{j'}^c)$. Here S_j, $S_{j'}$ are two different training sets drawn randomly from the data universe and S_j^c, $S_{j'}^c$ are their corresponding test sets taken to be the complement of S_j and $S_{j'}$ with respect to the data universe. We then obtain variance estimators of the generalization error for the k-fold cross validation estimator, and extend the results to the regression case. We also indicate how the results can be extended to the classification case.

The paper is organized as follows. Section 2 introduces the framework and discusses existing literature on the problem of variance estimation of the cross validation estimators of the generalization error. Section 3 presents the moment approximation method for developing the new estimator. Section 4 presents computer experiments and compares our estimator with the estimator proposed by Nadeau and Bengio (2003). Section 5 presents discussion and conclusions.

2. Framework and Related Work

In what follows we describe the framework within which we will work.

2.1 The Framework and the Cross Validation Estimator of the Generalization Error

Let data $X_1, X_2, \cdots, X_n$ be collected such that the data universe, $Z_1^n = \{X_1, X_2, \cdots, X_n\}$, is a set of independent, identically distributed observations which follow an unknown probability distribution, denoted by F. Let S represent a subset of size n_1, $n_1 < n$, taken from Z_1^n. This subset of observations is called a training set; on the basis of a training set a rule is constructed. The test set contains all data that do not belong in S, that is the test set is the set $S^c = Z_1^n \backslash S$, the complement of S with respect to the data universe Z_1^n. Denote by n_2 the number of elements in a test set, $n_2 = n - n_1$, $n_2 < n$.

Let $L : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ be a function, and assume that Y is a target variable and $\hat{f}(x)$ is a decision rule. The function $L(Y, \hat{f}(X))$ that measures the error between the target variable and the prediction rule is called a loss function.

As an example, consider the estimation of the sample mean. In this problem the learning algorithm uses $\hat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i = \overline{X}_{S_j}$ as a decision rule and $L(\overline{X}_{S_j}, X_i) = (\overline{X}_{S_j} - X_i)^2$, $X_i \in S_j^c$, the square error loss, as a loss function. Other typical choices of the loss function include the absolute error loss, $|\overline{X}_{S_j} - X_i|$ and the $0-1$ loss function mainly used in classification.

Our results take into account the variability in both training and test sets. The variance estimate of the cross validation estimator of the generalization error can be computed under the following cross validation schemes. The first is what we term as *complete random selection*. When this form of cross validation is used to compute the estimate of the generalization error of a learning method, the training sets, and hence the test sets, are randomly selected from the available data universe. In the *nonoverlapping test set selection* case, the data universe is divided into k nonoverlapping data subsets. Each data subset is then used as a test set, with the remaining data acting as a training set. This is the case of k-fold cross validation.

We now describe in detail the cross validation estimator of the generalization error whose variance we will study. This estimator is constructed under the complete random selection case.

Let A_j be a random set of n_1 distinct integers from $\{1,2,\cdots,n\}$, $n_1 < n$. Let $n_2 = n - n_1$ be the size of the corresponding complement set. Note here that n_2 is a fixed number and that $Card(A_j) = n_1$ is fixed. Let $A_1, A_2, \cdots, A_J$ be random index sets sampled independently of each other and denote by A_j^c, the complement of A_j, $j = 1,2,\cdots,J$. Denote also by $S_j = \{X_l : l \in A_j\}$, $j = 1,2,\cdots,J$. This is the training set obtained by subsampling Z_1^n according to the random index set A_j. Then the corresponding test set is $S_j^c = \{X_l : l \in A_j^c\}$. Now define $L(j,i) = L(S_j, X_i)$, where L is a loss function. Notice that L is defined by its dependence on the training set S_j and the test set S_j^c. This dependence on the training and test sets is through the statistics that are computed using the elements of these sets. The usual average test set error is then

$$\hat{\mu}_j = \frac{1}{n_2}\sum_{i \in S_j^c} L(j,i), \tag{2.1}$$

The cross validation estimator we will study is defined as

$$\tfrac{n_2}{n_1}\hat{\mu}_J = \frac{1}{J}\sum_{j=1}^{J} \hat{\mu}_j. \tag{2.2}$$

This version of the cross validation estimator of the generalization error depends on the value of J, the size of the training and test sets and the size of the data universe. The estimator has been studied by Nadeau and Bengio (2003). These authors provided two estimators of the variance of $\tfrac{n_2}{n_1}\hat{\mu}_J$. In the next section we review briefly the estimators presented by Nadeau and Bengio (2003) as well as other work on this subject. In a later section we will see that, when J is chosen appropriately, then the Nadeau and Bengio (2003) estimator is close to and performs similarly with the moment approximation estimator in some of the cases we study.

2.2 Related Work

Related literature for the problem of estimating the variance of the generalization error includes work by McLachlan (1972, 1973, 1974, 1976) and work by Nadeau and Bengio (2003) and Bengio and Grandvalet (2004). Here, we briefly review this work.

Let $S_{\hat{\mu}_j}^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\mu}_j - \frac{n_2}{n_1}\hat{\mu}_J)^2$ be the sample variance of $\hat{\mu}_j$, $j = 1,2,\cdots,J$. Then Nadeau and Bengio (2003) show that

$$E(S^2_{\hat{\mu}_j}) = \frac{Var(\frac{n_2}{n_1}\hat{\mu}_J)}{(\frac{1}{J} + \frac{\rho}{1-\rho})}, \tag{2.3}$$

where ρ is the correlation between $\hat{\mu}_j$ and $\hat{\mu}_{j'}$. Therefore, if ρ is known,

$$(\frac{1}{J} + \frac{\rho}{1-\rho})S^2_{\hat{\mu}_j}, \tag{2.4}$$

is an unbiased estimator of the $Var(\frac{n_2}{n_1}\hat{\mu}_J)$. Nadeau and Bengio (2003) observe that this estimator depends on the correlation ρ between the different $\hat{\mu}_j$s which is difficult to estimate. Thus, they propose an approximation to the correlation, $\hat{\rho} = \frac{n_2}{n}$, where n_2 is the cardinality of the test set. The final estimator of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$ is given as

$$(\frac{1}{J} + \frac{n_2}{n_1})S^2_{\hat{\mu}_j}. \tag{2.5}$$

Nadeau and Bengio (2003) note that the above suggested estimator is simple but it may have a positive or negative bias with respect to the actual $Var(\frac{n_2}{n_1}\hat{\mu}_J)$. That is, it will tend to overestimate or underestimate $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ according to whether $\hat{\rho} = \frac{n_2}{n} > \rho$ or $\hat{\rho} < \rho$. Therefore, this estimator is not exactly unbiased.

Nadeau and Bengio (2003) also suggested another estimator of the variance of the cross-validation estimator of the generalization error. This estimator is unbiased but overestimates the $Var(\frac{n_2}{n_1}\hat{\mu}_J)$. It is computed as follows. Let n be the size of the data universe and assume, without loss of generality, that n is even. Randomly split the data set into two, equal size, data subsets. Then compute the cross-validation estimator of the generalization error on these two data subsets. Notice that, the size of the training set is now $n_1' = [\frac{n}{2}] - n_2 < n_1$, smaller than the original size of the training set, but the test set size remains the same. Denote by $\hat{\mu}_1$ the estimator $\frac{n_2}{n_1'}\hat{\mu}_J$ computed on the first data subset and $\hat{\mu}_2$ the estimator $\frac{n_2}{n_1'}\hat{\mu}_J$ computed on the second data subset. To obtain an estimator of the variance of the cross validation estimator of the generalization error compute the sample variance of $\hat{\mu}_1$ and $\hat{\mu}_2$. The splitting process can be repeated M times and Nadeau and Bengio(2003) recommend $M = 10$. The proposed unbiased estimator is then given as

$$\frac{1}{2M}\sum_{m=1}^{M}(\hat{\mu}_{1,m} - \hat{\mu}_{2,m})^2. \tag{2.6}$$

This is an unbiased estimator of the $Var(\frac{n_2}{n_1'}\hat{\mu}_J)$.

Bengio and Grandvalet (2004) showed that there does not exist any unbiased and universal estimator of the variance of k-fold cross-validation that is valid under all distributions. Here, we derive estimators of the variance of the k-fold cross validation estimator of the generalization error that are almost unbiased. However, we also notice that our estimators do depend on the distribution of the errors and on the knowledge of the learning algorithm.

In a series of impressive papers McLachlan addressed the problem of estimation of the variance of the errors of misclassification of the linear discriminant function by developing a technique for deriving asymptotic expansions of the variances of the errors of misclassification of Anderson's classification statistic. McLachlan also established an asymptotic expansion of the expectation of the estimated error rate in discriminant analysis and obtained the distributions of the conditional error

rate and risk associated with Anderson's classification statistic in the context of the two-population discrimination problem. These derivations were carried out under the assumption of normality for the population distribution.

Our work has similarities with the work by McLachlan in the sense that we derive approximations to the moments of the distribution of the cross validation estimator of the generalization error and use these to obtain a variance estimator. However, we do not assume normality of the underlying mechanism that generated the data.

In what follows, we first present the method of moment approximation for obtaining an estimator of $Var(\frac{n_2}{n_1}\hat{\mu}_J)$. We then study the performance of this estimator and compare it with the Nadeau and Bengio (2003) estimator.

3. Moment Approximation Estimator for $Var(\frac{n_2}{n_1}\hat{\mu}_J)$

Recall that $\frac{n_2}{n_1}\hat{\mu}_J = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j = \frac{1}{J}\sum_{j=1}^{J}(\frac{1}{n_2}\sum_{i\in S_j^c}L(j,i))$. Therefore $\frac{n_2}{n_1}\hat{\mu}_J$ is a statistic. An estimator of $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ can thus be obtained by approximating the moments of the statistic $\frac{n_2}{n_1}\hat{\mu}_J$. A simple calculation shows that

$$Var(\tfrac{n_2}{n_1}\hat{\mu}_J) = \frac{1}{J^2}\sum_{j=1}^{J}Var(\hat{\mu}_j) + \frac{1}{J^2}\sum\sum_{j\neq j'}Cov(\hat{\mu}_j,\hat{\mu}_{j'}). \tag{3.1}$$

From the formula we see that if we can approximate the two terms of (3.1) then we can obtain an estimator for the variance of $\frac{n_2}{n_1}\hat{\mu}_J$. To achieve this goal, we need to estimate $E(\hat{\mu}_j)$, $E(\hat{\mu}_j^2)$ and $E(\hat{\mu}_j\hat{\mu}_{j'})$. In the following sections we will develop the theory that allows us to obtain the needed moment approximations. To illustrate the methodology clearly we treat separately the case of simple mean estimation and the regression case. We further treat separately the case where the loss function is differentiable from the case of non-differentiable loss functions.

3.1 The Sample Mean Case

We start by analyzing the case of the sample mean. Here, the loss function L depends on S_j through the statistics $\bar{X}_{S_j}$, the sample mean computed using the elements of S_j, and on S_j^c by elements $X_i \in S_j^c$. One of the reasons for presenting the sample mean case separately is because it illustrates clearly the contribution towards the estimator of $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ that is due to the variability among the different training and test sets. A second reason in favor of this case is because, under square error loss, we obtain a " golden standard" against which we can compare the new empirically computed variance estimator and the Nadeau and Bengio (2003) estimator. This "golden standard" is the exact theoretical value of the $Var(\frac{n_2}{n_1}\hat{\mu}_J)$. The obtained results show that the estimator of the variance of the cross validation estimator of the generalization error of the algorithms that use differentiable functions of the mean as loss functions, depends on the expectation of the random variables $Y = Card(S_j \bigcap S_{j'})$ and $Y^* = Card(S_j^c \bigcap S_{j'}^c)$.

Let the loss function $L(j,i) = L(\bar{X}_{S_j}; X_i)$ be differentiable. Below we list the conditions under which our theory holds.

Assumption 1. The distribution of $L(\bar{X}_{S_j}, X_i)$ does not depend on the particular realization of S_j and i.

Assumption 2. The loss function L as a function of $\bar{X}_{S_j}$ is such that its first four derivatives with respect to the first argument exist for all values of the variable that belongs in I, where I is an interval such that $P(v \in I) = 1$, and v indicates the first argument of the loss function.

Assumption 3. The fourth derivative of L is such that $|L^{(iv)}(\bar{X}_{S_j}; X_i)| \leq M(X_i)$, $E[M(X_i)] < \infty$.

Assumption 1 is also used by Nadeau and Bengio (2003, p. 244). Assumptions 2 and 3 are standard in the literature where approximations to the moments of a continuous, real function of the mean are discussed. See, for example Cramer (1946), Lehmman (1991) and Bickel and Doksum (2001). The boundedness of the fourth or some higher derivative is necessary for proposition 3.1 to hold.

Alternative conditions where stronger assumptions on the distributions of the data X_i and weaker conditions on the function L are imposed exist in the literature (Khan (2004)). Here L is a loss function and it seems reasonable to assume boundedness on some of its higher derivatives.

Proposition 3.1 offers an approximation of the expectation of $L(\bar{X}_{S_j}, X_i)$.

Proposition 3.1 Let $X_1, X_2, \cdots, X_n$ be independent, identically distributed random variables such that $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and finite fourth moment. Suppose that L satisfies assumptions 1, 2 and 3. Then

$$E[L(\bar{X}_{S_j}; X_i)] = E[L(\mu, X_i)] + \frac{\sigma^2}{2n_1} E[(L''(\mu, X_i))] + O(\frac{1}{n_1^2}),$$

where the remainder R_n is such that $E(R_n)$ is $O(\frac{1}{n_1^2})$, that is, there exists n_0 and $A < \infty$ such that $E(R_n) < \frac{A}{n_1^2}$,$\forall n > n_0$ and all μ. The prime indicates derivative with respect to the first argument of L.

Proof: We will use a conditional expectation argument. Write

$$E[L(\bar{X}_{S_j}; X_i)] = E_{S_j,i}\{E_{Z_1^n}[L(\bar{X}_{S_j}; X_i)|S_j, i]\}, \tag{3.2}$$

$j = 1, 2, \cdots, J$ and i indicates X_i and is such that $i \in S_j^c$.

