



Optimality of training/test size and resampling effectiveness in cross-validation[☆]

Georgios Afendras, Marianthi Markatou^{*}

Department of Biostatistics and Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, United States

ARTICLE INFO

Article history:

Received 8 December 2017

Received in revised form 18 July 2018

Accepted 19 July 2018

Available online 4 August 2018

Keywords:

Cross-validation estimator

Generalization error

Optimality

Resampling effectiveness

Training sample size

ABSTRACT

An important question in cross-validation (CV) is whether rules can be established to allow optimal sample size selection of the training/test set, for fixed values of the total sample size n . We study the cases of repeated train–test CV and k -fold CV for certain decision rules that are used frequently. We begin by defining the *resampling effectiveness* of repeated train–test CV estimators of the generalization error and study its relation to optimal training sample size selection. We then define optimality via simple statistical rules that allow us to select the optimal training sample size and the number of folds. We show that: (1) there exist decision rules for which closed form solutions of the optimal training/test sample size can be obtained; (2) in a broad class of loss functions the optimal training sample size equals half of the total sample size, independently of the data distribution and the data analytic task. We study optimal selection of the number of folds in k -fold CV and address the case of classification via logistic regression and support vector machines, substantiating our claims theoretically and empirically in both, small and large sample sizes. We contrast our results with standard practice in the use of CV.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Advances and accessibility to technology for capturing and storing large amounts of data has transformed many scientific fields. Yet, for many important scientific questions very large data sets are not available. In these cases, when learning algorithms are used, resampling techniques allow the estimation of the generalization error, an important aspect of algorithmic performance.

The generalization error is defined as the error an algorithm makes on cases that the algorithm has never seen before; it is important because it relates to the algorithm's prediction capabilities on independent data. The literature includes both, theoretical investigations of risk performance of machine learning algorithms as well as numerical comparisons.

Estimation of the generalization error can be achieved via the use of resampling techniques, and in particular via CV. Estimates of the generalization error are used in both, performance evaluation of computational algorithms as well as selection of an algorithm for a given problem (i.e. model selection). Important work in comparing classifiers, from the field of machine learning, includes Dietterich (1998); Demšar (2006); Garcia and Herrera (2008); Kohavi (1995). In the field of Biostatistics van de Wiel et al. (2009) develop an inference framework for the difference in errors between two prediction procedures. Similarly, model selection strategies for machine learning algorithms involve the numerical optimization of

[☆] Both authors acknowledge support provided by the Department of Biostatistics and the Jacobs School of Medicine and Biomedical Sciences, University at Buffalo (in the form of start-up package to the second author).

^{*} Corresponding author.

E-mail addresses: gafendra@buffalo.edu (G. Afendras), markatou@buffalo.edu (M. Markatou).

a criterion that is very often based on an estimate of the generalization error (Cawley and Talbot, 2010, p. 2087). In the statistics literature model selection via CV has been studied by Shao (1993); Yang (2006).

Carrying out inference on equality of generalization errors requires insight into the variance of the estimator of the generalization error. Nadeau and Bengio (2003) provide estimators for the variance of the repeated train–test CV estimator of the generalization error, while Bengio and Grandvalet (2003–2004) address estimators of variance in k -fold CV. Furthermore, Markatou et al. (2005) propose a moment approximation-based estimator of the same CV estimator of the generalization error, and compare this estimator with those provided by Nadeau and Bengio. Other relevant work on variance estimation includes Wang and Lindsay (2014, 2017); Fuchs et al. (2013); Markatou et al. (2011).

Selecting the size of the training set or the number of folds in k -fold CV and understanding the effect of this selection on the generalization error, its bias and its variance, is of interest to many areas of scientific investigation. Examples include pattern recognition and machine learning (Highleyman, 1962; Fukunaga and Hayes, 1989; Raudys and Jain, 1991; Guyon et al., 1998; Kearns, 1997), statistics (Berger and Pericchi, 2004; Cabras et al., 2015), remote sensing (Zhen et al., 2013; Jin et al., 2014), biostatistics and bioinformatics (Dobbin and Simon, 2005, 2011; Dobbin et al., 2008; Popovici et al., 2010; Shao et al., 2013) among others. Hall and Robinson (2009) show that “bagging CV” based on training samples of size $n/2$ significantly improves mean integrated squared error performance in kernel density bandwidth selection. The reason of the interest shown in the problem of training sample size selection is because it impacts the accuracy of the estimators of the generalization error. Arlot and Celisse (2010, p. 69, Sect. 10.2 and p. 70, Sect. 10.3) discuss explicitly the importance of this problem. Even if performance comparisons of algorithms are based on surrogate to the generalization error quantities, such as ranks, inaccurate estimates of the generalization error provide inaccurate surrogate quantities and therefore tests based on these quantities provide incorrect results. In this paper, we offer an analysis of the problem of training sample size selection when the total sample size is fixed and interest centers on carrying out inference on the generalization error. The analysis is complex, and for this reason we consider three kinds of problems to cover a good range of possible applications. These include the case where the decision rule is the sample mean, linear regression or ridge regression and classification via logistic regression or via support vector machines. The strategy we follow establishes a general framework that has the potential to address the issue of optimal selection of the training (and hence test) set sample size and selection of the number of folds in k -fold CV. Within this framework we refer to work by Burman (1989, 1990) to establish rules for optimal selection and study the role of the bias–variance trade-off. In our context, the number of predictors is considerably smaller than n .

Optimality is defined in terms of minimizing the variance of the test set error, and we show that this action is equivalent to minimizing the mean squared error of the CV estimators of the generalization error.

Our results are based on the following two observations. First, the sampling scheme used in repeated train–test and k -fold CV imposes a uniform distribution on the training sets; Lemma 1 in Appendix A provides all necessary tools for establishing our results. Secondly, in Section 2.3, Proposition 1 provides useful results that establish the mathematical foundations of this work. These results have not previously appeared in the literature.

We show that for large samples the variance term of the error estimate of the generalization error is highly relevant, while the bias term plays a secondary and often unimportant role. For small samples, both bias and variance terms are relevant. We present two novel methods for obtaining the resampling size J of the repeated train–test (RTT _{J}) CV estimator of the generalization error (Nadeau and Bengio, 2003; Markatou et al., 2005) that are based on the variance of the aforementioned estimator. We discuss a general framework that can be followed to obtain optimal training set sample size or optimal number of folds k of CV estimators of the generalization error for any classification algorithm and illustrate it using several decision rules. We study the impact of data distribution and elucidate the role of the loss function in selecting the optimal training set sample size and the number of folds in k -fold CV, by constructing and using in the respective problems two new loss functions (modified squared and double squared error loss). Associated empirical results are presented in Tables 6 and 7 of the online supplement, while the literature includes results mainly for the squared error loss.

The paper is organized as follows. Section 2 presents a brief discussion of relevant literature, establishes notation and offers a useful result that serves as a foundation of the proposed optimality rules. Section 3 presents the proposed optimality rules and results using various decision rules. Two methods for obtaining the resampling size of the RTT _{J} CV estimator are also discussed. Section 4 presents simulation results, while Section 5 contains discussion and recommendations.

2. Relevant work, notation and preliminaries

2.1. Relevant work

Interest in the problem of optimal partitioning of a fixed sample n in two sets, training and testing, dates back to 1962. Highleyman (1962) studies this problem in the context of designing a pattern recognition machine and defines the optimum partitioning of the total sample as that partitioning which minimizes the variance of an unbiased estimator of the error probability of the pattern recognition machine. Highleyman (1962) states that the same rule, that is the minimization of variance of the error of a classification rule, can be used with estimators of the error that can be corrected for bias. Similar work has been carried out by Larsen and Goutte (1999). These authors address the problem of splitting a fixed sample n into a training, a testing and a validation set using as criterion the mean squared error (MSE). Only the squared error loss is studied in the simple case of location parameter estimation.

The same problem appears also in the Biostatistics literature and in particular in the area of Bioinformatics. Dobbin and Simon (2011) consider the problem of developing genomic classifiers, empirically address the question of what proportion of

the samples should be devoted to the training set, and study the impact that this proportion has on the MSE of the prediction accuracy estimate. The context within which [Dobbin and Simon \(2011\)](#) address this question is the one where the number of predictors p is considerably larger than the number of samples n . For details see [Dobbin and Simon \(2011\)](#). Here, we note that there are fundamental differences between the work of these authors and our work. First, their setting of small n , large p is different than ours and secondly, even in this setting, they do not address the case of CV estimators.

The literature includes very few theoretical/methodological papers addressing directly the issue of sample splitting for performance evaluation. [Dietterich \(1998\)](#) constructs tests for comparing supervised classification learning algorithms by performing 5 replications of 2-fold CV (5×2 CV), effectively splitting the entire sample size into two equal parts. However, no mathematical justification for this half split is provided. [Dietterich \(1998\)](#) justifies this action by stating that it gives “large test sets and disjoint training sets”.

Empirical work indirectly provides information about sample splitting, for example in the context of comparison of different k -fold CV procedures ([Molinario et al., 2005](#)) or in the context of studying the performance of CV in small samples ([Airola et al., 2011](#)). [Braga-Neto and Dougherty \(2004\)](#) study repeated CV and bootstrap for a small-sample microarray classification and conclude that it is preferable to have small variance and some bias rather than unbiasedness and large variance ([Braga-Neto and Dougherty, 2004](#), Sect. 3, p. 376).

2.2. Notation

Next we set the notation we use in this paper and explicitly discuss the two estimators of the generalization error we work with.

Fix a positive integer n and consider the set $N = \{1, \dots, n\}$. Let $Z_i, i = 1, \dots, n$ be data collected such that the data universe, $\mathcal{Z}_N = \{Z_1, \dots, Z_n\}$, constitutes realizations from a set of independent variables. Let S be a subset of size $n_1, n_1 < n$, taken from $N, S^c \doteq N \setminus S$, the complement of S with respect to N . The subset of observations $\mathcal{Z}_S \doteq \{Z_i : i \in S\}$ is called a training set, used for the construction of a learning rule. The test set contains all data that do not belong to \mathcal{Z}_S ; it is defined as $\mathcal{Z}_{S^c} \doteq \mathcal{Z}_N \setminus \mathcal{Z}_S$, the complement of \mathcal{Z}_S with respect to \mathcal{Z}_N . Let n_2 denote the number of elements in the test set, $n_2 = n - n_1, n_2 < n$.

The generalization error of an algorithm is defined as

$$\mu^{(n)} \doteq \mathbb{E}[L(\mathcal{Z}_N, Z)],$$

where Z is an independent copy of the data Z_i, L is a loss function and the expectation is taken over everything that is random. That is, we take into account the variability in both, training and test set. Furthermore, let $S_j, j = 1, \dots, J$, be random sets such that

$$S = S_{n,n_1} \doteq \{S : S \subset N = \{1, \dots, n\}, \text{card}(S) = n_1\},$$

where $\text{card}(S) = \binom{n}{n_1}$; $S_j, j = 1, \dots, J$, are uniformly distributed on S , such that each S_j is independently sampled, with corresponding complement set (with respect N) S_j^c . The number J is called the *resampling size*. If \mathcal{Z}_{S_j} is defined similarly as above for all j , the usual average test set error is

$$\hat{\mu}_j \doteq \frac{1}{n_2} \sum_{i \in S_j^c} L(\mathcal{Z}_{S_j}, Z_i),$$

and is a function of both the training set S_j and the test set S_j^c . [Nadeau and Bengio \(2003\)](#) state the *repeated train–test CV estimator of the generalization error* as

$$\text{RTT}_J \doteq \frac{1}{J} \sum_{j=1}^J \hat{\mu}_j. \quad (1)$$

In the statistics literature, this method is called repeated learning–testing method ([Burman, 1989](#)).

