Penalized regression techniques for prediction: A case study for predicting tree mortality using remotely sensed vegetation indices

Article *in* Canadian Journal of Forest Research · January 2011

DOI: 10.1139/X10-180

CITATIONS
16

READS
292

3 authors, including:

Jan Verbesselt
Wageningen University & Research
120 PUBLICATIONS **5,059** CITATIONS

SEE PROFILE

Penalized regression techniques for prediction: a case study for predicting tree mortality using remotely sensed vegetation indices[1]

David C. Lazaridis, Jan Verbesselt, and Andrew P. Robinson

Abstract: Constructing models can be complicated when the available fitting data are highly correlated and of high dimension. However, the complications depend on whether the goal is prediction instead of estimation. We focus on predicting tree mortality (measured as the number of dead trees) from change metrics derived from moderate-resolution imaging spectroradiometer satellite images. The high dimensionality and multicollinearity inherent in such data are of particular concern. Standard regression techniques perform poorly for such data, so we examine shrinkage regression techniques such as ridge regression, the LASSO, and partial least squares, which yield more robust predictions. We also suggest efficient strategies that can be used to select optimal models such as 0.632+ bootstrap and generalized cross validation. The techniques are compared using simulations. The techniques are then used to predict insect-induced tree mortality severity for a *Pinus radiata* D. Don plantation in southern New South Wales, Australia, and their prediction performances are compared. We find that shrinkage regression techniques outperform the standard methods, with ridge regression and the LASSO performing particularly well.

Résumé : L'élaboration de modèles peut être compliquée lorsque les données à modéliser sont étroitement corrélées et de grande dimension. Cependant, les difficultés ne sont pas les mêmes si l'objectif est la prédiction plutôt que l'estimation. Nous nous concentrons sur la prédiction de la mortalité des arbres (mesurée par le nombre d'arbres morts) à partir de mesures de changement dérivées d'images satellitaires obtenues grâce au spectroradiomètre imageur à résolution moyenne. La grande dimensionnalité et la multicolinéarité inhérentes à de telles données sont particulièrement préoccupantes. Les techniques standards de régression sont peu performantes pour de telles données. Nous avons donc étudié les méthodes de régularisation de la régression, telles que la régression bornée, de type LASSO et des moindres carrés partiels, qui produisent des prédictions plus robustes. Nous proposons également des stratégies efficaces qui peuvent être utilisées pour choisir les meilleurs modèles, tels que 0,632+ l'auto-amorçage et la validation croisée généralisée. Les méthodes sont comparées à l'aide de simulations. Les méthodes sont ensuite utilisées pour prédire la sévérité de la mortalité induite par les insectes dans une plantation de *Pinus radiata* D. Don située dans le sud de l'État de la Nouvelle-Galles du Sud, en Australie et leurs performances prédictives sont comparées. Nous arrivons la conclusion que les méthodes de régularisation de la régression donnent de meilleurs résultats que les méthodes standards et que la régression bornée et de type LASSO sont particulièrement performantes.

[Traduit par la Rédaction]

Introduction

Insect-induced tree mortality causes significant timber losses in many regions of the world. It is important to be able to make timely predictions of tree mortality to guide forest and plantation management decisions, whether to diagnose forest health issues or to obtain predictions of the volume of resources.

In most Australian states, aerial surveillance is used to map the extent, incidence, and severity of tree mortality in plantations. However, aerial sketch maps and followup ground surveys are only conducted once a year due to the relatively high cost of aircraft hire and the paucity of experienced personnel. These infrequent observations are of little help if the goal is to detect changes in forest health sufficiently early that meaningful remedial action can be taken.

Satellite imagery has been used to detect tree mortality at the level of individual forest stands and trees (Coops et al. 2006; Wulder et al. 2006; White et al. 2007). For example, Wulder et al. (2006) provided a recent review of remote-sensing capacity for mapping mountain pine beetle (*Dendroctonus ponderosae*) damage at different spatial scales.

Received 12 January 2010. Accepted 12 September 2010. Published on the NRC Research Press Web site at cjfr.nrc.ca on 20 December 2010.

D.C. Lazaridis[2] and A.P. Robinson. Department of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Parkville, Victoria 3088, Australia.
J. Verbesselt. Remote Sensing Team, CSIRO Sustainable Ecosystems, Private Bag 10, Melbourne, Victoria 3169, Australia; Centre for Geo-Information, Wageningen University, Droevendaalsesteeg 3, 6708 PB, Wageningen, The Netherlands.