Now expand $L(\bar{X}_{S_j}; X_i)$ with respect to $\bar{X}_{S_j}$ around the mean μ to obtain:

$$\begin{aligned} L(\bar{X}_{S_j}; X_i) &= L(\mu, X_i) + L'(\mu, X_i)(\bar{X}_{S_j} - \mu) + \frac{1}{2} L''(\mu, X_i)(\bar{X}_{S_j} - \mu)^2 \\ &+ \frac{1}{6} L'''(\mu, X_i)(\bar{X}_{S_j} - \mu)^3 + \frac{1}{24} L^{(iv)}(\mu^*, X_i)(\bar{X}_{S_j} - \mu)^4. \end{aligned} \tag{3.3}$$

Denote by

$$R_n = L^{(iv)}(\mu^*, X_i)(\bar{X}_{S_j} - \mu)^4$$

and

$$E_{Z_1^n}\{R_n|S_j, i\} = E_{Z_1^n}\{L^{(iv)}(\mu^*, X_i)(\bar{X}_{S_j} - \mu)^4|S_j, i\}, \tag{3.4}$$

and since by assumption 1 the distribution of $L^{(iv)}(\mu^*, X_i)(\bar{X}_{S_j} - \mu)^4$ does not depend on the particular realization of S_j and i, we obtain

$$E_{S_j,i}\{E_{Z_1^n}[L^{(iv)}(\mu^*, X_i)(\bar{X}_{S_j} - \mu)^4|S_j, i]\} = E[L^{(iv)}(\mu^*, X_i)]E(\bar{X}_{S_j} - \mu)^4 \leq M \cdot E(\bar{X}_{S_j} - \mu)^4.$$

This is because by assumption 3 we have $E[L^{(iv)}(\mu^*, X_i)] \leq E[M(X_i)] < \infty$. Now Lemma A.5 of the appendix guarantees that $E(\bar{X}_{S_j} - \mu)^4$ is of order $1/n_1^2$. Thus, taking expectations in (3.3) and using (3.4) we obtain:

$$\begin{aligned} E[L(\bar{X}_{S_j}; X_i)] &= E_{S_j,i}\{E_{Z_1^n}[L(\mu, X_i)|S_j, i]\} + E_{S_j,i}\{E_{Z_1^n}[L'(\mu, X_i)(\bar{X}_{S_j} - \mu)|S_j, i]\} \\ &+ E_{S_j,i}\{E_{Z_1^n}[\frac{1}{2}L''(\mu, X_i)(\bar{X}_{S_j} - \mu)^2|S_j, i]\} \\ &+ E_{S_j,i}\{E_{Z_1^n}[\frac{1}{6}L'''(\mu, X_i)(\bar{X}_{S_j} - \mu)^3|S_j, i]\} + O(\frac{1}{n_1^2}). \end{aligned}$$

By assumption 1 the distribution of $L(\mu, X_i)$ does not depend on the particular realization of S_j and X_i. Thus

$$E_{S_j,i}\{E_{Z_1^n}[L(\mu, X_i)|S_j, i]\} = E_{Z_1^n}[L(\mu, X_i)].$$

Similar to the above arguments produce the approximation to the first moment given by

$$E[L(\bar{X}_{S_j}; X_i)] = E[L(\mu, X_i)] + \frac{\sigma^2}{2n_1}E[(L''(\mu, X_i))] + O(\frac{1}{n_1^2}).$$

Remark 1: Note that we do not impose distributional assumptions on the data. The only condition imposed is that samples come from distributions for which the fourth moment is finite. Many of the standard families of distributions satisfy this condition.

Remark 2: The requirement of the finiteness of the fourth moment for proposition 3.1 to hold implies limitations on the data sets on which this estimator can be computed. For example, it may be inappropriate to apply these methods to data sets which involve large variations, such as those from insurance and finance. On the other hand, the results apply to some thick tail distributions, such as the t-distribution with 5 or more degrees of freedom. The t_5-distribution, for example, is a thick tail distribution, for which the fourth moment exists.

The following proposition approximates the variance of the loss $L(\bar{X}_{S_j}, X_i)$.

Proposition 3.2 Let assumptions 1, 2 and 3 hold. If in addition the fourth derivative of $L^2(\bar{X}_{S_j}, X_i)$ is bounded, then

$$Var[L(\bar{X}_{S_j}; X_i)] = Var[L(\mu, X_i)] + \frac{\sigma^2}{n_1}\{E[(L'(\mu, X_i))^2] + Cov(L(\mu, X_i), L''(\mu, X_i))\} + O(1/n_1^2),$$

where the remainder term is $O(\frac{1}{n_1^2})$.

Proof: To obtain an expansion of the variance of $L(\bar{X}_{S_j}; X_i)$ apply proposition 1 to the function $L^2(\bar{X}_{S_j}; X_i)$ using the fact that

$$\begin{aligned} [L^2(\mu, X_i)]'' &= \frac{\partial^2}{\partial \mu^2}[L^2(\mu; X_i)] \\ &= 2(L'(\mu, X_i))^2 + 2L(\mu, X_i)L''(\mu, X_i). \end{aligned} \tag{3.5}$$

Then substituting the expansion for $L(\bar{X}_{S_j}, X_i)$ and using formula (3.5), proposition 1 and the formula of conditional variance we obtain:

$$Var[L(\bar{X}_{S_j}; X_i)] = Var[L(\mu, X_i)] + \frac{\sigma^2}{n_1}\{E[(L'(\mu, X_i))^2] + Cov(L(\mu, X_i), L''(\mu, X_i))\} + O(1/n_1^2).$$

To prove the above two propositions we use a series of lemmas that guarantee the rate of the remainder term. These lemmas are presented in the appendix.

We now present a theoretical example that verifies the approximations presented in propositions 1 and 2.

Example. Assume that $L(\bar{X}_{S_j}, X_i) = (\bar{X}_{S_j} - X_i)^2$, the square error loss that is widely used. An exact calculation of the expectation of $(\bar{X}_{S_j} - X_i)^2$ produces

$$E\{L(\bar{X}_{S_j}, X_i)\} = Var(\bar{X}_{S_j}) + Var(X_i) = \sigma^2 + \frac{\sigma^2}{n_1}.$$

On the other hand, if proposition 3.1 is used, we obtain:

$$E[L(\bar{X}_{S_j}, X_i)] = E(X_i - \mu)^2 + \frac{\sigma^2}{n_1} = \sigma^2 + \frac{\sigma^2}{n_1},$$

and the two formulas coincide. Notice that in the case of square error loss, the second derivative of the loss, with respect to μ, is bounded. The terms of order $1/n_1^2$ do not enter the formula as all higher order than two derivatives of the quadratic loss are 0. Thus, the approximation formula agrees with the exact computation.

We next turn to the variance formula. The exact computation is based on the formula

$$Var[L(\bar{X}_{S_j}, X_i)] = E_{S_j,i}\{Var_{Z_1^n}[(\bar{X}_{S_j} - X_i)^2 | S_j, i]\} + Var_{S_j,i}\{E_{Z_1^n}[(\bar{X}_{S_j} - X_i)^2 | S_j, i]\}. \quad (3.6)$$

Using this formula we obtain the exact variance as

$$Var[L(\bar{X}_{S_j}, X_i)] = 2\sigma^4 + \frac{4\sigma^4}{n_1} + \frac{2\sigma^4}{n_1^2}. \quad (3.7)$$

Using the formula given in proposition 3.2 we obtain that the approximate variance is

$$Var[L(j, i)] = 2\sigma^4 + \frac{4\sigma^4}{n_1} + O(\frac{1}{n_1^2}). \quad (3.8)$$

Comparing these two formulas we see that the variance approximation formula identifies all first order terms.

The following proposition establishes the approximation formula for the covariance terms that enter the computation of the variance of the cross validation estimators of the generalization error.

Proposition 3.3 Let S_j, $S_{j'}$ be two training sets drawn independently and at random from the data universe Z_1^n, and S_j^c, $S_{j'}^c$ the corresponding test sets. Let $X_i \in S_j^c, X_{i'} \in S_{j'}^c$, $D = S_j \bigcap S_{j'}$ and $Y = Card(D)$. Then, if $i \neq i'$

$$Cov[L(\bar{X}_{S_j}, X_i), L(\bar{X}_{S_{j'}}, X_{i'})] = \frac{\sigma^2}{n_1^2}E(Y)(E[L'(\mu, X_i)])^2 - \frac{\sigma^4}{4n_1^2}(E[L''(\mu, X_i)])^2 + O(\frac{1}{n_1^2}).$$

If $i = i'$,

$$\begin{aligned} Cov[L(\bar{X}_{S_j}, X_i), L(\bar{X}_{S_{j'}}, X_{i'})] &= Var(L(\mu, X_i)) + \frac{\sigma^2}{n_1}\{E[L(\mu, X_i)L''(\xi, i)] \\ &- E[L(\mu, X_i)]E[L''(\mu, X_i)]\} + \frac{\sigma^2}{n_1^2}E(Y)E[L'(\mu, X_i)]^2 \\ &- \frac{\sigma^4}{4n_1^2}\{E[L''(\mu, X_i)]\}^2 + O(\frac{1}{n_1^2}), \end{aligned}$$

where $E(Y)$ is the expectation of the random variable Y with respect to its distribution.

This proposition indicates that the variability due to random sampling of the training sets S_j is quantified by the expectation of the random variable $Y = Card(S_j \bigcap S_{j'})$, $j \neq j'$, $j, j' \in 1, 2, \cdots, J$. Since S_j, $S_{j'}$ are random sets of n_1 elements, Y is such that $max(0, 2n_1 - n) \leq Y \leq n_1$.

An additional random variable that enters the variance estimator of the cross validation estimator of the generalization error is $Y^* = Card(S_j^c \bigcap S_{j'}^c)$, the cardinality of the intersection of two different test sets. The following two lemmas derive the distribution of these two random variables.

Lemma 3.1 Let S_j and $S_{j'}$ be random sets of n_1 distinct elements from $Z_!^n$ and let $Y = Card(S_j \bigcap S_j^c)$, $\max(0, 2n_1 - n) \leq Y \leq n_1$. Then, the distribution of Y is

$$P(Y = y) = \frac{\binom{n_1}{y}\binom{n-n_1}{n_1-y}}{\binom{n}{n_1}},$$

a hypergeometric distribution.

Proof. We model the problem as the following $2 \times n$ table.

k	1	2	3	$\cdots$	n	Total
S_j	0	1	1	$\cdots$	0	n_1
$S_{j'}$	1	0	1	$\cdots$	0	n_1
	a_1	a_2	a_3	$\cdots$	a_n	$2n_1$

In the table we indicate whether the kth component of Z_1^n is sampled into the training set S_j or $S_{j'}$ by 1, otherwise we indicate it by 0. Denote by a_k the sum of the indicators for the kth component in the population Z_1^n over S_j and $S_{j'}$. Then

$$\begin{cases} a_1 + a_2 + \cdots + a_n = 2n_1 \\ 0 \leq a_i \leq 2 \qquad\qquad\qquad\qquad ,i = 1, \cdots, n. \end{cases}$$

Now, $P(Y = y)$ is equivalent to $P(\#\{a_i = 2\})$, $i = 1, \cdots, n$. Given $Y = y$, the number of $\{a_i = 1\}$ is $2n_1 - 2y$ and the number of $\{a_i = 0\}$ is $n - 2n_1 + y$. Since none of these three numbers could be negative, we obtain the domain of Y as $max(0, 2n_1 - n) \leq Y \leq n_1$. Recall also that S_j, $S_{j'}$ are sampled independently and each contains n_1 elements. Given $Y = y$, the distribution of the column totals is fixed; that is a_i can only take the values 0 ,1 or 2. The number of different tables with the same column totals is then $\binom{n}{y}\binom{n-y}{n_1-y}\binom{n-n_1}{n_1-y}$. and hence

$$P(Y = y) = \frac{\binom{n}{y}\binom{n-y}{n_1-y}\binom{n-n_1}{n_1-y}}{\binom{n}{n_1}\binom{n}{n_1}} = \frac{\binom{n_1}{y}\binom{n-n_1}{n_1-y}}{\binom{n}{n_1}},$$

the hypergeometric distribution.

Lemma 3.2 Let S_j and $S_{j'}$ be two training sets and S_j^c and $S_{j'}^c$ are their corresponding test sets. Let $Y^* = Card(S_j^c \bigcap S_{j'}^c)$, $0 \leq Y^* \leq n - n_1$. Then

$$P(Y^* = y) = \frac{\binom{n_1}{y-n+2n_1}\binom{n-n_1}{n-n_1-y}}{\binom{n}{n_1}} = \frac{\binom{n_2}{n_2-y}\binom{n-n_2}{n-n_2-(n_2-y)}}{\binom{n}{n-n_2}}.$$

Proof. From the proof of lemma 3.1 $P(Y^* = y) = P(\#\{a_i = 0\})$, $\{i = 1, \cdots, n\}$. Moreover, $Y^* = n - 2n_1 + Y$. Then, the result follows.

Theorem 3.1 provides the estimator of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$. We first state the theorem.

Theorem 3.1. The variance of the estimator of the generalization error $\frac{n_2}{n_1}\hat{\mu}_J$ is given as

$$Var(\tbinom{n_2}{n_1}\hat{\mu}_J) = \frac{1}{J}Var(\hat{\mu}_j) + \frac{J-1}{J}Cov(\hat{\mu}_j, \hat{\mu}_{j'}),$$

where

$$\begin{aligned} Var(\hat{\mu}_j) &= \frac{1}{n_2}[Var[L(\mu, X_i)] + \frac{\sigma^2}{n_1}\{E[(L'(\mu, X_i))^2] + Cov(L(\mu, X_i), L''(\mu, X_i))\}] \\ &+ \frac{n_2 - 1}{n_2}\frac{\sigma^2}{n_1}\{E[L'(\mu, X_i)]\}^2 + O(1/n_1^2), \end{aligned}$$

$$\begin{aligned} Cov(\hat{\mu}_j, \hat{\mu}_{j'}) &= (1 - \frac{E(Y^*)}{n_2^2})\Big[\frac{\sigma^2}{n_1^2}E(Y)(E[L'(\mu, X_i)])^2 - \frac{\sigma^4}{4n_1^2}(E[L''(\mu, X_i)])^2 + O(\frac{1}{n_1^2})\Big] \\ &+ \frac{E(Y^*)}{n_2^2}\Big[Var(L(\mu, X_i)) + \frac{\sigma^2}{n_1}\{E[L(\mu, X_i)L''(\mu, X_i)] - E[L(\mu, X_i)]E[L''(\mu, X_i)]\} \\ &+ \frac{\sigma^2}{n_1^2}E(Y)E[L'(\mu, X_i)]^2 - \frac{\sigma^4}{4n_1^2}\{E[L''(\mu, X_i)]\}^2 + O(\frac{1}{n_1^2})\Big], \end{aligned}$$

where $\mu = E_{Z_1^n}X_i$, $\sigma^2 = Var_{Z_1^n}(X_i)$.

The above formulas indicate clearly the dependence of $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ on the first moment of the random variables Y, Y^*. Since the distribution of Y and Y^* is known, we can substitute $E(Y), E(Y^*)$ by their corresponding values and simplify the above expressions. Because the distribution of Y, Y^* is hypergeometric $E(Y) = \frac{n_1^2}{n}$ and $E(Y^*) = \frac{n_2^2}{n}$. Then

$$\begin{aligned} Cov(\hat{\mu}_j, \hat{\mu}_{j'}) &= (1 - \frac{1}{n})\Big[\frac{\sigma^2}{n}(E[L'(\mu, X_i)])^2 - \frac{\sigma^4}{4n_1^2}(E[L''(\mu, X_i)])^2 + O(\frac{1}{n_1^2})\Big] \\ &+ \frac{1}{n}\Big[Var(L(\mu, X_i)) + \frac{\sigma^2}{n_1}\{Cov(L(\mu, X_i), L''(\mu, X_i))\} \\ &+ \frac{\sigma^2}{n}E[L'(\mu, X_i)]^2 - \frac{\sigma^4}{4n_1^2}(E[L''(\mu, X_i)])^2 + O(\frac{1}{n_1^2})\Big]. \end{aligned}$$

The final estimator of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$ is a plug-in estimator and it can be computed using theorem (3.1). We need to replace the unknown population mean μ and population variance σ^2 by their estimators, the sample mean and sample variance respectively. If it is not convenient to compute the sample variance and mean based on the data universe we may compute $\overline{X}_{S_j}$ and, if there are many different training sets, take as an estimator of the sample mean $\overline{X} = \frac{1}{J}\sum_{j=1}^{J}\overline{X}_{S_j}$. Moreover, $\hat{\sigma}_j^2 = \frac{1}{n_1-1}\sum_{l=1}^{n_1}(X_l - \overline{X}_{S_j})^2$, thus the variance estimate of the population variance will be $\hat{\sigma}^2 = \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_j^2$.