A second, commonly used, CV procedure is the k -fold CV. We split the entire data set of size n into k equal parts, where k is a divisor of n . The resulting k data sets $\mathcal{Z}_{S_1^c}, \dots, \mathcal{Z}_{S_k^c}, \text{card}(\mathcal{Z}_{S_l^c}) = n/k, l = 1, \dots, k$, correspond to training sets $\mathcal{Z}_{S_1}, \dots, \mathcal{Z}_{S_k}$ that are such that $\text{card}(\mathcal{Z}_{S_l^c}) = n(1 - 1/k), l = 1, \dots, k$. The average test set error is defined as $\hat{\mu}_j \doteq (k/n) \sum_{i \in S_j^c} L(\mathcal{Z}_{S_j}, Z_i)$ and the k -fold CV estimator of the generalization error is

$$\text{CV}_k \doteq \frac{1}{k} \sum_{j=1}^k \hat{\mu}_j. \quad (2)$$

In the cases where the prediction rule takes real values, [Efron \(1986\)](#) presents a wide class of loss functions, called q -class. Specifically, if q is an absolutely continuous and concave real function, the corresponding q -class loss functions have the form

$$^q L(\hat{\mu}, y) \doteq q(\hat{\mu}) + q'(\hat{\mu})(y - \hat{\mu}) - q(y),$$

where $q'(\cdot)$ is the almost sure derivative of the generator q . Commonly used loss functions, such as squared error loss and 0/1 loss belong to this class. In Section 3 we use the q -class of loss functions to elucidate the training sample size selection rules.

Remark 1. Braga-Neto and Dougherty (2004) and Airola et al. (2011) discuss the use of conditional expected performance in small samples. As discussed, for example in Dietterich (1998); Hastie et al. (2009); Schiavo and Hand (2000), the unconditional expected error and the conditional expected error correspond to addressing two different questions. We use the unconditional measure, averaging over all training sets.

2.3. Preliminaries

The following proposition shows that, in general, the quantities $\text{Var}(\hat{\mu}_j)$ and $\text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'})$ are independent of the indices j, j' . In the cases in which the realizations Z_i s are random variables, say X_i s, and the decision rule is the sample mean of each training set, \bar{X}_{S_j} , Nadeau and Bengio (2003) prove the aforementioned statement under the specific assumption that the distribution of $L(\bar{X}_{S_j}, X_i)$ does not depend on the particular realization of S_j, i , and we establish the result without the aforementioned assumption.

Proposition 1. Let n_1, n_2 be fixed, and let L be a loss function such that $\mathbb{E}[L^2(\mathcal{Z}_{S_j}, Z_i)]$ is finite for each realization of S_j and i , where S_j follow a uniform distribution (described in detail in the proof of the proposition). Then, the quantities $\text{Var}(\hat{\mu}_j)$ and $\text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'})$ are finite and do not depend on the indices j and j' , for both repeated train–test and k -fold CV.

Proof. First, $\mathbb{E}[L^2(\mathcal{Z}_{S_j}, Z_i)] < \infty$ gives $\text{Var}[L(\mathcal{Z}_{S_j}, Z_i)] < \infty$, and an application of Cauchy–Schwarz inequality gives that $\text{Cov}[L(\mathcal{Z}_{S_j}, Z_i), L(\mathcal{Z}_{S_j}, Z_{i'})]$ exists for each realization of S_j, i and i' . Thus, by definition of $\hat{\mu}_j$, $\text{Var}(\hat{\mu}_j) < \infty$. Again by Cauchy–Schwarz inequality we obtain that $\text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'})$ exists.

For both cases, repeated train–test and k -fold CV, due to sampling technique, the $\hat{\mu}_j$ s are exchangeable, which completes the proof. \square

Proposition 1 allows one to write

$$V \doteq \text{Var}(\hat{\mu}_j), \quad C \doteq \text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'}), \quad \rho \doteq \text{Corr}(\hat{\mu}_j, \hat{\mu}_{j'}) = C/V,$$

effectively obtaining

$$\text{Var}(\text{RTT}_J) = (V - C)/J + C, \quad \text{Var}(\text{CV}_k) = V/k + (1 - 1/k)C. \quad (3)$$

That is, the variance of both repeated train–test and k -fold CV estimators of the generalization error is a function of the variance of the average test set error and the covariance between two different average test set errors, which are constants with respect to j, j' .

3. Rules for optimal training sample size and optimal number of folds selection

By Cauchy–Schwarz inequality we obtain that $V - C \geq 0$, where the equality characterizes the trivial cases in which the test set error is a constant with probability 1 (Let $j \neq j'$ be two different indices and suppose that $V = C$. Then, in view of the proof of Proposition 1, $\mathbb{E}(\hat{\mu}_j - \hat{\mu}_{j'}) = 0$ and $\text{Var}(\hat{\mu}_j - \hat{\mu}_{j'}) = 2V - 2C = 0$; therefore, $\hat{\mu}_j = \hat{\mu}_{j'}$ with probability 1). So, in view of (3),

$$\text{Var}(\text{RTT}_J) \searrow C \geq 0, \quad \text{as } J \rightarrow \infty. \quad (4)$$

From (3) and (4) we obtain the following important result.

Theorem 1. The variance of RTT_J given by (1) satisfies the following double inequality:

$$\max\{C, V/J\} \leq \text{Var}(\text{RTT}_J) \leq V.$$

Proof. Using Cauchy–Schwarz inequality and (4) we get $0 \leq \rho \leq 1$. Also, (3) gives $\text{Var}(\text{RTT}_J) = [1/J + (1 - 1/J)\rho]V$. Finally, for each J fixed, the function $f(\rho) = 1/J + (1 - 1/J)\rho$, $0 \leq \rho \leq 1$, is increasing, and in view of (4) the proof is completed. \square

The importance of Theorem 1 is obvious in the definition of rules for selecting the sample size of the training set. Since the variance of the CV estimator of the generalization error is bounded above by the variance of the test error V , we can use V to create an optimal rule for selecting the training (and hence the test) sample size.

Our results indicate that the information provided by CV is increased when optimality is quantified via minimization of the variance of the average test set error (repeated train–test CV) or minimization of the variance of CV_k (k -fold CV).

The aforementioned discussion together with relation (3) leads to the following rules:

$$n_1^{\text{opt}} \doteq \arg \min_{n_1 \geq n/2} \{V\}, \quad (5)$$

$$k^{\text{opt}} \doteq \arg \min_k \{\text{Var}(\text{CV}_k)\}. \quad (6)$$

When (6) proposes as a solution $k = n$, leave-one-out CV (LOOCV), we note that the individual test set error has large variance. However, the average test set error, CV_n , has the smallest variance.

Remark 2. In Section 2 of the online supplement we show the equivalence between the proposed minimization rules (5), (6) and the rules based on MSE. Subsection 2.1 discusses in detail this equivalence, while Subsections 2.2 and 2.3 discuss this equivalence in the context of the various decision rules we discuss in this paper.

Remark 3. Rules (5), (6) are proposed for the case where the number of covariates, p , is much smaller than the sample size n . When $p \gg n$, direct minimization of the MSE is appropriate. Note that Dobbins and Simon (2011) address the question of how to split a high dimensional microarray gene expression sample into a training and test set algorithmically, via simulation.

3.1. The resampling size J

We now provide guidance for the selection of the resampling size J entering the construction of the repeated train–test CV estimator of the generalization error. In Section 3.1.1 we discuss two novel methods for selecting the resampling size J , the π -effectiveness and the r -reduction methods. Here, we note that the π -effectiveness and r -reduction, and hence the selection of J , are affected by the training set sample size. We illustrate these relationships in Section 4.

3.1.1. The π -effectiveness and r -reduction acceptable resampling size

We discuss two novel methods for selecting the resampling size J .

The quantities V and C in (3) depend on the sizes of the training and test set, n_1 and n_2 , as well as on F and L . Recall that an experimentalist selects the value of J for which there is no appreciable reduction in the variance of CV estimator of the generalization error. Taking into account the behavior of $\text{Var}(\text{RTT}_J)$ we give the following definition.

Definition 1. We define:

- (a) The *resampling effectiveness* of RTT_J by $\text{res eff}(\text{RTT}_J) \doteq C/\text{Var}(\text{RTT}_J)$.
- (b) The *reduction ratio* of RTT_J by $\text{red ratio}(\text{RTT}_J) \doteq [-\nabla \text{Var}(\text{RTT}_J)]/\text{Var}(\text{RTT}_J)$, where ∇ denote the backward difference with respect to J .

Notice that while $\text{res eff}(\text{RTT}_J)$ provides a comparison of the variance of RTT_J for a given J with the limiting variance C , the $\text{red ratio}(\text{RTT}_J)$ provides a comparison between the two variances for given J and $J - 1$. In this sense $\text{red ratio}(\text{RTT}_J)$ is a local measure of change in variance that is used to obtain J . This is what experimentalists are observing in order to set the value of J .

Now, we are in a position to give a minimum acceptable resampling size, either via the resampling effectiveness of RTT_J or via the corresponding reduction ratio. Observe that in non-trivial cases the resampling effectiveness of RTT_J takes a value in the interval $(0, 1)$; for the trivial cases this is assumed to be 1. From (3) it follows immediately that

$$\text{res eff}(\text{RTT}_J) = [1 + (1/\rho - 1)/J]^{-1}.$$

Fig. 1a shows the behavior of the resampling effectiveness of RTT_J for various J , ρ -values. Notice that when the correlation is low a larger resampling size is required in order to achieve high resampling effectiveness. As the correlation value increases the resampling size decreases; for example, when $\rho = .2$, $J = 36$ is sufficient to obtain a resampling effectiveness of 90%, while when $\rho = .3$, $J = 21$ obtains the same resampling effectiveness.

If π is the desired resampling effectiveness rate, then $\text{res eff}(\text{RTT}_J) \geq \pi$, and the minimum value of the resampling number J is

$$J_{\text{re}}(\pi) = \lfloor (1/\rho - 1)/(1/\pi - 1) \rfloor, \quad (7)$$

where $\lfloor x \rfloor$ denotes the upper integer part of x . We call the number $J_{\text{re}}(\pi)$ as the π -effectiveness minimum resampling size.

Note that, we define the resampling effectiveness (and thus select that value of J) as a number indicating percentage of the ratio $C/\text{Var}(\text{RTT}_J)$, which is between 0 and 1. The bigger the ratio $C/\text{Var}(\text{RTT}_J)$, the closer the $\text{Var}(\text{RTT}_J)$ is to the asymptotic value C of the variance of the CV estimator of the generalization error. That indicates that the value of π should be close to 1, i.e. .8, .9, .95 etc.