[1]This article is one of a selection of papers from Extending Forest Inventory and Monitoring over Space and Time.
[2]Corresponding author (e-mail: d.lazaridis@pgrad.unimelb.edu.au).

The authors showed that medium spatial resolution satellite data (e.g., Landsat) may be used for mapping beetle damage with 70%–75% overall accuracy and that high spatial resolution data (e.g, Quickbird) can achieve 71%–92% accuracy, depending on the level of infestation.

Recently, remote-sensing techniques have been developed that use temporal sequences of high spatial resolution images to map the locations of dead and dying trees (Goodwin et al. 2008; Wulder et al. 2008). However, infrequent coverage due to cloud cover and differences in viewing and illumination geometry limit the operational practicality of using high spatial resolution satellite data to assess forest health. Moreover, acquiring and processing such data over large areas can be prohibitively expensive (Wulder et al. 2008).

Coarse spatial resolution imagery with a high temporal resolution such as that provided by the advanced very high resolution radiometer (AVHRR), satellite pour l'observation de la terre (SPOT), and moderate-resolution imaging spectroradiometer (MODIS) can overcome the above mentioned issues by providing data that are available more frequently, have large coverage, and are much easier and cheaper to process.

Numerous studies show that coarse spatial resolution data can be successfully used to estimate insect-induced defoliation and mortality in forests (Fraser and Latifovic 2005; de Beurs and Townsend 2008). However, these studies have focused on mapping or estimating existing disturbances, whereas predicting future tree mortality is of critical importance for timely management decisions.

Verbesselt et al. (2009) derived change metrics for the normalized difference vegetation index (NDVI), normalized difference infrared index (NDII), and enhanced vegetation index (EVI) based on MODIS data to predict mortality severity for a *Pinus radiata* D. Don plantation in southern New South Wales, Australia. The question of interest was how well these change metrics could predict tree mortality and how many years of satellite data were required to obtain optimal estimates. Here, we shall focus only on change metrics derived from NDVI to predict mortality, as Verbesselt et al. (2009) showed that for the present study area, NDVI outperforms NDII- and EVI-derived change metrics at differentiating between areas of high and low tree mortality.

Our aim is to build a linear regression model in which we model our response variable (forest health) in terms of available predictor variables. Forest health can be represented using health metrics such as defoliation assessments, leaf area index, or forest inventory data. The response variable that we adopt for this study is insect-induced tree mortality, estimated as the square root of the number of dead tree crowns in each MODIS pixel determined via visual inspection of 15 cm imagery. Our predictor variables are the change metrics further described in "The MODIS data" section in the Materials and methods. Our model consists of change metrics derived from data ranging from February 2000 to February 2007 to predict the observed number of dead tree crowns in August 2007.

This research fits within an Australian forest health monitoring framework in which MODIS satellite data are used as a "first-pass" filter to identify regions and the timing of major change activity (Stone et al. 2008).

Analysis of remote-sensing data for prediction purposes can be complicated. Remote-sensing data sets are typically high dimensional, that is, the number of predictor variables is quite large relative to the number of observations, especially when ground-based observations such as climate and other data are included in the analysis. This leads to the problem of overfitting in which erroneous relationships between the predictors and the response are established. Also of concern is the potential high multicollinearity of the predictor variables. Standard regression techniques perform poorly in these situations, and thus, we will examine shrinkage regression techniques, which yield more robust predictions in the presence of high dimensionality and multicollinearity (Hastie et al. 2001).

The model selection techniques that we focus on are rather unsupervised, as the shrinkage regression and the model selection techniques that we consider readily lend themselves to this methodology. Unsupervised techniques for model selection give us a principled and unbiased method for determining which variables offer superior prediction performance or explanatory power. Expert knowledge can be used to determine which variables might or might not be included in a model due to external reasons, such as including them based on a theoretical or physical basis or removing them due to cost or other issues inherent to the data.

In reality, when a large number of predictor variables are available, a combination of expert knowledge and unsupervised techniques should be used.

The rest of the paper is structured as follows. In the next section, we briefly review popular regression techniques and model selection tools. In the Materials and methods, we describe how the data were gathered and prepared and outline the procedures that we use for model comparisons. In the Results, we report the fit of the various models for the MODIS and simulated data using the prescribed methods and determine which offer the best prediction properties. In the Discussion, we consider the effectiveness of these methods and discuss shortfalls and potential improvements that can be made to improve mortality prediction. The utility of these results for remote-sensing and forestry applications are also considered.