Example. In the case of square error loss the approximations to the variance of $\hat{\mu}_j$ and the $Cov(\hat{\mu}_j, \hat{\mu}_{j'})$ are given as:

$$Var(\hat{\mu}_j) = \frac{1}{n_2}Var[(X_i-\mu)^2] + \frac{4\sigma^4}{n_1 n_2} = \frac{1}{n_2}E[(x_i-\mu)^4] - \frac{\sigma^4}{n_2} + \frac{4\sigma^4}{n_1 n_2}, \tag{3.9}$$

$$Cov(\hat{\mu}_j, \hat{\mu}_{j'}) = (1-\frac{1}{n})(-\frac{\sigma^4}{n_1^2}) + \frac{1}{n}(\frac{4\sigma^4}{n} - \frac{\sigma^4}{n_1^2} + Var[(X_i-\mu)^2]). \tag{3.10}$$

If the data are from a $N(0,\sigma^2)$ then the moment approximation estimator of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$ is given by

$$\hat{\sigma}^4\{\frac{2(n_1+2)}{n_1 n_2}\frac{1}{J} + (\frac{J-1}{J})[\frac{2(n+2)}{n^2} - \frac{1}{n_1^2}]\},$$

where $\hat{\sigma}$ is the sample standard deviation. Thus the estimator of the variance $\frac{n_2}{n_1}\hat{\mu}_J$ is a multiple of the sample variance and the multiplication factor indicates the dependence of the estimator on n_1, n_2 and n.

Variance estimator of the k-fold CV estimator of the generalization error.

Here we present a variance estimator of the k-fold cross validation estimator of the generalization error of a learning algorithm. Notice that this is a special case of theorem 3.1. In k-fold cross validation the data universe is divided into k different non-overlapping test sets, each of which contains $\frac{n}{k}$ elements. The number of elements n_1, in any given training set, is then $n - \frac{n}{k} = \frac{(k-1)n}{k}$. Therefore, $Y = Card(S_j \bigcap S_{j'}) = \frac{(k-2)n}{k}$. Theorem 3.1 gives the approximations:

$$\begin{aligned} Var(\hat{\mu}_j) &= \frac{k}{n}[Var(L(\mu,X_i)) + \frac{\sigma^2}{n}(\frac{k}{k-1})\{E[(L'(\mu,X_i))^2] + Cov(L(\mu,X_i), L''(\mu,X_i))\}] \\ &+ \frac{n-k}{n}\frac{\sigma^2}{n}\frac{k}{k-1}\{E[L'(\mu,X_i)]\}^2 + O(1/n_1^2), \end{aligned}$$

and

$$Cov(\hat{\mu}_j, \hat{\mu}_{j'}) = \frac{\sigma^2}{n}\frac{k(k-2)}{(k-1)^2}(E[L'(\mu,X_i)])^2 - \frac{\sigma^4}{4n^2}(\frac{k}{k-1})^2(E[L''(\mu,X_i)])^2 + O(\frac{1}{n_1^2}).$$

Therefore, the variance estimate can be computed using relation (3.1), where $Var(\hat{\mu}_j)$ and $Cov(\hat{\mu}_j, \hat{\mu}_{j'})$ are replaced by their estimates. These can be obtained by replacing μ, σ^2 by their sample estimates using data from the training sets.

Now assume that the loss function used is square error. In this case, $L'(\mu, x_i) = 2(\mu - x_i)$ and $L''(\mu, x_i) = 2$. The formulas then for the variance of $\hat{\mu}_j$ and the covariance between different $\hat{\mu}_j$s simplify as follows:

$$Var(\hat{\mu}_j) = \frac{k}{n}\{Var[(X_i - \mu)^2] + \frac{4\sigma^4}{n}(\frac{k}{k-1})\}, \tag{3.11}$$

$$Cov(\hat{\mu}_j, \hat{\mu}_{j'}) = -\frac{\sigma^4}{n^2}(\frac{k}{k-1})^2, \quad j \neq j'. \tag{3.12}$$

Then $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ can be estimated by using formula (3.1) and replacing σ^2 and $Var[(X_i - \mu)^2]$ by the sample variance and an appropriate sample estimate for $Var[(X_i - \mu)^2]$. The final approximation of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$ is then

$$Var(\frac{n_2}{n_1}\hat{\mu}_J) = \frac{1}{n}\{Var[(X_i - \mu)^2]\} + \frac{3k\sigma^4}{(k-1)n^2} = \frac{1}{n}E[(X_i - \mu)^4] - \frac{\sigma^4}{n} + \frac{3k\sigma^4}{(k-1)n^2}.$$

A simple estimator of $E[(X_i - \mu)^4]$ can be computed from the training sample by taking the sample version of the above expectation, $\frac{1}{n_1}\sum_{i \in S_j}(X_i - \bar{X}_{S_j})^4$. To illustrate, if we further assume a normal population then $Var[(X_i - \mu)^2] = 2\sigma^4$ and the variance estimator of $\frac{n_2}{n_1}\hat{\mu}_J$ is given as

$$\frac{\hat{\sigma}^4}{n}(2 + \frac{3k}{n(k-1)}),$$

where $\hat{\sigma}$ is the sample standard deviation.

3.2 The Regression Case

The regression case is another case of fitting means. We consider here the problem of estimating the variance of the cross validation estimator of the generalization error $\frac{n_2}{n_1}\hat{\mu}_J$ in the case of regression. Therefore the data are realizations of random variables (Y_i, X_i), $i = 1, 2, \cdots, n$ such that $E(Y_i|X_i) = x_i^T\beta$. Notice that the explanatory variables here are treated as fixed; this formulation is known as the fixed design case. The vector of unknown parameters β is usually estimated by least squares; denote by $\hat{\beta}$ the least square estimator of β. Then for a new observation $(y_i, x_i) \in S_j^c$ denote by $\hat{y}_{i,S_j} = x_i^T\hat{\beta}_{S_j}$, where $\hat{\beta}_{S_j}$ indicates the estimator of β computed by using the data in the training set S_j. The loss function L is then dependent on $\hat{y}_{i,S_j}$ and y_i, that is $L(\hat{y}_{i,S_j}, y_i)$.

To derive the estimator of $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ we need to use the moment approximation method to obtain approximations for the moments of the statistic $\frac{n_2}{n_1}\hat{\mu}_J$. The idea is the same as in the case of simple mean estimation. That is, the loss function is expanded with respect to its first argument and evaluated at the point $E(Y_i|X_i) = x_i^T\beta_0$, where β_0 is the true parameter value. In other words, as before, the expansion is evaluated at the true mean.

We list now the assumptions under which our theory holds.

Assumption 1. If S_j is a training set with n_1 number of elements

$$\lim_{n_1 \to \infty} \frac{1}{n_1}(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1} = V$$

where V is finite and positive definite.

Assumption 2. Let x_{n_1k} denote the kth row of the design matrix $\mathbf{X}_{S_j}$. Then, for each $j = 1,2,\cdots,J$,

$$\max_{1\le k\le n_1} x_{n_1k}(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_{n_1k} \to 0$$

as $n_1 \to \infty$.

Notice that this condition is known as the generalized Noether condition.

Under the above conditions $\sqrt{n_1}(\hat{\beta}_{S_j} - \beta)$ converges in distribution to a $N(0,\sigma^2 V)$ random variable.

The following proposition establishes an approximation to the expectation of the loss function L.

Proposition 3.4: Suppose that assumptions 1 and 2 hold. Then

$$E[L(\hat{y}_{i,S_j},y_i)] = E[L(x_i^T\beta_0,y_i)] + \frac{\sigma^2}{2}E[L''(x_i^T\beta_0,y_i)]tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] + R_n,$$

where the remainder term is of order $O(\frac{1}{n_1^2})$, and the prime indicates derivative with respect to the first argument of the loss function.

Proof: First expand $L(\hat{y}_{i,S_j},y_i)$ with respect to the first argument to obtain:

$$\begin{aligned} L(\hat{y}_{i,S_j},y_i) &= L(x_i^T\beta_0,y_i) + L'(x_i^T\beta_0,y_i)x_i^T(\hat{\beta}_{S_j} - \beta_0) \\ &+ \frac{1}{2}L''(x_i^T\beta_0,y_i)(\hat{\beta}_{S_j} - \beta_0)^T x_ix_i^T(\hat{\beta}_{S_j} - \beta_0) + R_n, \end{aligned} \tag{3.13}$$

where R_n indicates the remainder term.
Now

$$\begin{aligned} E\{L(\hat{y}_{i,S_j},y_i)\} &= E_{S_j,i}\{E_{Z_1^n}[L(\hat{y}_{i,S_j},y_i)|S_j,i]\} \\ &= E_{S_j,i}\{E_{Z_1^n}[L(x_i^T\beta_0,y_i)|S_j,i]\} + E_{S_j,i}\{E_{Z_1^n}[L'(x_i^T\beta_0,y_i)x_i^T|S_j,i]E_{Z_1^n}[(\hat{\beta}_{S_j} - \beta_0)|S_j,i]\} \\ &+ \frac{1}{2}E_{S_j,i}\{E_{Z_1^n}[L''(x_i^T\beta_0,y_i)|S_j,i]E_{Z_1^n}[(\hat{\beta}_{S_j} - \beta_0)^T x_ix_i^T(\hat{\beta}_{S_j} - \beta_0)|S_j,i]\} \end{aligned}$$

But the expectation $E_{Z_1^n}[(\hat{\beta}_{S_j} - \beta_0)|S_j,i] = 0$ because $E_{Z_1^n}(\hat{\beta}_{S_j}|S_j,i) = E_{Z_1^n}(\hat{\beta}_{S_j}) = \beta_0$. Also since the distribution of $\hat{\beta}_{S_j}$ is asymptotically $N(\beta_0,\sigma^2(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1})$, under assumptions 1 and 2 we obtain:

$$\begin{aligned} E_{S_j,i}\{E_{Z_1^n}[(\hat{\beta}_{S_j} - \beta_0)^T x_ix_i^T(\hat{\beta}_{S_j} - \beta_0)|S_j,i]\} &= E_{Z_1^n}[(\hat{\beta}_{S_j} - \beta_0)^T x_ix_i^T(\hat{\beta}_{S_j} - \beta_0)] \\ &= \sigma^2\mathrm{tr}[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}], \end{aligned}$$

where $\sigma^2 = Var_{Z_1^n}(X_i)$, the variance of the sample, and $\mathrm{tr}(A)$ stands for the trace of the matrix A.
Therefore

$$E[L(\hat{y}_{i,S_j},y_i)] = E[L(x_i^T\beta_0,y_i)] + \frac{\sigma^2}{2}E[L''(x_i^T\beta_0,y_i)]tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] + R_n,$$

where the expectations are taken with respect to the distribution of the data. Moreover, R_n is of order $\frac{1}{n_1^2}$.

Proposition 3.5 establishes the approximation for the variance of $L(\hat{y}_{i,S_j}, y_i)$.

Proposition 3.5 Suppose that assumptions 1 and 2 hold. Then $Var(L(\hat{y}_{i,S_j}, y_i))$ can be approximated as follows:

$$\begin{aligned} Var\{L(\hat{y}_{i,S_j}, y_i)\} &= Var[L(x_i^T\beta_0, y_i)]\} + \sigma^2 tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}\{Cov(L(x_i^T\beta_0, y_i), \\ &\quad L''(x_i^T\beta_0, y_i)) + E[L'(x_i^T\beta_0, y_i)]^2\} + R_n, \end{aligned}$$

where $\sigma^2 = Var_{Z_1^n}(Y_i|X_i)$ and R_n is the remaining term of order $\frac{1}{n_1^2}$.

Proof: The proof is similar with that of proposition 3.2, in that we apply proposition 3.4 to $L^2(\hat{y}_{i,S_j}, y_i)$ and we use the fact that

$$[L^2(\hat{y}_{i,S_j}, y_i)]'' = 2L(\hat{y}_{i,S_j}, y_i)L''(\hat{y}_{i,S_j}, y_i) + 2[L'(\hat{y}_{i,S_j}, y_i)]^2,$$

where prime indicates derivative with respect to the first argument of the loss function.

Example. To verify the above approximations we use $L(\hat{y}_{i,S_j}, y_i) = (\hat{y}_{i,S_j} - y_i)^2$, the square error loss and the case of simple regression, that is

$$y_i = a + bz_i + \varepsilon_i = x_i^T\beta + \varepsilon_i,$$

where $x_i^T = (1, z_i)$, $\beta^T = (a, b)$ and $(y_i, x_i) \in S_j^c$. The notation $\hat{y}_{i,S_j}$ stands for $x_i^T\hat{\beta}_{S_j}$.
The exact expectation of $L(\hat{y}_{i,S_j}, y_i) = (x_i^T\hat{\beta}_{S_j} - y_i)^2$ is given as:

$$E[L(\hat{y}_{i,S_j}, y_i)] = \sigma^2 + \sigma^2 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i.$$

The approximate expectation is

$$E[L(\hat{y}_{i,S_j}, y_i)] = \sigma^2 + \sigma^2 tr(x_i x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}),$$

Because $tr(x_i x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}) = x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i$, the approximation to the expectation agrees with the exact computation. Similarly we can verify that the approximation of the variance produces the same result as the exact computation. To illustrate further the formulas assume that $y_i \sim N(x_i^T\beta, \sigma^2)$, then the exact calculation gives the variance of $L(\hat{y}_{i,S_j}, y_i)$,

$$Var(L(\hat{y}_{i,S_j}, y_i)) = 2\sigma^4 + 4\sigma^4 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i + 2\sigma^4(x_i(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i)^2.$$

The approximation is given by

$$Var(L(\hat{y}_{i,S_j}, y_i)) = 2\sigma^4 + 4\sigma^4 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i + O(\frac{1}{n_1^2}),$$

that is they agree up to first order terms.

To complete the variance approximation of the estimator $\frac{n_2}{n_1}\hat{\mu}_J$ we need an approximation of the covariance between $L(\hat{y}_{i,S_j},y_i)$ and $L(\hat{y}_{i',S_{j'}},y_{i'})$. The following proposition expresses the approximation of $Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i,S_{j'}},y_{i'}))$.