An alternative way to obtain J is as follows. From (3) we obtain

$$\text{red ratio}(\text{RTT}_J) = (1 - \rho) / [(J - 1) + (J - 1)^2 \rho]$$

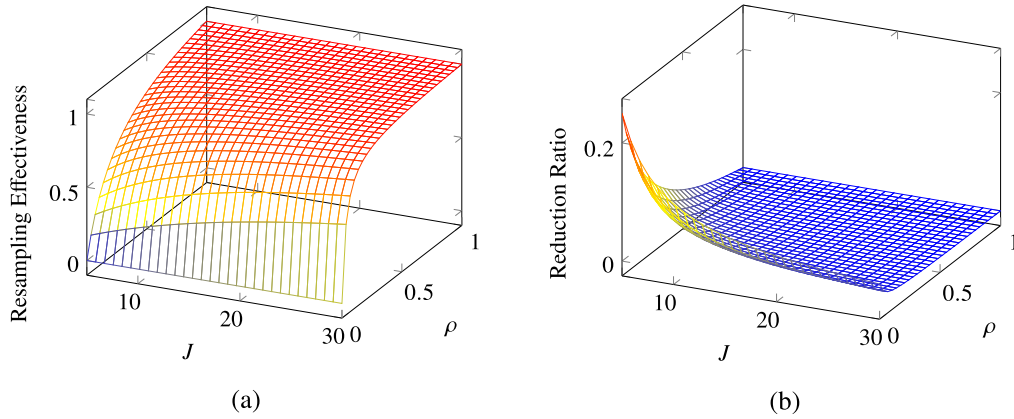


Fig. 1. The behavior of: (a) the resampling effectiveness of RTT_J and (b) the reduction ratio of RTT_J , for $J = 5, \dots, 30$ and $\rho \in [0, 1]$.

We then fix the desired reduction ratio to a value $r > 0$ and require that the reduction ratio of RTT_J should not exceed this value, that is, $\text{red ratio}(\text{RTT}_J) \leq r$. The minimum positive integer satisfying the preceding inequality is given by

$$J_{rr}(r) = \left\lceil 1 - 1/(2\rho) + \left\{ 1/(4\rho^2) + (1/\rho - 1)/r \right\}^{1/2} \right\rceil, \quad (8)$$

which we call the *r-reduction ratio minimum resampling size*. We define the reduction ratio $-\nabla \text{Var}(\text{RTT}_J)/\text{Var}(\text{RTT}_J)$ (and thus select that value of J) as a number indicating the relative reduction of the variance of the CV estimator, which is a positive number. The smaller this ratio, the closer $\text{Var}(\text{RTT}_J)$ is to the asymptotic value C of the variance of the CV estimator of the generalization error. That indicates that the value of r should be close to 0, i.e. .1, .05, .025, .01 etc.

Fig. 1b shows the reduction ratio curves as a function of the correlation ρ . Again, the smaller the value of the correlation between $\hat{\mu}_j, \hat{\mu}_{j'}$, the larger the value of J to obtain a desired variance reduction ratio.

Remark 4. Eqs. (7) and (8) depend on ρ that is generally unknown. This parameter needs to be estimated from the data. Nadeau and Bengio (2003) propose an approximation of the correlation coefficient given by $\hat{\rho} = n_2/n$, where n_2 is the cardinality of the test set. This estimator, although simple, is unreliable because it may overestimate or underestimate the true correlation. Markatou et al. (2005) suggest moment approximation estimators of V and C in certain cases that may be used to estimate the correlation ρ , and hence obtain the values $J_{re}(\pi)$ and $J_{rr}(r)$. Additionally, if the loss function belongs to Efron's q -class of loss functions, then the parameters can be easily estimated. See Section 3.2.1.3 for details.

Remark 5. The formalization of the selection of J brings additional insight into the analysis of factors affecting optimal estimation of the variance of the CV estimators of the generalization error. The formulas presented show the trade-off between the value of J and the correlation ρ of the test set errors. Furthermore, it is clear that as $J \rightarrow \infty$, the variance decreases. In the case of infinite computational resources, having established a rule to obtain the number of replications needed for optimally estimating the variance of CV estimators of the generalization error, saves computational resources. In the case of limited computational resources the existence of formal rules for obtaining J provides information about the degree of optimality we have available.

3.2. Closed form solutions

3.2.1. The case of sample mean

We begin by studying the simple case where the decision rule is the sample mean. Let X be a random variable the realizations of which are denoted by X_i . The variance formulas we use in our rules, and the conditions under which they are obtained, are given in Markatou et al. (2005, pp. 1131, 1138). For completeness reasons, we briefly summarize the results we use.

3.2.1.1. Training set sample size optimality in repeated train–test CV. Using results provided in Markatou et al. (2005, Theo. 3.1) we have

$$V \simeq A/n_1 + B/(n - n_1), \quad C \simeq (\alpha + \beta)/n + \gamma/n^2 + \delta/(n_1 n), \quad (9)$$

where the quantities that appear are defined as $\alpha \doteq \sigma^2 \mathbb{E}^2[L'_\mu(X)]$, $\beta \doteq \text{Var}[L_\mu(X)]$, $\gamma \doteq \sigma^2 \text{Var}[L'_\mu(X)]$, $\delta \doteq \sigma^2 \text{Cov}[L_\mu(X), L''_\mu(X)]$, $A \doteq \alpha + (\gamma + \delta)/n$ and $B \doteq \beta + (\gamma + \delta)/n$, with $L_\mu^{(i)}(x) \equiv d^i L(u, x)/du^i|_{u=\mu}$, $i = 0, 1, 2$.

By definition $\beta > 0$; hence, as n becomes large the parameter B takes positive values. Under this observation, using the approximation formula of V in (9) the solution to the optimization problem in (5) is given by Theorem 2.

Table 1

Relative efficiency of CV_k for squared error loss, normally distributed sample and various values of n and k . The same (qualitative) results hold for different sample distributions.

Relative efficiency of CV_k					
$n = 24, k = 2$ 1.073	$n = 30, k = 2$ 1.060	$n = 30, k = 3$ 1.029	$n = 40, k = 2$ 1.046	$n = 40, k = 4$ 1.015	$n = 50, k = 2$ 1.038
$n = 50, k = 5$ 1.009	$n = 100, k = 5$ 1.005	$n = 100, k = 10$ 1.002	$n = 150, k = 5$ 1.003	$n = 150, k = 10$ 1.001	$n = 150, k = 15$ 1.001

Theorem 2. The solution of the optimization problem (5) when V is given by (9) is

$$n_1^{\text{opt}} = \begin{cases} \lfloor n/2 \rfloor, & \text{if } A \leq B, \\ \min \left\{ n-1, \left\lfloor \frac{\sqrt{A}}{\sqrt{A} + \sqrt{B}} n \right\rfloor \right\}, & \text{if } A > B, \end{cases}$$

where A, B as in (9) and $\lfloor x \rfloor$ stands for the nearest integer of x .

Proof. Obviously, $n_1^{\text{opt}} = \arg \min_{t \in \{\lfloor n/2 \rfloor, \dots, n-1\}} \{g(t)\}$, where $g(t) = A/t + B/(n-t)$, $0 < t < n$. If $A \leq 0$ the desired result becomes trivial.

We will study the case $A > 0$. The derivative of g is $g'(t) = [Bt^2 - A(n-t)^2]/[t^2(n-t)^2]$, and it has the same sign with the numerator $N(t) = [(\sqrt{A} + \sqrt{B})t - \sqrt{An}][(\sqrt{B} - \sqrt{A})t + \sqrt{An}]$. If $A = B$, $N(t)$ is a linear polynomial with root $t_0 = n/2$, and takes negative values before t_0 and positive values after this. If $A \neq B$, $N(t)$ is a quadratic polynomial with two distinct roots $t_1 = \sqrt{An}/(\sqrt{A} + \sqrt{B}) \in (0, n)$ and $t_2 = \sqrt{An}/(\sqrt{A} - \sqrt{B})$. When $0 < A < B$ then $t_2 < 0$, and when $A > B$ then $t_2 > n$; for both cases we see that $N(t) < 0$ for all $t \in (0, t_1)$ and $N(t) > 0$ for all $t \in (t_1, n)$. By definition of n_1^{opt} and the monotonicity of g the proof is completed. \square

In practice, the numbers A and B (namely, the parameters α, β, γ and δ) are unknown and must be estimated. The estimation of these parameters however is unnecessary when, for the theoretical values A and B , $A \leq B$ or $A > B$ hold independently of the data distribution, as Example 2 indicates, see Appendix A.

3.2.1.2. Optimality of k in k -fold CV. In this case, $V \simeq k\alpha/[(k-1)n] + k\beta/n + k^2(\gamma + \delta)/[(k-1)n^2]$ and $C \simeq k(k-2)\alpha/[(k-1)^2n]$ (see Markatou et al., 2005). Thus, using (3), we obtain

$$\text{Var}(CV_k) \simeq (\alpha + \beta)/n + [(\gamma + \delta)/n^2][k/(k-1)]. \quad (10)$$

Proposition 2. The solution to optimization problem (6) where the approximation variance of CV_k is given by (10) is

$$k^{\text{opt}} = \begin{cases} \text{md}(n), & \text{if } \gamma + \delta \leq 0, \\ n, & \text{if } \gamma + \delta > 0, \end{cases}$$

where γ, δ as in (9), $\text{md}(n)$ is the minimum divisor of n that is greater than 1.

Proof. Since the function $k/(k-1)$ decreases in k , the result is obtained in view of relations (6) and (10). \square

Remark 6. Proposition 2 states that if $\gamma + \delta > 0$ then, the optimal value of k in terms of minimization of variance of CV_k is given by the LOOCV. However, one can replace LOOCV by k -fold CV, for large values of n because the quantity $(\gamma + \delta)/n^2 = O(1/n^2)$ and the ratio $k/(k-1)$ satisfy the inequality $1 < n/(n-1) \leq k/(k-1) \leq 2$. Therefore, for relatively small values of k the ratio $k/(k-1)$ takes values close to 1. For example, if $k \in \{5, \dots, 11\}$ the ratio $k/(k-1)$ is between 1.25 and 1.1, and selecting k between 5 and 10, i.e. $k = 10$, guarantees that the variance of the generalization error of CV_k is close to the minimum variance.

In practice, the parameters γ and δ are unknown and, usually, must be estimated. But, sometimes this estimation is unnecessary; for example, consider the squared error loss, then $\gamma + \delta = 4\sigma^4 > 0$ (see Example 2 in Appendix A), and thus the LOOCV is proposed, independently of the distribution of the data. Example 1 illustrates the comments in Remark 6 for the cases in which the LOOCV is proposed by Proposition 2.

Example 1. Let the data follow a $N(\mu, \sigma^2)$ distribution and the squared error loss is used. Then, $\alpha = 0, \beta = 2\sigma^4, \gamma = 4\sigma^4$ and $\delta = 0$; and thus, $\text{Var}(CV_k) = \sigma^4\{2/n + (4k)/[(k-1)n^2]\}$. Defining the relative efficiency of CV_k as the ratio of the variance of the estimator for k folds over its variance at $k^{\text{opt}} = n$ folds, we have that the relative efficiency is given by $\{2/n + 4k/[(k-1)n^2]\}/\{2/n + 4/[(n-1)n]\}$. Table 1 shows the specific values of this relative efficiency for various values of n and k , indicating that the relative efficiency is a function of both, sample size and number of folds. It approaches 1 when $n \geq 50$ and $k \geq 5$.