Background

In this section, we review various regression techniques in the context of overfitting and correlation among predictors. Our particular focus is in shrinkage regression, which is an estimation method that involves the use of penalties or constraints to "shrink" parameter estimates to avoid overfitting and can select a superior subset of predictors. By shrinking parameter estimates, shrinkage regression can reduce the impact that an unimportant variable has on a model and in some cases remove it completely.

The rationale for the selection of each of the techniques is briefly laid out as follows. Least-squares regression, coupled with stepwise model selection, is a standard model fitting and comparison algorithm for high-dimensional data sets. Ridge regression and the least absolute shrinkage and selection operator (LASSO) both penalize the parameter estimates to avoid overfitting; the LASSO also provides automated subset selection. Finally, partial least squares

(PLS) is a popular method for high-dimensional modeling in various fields (Wold et al. 2001; Carrascal et al. 2009).

For the shrinkage regression techniques, it is necessary to standardize the data set or design matrix before fitting. We transform all predictor variables (continuous and categorical) to have 0 mean and standard deviation equal to 1. This step is performed to ensure that the measurement scales of the estimates do not affect the fitting.

Least-squares and stepwise selection

Least-squares regression is used when we wish to fit a model such that a response variable Y is related to a number of explanatory or predictor variables X_1, X_2, ..., X_p. Although the formal details for least-squares estimation are well known, we introduce them briefly here to provide context for the following techniques.

The least-squares estimates of β are those that minimize the sum of the squared differences between the observed values and the predictions. For a linear model, this is given by

$$[1] \qquad \widehat{\beta}_{\mathrm{LS}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j \right)^2$$

where N is the number of observations in the data set. The least-squares estimates for β can be found analytically.

If the number of predictors p is large, then the least-squares approach may produce an overfitted model, which will have poor prediction performance for future observations. A popular method for producing parsimonious least-squares models is to use a "step" procedure. A step procedure is a method of iterative variable subset selection for which at each step, a multitude of models containing different subsets of predictors is compared against a nominated model. This comparison is normally done by comparing the models' F statistics or Akaike information criterion (AIC) (Akaike 1974) or Bayesian information criterion (BIC) (Schwarz 1978) values. While the procedure is easy to implement and appears intuitive, it may result in flawed models. Among other issues, stepwise procedures ignore the problem of multicollinearity of predictors and produce models whose R^2 are biased high. Harrell (2000, pp. 56–60) provided a good summary of the problems with using this method. Step procedures are generic and can be deployed in combination with other model-fitting techniques, although this rarely seems to happen.

Ridge regression

Ridge regression, introduced by Tikhonov (1943) and popularized by Hoerl and Kennard (1970*a*, 1970*b*), involves shrinking the regression coefficients by penalizing them or, equivalently, putting a constraint on their possible values. Formally, ridge regression involves minimizing the sum of squared errors including an L_2-norm penalty on the size of the parameter estimates. The coefficient estimates $\widehat{\beta}_{\mathrm{RR}}$ are given by

$$[2] \qquad \widehat{\beta}_{\mathrm{RR}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j \right)^2 + \lambda \sum_{j=1}^{p} \widehat{\beta}_j^2$$

where λ controls the amount of shrinkage. As with least squares, ridge regression estimates can be found analytically.

While it is easy to generate ridge regression solutions for a given λ, it is not obvious which particular value of λ to use, that is, we do not know beforehand how much shrinkage should be applied. Usually we find λ such that the prediction error is minimized on an independent data set or via cross-validation or bootstrapping (see "Model selection" section in the Materials and methods for more details).

Ridge regression results in lower variance for parameter estimates and increased prediction accuracy relative to ordinary least squares (Miller 2002). However, the parameter estimates are no longer unbiased, and the technique retains all parameters in the model, which can be undesirable when candidate predictor variables are expensive and we seek parsimonious solutions.

LASSO regression

The least absolute shrinkage and selection operator (LASSO) is a method for the estimation of linear models proposed by Tibshirani (1996) that involves minimizing the sum of squared errors with an L_1-norm penalty on the parameter estimates:

$$[3] \qquad \widehat{\beta}_{\mathrm{L}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j \right)^2 + \lambda \sum_{j=1}^{p} |\widehat{\beta}_j|$$

Again, λ controls the amount of shrinkage. One of the noted features of the LASSO is that, given enough shrinkage, it can set parameter estimates to be exactly 0, effectively removing them from the model. Thus, the LASSO performs automated feature selection as a part of the estimation procedure, resulting in parsimonious models. However, as with ridge regression, it is still not obvious how much shrinkage to apply. Strategies for selecting the optimal tuning parameter λ are discussed in the "Model selection" section.