Proposition 3.5. Suppose that assumptions 1 and 2 hold. Then for $j \neq j'$, $j,j' \in \{1,2,\cdots,J\}$ when $i \neq i'$

$$\begin{aligned} & Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i',S_{j'}},y_{i'})) = \sigma^2(E[L'(x_i^T\beta_0,y_i)])^2 x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_{i'} \\ + \quad & \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2 tr((x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T) \\ & (\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}). \end{aligned}$$

When $i = i'$,

$$\begin{aligned} & Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i',S_{j'}},y_{i'})) = Var(L(x_i^T\beta_0,y_i)) + \frac{\sigma^2}{2}\,Cov(L(x_i^T\beta_0,y_i),L''(x_i^T\beta_0,y_i)) \\ & (x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_i + x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i) \\ + \quad & \sigma^2(E[L'(x_i^T\beta_0,y_i)])^2 x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i \\ + \quad & \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2 tr((x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_ix_i^T) \\ & (\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}) \\ + \quad & \frac{\sigma^4}{4}Var(L''(x_i^T\beta_0,y_i))x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_ix_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_i. \end{aligned}$$

Proposition 3.6. Let S_j be a training set, $j = 1,2,\cdots,J$. Then for $i \neq i'$

$$\begin{aligned} & Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i',S_j},y_{i'})) = \sigma^2(E[L'(x_i^T\beta_0,y_i)])^2 tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\ + \quad & \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2 tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]. \end{aligned}$$

The proofs of Proposition 3.5 and Proposition 3.6 can be found in Appendix C.

Remark: If the loss is square error,

$$Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i',S_j},y_{i'})) = 2\sigma^4 tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]. \tag{3.14}$$

To estimate relationship (3.14) we only need to estimate σ. We estimate σ by the residual mean square error.

Under square error loss, we have

$$Var(\hat{\mu}_j) = \frac{1}{n_2^2}\sum_{i=1}^{n_2}\{2\sigma^4 + 4\sigma^4 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i\} + \frac{1}{n_2^2}\sum\sum_{i\neq i'} 2\sigma^4(x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_{i'})^2, \tag{3.15}$$

and

$$Cov(\hat{\mu}_j,\hat{\mu}_{j'}) = \frac{1}{n_2^2}\sum_{\substack{i\in S_j^c \\ i\neq i'}}\sum_{i'\in S_{j'}^c}\{2\sigma^4 tr\{(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}$$
$$(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}\}\}$$
$$\frac{1}{n_2^2}\sum_{\substack{i\in S_j^c \\ i=i'}}\sum_{i'\in S_{j'}^c}\{2\sigma^4+4\sigma^4 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1})x_i$$
$$+2\sigma^4 tr\{(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_i x_i^T)$$
$$(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}\}\} \tag{3.16}$$

The final estimate is obtained from relation (3.1) where $Var(\hat{\mu}_j)$ is estimated by using relation (3.15), $Cov(\hat{\mu}_j,\hat{\mu}_{j'})$ is estimated by using relation (3.16) and replacing σ^2 by an estimator of it. To obtain an estimator of σ^2, we fit the regression model and obtain $\hat{y}_i$. Then $\hat{\sigma}^2$ is the sample variance of the errors $\hat{\varepsilon}_i = y_i - \hat{y}_i$, that is the residual mean square.

Remark: Note that to derive the results above, we used as the distribution of the data the conditional distribution of Y given X, in effect treating X as fixed. Now, assume that instead of using the conditional distribution as the data distribution, we treat X as random and use the joint distribution of (X,Y). In this case, the data distribution is

$$f(x,y) = g(y - x^T\beta|x)k(x)$$

where $g(\cdot)$ is the distribution of the errors and $k(\cdot)$ is the distribution of the xs. We can then derive the formulas expressing the expectation, variance and covariance terms that are needed using the joint distribution of (X,Y). For example, $E(\hat{\beta}) = E_{(X,Y)}[(X^TX)^{-1}X^TY] = E_X\{E_{Y|X}[(X^TX)^{-1}X^TY|X]\} = \beta_0$, is still unbiased, and $Var(\hat{\beta}) = E_X\{Var_Y(\hat{\beta}|X)\} + Var_X\{E_Y(\hat{\beta}|X)\} = \sigma^2 E_X[(X^TX)^{-1}]$. Other adjustments that take into account the distribution of X are needed. These mainly concentrate on taking expectations, over X, of terms that are functions of the Xs, and can be easily computed from the data by using bootstrap. As an illustration, under square error loss, the formula in proposition 3.4 becomes $E[L(\hat{y}_{i,S_j},y_i)] = \sigma^2 + \sigma^2 E_X[tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]]$, where σ^2 is the variance of the error distribution.

4. Simulation Experiments.

We present here simulation experiments that illustrate the performance of the proposed estimators; moreover, we compare these estimators with the estimator proposed by Nadeau and Bengio (2003). The simulation experiments compare the proposed estimators with the Nadeau and Bengio estimator under two different error losses, the square error and the absolute error loss.

4.1 Square Error Loss

We will first describe the experimental setup for the simple mean case.

We generated data sets of size $n = 100$ from a $N(0,1)$ distribution in S-plus. For each different size n_1 of the training set S_j we randomly select n_1 data points from the available n and use S_j^c, the complement of S_j with respect to the generated data universe that contains 100 data points, as a test

set. We take J to be 15 (as recommended by Nadeau and Bengio, 2003), and 50. We then computed $S^2_{\hat{\mu}_j} = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\mu}_j - \frac{n_2}{n_1}\hat{\mu}_J)^2$ and the estimator of the variance of the generalization error, given as $(\frac{1}{J} + \frac{n_2}{n_1})S^2_{\hat{\mu}_j}$.

We also computed the moment approximation estimator given by expressions (3.9) and (3.10). Notice that we estimate σ^2 by using the sample variance, that is, $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. We also computed the variance estimator of $\frac{n_2}{n_1}\hat{\mu}_J$ using expression,

$$\frac{1}{J}\frac{1}{n_2}Var(X_i^2) + \frac{1}{J}\frac{4\sigma^4}{n_1 n_2} + \frac{J-1}{J}\{\frac{1}{n}(\frac{4\sigma^4}{n} - \frac{\sigma^4}{n_1^2} + Var(X_i^2) - (1-\frac{1}{n})\frac{\sigma^4}{n_1^2}\}.$$

The population variance σ^2 is estimated by using the sample variance averaged over 100 different data sets. The term $Var(X_i^2)$ is estimated as follows. Let $Z_i = X_i^2$, $i = 1, 2, \cdots, n$. We created a new data universe using Z_i and estimate $\hat{V}ar(Z_i) = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, over 100 different data sets.

Table 1 presents the results of the simulation. The first column of the table shows the size of the test set. The second column reports the value of the Nadeau and Bengio estimator, while the third column reports its variance. The variance is computed by simply taking the sample variance of the estimator that was computed over the 100 independent data sets. The fourth column of the table reports the value of the moment approximation estimator of the variance of the cross validation estimator of the generalization error, while the fifth column reports the sample variance of the moment approximation estimator.

n_2	NB	var(NB)	MA	var(MA)
10	0.0316	0.000310	0.0328	7.75e-06
15	0.0265	0.000241	0.0282	5.34e-05
20	0.0250	0.000179	0.0259	4.50e-05
25	0.0235	0.000213	0.0245	4.03e-05
30	0.0238	0.000145	0.0236	3.73e-05
35	0.0227	0.000175	0.0229	3.52e-05
40	0.0235	0.000188	0.0224	3.36e-05
45	0.0227	0.000122	0.0219	3.23e-05
50	0.0246	0.000236	0.0216	3.13e-05

Table 1: Simple mean case n=100, J=15. Nadeau-Bengio (NB) and moment approximation (MA) estimators of the variance of the cross validation estimator of the generalization error, and their sample variances. $J = 15$, and the results are averages over 100 independent data sets. The size of the data universe is 100.

We notice that the variance of the moment approximation estimator is at least one order of magnitude smaller than the variance of the Nadeau- Bengio estimator, thereby increasing the accuracy of the moment estimator.

Figure 1 plots the values of the Nadeau-Bengio and moment approximation estimate of the variance versus the sample size of the test set. Notice that the curve corresponding to the moment approximation is smooth. This is in contrast to the behavior of the Nadeau-Bengio estimator, which

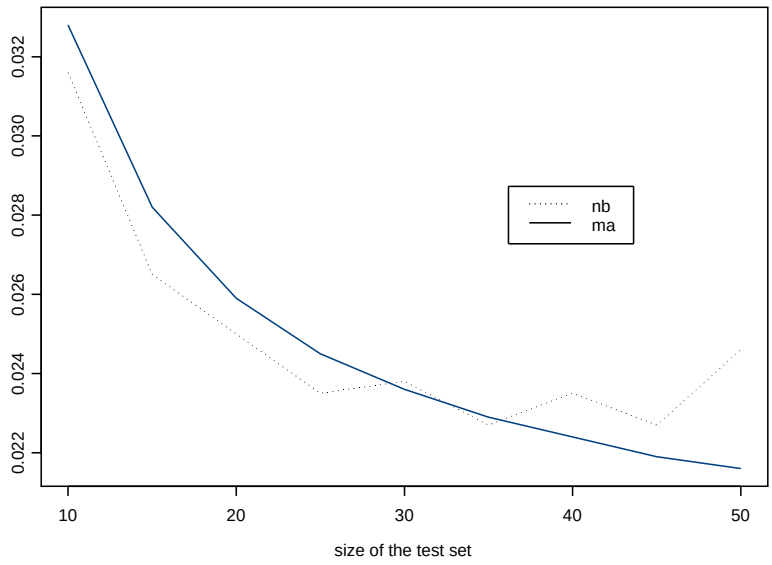

Figure 1: Simple mean case n=100, J=15

seems to fluctuate (this also is indicated by the value of the sample variance associated with the estimator and reported in table 1.)

n_2	NB	var(NB)	MA	var(MA)
10	0.0235	1.24e-04	0.0241	7.75e-06
15	0.0212	8.77e-05	0.0227	3.47e-05
20	0.0211	6.27e-05	0.0220	3.26e-05
25	0.0204	7.50e-05	0.0216	3.13e-05
30	0.0206	7.28e-05	0.0213	3.05e-05
35	0.0203	6.79e-05	0.0211	2.98e-05
40	0.0204	7.94e-05	0.0209	2.93e-05
45	0.0213	8.08e-05	0.0207	2.88e-05
50	0.0206	6.43e-05	0.0206	2.84e-05

Table 2: Simple mean case n=100, J=50. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance the cross validation estimator of the generalization error and their sample variances. $J = 50$, and the results are averages over 100 independent data sets. The size of the data universe is 100.

Table 2 presents the variance estimates of the CV estimators of the generalization error when $J = 50$. In this case we notice that the variance of the moment approximation estimator is about half of the variance of the Nadeau-Bengio estimator.

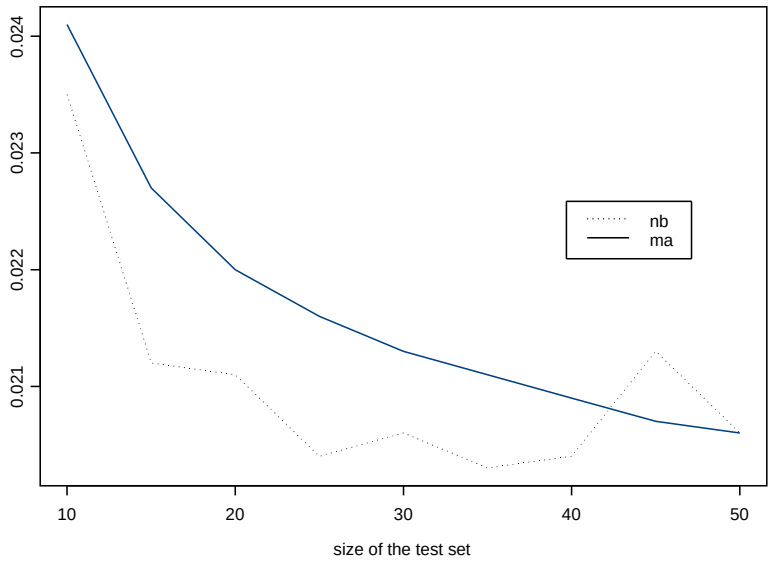

Figure 2: Simple mean case n=100, J=50

Figure 2 shows a plot of Nadeau-Bengio and moment approximation estimate of the variance as a function of the size of the test set. The larger variance of the Nadeau-Bengio estimator that was reported in table 2 can also be seen again in Figure 2.

Table 3 presents the values of the two variance estimators as well as their variance when the data universe has size $n = 1000$, for the case $J = 15$ and $J = 50$. We notice that the performance, in terms of variance, of the moment approximation estimator is, in both cases, superior to the performance of the Nadeau-Bengio estimator, always having variance that is smaller than the NB variance by one order of magnitude.

To address the problem of bias we computed the exact (and theoretical) value of the variance estimator of $\frac{n_2}{n_1}\hat{\mu}_J$. Therefore, we computed, using formula (3.1), $Var(\frac{n_2}{n_1}\hat{\mu}_J)$ under square error loss and under the assumption of a $N(0,1)$ distribution. The distributional assumption is used to obtain the theoretical value. This is done only for the purpose of comparison and in order to allow a bias computation to be carried out without having to estimate higher order moments. In practice, the distribution of the population from which the data arise is not known, and higher order moments need to be estimated from the data.

The exact theoretical value of $Var(\hat{\mu}_j)$ is

$$Var(\hat{\mu}_j) = \frac{2}{n_2}\{1 + \frac{2}{n_1} + \frac{n_2}{n_1^2}\}.$$

n_2	NB	var(NB)	MA	var(MA)
J=15				
100	0.00319	1.61e-06	0.00319	7.75e-06
150	0.00291	1.22e-06	0.00275	5.42e-08
200	0.00252	9.62e-07	0.00253	4.58e-08
250	0.00244	8.21e-07	0.00239	4.11e-08
300	0.00240	9.02e-07	0.00230	3.81e-08
350	0.00214	9.27e-07	0.00224	3.60e-08
400	0.00232	7.21e-07	0.00219	3.45e-08
450	0.00217	5.70e-07	0.00216	3.33e-08
500	0.00206	8.24e-07	0.00213	3.24e-08
J=50				
100	0.00241	3.20e-07	0.00235	5.90e-08
150	0.00225	2.82e-07	0.00222	3.54e-08
200	0.00225	3.68e-07	0.00215	3.33e-08
250	0.00216	2.43e-07	0.00211	3.21e-08
300	0.00213	1.96e-07	0.00209	3.12e-08
350	0.00216	2.83e-07	0.00207	3.07e-08
400	0.00211	2.70e-07	0.00205	3.02e-08
450	0.00218	2.36e-07	0.00204	2.99e-08
500	0.00206	2.18e-07	0.00203	2.96e-08

Table 3: Simple mean case n=1000, J=15 and J=50. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance of the cross validation estimator of the generalization error under random selection, and their sample variances. The size of the data universe is $n = 1000$ and $J = 15$ and 50.