3.2.1.3. Application to Efron's q -class of loss functions. The results of Sections 3.2.1.1 and 3.2.1.2 apply to the q -class of loss functions. We have

$${}^qL(\bar{X}_{S_j}, X_i) = q(\bar{X}_{S_j}) + q'(\bar{X}_{S_j})(X_i - \bar{X}_{S_j}) - q(X_i), \quad i \in S_j^c \quad \text{for all } j = 1, \dots, J;$$

where q is differentiable having at least up to five derivatives, and such that the fourth derivative of qL is bounded. The main gain is that the quantities α , β , γ and δ (therefore A and B too) can be computed easily. Observe that ${}^qL_{\mu}^{(i)}(x) \equiv d^i[{}^qL(\mu, x)]/d\mu^i = q^{(i+1)}(\mu)(x - \mu) - (i - 1)q^{(i)}(\mu)$, $i = 1, \dots, 4$. Hence, $\alpha = 0$, $\beta = [q'(\mu)]^2\sigma^2 + \text{Var}[q(X)] - 2q'(\mu)\text{Cov}[X, q(X)]$, $\gamma = [q''(\mu)]^2\sigma^4$, $\delta = q'(\mu)q'''(\mu)\sigma^4 - q'''(\mu)\text{Cov}[X, q(X)]\sigma^2$.

Remark 7. The function q is a known function, but the parameters are unknown (unless the distribution F is known). We estimate these parameters from the sample values X_1, \dots, X_n by replacing μ , σ^2 , $\text{Var}[q(X)]$ and $\text{Cov}[X, q(X)]$ with $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$, $\hat{\text{Var}}_q = (n-1)^{-1}\sum_{i=1}^n [q(X_i) - \bar{q}]^2$ and $\hat{\text{Cov}}_{X,q} = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})[q(X_i) - \bar{q}]$ respectively, where $\bar{q} = n^{-1}\sum_{i=1}^n q(X_i)$.

Remark 8. Because $\alpha = 0$ and $\beta > 0$ for each loss function in the q -class, Theorem 2 guarantees that $n_1^{\text{opt}} = \lfloor n/2 \rfloor$, for all data distributions and all loss functions that belongs to the q -class. Further, $V^{\text{opt}} \simeq 2\beta/n$ and $C^{\text{opt}} \simeq \beta/n$. Hence, the optimal value of the correlation coefficient is $\rho^{\text{opt}} \simeq 1/2$. Since C is unaffected by the splitting, for training sample size n_1 we have $\rho = C/V = C^{\text{opt}}/V \leq C^{\text{opt}}/V^{\text{opt}} = 1/2$. Therefore, the choice $n_1 = n_1^{\text{opt}}$ leads to the optimized values of resampling effectiveness and reduction ratio of RTT_J (see Section 3.1), see Fig. 1.

3.2.2. The regression case

In the regression case we solve the problem of optimal training sample size identification under squared error loss. Furthermore, we offer a solution under both, normality of the errors and under relaxation of the normality assumption. In the second case we require $\mathbb{E}(|\varepsilon_i|^{4+\epsilon}) < \infty$ for some $\epsilon > 0$. The results presented below are different from the results presented in Markatou et al. (2005) in the following aspects: (1) the analysis presented in Markatou et al. (2005) is conditional on the training sample S_j ; here, we use the distribution of S_j in light of Proposition 1; (2) we study the case of k -fold CV; and (3) we present closed form expressions of the expectation, variance and covariance between two test set errors.

We first derive expressions for the quantities V , C entering the computation of the variance of RTT_J by exploiting relation (11).

Let $Z_i = (y_i, \mathbf{x}_i)$, $i = 1, \dots, n$, be random variables with $\mathbf{x}_i = (1, x_{i,1}, \dots, x_{i,p-1})^T \in \mathbb{R}^p$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T \in \mathbb{R}^p$ is the parameter vector and $\mathbf{y} = (y_1, \dots, y_n)^T$ is such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ is a vector of errors, $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, \mathbf{I}_n is the $n \times n$ identity matrix. Let a training set \mathcal{Z}_{S_j} be of size n_1 ; then, \mathbf{X}_{S_j} is the $n_1 \times p$ matrix according to S_j and $\hat{\boldsymbol{\beta}}_{S_j}$ indicates the estimator of $\boldsymbol{\beta}$ computed by using the data in the training set \mathcal{Z}_{S_j} .

Assumption. If S_j is the index set of a training set with n_1 indices, then

$$\lim_{n_1 \rightarrow \infty} \frac{1}{n_1} \mathbf{X}_{S_j}^T \mathbf{X}_{S_j} = \mathbf{V}^{-1}, \quad (11)$$

where \mathbf{V}^{-1} is finite and positive definite.

Under the above conditions,

$$\sqrt{n_1} (\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{V}), \quad \text{as } n_1 \rightarrow \infty. \quad (12)$$

For each $S_j = \{i_1, \dots, i_{n_1}\} \subset \{1, \dots, n\}$ we define the $n_1 \times n$ matrix $\mathbf{E}_{S_j} \doteq (\mathbf{e}_{i_1}^T, \dots, \mathbf{e}_{i_{n_1}}^T)^T$, where \mathbf{e}_i is the i th element of the usual basis of \mathbb{R}^n , and the $n \times n$ diagonal matrix

$$\mathbf{I}_{S_j} \doteq \mathbf{E}_{S_j}^T \mathbf{E}_{S_j} = \text{diag}(\mathbb{1}_{\{1 \in S_j\}}, \dots, \mathbb{1}_{\{n \in S_j\}}).$$

Observe that $\text{rank}(\mathbf{I}_{S_j}) = \text{card}(S_j)$ and $\mathbf{I}_{S_j} \mathbf{I}_{S_{j'}} = \mathbf{I}_{S_j \cap S_{j'}}$; also, $\mathbf{X}_{S_j} = \mathbf{E}_{S_j} \mathbf{X}$ and $\mathbf{y}_{S_j} = \mathbf{E}_{S_j} \mathbf{y}$. Thus, (11) is reformulated as:

$$[\text{card}(S_j)]^{-1} \mathbf{X}^T \mathbf{I}_{S_j} \mathbf{X} \rightarrow \mathbf{V}^{-1}, \quad \text{as } \text{card}(S_j) \rightarrow \infty. \quad (13)$$

Generally, the matrix \mathbf{V} is unknown; but it is estimated easily by $n(\mathbf{X}^T \mathbf{X})^{-1}$. Hereafter, we use this estimation as the true matrix \mathbf{V} .

In the analysis that follows, we restrict ourselves to the case $\mathbb{E}(Y_i | \mathbf{X}_i = \mathbf{x}_i)$, i.e. the explanatory variables are treated as fixed. This formulation is known as *the fixed design case*.

3.2.2.1. Squared error loss. Using the above results, we state and prove Propositions 4–6, see Appendix B. These propositions offer closed form expressions for the expectation, variance and covariance of average test set errors for both repeated

train–test and k -fold CV. Based on these results we prove the following proposition that offers closed form solutions to the optimization problems (5) and (6).

Proposition 3.

(a) Under normality of errors,

- i. in the repeated train–test CV case $n_1^{\text{opt}} = \lfloor n/2 \rfloor$;
- ii. in the k -fold case $k^{\text{opt}} = n$, that is, the LOOCV is proposed.

(b) For general error distributions with $\mathbb{E}(\varepsilon_i^{4+\epsilon}) < \infty$ for some $\epsilon > 0$, and in the case of repeated train–test CV $n_1^{\text{opt}} = \lfloor n/2 \rfloor$.

Proof. (a) i. We omit the term $O(1/n^2)$ of $\text{Var}(\hat{\mu}_j)$ in Proposition 4 and consider the function $g(t) = 2/(n-t) + (4p+3\theta)/[t(n-t)]$, $0 < t < n$. As in proof of Theorem 2, we find that g decreases up to $[(4p+3\theta)/n]^{1/2}/\{[(4p+3\theta)/n]^{1/2} + [2+(4p+3\theta)/n]^{1/2}\}n \in (0, n/2)$ and increases after this, thus, $n_1^{\text{opt}} = \lfloor n/2 \rfloor$, which does not depend on the parameters or the observations.

ii. From Proposition 6, since $\text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'})$ is a $O(1/n^2)$ quantity, omitting the terms $O(1/n^3)$ we get $\text{Var}(\hat{\mu}_{k\text{-fold}})/\sigma^4 = 2/n + [k/(k-1)][p^2 + 4(n-1)p + (2n-3)\theta]/[n^2(n-1)] + [k/(k-1)^2][2p/n^2] + [k^3/(k-1)^3][2(p-\theta)/n^2]$. Observe that the quantities $[p^2 + 4(n-1)p + (2n-3)\theta]/[n^2(n-1)]$, $2p/n^2$ and $2(p-\theta)/n^2$ are positive and independent on k , and the quantities $k/(k-1)$, $k/(k-1)^2$ and $k^3/(k-1)^3$ are decreasing functions of $k \in \{2, \dots, n\}$. Thus, $k^{\text{opt}} = n$.

(b) From Proposition 5, using the same arguments as in proof of Theorem 2, the optimal value of n_1 in (5) follows. \square

Remark 9. Similarly to the case where the decision rule was \bar{X}_{S_j} , for the optimal value of $n_1 = \lfloor n/2 \rfloor$, $V^{\text{opt}} = 4\sigma^4/n + O(1/n^2)$ and $C^{\text{opt}} = 2\sigma^4/n + O(1/n^2)$, hence, $\rho^{\text{opt}} \simeq 1/2$. Therefore, the resampling size J can be chosen either via specification of the resampling effectiveness or specification of the reduction ratio of RTT_J (see Section 3.1), for a given π or r , by the following relations, see (7) and (8),

$$J_{\text{re}}(\pi) = \lfloor \pi/(1-\pi) \rfloor, \quad \text{or} \quad J_{\text{rr}}(r) = \lfloor (1+1/r)^{1/2} \rfloor.$$

Remark 10. The above results can be extended to a subclass of the q -loss functions class; this subclass contains differentiable functions that can be expressed as functions of the errors.

3.2.3. The ridge regression case

In the case of ridge regression the estimates of β , based on the elements of \mathcal{Z}_{S_j} , are obtained by minimizing

$$\sum_{i \in S_j} (y_i - \mathbf{x}_i^T \beta) + \lambda \sum_{j=1}^p \beta_j^2.$$

The form of the estimator of β is given as (see Hoerl and Kennard, 1970, p. 57)

$$\tilde{\beta}_{S_j} = (\mathbf{X}_{S_j}^T \mathbf{X}_{S_j} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_{S_j}^T \mathbf{y}_{S_j} = \mathbf{Z}_{S_j} \hat{\beta}_{S_j}, \quad \lambda \geq 0,$$

where $\mathbf{Z}_{S_j} = [\mathbf{I}_p + \lambda(\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1}]^{-1}$. Let Assumption (11) hold (for a more general format, see Afendras and Markatou, 2016) and λ be a function of n_1 . Suppose now that either $\lambda/n_1 \rightarrow 0$ or $\lambda/n_1 \rightarrow c \in (0, \infty)$ or $\lambda/n_1 \rightarrow \infty$ as $n_1 \rightarrow \infty$; then, $\mathbf{Z}_{S_j} \rightarrow \mathbf{A}$, where $\mathbf{A} = \mathbf{I}_p$ or $[\mathbf{I}_p + c\mathbf{V}]^{-1}$ or $(0)_{p \times p}$, respectively. Using multivariate Slutsky's theorem, we get $\sqrt{n_1}(\tilde{\beta}_{S_j} - \mathbf{Z}_{S_j}\beta) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{A} \mathbf{V} \mathbf{A}^T)$.

Under suitable assumptions on the error distribution, we can have that the moments of random vector $\sqrt{n_1}(\tilde{\beta}_{S_j} - \mathbf{Z}_{S_j}\beta)$ converge to the corresponding moments of its asymptotic distribution; see Afendras and Markatou (2016). Therefore, if the squared error loss is used, we can have close form solutions for the optimization problems (5) and (6), as in Proposition 3.