Unlike the case for least squares or ridge regression, a closed-form analytic solution for the LASSO estimates does not exist. Iterative methods have been developed such as the local linearization and active set methods (Osborne et al. 2000*a*, 2000*b*) and the least angle regression algorithm (Efron et al. 2004).

PLS regression

PLS is a regression technique that was developed by Wold (2001) for modeling complex economics problems. PLS has become popular in many fields of research due to its ability to handle high-dimensional and highly multicollinear data sets. It also has ability to handle data sets with missing features. Wold et al. (2001) provided a useful overview of the PLS technique.

PLS has been used extensively with hyperspectral data (Coops et al. 2003; Smith et al. 2003; Asner and Martin 2008). Wolter et al. (2008, 2009) demonstrated the capability of this approach to estimate forest structural parameters using broadband satellite sensor data. PLS regression is convenient in this regard, as it allows simultaneous modeling of multiple continuous predictor variables, it does not make unrealistic assumptions about measurement errors, as in ordi-

nary least-squares regression (Wolter et al. 2009), and it addresses the problem of multicollinearity among multiple predictor and response variables. The method is similar to principal components analysis/regression (PCA/PCR) (Hastie et al. 2001).

We briefly sketch the algorithm. The PLS method generates Z, a new set of predictor variables, from linear combinations of the standardized observed variables X and the centred response variable Y. The linear combinations are created sequentially according to the following process. First, form a new predictor Z_1 as the weighted sum of the standardized predictors X, where the weights are the regression coefficients of each predictor fit singly against the response Y. Then create a new response variable, which is the residuals of the regression of Z_1 against Y, and a new set of predictor variables, which is that part of X that is orthogonal to Z_1. Repeat this process until p new predictors have been formed. These new variables are then used as the predictor variables, added in order of creation, for the regression model.

As with the other shrinkage techniques, the number of predictors to be retained has to be determined as a part of the model-fitting process (see "Model selection" section). When the number of components (A) equals the total number of parameters or predictor variables, p, PLS will generate the standard least-squares solution. By reducing A, the resulting estimates are shrunk in magnitude but retained in the model; thus, PLS models do not provide parsimonious solutions.

Materials and methods

The MODIS data

The data used for this study comprise 62 observations, each corresponding to a 250 m × 250 m area covered by a MODIS pixel. The number of dead tree crowns in each pixel was determined via visual inspection of 15 cm resolution imagery taken with an ADS40 digital camera (Leica Geosystems) in August 2007.

MODIS data for each pixel were obtained as 16-day composite images (MOD13Q1 collection 5) that were generated using a constrained view angle maximum NDVI value compositing (CV-MVC) technique (Huete et al. 2002). The images have been radiometrically, geometrically, and atmospherically corrected. The temporal coverage of the imagery is from February 2001 until the start of 2007. The images are projected in the global sinusoidal grid projection system and have a spatial resolution of 231.7 m.

The MOD13Q1 data include red (620–670 nm) and near-infrared (NIR) (841–876 nm) reflectance bands that are used to derive NDVI as follows:

$$[4] \qquad \mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{red}}{\mathrm{NIR} + \mathrm{red}}$$

Verbesselt et al. (2009) derived change metrics to capture the change in stand vigor related to crown phenology between winter and summer for a given vegetation index:

$$[5] \qquad \mathrm{Change} = \frac{\mathrm{VI}_{\mathrm{winter}} - \mathrm{VI}_{\mathrm{summer}}}{\mathrm{VI}_{\mathrm{winter}}}$$

where $\mathrm{VI}_{\mathrm{summer}}$ is the median vegetation index value during the (Southern Hemisphere) summer months, defined to be December through to February, and $\mathrm{VI}_{\mathrm{winter}}$ is the median VI value during the winter months, which are defined as June through to August immediately preceding that summer. For example, C_{67} represents the change metric for winter in 2006 and summer ending in 2007. Six MODIS images were available for each summer and winter season within each pixel. The median vegetation index for each of these six images was derived to reduce short-term variability that is not attributable to changes in vegetation cover (such as subpixel clouds, surface moisture, haze, and low geometric precision). This analysis examined change metrics that have been derived for years 2001 until the beginning of 2007. A more detailed description of how the data were prepared can be found in Verbesselt et al. (2009).