Using theorem 3.1 the approximation to the value of $Var(\hat{\mu}_j)$ is

$$Var(\hat{\mu}_j) = \frac{2}{n_2}\{1 + \frac{2}{n_1} + O(\frac{1}{n_1^2})\}.$$

The same theorem provides the approximation to $Cov(\hat{\mu}_j, \hat{\mu}_{j'})$ as follows:

$$Cov(\hat{\mu}_j, \hat{\mu}_{j'}) = \frac{2}{n}(1 + \frac{2}{n}) + O(\frac{1}{n_1^2}).$$

The exact theoretical computation of the covariance provides us with the formula

$$Cov(\hat{\mu}_j, \hat{\mu}_{j'}) = \frac{2}{n}(1 + \frac{2}{n}) + \frac{2}{n_1}(\frac{1}{n_1} - \frac{1}{n}).$$

Using these expressions we computed the exact value of the variance of $\frac{n_2}{n_1}\hat{\mu}_J$ for the square error loss. This computation allows us to get a sense of the bias of the moment approximation and Nadeau-Bengio estimators. Table 4 presents the results for the case where the data universe is 100

n_2	Exact Variance	Bias of MA estimator	Bias of NB estimator
10	0.0327	0.0001	-0.0011
15	0.0282	0	-0.0017
20	0.0259	0	-0.0009
25	0.0246	-0.0001	-0.0011
30	0.0237	-0.0001	0.0001
35	0.0232	-0.0003	-0.0005
40	0.0227	-0.0003	0.0008
45	0.0223	-0.0004	0.0004
50	0.0222	-0.0006	0.0024

Table 4: Bias of MA and NB estimators. Bias of MA of NB estimators for the case of the simple mean. The data universe has size 100, J=15. The bias is calculated as the expectation of the estimator minus the exact value.

and $J = 15$. We observe that the moment approximation estimator has a very small bias, consistently smaller than the bias of the Nadeau-Bengio estimator. Notice that when the sizes of the training and test sets are equal ($n_1 = n_2 = 50$) the bias of the Nadeau-Bengio estimator is four times higher, in absolute value, than that of the moment approximation estimator.

At this point, we remind the reader that the Nadeau-Bengio estimator given in (2.5) is generally applicable. The proposed estimators take advantage of information about the data and the learning algorithm. Hence, it is not completely surprising that they perform better than the Nadeau Bengio estimator in terms of variance and bias.

For comparison reasons, after a referee's suggestion, we computed the second estimator proposed by Nadeau and Bengio(2003) and given by (2.6). Table 5 presents the values of the estimators of the variance given by (2.5) and (2.6) and the moment approximation estimator. Expressions (3.9) and (3.10) were used to obtain the needed variance and covariance terms. The size of the data universe is 50, 100, 500 and 1000, the size of the test set is taken to be 10, 20, 100 and 200 and J is either 15 or 50. ¿From table 5 we see that the estimator given by (2.6) is indeed conservative; its value is almost twice as big as the value of either the cheap to compute Nadeau and Bengio estimator given by (2.5) and the moment approximation estimator. It is interesting to notice that, when the training set size is the same with the training set size used to compute (2.5) and the moment approximation estimator, the value of (2.6) is comparable to the value of the other two estimators. This observation indicates the importance of the size of the training set in the computation of the variance of the cross-validation estimators of the generalization error.

To exemplify the fact that the framework we propose allows one to compute the variance estimator of the k-fold cross validation estimator of the generalization error we computed the variance of leave-one-out cross validation (LOOCV) estimator of the generalization error, the 4-fold, the 5-fold and the 10-fold in the case of square error loss and when the data universe consisted of 100 data points generated from a N(0,1) distribution. The case was prediction of simple mean. We did the same when the data universe consisted of 1000 normal data points. Table 6 presents the moment approximation variance estimators together with their variance and the corresponding NB estimators.

Sample Size	Training Set Size	J	NB	MA	NB(Conserv.)
50	10	15	0.0539	0.0537	0.0988
100	10	15	0.0314	0.0328	0.0542
50	20	15	0.0458	0.0462	0.1213
100	20	15	0.0257	0.0259	0.0456
50	10	50	0.0443	0.0456	0.0836
100	10	50	0.0236	0.0241	0.0420
50	20	50	0.0421	0.0430	0.1131
100	20	50	0.0218	0.0220	0.0467
500	100	15	0.0052	0.0051	0.0081
1000	100	15	0.0032	0.0032	0.0050
500	200	15	0.0044	0.0044	0.0082
1000	200	15	0.0025	0.0025	0.0041
500	100	50	0.0044	0.0043	0.0078
1000	100	50	0.0023	0.0023	0.0040
500	200	50	0.0042	0.0041	0.0081
1000	200	50	0.0022	0.0022	0.0040

Table 5: Comparison among three estimators. Values of NB, MA and the conservative NB estimates for the case of the simple mean. The universe sample size is 50, 100, 500 and 1000.

	k-fold	MA	Variance	NB	Variance
n=100	4-fold	0.02096	0.00003302	0.0417	0.001262
	5-fold	0.02093	0.00003293	0.04516	0.0009909
	10-fold	0.02089	0.0000328	0.04426	0.0005567
	LOOCV	0.02086	0.0000327	0.04141	0.0002177
n=1000	4-fold	0.002	3.02E-08	0.00423	1.308E-05
	5-fold	0.002	3.02E-08	0.00412	8.60E-06
	10-fold	0.002	3.02E-08	0.00405	3.74E-06
	LOOCV	0.002	3.02E-08	0.00398	2.00E-07

Table 6: Variance estimators for k-fold CV. Moment approximation and Nadeau-Bengio variance estimators for k-fold cross-validation estimators of the generalization error and their variances.

When the data universe is 100 the 4-fold cross validation divides it into 4 non-overlapping test sets each containing 25 data points. Similarly, we define 5-fold and 10-fold cases. We notice that the variance estimation of LOOCV is not appreciably better than that of the other cross validation estimators. In fact, the slight advantage of the LOOCV diminishes when the data universe is large and the size of the test set becomes large. For illustration purposes we present the NB estimator and its variance. The value of the NB estimator is twice as large as the value of the moment

approximation estimator. However, note that Nadeau and Bengio (2003) do not discuss the case of k-fold cross validation.

K	MSE	Var	Bias
4	0.02123	0.02098	0.002283
10	0.02099	0.02091	0.002224

Table 7: Comparison between 10-fold and 4-fold Cross Validation Under Simple Mean Case. MA estimator is used to estimate the variance of the cross -validation estimator of the generalization error. The results reported in the table are averages over 100 different data sets.

To understand the effect of the loss function in the performance of the methods we used the mean squared error (MSE) to compare the estimators as well as their variance. Table 7 presents the values of the MSE and the variance, as well as the bias for the 4-fold and 10-fold estimators of variance for the simple mean case. We see that the reduction in variance between the 4-fold and 10-fold CV variance estimator is not appreciably different. This difference is more pronounced when the corresponding MSE are compared. Overall it appears that the 10-fold cross validation differs from the 4-fold cross validation an order of magnitude less when the comparison between the two is made on the basis of variance than when the comparison is made on the basis of MSE.

4.2 Absolute Error Loss

The previous theory was developed for loss functions that are differentiable. One loss that is not differentiable at the mean is the absolute error loss. However, we are able to apply the above theory in the case of the absolute error loss because we can replace $|\bar{X}_{S_j} - X_i|$ by the equivalent function $\sqrt{(\bar{X}_{S_j} - X_i)^2 + d}$, where d is a small positive number. The function $[(\bar{X}_{S_j} - X_i)^2 + d]^{1/2}$ replaces the absolute error loss and is differentiable everywhere. We use $d = \frac{1}{n}$ and $\frac{1}{n_2}$ and computed the Nadeau-Bengio estimate and the moment approximation estimate for the sizes of the data universe of 100 and 500. Notice that the Nadeau-Bengio estimate was computed using $L(\bar{X}_{S_j}, X_i) = |X_i - \bar{X}_{S_j}|$, while the moment approximation estimator uses the loss function $L(\bar{X}_{S_j}, X_i) = [(X_i - \bar{X}_{S_j})^2 + d]^{1/2}$, which is almost the same with the absolute error loss. We generate data from a $N(0,5)$ distribution in S-plus and used $J = 15$.

Table 8 shows the values of the Nadeau-Bengio and moment approximation estimators together with their sample variances. Notice that $d = \frac{1}{n}$ was used in the first computation of the moment approximation estimator, where n is the size of the data universe, and $d = \frac{1}{n_2}$, where n_2 is the size of the test set was used in the second computation. The table reports results that are averaged over 100 different data sets.

The first observation we make is that the effect of d on the moment approximation estimator and its sample variance is almost undetectable, as the values of the estimator and its sample variance (averaged over 100 different data sets) do not change with d being $\frac{1}{n}$ or $\frac{1}{n_2}$. Secondly, we see that the variance of the Nadeau-Bengio estimator is larger than the variance of the moment approximation estimator by one order of magnitude.

n_2	NB estimator	var(NB)	MA estimator	var(MA)
$d = \frac{1}{n}$				
10	0.0287	1.16e-04	0.0293	1.43e-05
15	0.0271	1.25e-04	0.0252	1.06e-05
20	0.0256	7.93e-05	0.0231	8.93e-06
25	0.0224	7.72e-05	0.0219	7.98e-06
30	0.0218	8.77e-05	0.0210	7.36e-06
35	0.0207	7.53e-05	0.0204	6.92e-06
40	0.0208	6.52e-05	0.0199	6.58e-06
45	0.0191	6.14e-05	0.0194	6.30e-06
50	0.0205	7.07e-05	0.0191	6.06e-06
$d = \frac{1}{n_2}$				
10	0.0287	1.16e-04	0.0291	1.43e-05
15	0.0271	1.25e-04	0.0251	1.06e-05
20	0.0256	7.93e-05	0.0231	8.92e-06
25	0.0224	7.72e-05	0.0218	7.97e-06
30	0.0218	8.77e-05	0.0210	7.36e-06
35	0.0207	7.53e-05	0.0204	6.91e-06
40	0.0208	6.52e-05	0.0199	6.58e-06
45	0.0191	6.14e-05	0.0194	6.30e-06
50	0.0205	7.07e-05	0.0191	6.06e-06

Table 8: Absolute Error Loss Case n=100, J=15. Nadeau-Bengio (NB) and moment approximation (MA) estimators and their corresponding variance estimates. Data are N(0,5) and J=15. The loss function is absolute error.

Table 9 presents the Nadeau-Bengio and moment approximation estimators but now the value of $J = 50$. Notice that, in contrast with the square error loss case, the Nadeau-Bengio estimator has a higher variance than the moment approximation estimator. Its variance is still an order of magnitude higher than the variance of the moment approximation estimator.

Table 10 presents the two estimators and their corresponding sample variances when the size of the data universe is 500. The population is still N(0,5) and $d = 1/n$. Notice that for $J = 15$ the NB estimate has larger, by two orders of magnitude, variance than the moment approximation estimator, while $J = 50$ it still maintains a larger than the moment approximation estimator variance, only this time by one order of magnitude.

4.3 Regression

In the regression case the data generation was done as follows. The model adopted was simple regression, that is $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, 2, \cdots, n$, where ε_i are independent, mean 0 and variance 1, normal random variables. The parameters α, β were set to equal 2 and 3 respectively. The explanatory variable was generated from a uniform distribution with range [0,10]. Finally, we generated the errors from a N(0,1) distribution and $y_i = 2 + 3x_i + \varepsilon_i$, $i = 1, 2, \cdots, 100$. We generated 100 different

n_2	NB	var(NB)	MA	var(MA)
$d = \frac{1}{n}$				
10	0.0208	3.18e-05	0.0216	7.77e-06
15	0.0209	2.48e-05	0.0203	6.89e-06
20	0.0206	2.92e-05	0.0197	6.46e-06
25	0.0189	2.30e-05	0.0193	6.19e-06
30	0.0199	2.41e-05	0.0190	6.00e-06
35	0.0191	2.45e-05	0.0187	5.86e-06
40	0.0192	2.49e-05	0.0185	5.73e-06
45	0.0188	3.16e-05	0.0184	5.61e-06
50	0.0195	2.56e-05	0.0182	5.50e-06
$d = \frac{1}{n_2}$				
10	0.0208	3.18e-05	0.0214	7.75e-06
15	0.0209	2.48e-05	0.0202	6.88e-06
20	0.0206	2.92e-05	0.0196	6.45e-06
25	0.0189	2.30e-05	0.0192	6.19e-06
30	0.0199	2.41e-05	0.0190	6.00e-06
35	0.0191	2.45e-05	0.0187	5.86e-06
40	0.0192	2.49e-05	0.0185	5.73e-06
45	0.0188	3.16e-05	0.0183	5.61e-06
50	0.0195	2.56e-05	0.0182	5.50e-06

Table 9: Absolute Error Loss Case n=100, J=50. Nadeau-Bengio (NB) and moment approximation (MA) estimators and their sample variance. Data are $N(0,5)$ and J=50. The loss function is absolute error.

data sets; for each data set, and for each value of n_2, n_1 we computed the Nadeau-Bengio and the moment approximation estimator and then average those over the 100 different data sets.

Tables 11 and 12 present the two estimators together with their corresponding sample variances and for values of J equal to 15 and 50. Notice that the moment approximation estimator has variance that is at least one order of magnitude smaller than the variance of Nadeau-Bengio estimator.

Table 13 computes the NB and moment approximation variance estimators of the generalization error when the size of the data universe is 500. We see that the moment approximation estimator still maintains a variance of an order of magnitude lower than the NB estimator.

We also computed the variance estimators for k-fold cross validation estimators of the generalization error in the regression case. Table 14 shows the value of the moment approximation and Nadeau-Bengio estimator and their sample variances computed over 100 different data sets of size 100.

Again, the advantage of LOOCV in this case is questionable. Moreover, given the fact that 4-fold cross validation saves a lot of computing time it seems to be preferable to use (recall that 4-fold CV assigns 25% of the data points in the test set).

n_2	NB	var(NB)	MA	var(MA)
$J = 15$				
50	0.00651	8.31e-06	0.00588	1.09e-07
75	0.00527	3.19e-06	0.00506	8.04e-08
100	0.00455	3.23e-06	0.00465	6.79e-08
125	0.00459	2.62e-06	0.00440	6.09e-08
150	0.00428	3.10e-06	0.00424	5.64e-08
175	0.00420	2.55e-06	0.00412	5.33e-08
200	0.003971	2.41e-06	0.00403	5.10e-08
225	0.00390	1.83e-06	0.00396	4.92e-08
250	0.00361	2.03e-06	0.00390	4.78e-08
$J = 50$				
50	0.00456	1.05e-06	0.00433	5.90e-08
75	0.00402	6.61e-07	0.00409	5.25e-08
100	0.00406	9.03e-07	0.00396	4.93e-08
125	0.00404	7.45e-07	0.00389	4.75e-08
150	0.00396	7.16e-07	0.00384	4.62e-08
175	0.00388	8.07e-07	0.00380	4.54e-08
200	0.00377	5.23e-07	0.00377	4.47e-08
225	0.00377	5.67e-07	0.00375	4.41e-08
250	0.00365	6.26e-07	0.00373	4.36e-08

Table 10: Absolute Error Loss Case n=500, $d = \frac{1}{n}$. Nadeau-Bengio (NB) and moment approximation (MA) estimators and their sample variance. The size of the data universe is 500.