3.2.4. Classification via logistic regression

To illustrate the difficulty of obtaining a close form solution for the training set sample size in general, when other than the squared error loss functions are used, we discuss the case of classification via logistic regression.

Hastie et al. (2009) formalize logistic regression as a linear classification method, where y is the label of the data and \mathbf{x} is a feature vector. The loss function here is the logistic loss given as $\log(1 + e^{-y p_i})$, where y is a label and p_i is the algorithm prediction, such that

$$\text{logit}(p_i) = \mathbf{x}_i^T \beta + \varepsilon_i, \quad p_i = \mathbb{P}(y_i = 1 | \mathbf{x}_i).$$

In this case the minimization problem (5) does not have a closed-form solution for obtaining the optimal value of n_1 , and the problem is reduced to numerical optimization. The analysis of this case is presented in Appendix B.2. Our experimental results, as well as implementation of Algorithm 1 (see Appendix B.2) designed to obtain numerically the solution of the associated optimization problem, indicate that the optimal training sample size is $n_1 = \lfloor n/2 \rfloor$, see Table 12*.

Table 2

Loss functions: squared, q -class with $q(t) = -(1+t^2)^{1/2}$, approximate absolute (where $d > 0$), modified squared and double squared, and their first two derivatives with respect to μ .

Name	$L_\mu(x)$	$L'_\mu(x)$	$L''_\mu(x)$
Squared:	$(x - \mu)^2$	$-2(x - \mu)$	2
q -class:	$-(1 + \mu^2)^{1/2} - \frac{\mu(x - \mu)}{(1 + \mu^2)^{1/2}} + (1 + x^2)^{1/2}$	$\frac{\mu - x}{(\mu^2 + 1)^{3/2}}$	$\frac{-2\mu^2 + 3x\mu + 1}{(\mu^2 + 1)^{5/2}}$
Approximated absolute:	$((x - \mu)^2 + d)^{1/2}$	$\frac{\mu - x}{((x - \mu)^2 + d)^{1/2}}$	$\frac{((\mu - x)^2 + d)^{3/2}}{d}$
Modified squared:	$(x - \mu)^2 + \mu^2$	$4\mu - 2x$	4
Double squared:	$(x^2 - \mu^2)^2$	$-4\mu(x^2 - \mu^2)$	$-4(x^2 - \mu^2) + 8\mu^2$


3.2.5. Support vector machine for classification

In the case of support vector machine the corresponding loss function is the ϵ -insensitive loss function, given by $|\xi|_\epsilon \doteq (|\xi| - \epsilon)^+$, where ξ is a random variable and $(x)^+$ denotes the positive part of x (see Smola and Schölkopf, 2004, p. 200). This loss function is more complicated than the logistic 0/1 loss function. Thus, it appears that there is no closed form solution in this problem. One may study the problem algorithmically as it was done in our classification via logistic regression case.

4. Simulation study

In this section we present simulation results using a variety of distributions and loss functions with the goal of illustrating empirically our theoretical results. We discuss the empirical performance of our rules organizing the presentation according to the cases studied above, and we note that general simple recommendations about optimal selection of training sample size are possible for the cases studied here. The online supplement provides associated results for samples as small as 10 observations.

4.1. Sample mean

Using the version of  Ri386 3.1.2 on a DELL Latitude E7240 PC we simulated 10^4 samples of size $n = 60, 100, 301, 750, 1501, 5000$ from distributions that can be categorized into symmetric with a variety of tail behaviors (normal, $U(-1, 1)$, t_{12} , t_6) and asymmetric ($\exp(1)$, log-normal, Pareto(15), Pareto(6)). The normal distribution is central in statistics, while the two t -distributions exhibit heavier than the normal, tail behavior. The log-normal distribution is used in biostatistics in biomarker studies, while the Pareto distribution is a power law distribution used in the description of social, geophysical, scientific, actuarial and many other observable phenomena. The selection of t_6 and Pareto(6) distributions is not arbitrary. Both the t_6 and Pareto(6) distributions possess less than six moments, and for our theory to apply we require the existence of up to six moments. Thus, these selections reflect the performance of the methods in limit cases. To illustrate the effect the choice of loss function has on the size of training set and on the choice of the number of folds k , we use the loss functions presented in Table 2.

Our results indicate that for the q -class of loss functions the optimal training sample size is $\lfloor n/2 \rfloor$, independent of the data distribution. However, one can construct loss functions, such as the modified squared error loss given in Table 2, for which the optimal training sample size is not $\lfloor n/2 \rfloor$. Notice that, the modified squared error loss function does not belong to the q -class. Tables 4–10, of the online supplement, corroborate our theoretical results for a wide variety of sample sizes and distributions.

When CV estimators of the generalization error are used the usual recommendation made is to use 70%–80% of the data for training and the remaining for testing. Figure 1 (see online supplement) plots the relative efficiency of RTT_J against the ratio n_1/n , for $J = 1$ (it corresponds to $\text{Var}(\hat{\mu}_J)$), 10 and 15. The relative efficiency of RTT_J is defined as the ratio of the variance of RTT_J for any given data splitting over the variance of RTT_J at the optimal data splitting. Figure 1 clearly shows that, for all sample sizes, all loss functions with $n_1^{\text{opt}} = \lfloor n/2 \rfloor$, and all distributions, the popular recommendations in terms of splitting exhibit substantial increase in $\text{Var}(\text{RTT}_J)$. For example, when 75% of the total sample is used for training, the increase in $\text{Var}(\text{RTT}_{10})$ based on the squared error loss and $n = 24$ is 17.3%, while when 80% (or 90%) of the total sample is used for training this increase is 24.4% (or 88.6%). If $J = 15$, the increase in $\text{Var}(\text{RTT}_{15})$ based on a training set that equals 75% of the total sample, squared error loss and $n = 24$ is 11.86%; it is 16.8% (or 61.2%) when 80% (or 90%) of the sample is used for training.

Notice that $\hat{\text{Var}}(\text{RTT}_\infty) = \hat{C}$ is almost unaffected by the splitting mechanism across all loss functions, independently of data distribution. This provides further justification of our selection of the optimality rule (5), indicating that (5) is equivalent to minimizing $V - C$ and to minimizing $\text{Var}(\text{RTT}_J)$ for all J (see also relation (9)).

Furthermore, $\rho^{\text{opt}} = .5$ when $n_1^{\text{opt}} = \lfloor n/2 \rfloor$ (both theoretically and empirically), so that $C^{\text{opt}} = .5V^{\text{opt}}$, and the limited variance C is almost unaffected by the splitting. In view of Figure 1, $V(.75) \simeq 1.9V^{\text{opt}} \simeq 3.8C^{\text{opt}} \simeq 3.8C(.75)$, while

Table 3

The resampling effectiveness and reduction ratio of RTT_J , $J = 10, 15$, for loss functions which have $n_1^{\text{opt}} = \lfloor n/2 \rfloor$, in the sample mean case.

n_1	Resampling effectiveness		Reduction ratio	
	$J = 10$	$J = 15$	$J = 10$	$J = 15$
.50n	90%	94%	.010	.0045
.75n	78%	84%	.025	.0114
.80n	74%	81%	.029	.0137
.85n	66%	74%	.038	.0185
.90n	53%	63%	.053	.0268

$V(.80) \simeq 4.6C(.80)$, $V(.85) \simeq 6.2C(.85)$ and $V(.90) \simeq 10C(.90)$, where $V(\pi)$ denotes the variance of the test set error $\hat{\mu}_j$ for training sample size $n_1 = \pi n$. The notations $C(\pi)$ and $\rho(\pi)$ have similar interpretation. Thus, $\rho^{\text{opt}} \simeq .5$, $\rho(.75) \simeq .26$, $\rho(.80) \simeq .22$, $\rho(.85) \simeq .16$, and $\rho(.90) \simeq .1$. The relationship of these results to the resampling effectiveness and reduction ratio of RTT_J are shown in Table 3 for $J = 10, 15$ and when the training sample size is .5n ($= n_1^{\text{opt}}$), .75n, .80n, .85n and .90n. These results show that when the training sample size is not optimal then we need to increase the resampling size to obtain acceptable levels of reduction ratio and/or resampling effectiveness, thereby increasing the computational burden of the procedure.

4.2. Regression and classification via logistic regression

We empirically now study the variance of generalization error in the cases of linear regression and classification via logistic regression.

Data were generated as $y_i = 1 + X_1 + X_2 - X_3 + X_4 + \varepsilon_i$, $i = 1, 2, \dots, n$, where the sample size $n = 40, 60, 100$ and 200; the error vector ε_n follows $N_n(\mathbf{0}, \mathbf{I}_n)$ distribution. The four covariates were generated only once as follows. The first covariate X_1 is a binary variable generated from Bernoulli(.6) distribution, the second variable is generated from Poisson(2) distribution, the third variable is generated from the $U(0, 5)$ distribution and variable X_4 is generated from the $U(0, 3)$ distribution. The squared error loss is used and the number of Monte Carlo repetitions equals 5000. Table 11* and Figure 4* of the online supplement, present relevant results; notice that, both the table and figure, indicate an increase in the variance of the test set error, $\text{Var}(\hat{\mu}_j)$, as well as in $\text{Var}(\text{RTT}_J)$ with $J = 10, 15$, when the training sample size moves away from the optimal value of $n_1 = \lfloor n/2 \rfloor$. This is true for all sample sizes.

Table 12* (see online supplement) presents simulation results for the case of classification via logistic regression. Two models were fitted, one containing four covariates and a second one with nine covariates. In both cases the errors were normally distributed with mean vector $\mathbf{0}$ and identity variance–covariance matrix. The covariates are generated only once as follows: X_1 is Bernoulli(.6), X_2 follows Poisson(2), X_3 is generated from a discrete Uniform on the set $\{0, 1, 2, 3, 4\}$ and X_4 is Uniform on the set $\{0, 1, 2, 3\}$. The remaining covariates are also generated once as follows: $X_5 \sim \text{Negative Binomial}(3, .3)$, $X_6 \sim U(0, 1)$, $X_7 \sim U(-1, 1)$ and $\mathbf{X}_{8,9}$ from a bivariate normal distribution with $\mathbf{0}$ mean vector and variance–covariance matrix $\Sigma = \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}$. Also, $X_1, \dots, X_7, \mathbf{X}_{8,9}$ are independent. The total sample size n used is 60, 100, 200, 300 and 500. Table 12 presents the relative efficiency of the test set estimate $\hat{\mu}_j$ for various values of the training sample size n_1 . It can be seen from the table that $\hat{\mu}_j$ is most efficient when $n_1 = \lfloor n/2 \rfloor$. For example, when $n = .75n$, the sample size is 100 and we have 4 covariates the $\text{Var}(\hat{\mu}_j)$ is 2.34 times greater than the variance of $\hat{\mu}_j$ computed under $n_1 = \lfloor n/2 \rfloor$. The results indicate that, in both cases, $\text{Var}(\hat{\mu}_j)$ is minimized when the training sample size is equal to $\lfloor n/2 \rfloor$.

Results from simulations using the bias-corrected procedures are discussed in our online supplement.

5. Discussion and recommendations

In this paper we address the problem of optimal selection of the size of training set and the number of folds for accurate estimation of the generalization error of prediction algorithms and present closed form solutions for a variety of decision rules. We study two types of CV estimators of the generalization error. These are repeated train–test CV and k -fold CV estimators. We explicitly establish the equivalence of the rules we propose with those based on MSE, and connect our rules with two novel methods for selecting the resampling size in the repeated train–test CV method.