Our goal was to predict tree mortality. To test the performance of the regression strategies, we fitted a range of models on our MODIS-derived data and calculated the prediction error using the 0.632+ bootstrap (see Appendix A for details). Our response variable was tree mortality defined as the square root of dead tree count measured in August 2007 and our predictor variables were change metrics derived from MODIS data spanning 2001–2007. The full model that we considered took the form

$$[6] \qquad \text{Tree mortality} \sim \mu + C_{67} + C_{56} + C_{45} + C_{34} + C_{23} + C_{12} + \epsilon$$

by which we mean the conditional distribution of tree mortality given the noted predictors, with an intercept term included. It is assumed that $\epsilon \sim N(0, \sigma^2)$. Figure 1 shows the pairwise relationships between the response variable and the predictors. The high-multicollinearity of the predictors is evident.

As we are dealing with count data in a standard linear model approach, we see two potential issues: first, the possibility of a negative number being predicted for our response, and second, nonconstant variance or the possibility of a relationship between the mean and variance. Fortunately, neither of these appears to be a concern.

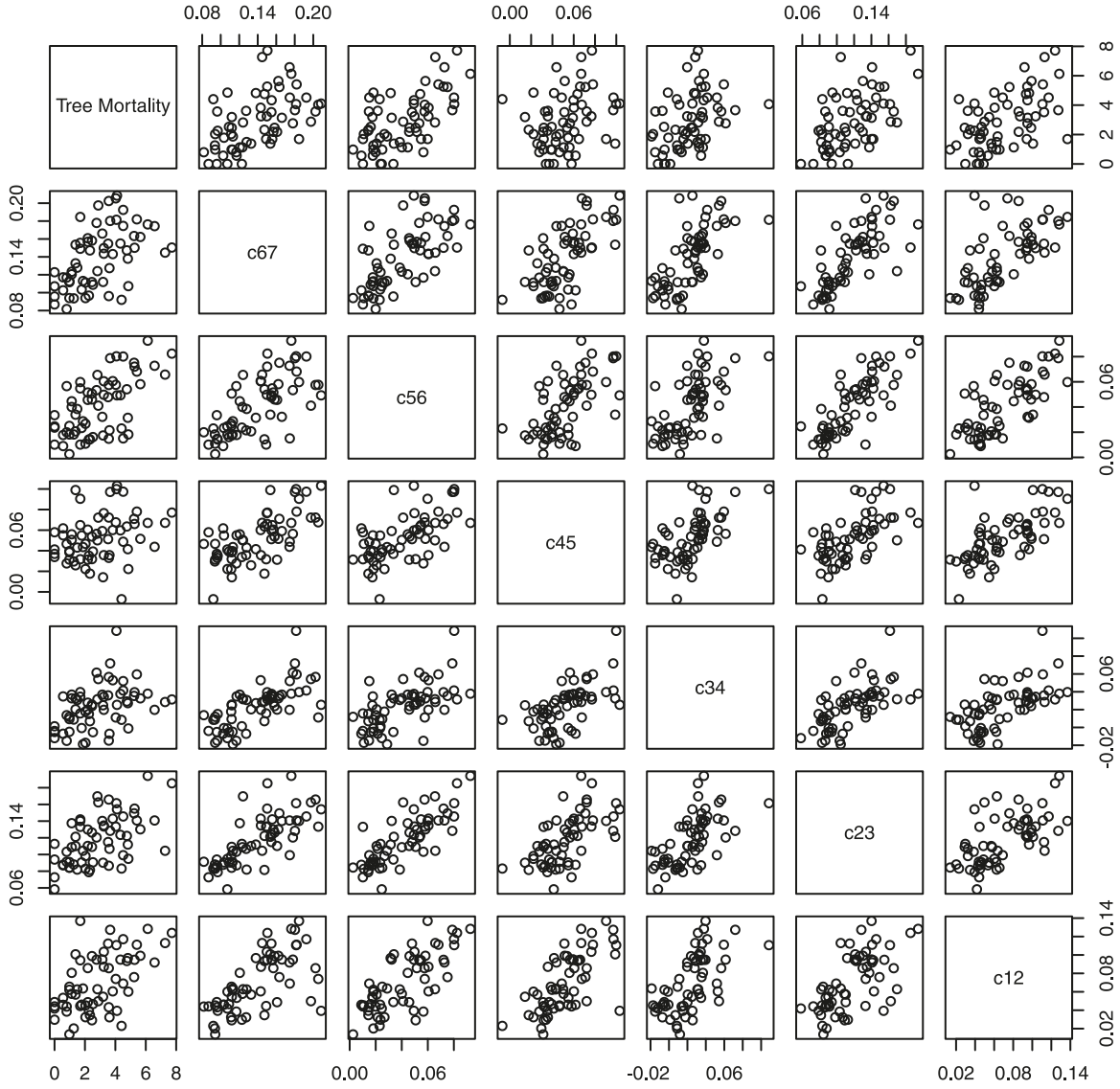
Model selection

The two challenges for model selection are as follows. First, we must choose the best model among a family of competing models, that is, to resolve the question of which terms to retain in stepwise procedures, how much shrinkage for ridge regression, or how many components to retain for PLS. Second, we must compare the results between families, that is, compare the best-fitting ridge regression model with the best-fitting PLS model.

Each of the modeling strategies requires the selection of a model from competing models. An objective function must be chosen to compare the models. The modeling strategies differ in what they choose to optimize, so a universal objective function for choosing the best model within a strategy is unnecessary and undesirable. Here, we outline the objective functions nominated for each modeling strategy.

We adopted AIC as the objective function in least-squares regression and stepwise model selection. For ridge regression and the LASSO, we adopted generalized cross-valida-

Fig. 1. Pairwise plots of our remote-sensed data set. The response, tree morality, and our change metric predictor variables.



tion (GCV) (Golub et al. 1979) and nonlinear GCV (nGCV) (Fu 2005) for our objective functions:

$$[7] \qquad \mathrm{GCV} = \frac{\sum_{i=1}^{N}\left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j\right)^2}{N(1 - p/N)^2}$$

$$[8] \qquad \mathrm{nGCV} = \frac{\sum_{i=1}^{N}\left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j\right)^2}{N(1 - ps/N)^2}$$

where p is the number of parameters in the full model and