4.4 Classification

In this section we briefly indicate how these results can possibly be extended to the classification case. We present some ideas that appear promising in treating this case and a very limited simulation experiment in the simplest case, where the prediction rule is based on the mean of the training set. The results presented here are promising; however, we would like to stress that a more detailed study than the one presented here, is required to understand the performance of these methods in classification.

Recall that a central requirement on the loss function is to be differentiable. In the classification case the loss function is an indicator function and hence it is discontinuous at one point. The idea is to replace the discontinuous function by a continuous, differentiable function that is close to the original loss function. We approximate therefore the indicator function by a polynomial of order 3. Let the data be (x_i, g_i), $i = 1, \cdots, n$, where x_i indicates the data value, and g_i indicates the group membership. Assume that there are only two groups in the population; then $g_i = 1$ if x_i belongs in group 1 and $g_i = 2$ if x_i belongs in group 2. Moreover, assume that group 1 has smaller mean than group 2. The prediction rule we use states that if $\bar{X}_{S_j} - X_k > 0$ then X_k belongs in group 1, otherwise it belongs in group 2. Therefore, $\hat{g}_k$ is either 1 or 2 depending on whether $\bar{X}_{S_j} - X_k$ is greater than 0 or less than or equal to 0. The loss function is then $I(g_k \neq \hat{g}_k)$.

n_2	NB	var(NB)	MA	var(MA)
10	0.0327	0.000493	0.0326	1.14e-04
15	0.0293	0.000366	0.0284	8.44e-05
20	0.0259	0.000184	0.0260	7.21e-05
25	0.0242	0.000199	0.0247	6.29e-05
30	0.0235	0.000168	0.0238	5.74e-05
35	0.0226	0.000176	0.0232	5.66e-05
40	0.0235	0.000144	0.0227	5.35e-05
45	0.0249	0.000255	0.0223	5.16e-05
50	0.0233	0.000142	0.0221	5.06e-05

Table 11: Regression case n=100, J=15. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance of the cross validation estimator of the generalization error and their sample variances in the regression case. The value of J is 15, and the results are averages over 100 independent data sets. The size of the data universe is 100.

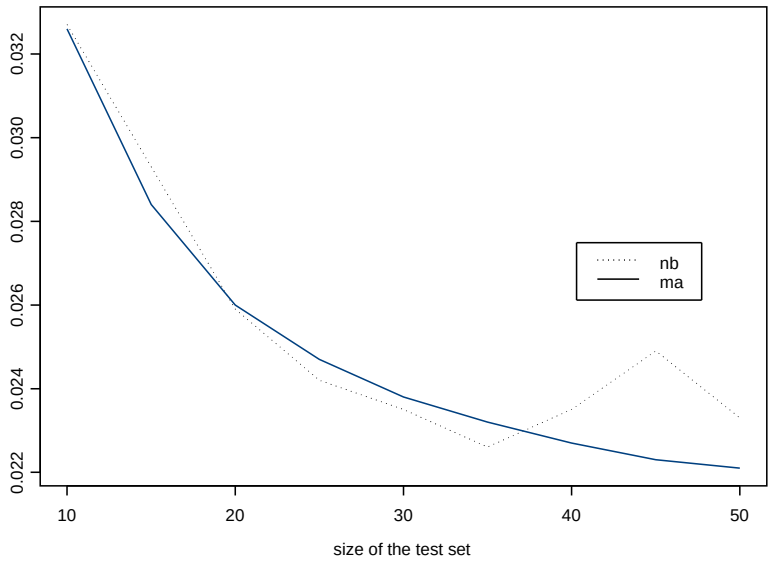

Figure 3: Regression case n=100, J=15

We can write this loss function as a function of $z_k = \bar{x}_{S_j} - x_k$, $\delta_k = I(g_k = 1)$ and two continuous differentiable functions L_{k1} and L_{k2}. Thus

$$I(g_k \neq \hat{g}_k) = \delta_k L_{k1} + (1 - \delta_k) L_{k2},$$

n_2	NB	var(NB)	MA	var(MA)
10	0.0253	1.84e-04	0.0242	6.00e-05
15	0.0233	1.29e-04	0.0229	5.41e-05
20	0.0228	1.24e-04	0.0222	5.06e-05
25	0.0223	1.15e-04	0.0218	4.92e-05
30	0.0219	1.07e-04	0.0215	4.79e-05
35	0.0222	1.10e-04	0.0213	4.70e-05
40	0.0215	1.00e-04	0.0212	4.63e-05
45	0.0231	1.31e-04	0.0211	4.60e-05
50	0.0231	9.56e-05	0.0210	4.54e-05

Table 12: Regression case n=100, J=50. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance of the cross validation estimator of the generalization error and their sample variances in the regression case. The value of J is 50, and the results are averages over 100 independent data sets. The size of the data universe is 100.

n_2	NB	var(NB)	MA	var(MA)
50	0.00653	7.64e-06	0.00643	8.94e-07
75	0.00563	4.80e-06	0.00555	6.71e-07
100	0.00498	4.10e-06	0.00511	5.92e-07
125	0.00470	3.86e-06	0.00483	5.02e-07
150	0.00495	4.35e-06	0.00464	4.54e-07
175	0.00469	3.57e-06	0.00452	4.32e-07
200	0.00450	2.42e-06	0.00443	4.16e-07

Table 13: Regression case n=500, J=15. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance of the cross validation estimator of the generalization error and their sample variances in the regression case. The value of J is 15, and the results are averages over 100 independent data sets. The size of the data universe is 500.

where

$$Ł_{k1} = \begin{cases} 1 & ,z_k < -h \\ \frac{2}{h^3}z_k^3 + \frac{3}{h^2}z_k^2 & ,-h \le z_k < 0 \\ 0 & ,z_k \ge 0 \end{cases}$$

$$Ł_{k2} = \begin{cases} 0 & ,z_k < 0 \\ -\frac{2}{h^3}z_k^3 + \frac{3}{h^2}z_k^2 & ,0 \le z_k < h. \\ 1 & ,z_k \ge h \end{cases}$$

The needed terms then can be easily computed. For example, we can compute expectation of the above loss function as

$$E\{E(\delta_k L_{k1} + (1-\delta_k)L_{k2}|\delta_k)\} = P(\delta_k = 1)E(L_{k1}|\delta_k = 1) + P(\delta_k = 0)E(L_{k2}|\delta_k = 0)$$

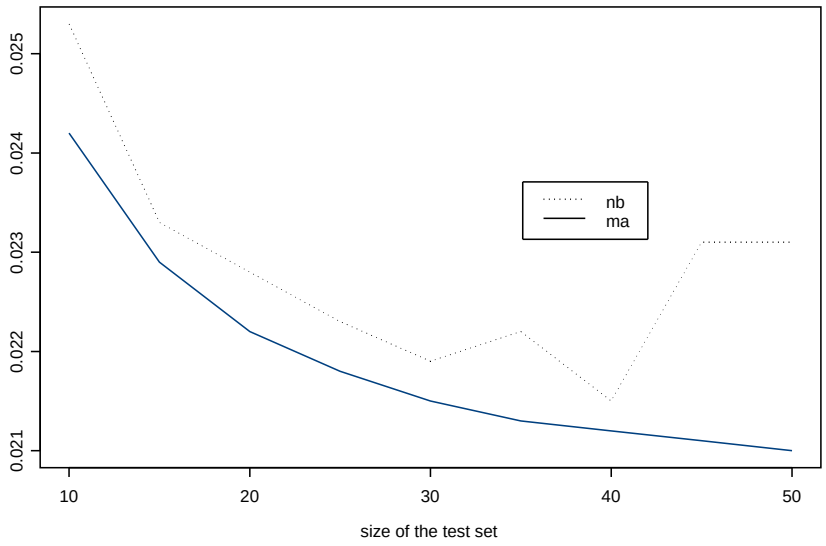

Figure 4: Regression case n=100, J=50

k-fold	MA	Variance	NB	Variance
4-fold	0.02132	0.0000357	0.04854	0.00227
5-fold	0.02135	0.0000358	0.04634	0.00121
10-fold	0.02138	0.0000359	0.04493	0.00062
LOOCV	0.02139	0.0000359	0.04323	0.00023

Table 14: Variance estimators in regression. Variance estimators of k-fold cross-validation estimator of the generalization error and their sample variances, in regression.

and the terms $P(\delta_k = 1)$, $P(\delta_k = 0)$ are computed from the data. Similarly, we can compute from the data all terms that involve variance and covariance terms.

Table 15 presents the results obtained from a small scale simulation. Data were generated in *Splus* from two groups of normal distributions; these were $N(3,1)$ and $N(1,1)$. Group membership is assigned by generating a Bernoulli(0.6) random variable. If the value of 1 is obtained then the data point is generated from a $N(1,1)$ distribution, otherwise it is generated from a $N(3,1)$. The training set used 80% of the available data points. For example, when $n = 200$ the training set contains 160 elements and thus $n_2 = 40$. The value of h in constructing the L_{k1}, L_{k2} functions was taken to be 0.1.

Table 15 shows the moment approximation variance estimator and NB estimator for various values of the data universe. For illustration reasons we present the values of the MA estimator for both cases when normality is assumed and when is not. We see that the moment approximation estimator (computed without any distributional assumption) is very competitive.

Table 15: Simple Classification Example

n	MA.Free	MA.Normal	NB
200	0.0008355	0.0008275	0.0009240
2000	0.00008593	0.00008273	0.00010028
20000	0.000008603	0.000008299	0.000008815

Table 15. Moment approximation (MA) and Nadeau-Bengio (NB) estimators of the variance of the cross validation estimator of the generalization error and their sample variances in the simple classification case. The value of J is 15, MA.Free denote the MA estimator without distribution assumption and MA.Normal denote the MA estimator under normal distribution. The results are averages over 100 independent data sets. 80 percent of the data are used as training data; h used here is 0.1

5. Discussion and Conclusion

We presented a method for deriving variance estimators of the cross validation estimator of the generalization error in the cases of smooth loss functions and the absolute error loss. The approximation we propose illustrates clearly the role of the training and test sets in the estimation of the variance of the generalization error. We also provide a unifying framework, under which we can obtain variance estimators of the estimators of the generalization error for both, complete random sampling and non-random test set selection.

We compared the moment approximation estimators with an estimator proposed by Nadeau and Bengio (2003). The results indicate that the moment approximation estimators perform better in terms of both, variance and bias, than the Nadeau and Bengio (2003) estimator. The new estimators use additional information from both the data and the learning algorithm. On the other hand, the Nadeau and Bengio estimator is computationally simpler than the moment approximation estimator for general loss functions, as it does not require the computation of the derivatives of the loss function. In the case of non-random test set selection, the Nadeau-Bengio estimator is not appropriate to use. The moment approximation estimator in this case is a reasonable estimator and can be computed. It is interesting to notice that the results indicate against use of the leave-one-out cross validation (LOOCV). Its slight advantage is terms of variance, over the other forms of cross-validation quickly diminishes as the size of the universe, and hence the size of the test set of other cross validation schemes increases. Overall, a test set that use 25% of the available data seems to be a reasonable compromise in selecting among the various forms of k-fold cross validation.

We presented results for general differential loss functions and for absolute error loss. We also indicated possible extensions of this methodology to the classification problem and discussed briefly a very simple version of the classification problem. An extensive study of this problem will be the subject of a different paper. Finally, we would like to indicate here that the methods presented here can similarly apply to SVM loss function as well as the kernel regression.

Acknowledgments

The first author would like to acknowledge support for this project from the National Science Foundation (NSF grants DMS-0072319 and DMS-0504957) . The last author would like to acknowledge

the support from the National Library of Medicine (R01-LM08910), NIH. The authors would like to thank two referees for their constructive suggestions that improve the presentation of the paper.

Appendix A.

Here we present a series of lemmas that guarantee that the remainder term in the approximations for the case of sample mean.

Before we state these we need the following definitions.

Definition 1. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. We say that a random variable X belongs in the $\mathcal{L}_p$ space if $E|X|^p < \infty$, $p > 0$.

Definition 2. A sequence of random numbers R_n is said to be $O(1/k_n)$ if $\exists\, M$ and n_0 such that $|k_n R_n| < M$, $\forall n > n_0$, or, equivalently, $k_n R_n$ is bounded.

Lemma A.1 Let X, Y be independent random variables and $X+Y \in \mathcal{L}_r$ for some $r \in (0,\infty)$. Then $X \in \mathcal{L}_r$ and $Y \in \mathcal{L}_r$.

Proof. For a large $\lambda_0 > 0$, $\forall \lambda > \lambda_0$

$$\begin{aligned} P(|X| > \lambda) &\leq 2P(|X| > \lambda,\ |Y| < \frac{\lambda}{2}) \\ &\leq 2P(|X+Y| > \frac{\lambda}{2}), \end{aligned}$$

If $E|X|^r < +\infty$, then $E|X|^r = \int_0^\infty P[|X|^r > \lambda] d\lambda$. Hence, if $X+Y \in \mathcal{L}_r$,

$$\begin{aligned} \int_{\lambda \geq \lambda_0} P(|X|^r > \lambda) &= \int_{\lambda \geq \lambda_0} P(|X| > \lambda^{\frac{1}{r}}) d\lambda \\ &\leq 2 \int_{\lambda \geq \lambda_0} P(|X+Y| > \frac{\lambda^{\frac{1}{r}}}{2}) d\lambda \\ &= 2 \int_{\lambda \geq \lambda_0} P(|X+Y|^r > \frac{\lambda}{2^r}) d\lambda < \infty. \end{aligned}$$

Thus, $E|X|^r < \infty$. The proof for $E|Y|^r < \infty$ is similar.

Lemma A.2 If $0 < r' < r$ and $E|X|^r < \infty$, then $E|X|^{r'} < \infty$.

Proof. Write

$$(E|X|^{r'})^{r/r'} \leq E(|X|^{r'})^{r/r'} = E|X|^r < \infty,$$

and the proof is obtained by Jensen's inequality.

Lemma A.3 If $E|\bar{X}_n|^p < \infty$, then $E|X_1|^p < +\infty$, where $p \in \mathcal{Z}^+$, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean.

Proof. We will use transfinite induction. For $n = 1$ and $n = 2$, it is trivial since $\bar{X}_n = X_1$. For $n = 2$, $\bar{X}_2 = \frac{1}{2}(X_1 + X_2)$ and use lemma 1 to obtain the result, relying on the fact that X_1, X_2 are identically distributed. Suppose now that for $n \leq k-1$ the result holds. We will prove it true for $n = k$. Write

$$E(|\bar{X}_k|^p) = E(|\frac{1}{k}\sum_{i=1}^k X_i|^p) = E(|\frac{1}{k}(\sum_{i=1}^{k-1} X_i + X_k)|^p).$$

Thus

$$E|\bar{X}_k|^p = E|\frac{1}{k}X_k + \frac{k-1}{k}\bar{X}_{k-1}|^p,$$

and using lemma 1, we obtain $E(|X_k|) < \infty$.