There are a number of practical implications of our results which we now describe. Our results indicate that the optimal training sample size selection is a complex problem and depends primarily on the loss function that is used, as well as on the data distribution and the decision rule. Despite the complexity of the analysis, we provide a general framework that can be used with any decision rule under consideration. Specifically, we demonstrate that: (a) the bias term of CV estimators of the generalization error in small samples is affected by the training set sample size selection, while in large samples bias plays an insignificant role; (b) the variance term is affected similarly by the training set sample size selection in both, small and large samples. To deal with the differential impact outlined in (a) we recommend the use of bias-corrected estimators provided by Burman (1989). In the case of small samples, bias can be substantial. Empirical, as well as theoretical results, show that, in general, estimators of the generalization error are almost unbiased and the selection of $n_1 = \lfloor n/2 \rfloor$ almost

universally provides the minimum value of the absolute bias. Therefore, we recommend the bias-corrected RTT_J^* estimators for use with small samples.

We illustrate the effect of popular training set sample size selections, such as $.75n$ or $.80n$, have on the $\text{Var}(\hat{\mu}_j)$ and $\text{Var}(\text{RTT}_J)$, $J = 10, 15$. We found that as the training sample size increases away from its optimal value the aforementioned variances increase substantially. To decrease those we need to increase the resampling size J from 15 to a value that achieves acceptable reduction ratio and/or resampling effectiveness, thereby increasing the computational cost. These results are general, and hold for the general class of q -loss functions. For the case of classification via logistic regression, the results carry over. Both, extensive simulation experiments and the application of Algorithm 1 presented in Appendix B.2, indicate that the optimal value of the training set is $\lfloor n/2 \rfloor$.

The selection of the resampling size J of a repeated train–test CV estimator of the generalization error is important, as it contributes to the variance reduction of this estimator. We propose two methods of selecting J and exemplify their use. Our analysis indicates that, when the correlation between two different test sets is moderate, the resampling size $J \geq 14$ provides resampling effectiveness greater than or equal to .85. The higher the desired resampling effectiveness, the higher the value of J . Similar results hold when the reduction ratio forces the variance of RTT_J closer to its asymptotic value, and provides for larger values of J . Our work shows that the correlation coefficient ρ is affected by the training sample size and since J is a function of ρ it is also affected by the splitting of the total sample. The effect of training sample size n_1 is to decrease the correlation as n_1 moves away from its optimal value, requiring larger and larger values of J to achieve small values of $\text{Var}(\text{RTT}_J)$. Similar results are obtained for selecting k^{opt} in k -fold CV.

Appendix A. Useful theoretical results

We present statements of results needed to establish the mathematical foundations of our work.

Example 2. Assume that the data are from a population with finite eighth moment and the loss function is squared error, that is $L(X_{S_j}, X_i) = (X_i - \bar{X}_{S_j})^2$, $i \in S_j^c$. Then, $L_\mu(X) = (X - \mu)^2$ with derivatives $L'_\mu(X) = -2(X - \mu)$ and $L''_\mu(X) = 2$. We compute $\alpha = 0$, $\beta = \mu_4 - \sigma^4 > 0$, $\gamma = 4\sigma^4 > 0$ and $\delta = 0$, where $\sigma^2 = \text{Var}(X)$ and $\mu_4 = \mathbb{E}(X - \mu)^4$. Therefore, $A = 4\sigma^4/n < \mu_4 - \sigma^4 + 4\sigma^4/n = B$; and so, $n_1^{\text{opt}} = \lfloor n/2 \rfloor$, independently of the distribution of the data.

Lemma 1. Let S_j and $S_{j'}$ be two index sets of size $n_1 < n$ (and $n_2 = n - n_1$) as in repeated train–test CV; and let us consider the fixed indices $i \neq i' \neq i'' \in N = \{1, \dots, n\}$. Then,

- (a) $\mathbb{P}(\mathbb{1}_{\{i \in S_j^c\}} = 1) = n_2/n$.
- (b) $\mathbb{P}(\mathbb{1}_{\{i, i' \in S_j^c\}} = 1) = n_2(n_2 - 1)/[n(n - 1)]$ and $\mathbb{P}(\mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}\}} = 1) = n_1 n_2 / [n(n - 1)]$.
- (c) $\mathbb{P}(\mathbb{1}_{\{i, i' \in S_j^c, i'' \in S_{j'}\}} = 1) = n_1 n_2 (n_1 - 1) / [n(n - 1)(n - 2)]$.
- (d) $\mathbb{E}(\mathbb{1}_{\{i \in S_j^c\}}) = n_2/n$ and $\text{Var}(\mathbb{1}_{\{i \in S_j^c\}}) = n_1 n_2 / n^2$.
- (e) $\mathbb{E}(\mathbb{1}_{\{i, i' \in S_j^c\}}) = n_2(n_2 - 1)/[n(n - 1)]$ and $\text{Cov}(\mathbb{1}_{\{i \in S_j^c\}}, \mathbb{1}_{\{i' \in S_j^c\}}) = -n_1 n_2 / [n^2(n - 1)]$.
- (f) $\mathbb{E}(\mathbb{1}_{\{i \in S_j^c \cap S_{j'}^c\}}) = \mathbb{E}(\mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c\}}) = n_2^2/n^2$ and $\text{Cov}(\mathbb{1}_{\{i \in S_j^c\}}, \mathbb{1}_{\{i \in S_{j'}^c\}}) = \text{Cov}(\mathbb{1}_{\{i \in S_j^c\}}, \mathbb{1}_{\{i' \in S_{j'}^c\}}) = 0$.
- (g) $\mathbb{E}(\mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c\}}) = \mathbb{E}(\mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c, i \in S_{j'}\}}) = n_1 n_2^2 / [n^2(n - 1)]$.
- (h) $\mathbb{E}(\mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c, i'' \in S_{j'}\}}) = n_1^2 n_2^2 / [n^2(n - 1)^2]$.
- (i) $\mathbb{E}(\text{card}(S_j \cap S_{j'}) \mathbb{1}_{\{i \in S_j^c \cap S_{j'}^c\}}) = n_1^2 n_2^2 / [n^2(n - 1)]$ and $\mathbb{E}(\text{card}(S_j \cap S_{j'}) \mathbb{1}_{\{i \in S_j^c \cap S_{j'}\}}) = n_1^2 n_2^2 [n_1(n_1 - 1) + n_2 - 1] / [n^2(n - 1)(n - 2)]$.
- (j) $\mathbb{E}(\text{card}^2(S_j \cap S_{j'}) \mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c\}}) = n_1^2 n_2^2 [(n - 2)^2 + (n - 3)(n_1 - 1)^2] / [n^2(n - 1)^2(n - 2)]$.
- (k) $\mathbb{E}(\text{card}(S_j \cap S_{j'}) \mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c, i \in S_{j'}\}}) = \mathbb{E}(\text{card}(S_j \cap S_{j'}) \mathbb{1}_{\{i \in S_j^c, i' \in S_{j'}^c, i' \in S_{j'}\}}) = n_1^2 n_2^2 (n_1 - 1) / [n^2(n - 1)^2]$.

Proof. See Online Supplement. \square

In view of (5), we prove the following general result. This theorem provides expressions for the expected value and variance of $\hat{\mu}_j$ that enter the specification of the optimality rule.

Theorem 3. Let L be a loss function, S_j an index set of size $n_1 < n$ ($n_2 = n - n_1$) and $i, i' \in S_j^c$. If the asymptotic values of $\mathbb{E}[L(\hat{y}_{S_j, i}, y_i) | S_j, i]$, and $\mathbb{E}[L(\hat{y}_{S_j, i}, y_i) L(\hat{y}_{S_j, i'}, y_{i'}) | S_j, i, i']$ do not depend on the particular realization of S_j , say e_i and $e_{i, i'}$ respectively, then, the asymptotic value of $\mathbb{E}(\hat{\mu}_j)$ and $\text{Var}(\hat{\mu}_j)$ are

$$\mathbb{E}(\hat{\mu}_j) = \frac{1}{n} \sum_{i=1}^n e_i,$$

$$\text{Var}(\hat{\mu}_j) = \frac{1}{n^2 n_2} \left\{ \sum_{i=1}^n (n e_{i, i} - n_2 e_i^2) + \frac{2}{n-1} \sum_{1 \leq i < i' \leq n} [n(n_2 - 1) e_{i, i'} - (n-1) n_2 e_i e_{i'}] \right\}.$$

Proof. See Online Supplement. \square

Appendix B. Analysis for the regression case

B.1. Linear regression

Let $S_j, S_{j'}$ be given training sets, $i \in S_j^c, i' \in S_{j'}^c$ and squared error loss such that $L_s(\hat{y}_{S_j,i}, y_i) = (\hat{y}_{S_j,i} - y_i)^2 = (\mathbf{x}_i^T \hat{\beta}_{S_j} - y_i)^2$ and $L_s(\hat{y}_{S_{j'},i'}, y_{i'}) = (\mathbf{x}_{i'}^T \hat{\beta}_{S_{j'}} - y_{i'})^2$.

Define the quantities

$$\begin{aligned} C_1 &\doteq \sum_{i=1}^n \mathbf{x}_i^T \mathbf{V} \mathbf{x}_i, & C_3 &\doteq \sum_{i \neq i'} \sum_{i'=1}^n (\mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'})^2 = \sum_{i=1}^n \sum_{i'=1}^n (\mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'})^2 - C_2, \\ C_2 &\doteq \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{V} \mathbf{x}_i)^2, & C_4 &\doteq \sum_{i \neq i'} \sum_{i'=1}^n \mathbf{x}_i^T \mathbf{V} \mathbf{x}_i \mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_{i'} = C_1^2 - C_2. \end{aligned}$$

Write the $n \times n$ matrix

$$n\mathbf{H} = n\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{X} \mathbf{V} \mathbf{X}^T = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \mathbf{V} (\mathbf{x}_1, \dots, \mathbf{x}_n) = (\mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'}),$$

where $\mathbf{H} = (h_{ii'})$ is the hat matrix of regression, and observe that $C_1 = n \operatorname{tr}(\mathbf{H})$, $C_2 = n^2 \sum_{i=1}^n h_{ii}^2$ and $C_3 = n^2 2 \sum_{1 \leq i < i' \leq n} h_{ii'}^2$. It is well known that $\operatorname{tr}(\mathbf{H}) = \operatorname{rank}(\mathbf{X}) = p$, $0 \leq \sum_{i=1}^n h_{ii}^2 \leq p$ and $\sum_{i=1}^n \sum_{i'=1}^n h_{ii'}^2 = p$ (see, for example, [Chatterjee and Hadi, 1988](#)). Setting the parameter $\theta \doteq \sum_{i=1}^n h_{ii}^2 \in [0, p]$, we have that

$$C_1 = np, \quad C_2 = n^2 \theta, \quad C_3 = n^2(p - \theta), \quad C_4 = n^2(p^2 - \theta). \quad (14)$$

B.1.0.1. Normal Error Distribution. Assume the errors in the linear regression model are normally distributed, that is $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Let $\hat{\beta}_{S_j}, \hat{\beta}_{S_{j'}}$ be the ordinary least square (OLS) estimators computed on the training sets $S_j, S_{j'}$. It is well known that $\hat{\beta}_{S_j} = (\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1} \mathbf{X}_{S_j}^T \mathbf{y}_{S_j}$ (and similarly for the index j'), with $\mathbf{X}_{S_j} = \mathbf{E}_{S_j} \mathbf{X}$ and $\mathbf{y}_{S_j} = \mathbf{E}_{S_j} \mathbf{y}$ (and similarly for the index j') are defined above as the product of \mathbf{E}_{S_j} and \mathbf{X} . This allows one to write $\hat{\beta}_{S_j} = (\mathbf{X}^T \mathbf{I}_{S_j} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}_{S_j} \mathbf{y}$ and $\hat{\beta}_{S_{j'}} = (\mathbf{X}^T \mathbf{I}_{S_{j'}} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{I}_{S_{j'}} \mathbf{y}$.