$$[9] \qquad s = \left(\sum_{j=1}^{p}|\widehat{\beta}_j|^q\right)^{1/q}\left(\sum_{j=1}^{p}|\widehat{\beta}_j^{LS}|^q\right)^{-1/q}$$

is a shrinkage rate for which q represents the magnitude of the norm for the penalty on the parameter estimates, i.e., q = 1 for the LASSO and q = 2 for ridge regression.

PLS models are compared using the 0.632+ bootstrap estimates of prediction error. The 0.632+ bootstrap is a variant of Efron's (1979) bootstrap proposed by Efron and Tibshirani (1997) and can be thought of as a relatively efficient cross-validation equivalent. We describe the algorithm in Appendix A.

In comparing families of models, the goal of model selection is to find a model that offers the best predictive accuracy. A commonly used metric for calculating prediction accuracy is known as the mean squared error (MSE) and can be calculated as follows:

$$[10] \qquad \mathrm{MSE} = \sum_{i=1}^{N}\left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j\right)^2$$

We note that when the same data are used to test the model as were used to fit the model, then the metric of model quality may be optimistic and the model is unlikely to perform

as well for any new data for which we may wish to make predictions. We corrected for this phenomenon during the analysis of the mortality data using the 0.632+ bootstrap. We did not need to include this step for the simulations because the simulations were based on scenarios for which the true values were known.

Simulation studies

We performed a series of simulations to compare the regression strategies. We simulated a variety of different data sets to highlight the differences in performance among the strategies when used in different situations. We have focused on situations in which multicollinearity is high and in which some or many of the predictor variables have no relationship to the response. Because we usually do not know which variables are relevant, it is important for the modeling technique to provide guidance on variable selection to reduce the impact of erroneous associations and overfitting.

For $k = 100$ replications, simulation data sets were generated as follows. To simulate a situation with N observations and p predictors, a $3N \times 20$ matrix of uniformly randomly distributed variables was generated and a 20×20 correlation matrix was created:

$$\begin{pmatrix} 1 & 0.9 & 0.8 & \dots & -0.8 & -0.9 \\ 0.9 & 1 & 0.9 & \dots & -0.7 & -0.8 \\ 0.8 & 0.9 & 1 & \dots & -0.6 & -0.7 \\ \vdots & & & \ddots & & \vdots \\ -0.8 & -0.7 & -0.6 & \dots & 1 & 0.9 \\ -0.9 & -0.8 & -0.7 & \dots & 0.9 & 1 \end{pmatrix}$$

Using a Cholesky decomposition, we generated a data matrix X with the above correlation structure from our uncorrelated data. A subset of p columns are then extracted at random from X and are used as our new design matrix X^*. This is done to ensure a different correlation matrix with every replication.