Lemma A.4 Let $n > 2k$ and $a_1, a_2, \cdots, a_n$ be such that

$$\left.\begin{array}{l} a_1 + a_2 + \cdots + a_n = 2k \\ a_i \in Z, a_i \geq 0, a_i \neq 1 \end{array}\right\} \quad (1)$$

$$\left.\begin{array}{l} a_1 + a_2 + \cdots + a_n = 2k-1 \\ a_i \in Z, a_i \geq 0, a_i \neq 1 \end{array}\right\} \quad (2)$$

Then the number of solutions for (1) and (2), denoted by $A_n(2k)$ and $A_n(2k-1)$ respectively, satisfy $A_n(2k) = O(n^k)$, and $A_n(2k-1) = O(n^{k-1})$.

Proof. The maximal order of the $A_n(2k)$ comes from the $\{(2,\cdots,2),(0,\cdots,0)\}$, where $(2,\cdots,2)$ is a k-tuple. There are $\binom{n}{k}$ solutions for (1) of this form. The order is $O(n^k)$, because

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = O(n^k).$$

The maximal order of the $A_n(2k-1)$ comes from the $\{(2,\cdots,2,3),(0,\cdots,0)\}$, where the k-tuple $(2,2,\cdots,2,3)$ has $k-1$ elements equal to 2. There are $\binom{n}{k-1}$ solutions of (2) of this form. The order is $O(n^{k-1})$ because

$$\binom{n}{k-1} = \frac{n(n-1)\cdots(n-k+2)}{(k-1)!} = O(n^{k-1}).$$

Lemma A.5 Let $X_1, X_2, \cdots, X_n$ be independent identically distributed random variables with $E(X_i) = \mu$, and k is a positive integer. Then $E(\bar{X}-\mu)^{2k-1}$ and $E(\bar{X}-\mu)^{2k}$, if they exist, are both $O(1/n^k)$.

Proof. Without loss of generality, we suppose $E(X) = \mu = 0$, then

$$E(\bar{X}_n^{2k}) = \frac{1}{n^{2k}}E(\sum_{i=1}^{n} X_i)^{2k} = \frac{O(n^k)}{n^{2k}} = O(n^{-k}),$$

$$E(\bar{X}_n)^{2k-1} = \frac{1}{n^{2k-1}}E(\sum_{i=1}^{n} X_i)^{2k-1} = \frac{O(n^{k-1})}{n^{2k-1}} = O(n^{-k}).$$

Appendix B.

Here we present the set up we use for the linear regression case and lemmas that guarantee the validity of the obtained results.

The Gauss-Markov set up for a linear model defines $y_i = x_i^T\beta + \varepsilon_i$, where $y_1, y_2, \cdots, y_n$ are observable response variables and $X = (x_{ij})$ is an $n_1 \times p$ matrix of known constants. Moreover $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are unobservable random variables that follow a probability distribution F, and are

such that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent. The least square solution is $\hat{\beta} = (X^T X)^{-1} X^T Y$, where Y is an $n_1 \times 1$ vector, so that $E\hat{\beta} = \beta$ and $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

Consider an arbitrary linear combination $U_n = \lambda^T (\hat{\beta} - \beta)$, $\lambda \in \mathbb{R}^P$. Then $U = \lambda^T (X^T X)^{-1} X^T \varepsilon_i$ with $c = \lambda^T (X^T X)^{-1} X^T$. To obtain the asymptotic distribution of U all is needed is to verify that c satisfies the regularity condition of Hajek-Sidak central limit theorem.

We need first the following definition.

Definition(Convergence in distribution). A sequence $\{T_n\}$ of random variables with distributions $\{F_n\}$ is said to converge in distribution (or in law) to a (possible degenerate) random variable T with a distribution function F, if for every $\varepsilon > 0$, there exists $n_0 = n_0(\varepsilon)$, $n_0 \in \mathbf{Z}^+$ such that at every point of continuity x of F

$$|F_n(x) - F(x)| < \varepsilon,$$

for all $n \geq n_0$.

Hajek-Sidak Central Limit Theorm(Sen and Singer, 1993). Let $\{Y_n\}$ be a sequence of independent, identically distributed random variables with mean μ and variance σ^2 finite; let $\{C_n\}$ be a sequence of real vectors. Then if $C_n = (c_{n1}, c_{n2}, \cdots, c_{nn})^T$ and

$$\frac{max_{1 \leq i \leq n} c_{ni}^2}{\sum_{i=1}^n c_{ni}^2} \to 0, \text{ as } \quad n \to +\infty$$

it follows that

$$Z_n = \frac{\sum_{i=1}^n c_{ni}(Y_i - \mu)}{\sqrt{\sigma^2 \sum_{i=1}^n c_{ni}^2}} \xrightarrow{D} Z$$

where Z is a $N(0,1)$ random variable.

The following theorem completes the proof of the asymptotic distribution of the least squares estimator.

Cramer-Wold Theorem (Sen and Singer, 1993). Let $X_1, X_2, \cdots$ be random vectors in $\mathbf{R}^p$; then $X_n \xrightarrow{D} X$ if and only if, for every fixed $\lambda \in \mathbf{R}^p$ we have $\lambda^T X_n \xrightarrow{D} \lambda^T X$.

Remark: We note here that the generalized Noether condition (assumption 2) can be modified to extend the asymptotic normality result to the heteroscedastic model, that is, the model where $E(\varepsilon) = \sigma_i^2$, $i = 1, 2, \cdots, n_1$. Also notice that the normality of the least squares estimators is not obtained under normality of the errors. Assumptions 1 and 2 of section 3.2 together with the finiteness of the second moment of the, otherwise unknown, error distribution suffices for these results to hold.

The following lemmas that are listed without proof are used to arrive at the given form of the covariance terms.

Lemma B.1 Let U be distributed as a $N(0,V)$ random variable. Then

$$Var(U^T AU) = 2tr(AV)^2$$

where A is a known matrix.

Lemma B.2 Let U be distributed as a $N(\mu,V)$ random variable. Then

(i) $E(U^TAU) = tr(AV) + \mu^T A\mu$,

(ii) $Cov(U, U^TAU) = 2VA\mu$,

(iii) $Cov(U^TPU, U^TQU) = 2tr[PVQV] + 4\mu^T PVQ\mu$.

The following lemmas are used in establishing the equivalence of the different cases in the computation of the covariance terms. The first lemma, the well-know Holder's inequality, is stated without proof.

Lemma B.3 Denote by $||X||_p = E^{1/p}(|X|^p)$, $p > 0$, where X is a random variable, the p-norm of X. Then, if X, Y are measurable functions on a probability space for $p > 1$, $p' > 1$, $\frac{1}{p} + \frac{1}{p'} = 1$

$$E|XY| \leq ||X||_p \cdot ||Y||_{p'}.$$

The special case where $p = p' = 2$ is known as Schwarz's inequality.

Lemma B.4 Let S_j, $S_{j'}$ be training sets and S_j^c, $S_{j'}^c$ their corresponding test sets. Assume that for $(y_i, x_i) \in S_j^c$, $(y_i, x_i) \in S_j$, for some $i \in \{1, 2, \cdots, n_2\}$. Assume that $E([L'(x_{i'}^T\beta_0, y_{i'})]^2) < \infty$, and $E[L^4(x_{i'}^T\beta_0, y_{i'})] < \infty$,

$$\begin{aligned} sup_{||\hat{\beta}_{S_{j'}} - \beta_0|| \leq k/\sqrt{n_1}} \quad & |E[L(x_i\beta_0, y_i)L'(x_{i'}^T\beta_0, y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \\ - \quad & E[L(x_i\beta_0, y_i)]E[L'(x_{i'}^T\beta_0, y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]| = o(1) \end{aligned}$$

Proof Write

$$\begin{aligned} & |E[L(x_i\beta_0, y_i)L'(x_{i'}^T\beta_0, y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] - E[L(x_i\beta_0, y_i)]E[L'(x_{i'}^T\beta_0, y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]| \\ & \leq |E[L(x_i\beta_0, y_i)L'(x_{i'}^T\beta_0, y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]| + |E[L(x_i\beta_0, y_i)]E[L'(x_{i'}^T\beta_0, y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \\ & \leq E\{|L(x_i\beta_0, y_i)x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)||L'(x_{i'}^T\beta_0, y_{i'})|\} + |E[L(x_i\beta_0, y_i)]||E[L'(x_{i'}^T\beta_0, y_{i'})]||E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \end{aligned}$$

Using lemma A2.3 and the fact that $E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] = 0$ the above relationship becomes:

$$\begin{aligned} & |E[L(x_i\beta_0, y_i)L'(x_{i'}^T\beta_0, y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] - E[L(x_i\beta_0, y_i)]E[L'(x_{i'}^T\beta_0, y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]| \\ & \leq \sqrt{E([L(x_{i'}^T\beta_0, y_{i'})]^2)}\sqrt{E[L^2(x_i^T\beta_0, y_i)(\hat{\beta}_{S_{j'}} - \beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]} \end{aligned}$$

Apply once more Lemma A2.3 on

$$\begin{aligned} & E[L^2(x_i^T\beta_0, y_i)(\hat{\beta}_{S_{j'}} - \beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \\ & \leq \sqrt{E[L^4(x_i^T\beta_0, y_i)}\sqrt{E[(\hat{\beta}_{S_{j'}} - \beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)]} \end{aligned}$$

Thus

$$
\begin{aligned}
&sup_{||\hat{\beta}_{S_{j'}}-\beta_0||\leq k/\sqrt{n_1}}|E[L(x_i\beta_0,y_i)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]\\
&-E[L(x_i\beta_0,y_i)]E[L'(x_{i'}^T\beta_0,y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|\\
&\leq sup_{||\hat{\beta}_{S_{j'}}-\beta_0||\leq k/\sqrt{n_1}}M\cdot\sqrt[4]{E[(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]^2}\\
&\leq M\cdot\sqrt[4]{[(\sum_{l=1}^{p}x_{i',l})^2\frac{k^2}{p^2n_1}]^2}\\
&\leq \frac{1}{\sqrt{n_1}}[\frac{M_k}{p}(\sum_{l=1}^{p}x_{i',l})]\\
&\leq \frac{c}{\sqrt{n_1}}
\end{aligned}
$$

where $c=\frac{M_k}{p}(\sum_{l=1}^{p}x_{i',l})<\infty$.

Lemma B.5 Let S_j,$S_{j'}$ be two training sets and S_j^c, $S_{j'}^c$ be their corresponding test sets. Under the assumption that $E[L''(x_i^T\beta_0,y_i)]$ is finite and for some $(y_i,x_i)\in S_j^c$, $(y_i,x_i)\in S_{j'}$.

$$
\begin{aligned}
&sup_{||\hat{\beta}_{S_{j'}}-\beta_0||\leq k/\sqrt{n_1}}|E[L(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]\\
&-E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]E[(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|=o(1)
\end{aligned}
$$

Proof. Write

$$
\begin{aligned}
&|E[L(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]\\
&-E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]E[(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|\\
&\leq|E[L(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|\\
&+|E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]E[(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|
\end{aligned}
$$

The first term of the above relationship gives:

$$
\begin{aligned}
&|E[L(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|\\
&\leq E|\{L(x_i^T\beta_0,y_i)(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)\}L''(x_{i'}^T\beta_0,y_{i'})|\\
&\leq\sqrt{E[L''(x_{i'}^T\beta_0,y_{i'})^2]}\sqrt{E[L^2(x_i^T\beta_0,y_i)((\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)^2)]}\\
&\leq\frac{c}{n_1}
\end{aligned}
$$

where c is a constant. The second term is

$$
\begin{aligned}
&|E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]E[(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]|\\
&\leq E[|L(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})|]E[x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]^2\\
&\leq|E[|L(x_i^T\beta_0,y_i)|]E[|L''(x_{i'}^T\beta_0,y_{i'})|]\frac{c_1}{n_1}\\
&\leq\frac{c^*}{n_1}
\end{aligned}
$$

where c^* is a constant. Thus the lemma is proved. Similarly we can prove that the terms, in the computation of covariance, where $(y_i, x_i) \in S_{j'}$ and/or $(y_{i'}, x_{i'}) \in S_j$ can be replaced and treated as the case where $(y_i, x_i) \notin S_{j'}$ and/or $(y_{i'}, x_{i'}) \notin S_j$ in the neighborhood of the true value of β_0.

Lemma B.6 Suppose

$$x = \begin{pmatrix} u \\ v \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

where u is $q \times 1$ vector, v is $(p-q) \times 1$ vector, a is a known $q \times 1$ vector, B is known $(p-q) \times (p-q)$ matrix.
Then

$$E(a^T u v^T B v) = 0.$$

Proof: Using conditional probability argument, we have

$$\begin{aligned}
&E(a^T u v^T B v) = E_u\{E_{v|u}[a^T u v^T B v]\} \\
&= E_u\{a^T u E_{v|u}[v^T B v]\} \\
&= E_u\{a^T u[tr(B\Sigma_{22\cdot 1}) - (\Sigma_{21}\Sigma_{11}^{-1}u)^T B(\Sigma_{21}\Sigma_{11}^{-1}u)]\} \\
&= E_u\{a^T u(\Sigma_{21}\Sigma_{11}^{-1}u)^T B(\Sigma_{21}\Sigma_{11}^{-1}u)\} \\
&= E_u\{a^T u u^T \Sigma_{11}^{-1}\Sigma_{12} B(\Sigma_{21}\Sigma_{11}^{-1}u)\} \\
&= E_u\{a^T u u^T C u)\} \\
&= a^T E_u\{u u^T C u\} \\
&= a^T\{Cov(u, u^T C u) + E u E(u^T C u)\} \\
&= a^T 2\Sigma_{22} C \cdot 0 + 0 \\
&= 0
\end{aligned}$$

where $c = \Sigma_{11}^{-1}\Sigma_{12} B \Sigma_{21}\Sigma_{11}^{-1}$. We use the property that if x is $N(\mu, V)$, then $cov(x, x^T A x) = 2vA\mu$.

Appendix C.

Proof of Proposition 3.5: To obtain the approximation given above we need first an approximation for the product $L(\hat{y}_{i,S_j},y_i)L(\hat{y}_{i',S_{j'}},y_{i'})$. Using expansion (3.13) we obtain:

$$\begin{aligned}
L(\hat{y}_{i,S_j},y_i)L(\hat{y}_{i',S_{j'}},y_{i'}) &= L(x_i\beta_0,y_i)L(x_{i'}^T\beta_0,y_{i'})+L(x_i\beta_0,y_i)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_j}-\beta_0)\\
&+ \frac{1}{2}L(x_i\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)\\
&+ L'(x_i\beta_0,y_i)x_i^T(\hat{\beta}_{S_j}-\beta_0)L(x_{i'}^T\beta_0,y_{i'})\\
&+ L'(x_i\beta_0,y_i)x_i^T(\hat{\beta}_{S_j}-\beta_0)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)\\
&+ \frac{1}{2}L(x_i^T\beta_0,y_i)x_i^T(\hat{\beta}_{S_j}-\beta_0)L'(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_{j'}}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)\\
&+ \frac{1}{2}L(x_{i'}^T\beta_0,y_{i'})L''(x_i^T\beta_0,y_i)(\hat{\beta}_{S_j}-\beta_0)x_ix_i^T(\hat{\beta}_{S_j}-\beta_0)\\
&+ \frac{1}{2}L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)L''(x_i^T\beta_0,y_i)(\hat{\beta}_{S_j}-\beta_0)x_ix_i^T(\hat{\beta}_{S_j}-\beta_0)\\
&+ \frac{1}{4}L''(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_j}-\beta_0)x_ix_i^T(\hat{\beta}_{S_j}-\beta_0)\\
&\quad (\hat{\beta}_{S_{j'}}-\beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)+R_n. \qquad (1)
\end{aligned}$$

We need the expectation, over everything random, of relationship (1). Assume first that $i\neq i'$. Recall that$(y_i,x_i)\in S_j^c$ and $(y_{i'},x_{i'})\in S_{j'}^c$ and (y_i,x_i) is independent of $(y_{i'},x_{i'})$. Then the first term of the above expansion is

$$E[L(x_i\beta_0,y_i)]E[L(x_{i'}^T\beta_0,y_{i'})]=(E[L(x_i\beta_0,y_i)])^2. \qquad (2)$$

(If $L(x_i^T\beta,y_i)=(x_i^T\beta_0-y_i)^2=\varepsilon_i^2$ and the $E(\varepsilon_i^2)=\sigma^2$).