Let the bivariate random vector

$$\xi_{j,j',i,i'} = (\xi_{j,i}, \xi_{j',i'})^T \doteq (\mathbf{x}_i^T \hat{\beta}_{S_j} - y_i, \mathbf{x}_{i'}^T \hat{\beta}_{S_{j'}} - y_{i'})^T.$$

Then, as n_1 becomes large

$$\begin{aligned} \xi_{j,j',i,i'} &\stackrel{d}{\approx} N_2 \left(\mathbf{0}, \Sigma_{j,j',i,i'}^\xi \right), \text{ where} \\ \Sigma_{j,j',i,i'}^\xi &\doteq \sigma^2 \begin{pmatrix} \frac{\mathbf{x}_i^T \mathbf{V} \mathbf{x}_i}{n_1} + 1 & \frac{\operatorname{card}(S_j \cap S_{j'})}{n_1^2} \mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'} - \frac{\mathbb{1}_{\{i' \in S_j\}} + \mathbb{1}_{\{i \in S_{j'}\}}}{n_1} \mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'} + \mathbb{1}_{\{i=i'\}} \\ * & \frac{\mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_{i'}}{n_1} + 1 \end{pmatrix}; \end{aligned} \quad (15)$$

and the moments of $\xi_{j,j',i,i'}$ are approximated by the associate moments of the approximated distribution.

Proposition 4. Under normality of errors, (11) and squared error loss

$$\begin{aligned} \mathbb{E}(\hat{\mu}_j) &= \left(1 + \frac{p}{n_1} \right) \sigma^2, \\ \operatorname{Var}(\hat{\mu}_j) &= \left\{ \frac{2}{n_2} + \frac{4p + 3\theta}{n_1 n_2} \right\} \sigma^4 + O(1/n^2), \quad \operatorname{Cov}(\hat{\mu}_j, \hat{\mu}_{j'}) = \frac{2\sigma^4}{n} + O(1/n^2), \end{aligned}$$

where $\theta = \sum_{i=1}^n h_{ii}^2$, $\theta \in [0, p]$, and h_{ii} is the i th diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

Proof. See Online Supplement. \square

We now relax the assumption of normality of errors by requiring a much weaker condition. That is, we require that there exists an $\epsilon > 0$ such that $\mathbb{E}(|\varepsilon_i|^{4+\epsilon}) < \infty$, effectively requiring the existence of moments of order $4 + \epsilon$. To relax the assumption of normality of the errors we need to guarantee the convergence of the moments of the estimators of β to the moments of the corresponding asymptotic distribution. [Afendras and Markatou \(2016\)](#) address this issue, as well as the associated rate of convergence.

Lemma 2. Define the random vector $\mathbf{W}_{i,i'} = (W_i, W_{i'})^T$, where $W_i = \sqrt{n_1} \mathbf{x}_i^T (\hat{\beta}_{S_j} - \beta)$, $W_{i'} = \sqrt{n_1} \mathbf{x}_{i'}^T (\hat{\beta}_{S_{j'}} - \beta)$, S_j is the index set of a training set and $i, i' \in S_j^c$. Then

$$\mathbf{W}_{i,i'} \xrightarrow{d} N_2(\mathbf{0}, \Sigma_{i,i'}^{\mathbf{W}}), \quad \text{as } n_1 \rightarrow \infty, \quad \text{where } \Sigma_{i,i'}^{\mathbf{W}} = \sigma^2 \begin{pmatrix} \mathbf{x}_i^T \mathbf{V} \mathbf{x}_i & \mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'} \\ \mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_i & \mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_{i'} \end{pmatrix}.$$

Proof. It follows easily by (12) and *delta*-method. \square

The following proposition provides the needed expressions for computing the variance of the CV generalization error estimate and the optimal training sample size.

Proposition 5. Under the assumption that $\mathbb{E}(|\varepsilon_i|^{4+\epsilon}) < \infty$, $\epsilon > 0$, and in the case of squared error loss,

$$\text{Var}(\hat{\mu}_j) = \frac{\mu_4 - \sigma^4}{n_2} + \frac{(4p + 3\theta)\sigma^4}{n_1 n_2} + O(1/n^2),$$

where $\mu_4 = \mathbb{E}(\varepsilon_i^4)$.

Proof. See Online Supplement. \square

B.1.0.2. k-fold case. In the k -fold CV case $n_1 = (1 - 1/k)n$ and $n_2 = n/k$. Then,

Proposition 6. Under normality of errors, (11) and squared error loss

$$\begin{aligned} \mathbb{E}(\hat{\mu}_j) &= \left(1 + \frac{kp}{(k-1)n}\right) \sigma^2, \\ \frac{1}{k} \text{Var}(\hat{\mu}_j) &= \left\{ \frac{2}{n} + \frac{k}{k-1} \frac{4p+3\theta}{n^2} + \frac{k}{(k-1)^2} \frac{2p}{n^2} \right\} \sigma^4 + O(1/n^3) = \frac{2\sigma^4}{n} + O(1/n^2), \\ \frac{k-1}{k} \text{Cov}(\hat{\mu}_j, \hat{\mu}_{j'}) &= \left\{ \frac{k}{k-1} \frac{p^2 - n\theta}{n^2(n-1)} + \frac{k^3}{(k-1)^3} \frac{2(p-\theta)}{n^2} \right\} \sigma^4 = O(1/n^2). \end{aligned}$$

Proof. See Online Supplement. \square

Remark 11. Consider the case of linear regression under squared error loss (see Section 3.2.2.1) and, in addition, assume that the ε_i s are iid from $N(0, \sigma^2)$. Then, we prove that $\mathbb{E}(\text{RTT}_j) = (1 + p/n_1)\sigma^2$ for the repeated train–test CV procedure, and $\mathbb{E}(\hat{\mu}_{k\text{-fold}}) = \{1 + kp/[(k-1)n]\}\sigma^2$ for the k -fold CV procedure, see Propositions 4 and 6 respectively. To obtain the bias of these estimators we need to compute the expected value of $L(\mathcal{Z}_N, Z)$, where $Z = (y, \mathbf{x})$ is an unobserved value which is independent of the universe data set \mathcal{Z}_N . Specifically, $\mathbb{E}[L(\mathcal{Z}_N, Z)] = \mathbb{E}[(\mathbf{x}^T \hat{\boldsymbol{\beta}} - y)^2]$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of the parameter vector $\boldsymbol{\beta}$ based on \mathcal{Z}_N . This, expected value for both cases, repeated train–test and k -fold CV, is $[\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} + 1]\sigma^2$. This value may not be used for some designs, because it contains the unobserved/unknown value \mathbf{x} which sometimes cannot be estimated; and thus, the bias of the estimators cannot be estimated also.

B.2. Classification via logistic regression

Given S_j and $i, i' \in S_j^c$ standardize the components of $\mathbf{W}_{i,i'}$ in Lemma 2 as $\Psi_i = W_i/[\sigma(\mathbf{x}_i^T \mathbf{V} \mathbf{x}_i)^{1/2}]$, $\Psi_{i'} = W_{i'}/[\sigma(\mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_{i'})^{1/2}]$ and consider the random vector $\boldsymbol{\Psi}_{i,i'} = (\Psi_i, \Psi_{i'})^T$. Then,

$$\boldsymbol{\Psi}_{i,i'} \xrightarrow{d} N_2(\mathbf{0}, \boldsymbol{\Sigma}_{i,i'}^{\boldsymbol{\Psi}}), \quad \text{where } \boldsymbol{\Sigma}_{i,i'}^{\boldsymbol{\Psi}} = \begin{pmatrix} 1 & \rho_{i,i'} \\ \rho_{i,i'} & 1 \end{pmatrix}, \quad \text{with } \rho_{i,i'} = \frac{\mathbf{x}_i^T \mathbf{V} \mathbf{x}_{i'}}{(\mathbf{x}_i^T \mathbf{V} \mathbf{x}_i \mathbf{x}_{i'}^T \mathbf{V} \mathbf{x}_{i'})^{1/2}}.$$

Also, define

$$\zeta_i = \frac{\mathbf{x}_i^T \boldsymbol{\beta}}{\sigma(\mathbf{x}_i^T \mathbf{V} \mathbf{x}_i)^{1/2}}, \quad i = 1, \dots, n.$$

The decision rule for classification is then given as follows. For a training set \mathcal{Z}_{S_j} , from the model we estimate the probability p_i for each $i \in S_j^c$, say $\hat{p}_{S_j,i}$ its estimator. After, we estimate the value y_i as

$$\hat{y}_{S_j,i} = \mathbb{1}_{\{\hat{p}_{S_j,i} \geq 1/2\}} = \mathbb{1}_{\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{S_j} \geq 0\}} = \mathbb{1}_{\{\Psi_i \geq -\sqrt{n_1} \zeta_i\}}.$$

The loss function is $L_{0/1}(\hat{y}_{S_j,i}, y_i) = \mathbb{1}_{\{\hat{y}_{S_j,i} \neq y_i\}}$ (see McAllester, 2007). We compute the expected values $\mathbb{E}[L_{0/1}(\hat{y}_{S_j,i}, y_i)|S_j, i] = \mathbb{E}[L_{0/1}^2(\hat{y}_{S_j,i}, y_i)|S_j, i]$ and $\mathbb{E}[L_{0/1}(\hat{y}_{S_j,i}, y_i)L_{0/1}(\hat{y}_{S_j,i'}, y_{i'})|S_j, i, i']$, see Online Supplement, which are given as

$$\begin{aligned} e_i &= e_{i,i} = \Phi(-\sqrt{n_1} \zeta_i) p_i + \Phi(\sqrt{n_1} \zeta_i) (1 - p_i), \\ e_{i,i'} &= \Phi_{2,\rho_{i,i'}}(-\sqrt{n_1} \zeta_i, -\sqrt{n_1} \zeta_{i'}) p_i p_{i'} + \Phi_{2,-\rho_{i,i'}}(-\sqrt{n_1} \zeta_i, \sqrt{n_1} \zeta_{i'}) p_i (1 - p_{i'}) \\ &\quad + \Phi_{2,-\rho_{i,i'}}(\sqrt{n_1} \zeta_i, -\sqrt{n_1} \zeta_{i'}) (1 - p_i) p_{i'} + \Phi_{2,\rho_{i,i'}}(\sqrt{n_1} \zeta_i, \sqrt{n_1} \zeta_{i'}) (1 - p_i) (1 - p_{i'}), \end{aligned} \quad (16)$$

where Φ is the cumulative distribution function of standard normal distribution and $\Phi_{2,\rho}$ is the cumulative distribution function of $N_2\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

Theorem 3 gives a formula to calculate the variance of $\hat{\mu}_j$. But, in classification via logistic regression case, the minimization process of (5) does not give a close form for the optimal value of n_1 , it is reduced to a numerical minimization process. Computing the quantities p_i , ζ_i and $\rho_{i,i'}$, via **Theorem 3** and (16), we calculate the variance of $\hat{\mu}_j$ for $n_1 = \lfloor n/2 \rfloor, \dots, n-1$. The value of n_1 which gives the minimum value of the variance is the optimal choice of n_1 .