Our responses Y were then obtained from

$$[11] \qquad Y = X^*\beta + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

We apply various different β with different levels of p to generate a variety of modeling scenarios.

Sixteen modeling scenarios were considered, half with $N = 30$ and half with $N = 60$. For each of these, we then considered situations with a moderate ($p = 10$) and high ($p = 20$) number of predictor variables. Four different scenarios were generated with varying characteristics for each combination of N and p. Some had large numbers of important predictors, some with combinations of small and large important predictors, and others that contained some or many irrelevant predictors.

We present here four modeling scenarios with the following properties: A:

$$N = 30, \beta = (\underbrace{1, \dots, 1}_{10}), \sigma = 4$$

B:

$$N = 30, \beta = (\underbrace{0, \dots, 0}_{5}, \underbrace{10, \dots, 10}_{5}, \underbrace{0, \dots, 0}_{10}), \sigma = 10$$

C:

$$N = 60, \beta = (\underbrace{1, \dots, 1}_{10}), \sigma = 5$$

D:

$$N = 60, \beta = (\underbrace{1, \dots, 1}_{10}, \underbrace{10, \dots, 10}_{10}), \sigma = 25$$

Informally, the differences between the scenarios were as follows: A and C, small number of equally important predictors; B, small number of important predictors and numerous irrelevant predictors; D, small number of important predictors and small number of less important predictors.

For each replication, the simulated data were divided into three equally sized independent sets: a training set, a validation set, and a test set. For each regression method, models were generated on the training set. The fitted models were then tested on the validation set. The model that minimized prediction error on the validation set was then used to make predictions on the test set. The R^2 on the test set was then recorded:

$$[12] \qquad R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}}$$

where y_i represents response i on the test set, $\bar{y}$ represents the mean of the response of the test set, and $\hat{y}_i$ represents the prediction generated from training set for observation i.

The modeling was performed in the statistical environment R 2.7.2 (R Development Core Team 2008). The "lm" and "step" functions were used to generate the least squares models. "lm.ridge" from the MASS package (Venables and Ripley 2002) was used to create ridge regression models. The "lasso2" package (Lokhorst et al. 2009) was used to create LASSO models and the "pls" package (Wehrens and Mevik 2007) was used for PLS models. All scripts are available from the first author.

Results

Simulation results

Table 1 shows a subset of the simulation results. The table reports median R^2 values for the models computed from the independent test set and the standard deviation over the 100 simulation replications. Figure 2 shows a boxplot of the simulation results for each of the data scenarios.The R^2 is the amount of variance explained by the model. Ideally, we want our modeling techniques to give us a high R^2 without largely varying performance over the replications.

The results suggest that the performance of the techniques depends largely on the type of data set being used, that is, it appears that no modeling technique is superior to all others in all situations. Standard least-squares and stepwise methods performed worst in all scenarios, having the lowest prediction power and the most wildly varying results. The

Table 1. Simulation results (median R^2 and standard deviation) for the simulated data sets over 100 replications.

	Scenario A		Scenario B		Scenario C		Scenario C	
	R^2	SD	R^2	SD	R^2	SD	R^2	SD
Least squares	0.53	0.22	0.45	0.42	0.50	0.15	0.82	0.06
Stepwise	0.57	0.18	0.59	0.32	0.51	0.15	0.83	0.06
Ridge regression	0.63	0.11	0.76	0.12	0.55	0.11	0.85	0.05
LASSO	0.60	0.12	0.78	0.12	0.53	0.12	0.85	0.05
PLS	0.62	0.11	0.73	0.14	0.55	0.12	0.85	0.05

Fig. 2. Boxplots showing the distribution of R^2 values for each of the modeling techniques for each of the data scenarios over 100 replications.