We need now

$$E\{L(x_i^T\beta_0,y_i)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)\} \qquad (3)$$

Notice that all expectations here are conditional on X, that is, we treat the fixed design case. To evaluate this expectation we need to distinguish between two cases. The first corresponding to $(y_i,x_i)\notin S_{j'}$. In this case (3) equals 0. The second corresponds to $(y_i,x_i)\in S_{j'}$. Lemma B.4 of the appendix proves that (3) can be replaced by

$$E[L(x_i^T\beta_0,y_i)]E[L'(x_{i'}^T\beta_0,y_{i'})]E[x_{i'}^T(\hat{\beta}_{S_{j'}}-\beta_0)]=0. \qquad (4)$$

Therefore the second term is 0. Similarly, the expectation of the third term is

$$\frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}], \qquad (5)$$

in both cases, when $(y_i,x_i)\notin S_{j'}$ and when $(y_i,x_i)\in S_{j'}$.

The expectation of the fourth term of relationship (1) is 0. To evaluate the expectation of the fifth term we distinguish four cases:(i) $(y_i,x_i)\notin S_{j'}$ and $(y_{i'},x_{i'})\notin S_j$, (ii) $(y_i,x_i)\notin S_{j'}$ and $(y_{i'},x_{i'})\in S_j$,

(iii) $(y_i, x_i) \in S_{j'}$ and $(y_{i'}, x_{i'}) \notin S_j$, (iv) $(y_i, x_i) \in S_{j'}$ but $(y_{i'}, x_{i'}) \in S_j$. Lemma B.6 of the appendix allows in case (ii), (iii) and (iv), the replacement of the correct value of the expectation by the value obtained from expression (6) given below. Thus, the expectation of the fifth term is:

$$E[L'(x_i^T\beta_0, y_i)]E[L'(x_{i'}^T\beta_0, y_{i'})]x_i^T Cov(\hat{\beta}_{S_j}, \hat{\beta}_{S_{j'}})x_{i'}. \tag{6}$$

Since $S_j \bigcap S_{j'} \neq \emptyset$, and assuming the X_{S_j}, $X_{S_{j'}}$ have that upper $k \times p$ part common, relationship (.6) can be written as

$$\sigma^2(E[L'(x_i^T\beta_0, y_i)])^2 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_{i'},$$

where X_1 is of dimension $k \times p$, $k = Card(S_j \bigcap S_{j'})$, and σ^2 is the population variance. To compute the expectation of the sixth term we again distinguish between case (i), (ii), (iii) and (iv) as above. However, all cases reduce to the case (i). For this expectation we have from lemma B.6,

$$\frac{1}{2}(E[L'(x_i^T\beta_0, y_i)])E[L''(x_{i'}^T\beta_0, y_{i'})]E[x_i^T(\hat{\beta}_{S_j} - \beta_0)(\hat{\beta}_{S_{j'}} - \beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] = 0. \tag{7}$$

For the expectation of the seventh term we distinguish two cases: (i) $(y_{i'}, x_{i'}) \notin S_j$ and (ii) $(y_{i'}, x_{i'}) \in S_j$. Both cases can be treated using the following expression for the expectation of the seventh term:

$$\begin{aligned} & \frac{\sigma^2}{2}E[L(x_{i'}^T\beta_0, y_{i'})](E[L''(x_i^T\beta_0, y_i)])E[(\hat{\beta}_{S_j} - \beta_0)x_i x_i^T(\hat{\beta}_{S_j} - \beta_0)] \\ = \quad & \frac{\sigma^2}{2}E[L(x_{i'}^T\beta_0, y_{i'})](E[L''(x_i^T\beta_0, y_i)])tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]. \end{aligned} \tag{8}$$

The expectation of the eighth term is treated as the expectation of the sixth term, therefore it is given by relationship (7). For the expectation of last term we distinguish the four different cases that are listed above. In this case again all different cases can be treated as case (i). Therefore the expectation of the ninth term is

$$\frac{1}{4}E[L''(x_i^T\beta_0, y_i)]^2 E[(\hat{\beta}_{S_j} - \beta_0)^T x_i x_i^T(\hat{\beta}_{S_j} - \beta_0)(\hat{\beta}_{S_{j'}} - \beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \tag{9}$$

But

$$\begin{aligned} & E[(\hat{\beta}_{S_j} - \beta_0)^T x_i x_i^T(\hat{\beta}_{S_j} - \beta_0)(\hat{\beta}_{S_{j'}} - \beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_{j'}} - \beta_0)] \\ = \quad & 2tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}] \\ & + \sigma^4 tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \cdot tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}]. \end{aligned} \tag{10}$$

Therefore the covariance is given as

$$\begin{aligned} & Cov(L(\hat{y}_{i,S_j}, y_i), L(\hat{y}_{i',S_{j'}}, y_{i'})) = \sigma^2(E[L'(x_i^T\beta_0, y_i)])^2 x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_{i'} \\ + \quad & \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0, y_i)])^2 tr((x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T) \\ & (\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}). \end{aligned}$$

Note that, when L is the square error loss the covariance is given as

$$2\sigma^4 tr\{(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}\}.$$

When $i = i'$, the covariance is given as

$$\begin{aligned}
& Cov(L(\hat{y}_{i,S_j},y_i),L(\hat{y}_{i',S_{j'}},y_{i'})) = Var(L(x_i^T\beta_0,y_i)) + \frac{\sigma^2}{2}\,Cov(L(x_i^T\beta_0,y_i),L''(x_i^T\beta_0,y_i)) \\
& (x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_i + x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i) \\
+ \quad & \sigma^2(E[L'(x_i^T\beta_0,y_i)])^2 x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1})x_i \\
+ \quad & \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2 tr((x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_i x_i^T) \\
& (\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}) \\
+ \quad & \frac{\sigma^4}{4}Var(L''(x_i^T\beta_0,y_i))x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}x_i.
\end{aligned}$$

Note that, when L is the square error loss the covariance is given as

$$\begin{aligned}
& 2\sigma^4 + 4\sigma^4 x_i^T(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1})x_i \\
& +2\sigma^4 tr\{(x_i x_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(\mathbf{X}_1^T\mathbf{X}_1)(\mathbf{X}_{S_{j'}}^T\mathbf{X}_{S_{j'}})^{-1}\}
\end{aligned}$$

Proof of Proposition 3.6: Write:

$$\begin{aligned}
L(\hat{y}_{i,S_j},y_i)L(\hat{y}_{i',S_j},y_{i'}) &= L(x_i\beta_0,y_i)L(x_{i'}^T\beta_0,y_{i'}) + L(x_i\beta_0,y_i)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ \frac{1}{2}L(x_i\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_j}-\beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ L'(x_i\beta_0,y_i)L(x_{i'}^T\beta_0,y_{i'})x_i^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ L'(x_i\beta_0,y_i)x_i^T(\hat{\beta}_{S_j}-\beta_0)L'(x_{i'}^T\beta_0,y_{i'})x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ \frac{1}{2}L'(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})x_i^T(\hat{\beta}_{S_j}-\beta_0)(\hat{\beta}_{S_j}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ \frac{1}{2}L(x_{i'}^T\beta_0,y_{i'})L''(x_i^T\beta_0,y_i)(\hat{\beta}_{S_j}-\beta_0)x_i x_i^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ \frac{1}{2}L'(x_{i'}^T\beta_0,y_{i'})L''(x_i^T\beta_0,y_i)x_{i'}^T(\hat{\beta}_{S_j}-\beta_0)(\hat{\beta}_{S_j}-\beta_0)x_i x_i^T(\hat{\beta}_{S_j}-\beta_0) \\
&+ \frac{1}{4}L''(x_i^T\beta_0,y_i)L''(x_{i'}^T\beta_0,y_{i'})(\hat{\beta}_{S_j}-\beta_0)x_i x_i^T(\hat{\beta}_{S_j}-\beta_0) \\
& \quad (\hat{\beta}_{S_j}-\beta_0)^T x_{i'}x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) + R_n. \qquad (11)
\end{aligned}$$

We need to evaluate the expectation of relation (11). We have

$$
\begin{aligned}
& E\{L(\hat{y}_{i,S_j},y_i)L(\hat{y}_{i',S_j},y_{i'})\} \\
= \;& (E[L(x_i^T\beta_0,y_i)])^2+\frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \sigma^2E[L'(x_i^T\beta_0,y_i)])^2tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{1}{2}E[L'(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]E[x_i^T(\hat{\beta}_{S_j}-\beta_0)(\hat{\beta}_{S_j}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_j}-\beta_0)] \\
+ \;& \frac{1}{2}E[L'(x_{i'}^T\beta_0,y_{i'})]E[L''(x_i^T\beta_0,y_i)]E[x_{i'}^T(\hat{\beta}_{S_j}-\beta_0)(\hat{\beta}_{S_j}-\beta_0)x_ix_i^T(\hat{\beta}_{S_j}-\beta_0)] \\
+ \;& \frac{1}{4}(E[L''(x_i^T\beta_0,y_i)])^2E[(\hat{\beta}_{S_j}-\beta_0)x_ix_i^T(\hat{\beta}_{S_j}-\beta_0)(\hat{\beta}_{S_j}-\beta_0)x_{i'}x_{i'}^T(\hat{\beta}_{S_j}-\beta_0)].
\end{aligned}
$$

Now,

$$
\begin{pmatrix} x_i^T(\hat{\beta}_{S_j}-\beta_0) \\ x_{i'}^T(\hat{\beta}_{S_j}-\beta_0) \end{pmatrix} \sim MVN(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\Sigma) \tag{12}
$$

where

$$
\Sigma=\sigma^2\begin{pmatrix} x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i & x_i^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_{i'} \\ x_{i'}^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_i & x_{i'}^T(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}x_{i'} \end{pmatrix}.
$$

Notice here that we do not assume normality of the errors. The assumption of normality for the error distribution is too restrictive. Instead, assumptions A1 and A2 establish the asymptotic distribution of the least squares estimators as the size of the training set n_1 becomes larger and larger. That guarantees that (12) holds. Therefore,

$$
\begin{aligned}
& E\{L(\hat{y}_{i,S_j},y_i)L(\hat{y}_{i',S_j},y_{i'})\} \\
= \;& (E[L(x_i^T\beta_0,y_i)])^2+\frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \sigma^2E[L'(x_i^T\beta_0,y_i)])^2tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})]tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^4}{4}(E[L''(x_i^T\beta_0,y_i)])^2tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
= \;& (E[L(x_i^T\beta_0,y_i)])^2+\frac{\sigma^2}{2}E[L(x_i^T\beta_0,y_i)]E[L''(x_{i'}^T\beta_0,y_{i'})](tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}])+\sigma^2E[L'(x_i^T\beta_0,y_i)])^2tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^4}{2}(E[L''(x_i^T\beta_0,y_i)])^2tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}] \\
+ \;& \frac{\sigma^4}{4}(E[L''(x_i^T\beta_0,y_i)])^2tr[(x_ix_i^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}]tr[(x_{i'}x_{i'}^T)(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}].
\end{aligned}
$$

Therefore,

$$\begin{aligned} & Cov(L(\hat{y}_{i,S_j}, y_i), L(\hat{y}_{i',S_j}, y_{i'})) = \sigma^2 (E[L'(x_i^T \beta_0, y_i)])^2 tr[(x_{i'} x_{i'}^T)(\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1}] \\ + \quad & \frac{\sigma^4}{2} (E[L''(x_i^T \beta_0, y_i)])^2 tr[(x_i x_i^T)(\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1}(x_{i'} x_{i'}^T)(\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1}]. \end{aligned}$$

References

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

P. J. Bickel and K. A. Doksum. *Mathematical Statistics*, Prentice Hall, 2001.

L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**: 2350-2383, 1996.

H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 19th Printing, 1999.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, **10**: 1895-1923, 1998.

B. Efron. Estimating the error rate of a predication rule: Improvement on cross-validation. *Journal of the American Statistical Association*, **78**: 316-331, 1983.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

B. Efron. The Estimation of Prediction Error: Covariance Penalties and Cross-Validation. *Journal of the American Statistical Association*, **99**: 619-632, 2004.

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, 2001.

K. Hitomi and M. Kagihara. Calculation methods for nonlinear dynamic least absolute deviations estimator. *Journal of the Japan. Statist. Society*, **31**: 39-51, 2001.

A. D. Ioffe and J-P. Penot. Limiting subhessians, limiting subjets and their calculus. *Transactions of the American Mathematical Society* , **349**: 789-807, 1997.

G. M. James. Variance and bias for general loss functions. *Machine Learning*, **51**: 115-135, 2003.

M. Kearns. A bound on the error of cross validation with consequences for the training-test split. *In Advances in Neural Information Processing Systems*, **8**: 183-189, 1996.

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross validation. *Neural Computation*, **11**: 1427-1453, 1999.

R. A. Khan. Approximation of the expectation of a function of the sample mean. *Statistics*, **38**: 117-122, 2004.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *In The International Joint Conference on Artificial Intelligence*, 1137-1143, 1995.

E. L. Lehmann. *Theory of Point Estimation*. Wiley and Sons, 1983.

G. J. McLachlan. An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function. *Australian Journal of Statistics*, **14**: 68-72, 1972.

G. J. Mclachlan. An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Australian Journal of Statistics*, **15**: 210-214, 1974.

G. J. McLachlan. The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika*, **61**: 131-135, 1974.

G. J. McLachlan. The bias of the apparent error rate in discriminant analysis. *Biometrika*, **63**: 239-244, 1976.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, **52**: 239-281, 2003.

R. R. Picard and R. D. Cook. Cross validation of regression models. *Journal of the American Statistical Association*, **79**: 575-583, 1984.

J. Piper. Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters* , **13** : 685-692, 1992.

E. Ronchetti and L. Ventura. Between stability and higher order asymptotics .*Statistics and Computing*, **11**: 67-73, 2001.

P. K. Sen and J. M. Singer. *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman and Hall, 1993.