The parameters σ^2, β are unknown and we estimate those before the minimization process begins using the entire data set. We also compute \mathbf{V} as $\hat{\mathbf{V}} = n(\mathbf{X}^T \mathbf{X})^{-1}$.

Algorithm 1 Optimal size of training set in classification via logistic regression.

- 1: Compute the matrix $\mathbf{V} = n(\mathbf{X}^T \mathbf{X})^{-1}$ and via logistic regression estimate the parameters β and σ^2 , say $\hat{\beta}$ and $\hat{\sigma}^2$ respectively, using the entire sample.
- 2: Compute the following probabilities and quantities

$$p_i = \Pr(y_i = 1 | \mathbf{x}_i), \quad \hat{\zeta}_i = \frac{\mathbf{x}_i^t \hat{\beta}}{\hat{\sigma}(\mathbf{x}_i^t \mathbf{V} \mathbf{x}_i)^{1/2}}, \quad \hat{\rho}_{i,i'} = \frac{\mathbf{x}_i^t \mathbf{V} \mathbf{x}_{i'}}{(\mathbf{x}_i^t \mathbf{V} \mathbf{x}_i \mathbf{x}_{i'}^t \mathbf{V} \mathbf{x}_{i'})^{1/2}}, \quad i, i' = 1, \dots, n.$$

▷ Each $\hat{\rho}_{i,i'}$ is between -1 and 1 .

- 3: For $n_1 = \lfloor n/2 \rfloor, \dots, n-1$

- a. Compute the quantities e_i , $e_{i,i}$ and $e_{i,i'}$ of **Theorem 3**

$$\begin{aligned} e_i &= e_{i,i} = \Phi(-\sqrt{n_1} \hat{\zeta}_i) p_i + \Phi(\sqrt{n_1} \hat{\zeta}_i) (1 - p_i), \\ e_{i,i'} &= \Phi_{2,\hat{\rho}_{i,i'}}(-\sqrt{n_1} \hat{\zeta}_i, -\sqrt{n_1} \hat{\zeta}_{i'}) p_i p_{i'} + \Phi_{2,-\hat{\rho}_{i,i'}}(-\sqrt{n_1} \hat{\zeta}_i, \sqrt{n_1} \hat{\zeta}_{i'}) p_i (1 - p_{i'}) \\ &\quad + \Phi_{2,-\hat{\rho}_{i,i'}}(\sqrt{n_1} \hat{\zeta}_i, -\sqrt{n_1} \hat{\zeta}_{i'}) (1 - p_i) p_{i'} + \Phi_{2,\hat{\rho}_{i,i'}}(\sqrt{n_1} \hat{\zeta}_i, \sqrt{n_1} \hat{\zeta}_{i'}) (1 - p_i) (1 - p_{i'}). \end{aligned}$$

- b. Using **Theorem 3**, compute $V(n_1) = \text{Var}(\hat{\mu}_j)$.

- 4: Set $n_1^{\text{opt}} = \arg \min_{n_1 = \lfloor n/2 \rfloor, \dots, n-1} \{V(n_1)\}$.
-

Appendix C. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jspi.2018.07.005>.

References

- Afendras, G., Markatou, M., 2016. Uniform integrability of the OLS estimators, and the convergence of their moments. *TEST* 25 (4), 775–784. <http://dx.doi.org/10.1007/s11749-016-0498-y>.
- Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., Salakoski, T., 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput. Statist. Data Anal.* 55 (4), 1828–1844. <http://dx.doi.org/10.1016/j.csda.2010.11.018>.
- Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4, 40–79. <http://dx.doi.org/10.1214/09-SS054>.
- Bengio, Y., Grandvalet, Y., 2003–2004. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* 5, 1089–1105.
- Berger, J.O., Pericchi, L.R., 2004. Training samples in objective Bayesian model selection. *Ann. Statist.* 32 (3), 841–869. <http://dx.doi.org/10.1214/009053604000000238>.
- Braga-Neto, U.M., Dougherty, E.T., 2004. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20 (3), 374–380. <http://dx.doi.org/10.1093/bioinformatics/btg419>.
- Burman, P., 1989. A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika* 76 (3), 503–514. <http://dx.doi.org/10.1093/biomet/76.3.503>.
- Burman, P., 1990. Estimation of optimal transformations using v -fold cross validation and repeated learning-testing methods. *Sankhya A* 52 (3), 314–345.
- Cabras, S., Castellanos, M.E., Perra, S., 2015. A new minimal training sample scheme for intrinsic Bayes factors in censored data. *Comput. Statist. Data Anal.* 81, 52–63. <http://dx.doi.org/10.1016/j.csda.2014.07.012>.
- Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.
- Chatterjee, S., Hadi, A.S., 1988. Sensitivity Analysis in Linear Regression. In: *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, John Wiley & Sons, Inc., New York, p. xvi+315. <http://dx.doi.org/10.1002/9780470316764>.
- Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* 7, 1–30.
- Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10, 1895–1923.
- Dobbin, K.K., Simon, R.M., 2005. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* 6, 27–38. <http://dx.doi.org/10.1093/biostatistics/kxh015>.
- Dobbin, K.K., Simon, R.M., 2011. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med. Genomics* 4 (1), 31.
- Dobbin, K.K., Zhao, Y., Simon, R.M., 2008. How large a training set is needed to develop a classifier for microarray data? *Clin. Cancer Res.* 14 (1), 108–114. <http://dx.doi.org/10.1158/1078-0432.ccr-07-0443>.
- Efron, B., 1986. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* 81 (394), 461–470. URL [http://links.jstor.org/sici?sici=0162-1459\(198606\)81:394<461:HBITAE>2.0.CO;2-T&origin=MSN](http://links.jstor.org/sici?sici=0162-1459(198606)81:394<461:HBITAE>2.0.CO;2-T&origin=MSN).

- Fuchs, M., Hornung, R., De Bin, R., Boulesteix, A.L., 2013. A U-statistic estimator for the variance of resampling-based error estimators. *arXiv:1310.8203*.
- Fukunaga, K., Hayes, R., 1989. Estimation of classifier performance. *IEEE Trans. Pattern Anal. Mach. Intell.* 11 (10), 1087–1101. <http://dx.doi.org/10.1109/34.42839>.
- Garcia, S., Herrera, F., 2008. An extension on “statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons. *J. Mach. Learn. Res.* 9, 2677–2694.
- Guyon, I., Makhoul, J., Schwartz, R., Vapnik, V., 1998. What size test set gives good error rate estimates? *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (1), 52–64. <http://dx.doi.org/10.1109/34.655649>.
- Hall, P., Robinson, A.P., 2009. Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika* 96 (1), 175–186. <http://dx.doi.org/10.1093/biomet/asn068>.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. In: Springer Series in Statistics, Springer-Verlag, New York, p. XXII, 745. <http://dx.doi.org/10.1007/978-0-387-84858-7>.
- Highleyman, H.W., 1962. The design and analysis of pattern recognition experiments. *Bell Syst. Tech. J.* 41 (1), 723–744. <http://dx.doi.org/10.1002/j.1538-7305.1962.tb02426.x>.
- Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1), 55–67. <http://dx.doi.org/10.1080/00401706.1970.10488634>.
- Jin, H., Stehman, S.V., Mountrakis, G., 2014. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. *Int. J. Remote Sens.* 35 (6), 2067–2081. <http://dx.doi.org/10.1080/01431161.2014.885152>.
- Kearns, M., 1997. A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. *Neural Comput.* 9 (5), 1143–1161. <http://dx.doi.org/10.1162/neco.1997.9.5.1143>.
- Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in: *IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137–1143.
- Larsen, J., Goutte, C., 1999. On optimal data split for generalization estimation and model selection, in: *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing IX*, 225–234. <http://dx.doi.org/10.1109/NNSP.1999.788141>.
- Markatou, M., Dimova, R., Sinha, A., 2011. A comparison of estimators for the variance of cross-validation estimators of the generalization error of computer algorithms. In: *Nonparametric Statistics and Mixture Models*. World Sci. Publ., Hackensack, NJ, pp. 226–251. http://dx.doi.org/10.1142/9789814340564_0014.
- Markatou, M., Tian, H., Biswas, S., Hripcsak, G., 2005. Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.* 6, 1127–1168. <http://www.jmlr.org>.
- McAllester, D., 2007. Statistical methods for artificial intelligence. In: *Lecture Notes for TTIC103*, Toyota Technological Institute at Chicago. URL <http://ttic.uchicago.edu/~dmcalleser/ttic101-06/lectures/genreg/genreg.pdf>.
- Molinari, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21 (15), 3301–3307. <http://dx.doi.org/10.1093/bioinformatics/bti499>.
- Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. *Mach. Learn.* 52 (3), 239–281. <http://dx.doi.org/10.1023/A:1024068626366>.
- Popovici, V., Chen, W., Gallas, B.G., Hatzis, C., Shi, W., Samuelson, F.W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., Hess, K.R., Valero, V., Booser, D., Delorenzi, M., Hortobagyi, G.N., Shi, L., Symmans, W.F., Pusztai, L., 2010. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* 12 (1), R–5. <http://dx.doi.org/10.1186/bcr2468>.
- Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (3), 252–264. <http://dx.doi.org/10.1109/34.75512>.
- Schiavo, R.A., Hand, D.J., 2000. Ten more years of error rate research. *Internat. Statist. Rev.* 68 (3), 295–310. <http://dx.doi.org/10.1111/j.1751-5823.2000.tb00332.x>.
- Shao, J., 1993. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88 (422), 486–494. URL: [http://links.jstor.org/sici?sici=0162-1459\(199306\)88:422<486:LMSBC>2.0.CO;2-C&origin=MSN](http://links.jstor.org/sici?sici=0162-1459(199306)88:422<486:LMSBC>2.0.CO;2-C&origin=MSN).
- Shao, L., Fan, X., Cheng, N., Wu, L., Cheng, Y., 2013. Determination of minimum training sample size for microarray-based cancer outcome prediction—an empirical assessment. *PLOS One* 8 (7), e68579. <http://dx.doi.org/10.1371/journal.pone.0068579>.
- Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. *Stat. Comput.* 14 (3), 199–222. <http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88>.
- van de Wiel, M.A., Berkhof, J., van Wieringen, W.N., 2009. Testing the prediction error difference between 2 predictors. *Biostatistics* 10 (3), 550–560. <http://dx.doi.org/10.1093/biostatistics/kxp011>.
- Wang, Q., Lindsay, B., 2014. Variance estimation of a general U-statistic with application to cross-validation. *Statist. Sinica* 24 (3), 1117–1141.
- Wang, Q., Lindsay, B., 2017. Pseudo-kernel method in U-statistic variance estimation with large kernel size. *Statist. Sinica* 27 (3), 1155–1174.
- Yang, Y., 2006. Comparing learning methods for classification. *Statist. Sinica* 16 (2), 635–657.
- Zhen, Z., Quackenbush, L.J., Stehman, S.V., Zhang, L., 2013. Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification. *Int. J. Remote Sens.* 34 (19), 6914–6930. <http://dx.doi.org/10.1080/01431161.2013.810822>.