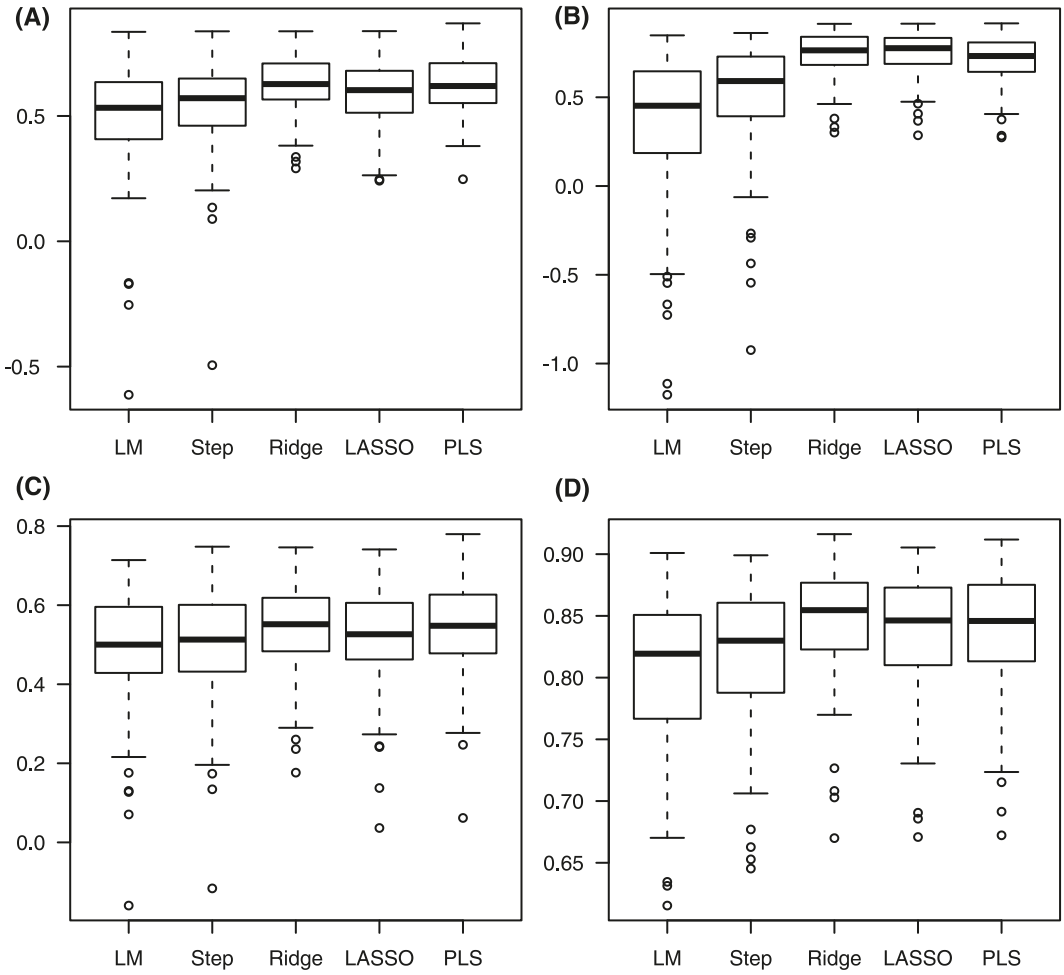


Table 2. Prediction errors for the square root of the number of dead trees per pixel estimated by the 0.632+ bootstrap.

Regression technique	Model selection technique	Prediction error
Least squares	Full model	2.34
	Stepwise	2.39
PLS	0.632+ bootstrap	2.24
Ridge regression	GCV	2.22
	nGCV	2.18
	0.632+ bootstrap	2.13
LASSO	GCV	2.30
	nGCV	2.26
	0.632+ bootstrap	2.23

shrinkage techniques performed similarly in these scenarios, with ridge regression and PLS performing well when the true regression coefficients were all nonzero (scenarios C and D in Table 1). As expected, the LASSO performed best when the true coefficients contained many zeros (scenario B in Table 1).

Analysis of MODIS data

Table 2 gives the 0.632+ bootstrap prediction errors in terms of the square root of the number of dead trees for least-squares, stepwise, ridge, LASSO, and PLS regression models. The shrinkage regression methods far outperform the standard least-squares methods. We observe the ridge regression in general performs better than the LASSO. For both methods, tuning parameter selection obtained by minimizing the 0.632+ prediction error resulted in the least error, but this is expected, as we are measuring performance with the same tool that we used to select the model. Otherwise, model selection via nGCV performed almost as well, outperforming the traditional GCV selection strategy for both regression techniques. PLS also performs competitively. The optimal ridge regression and PLS models retained all seven predictors (including the intercept), while the optimal LASSO model retained six.

Discussion

Our results show that the shrinkage regression techniques offer superior performance to the standard least-squares and stepwise model-building procedures for the scenarios that we have considered. Furthermore, among the three modern techniques, ridge regression generally performed the best. This was also reflected in the results from the analysis of the MODIS data.

Ordinary least-squares regression usually results in highly variable coefficient estimates when the data contain highly correlated predictors. It also suffers from overfitting when the number of predictors is large relative to the number of observations. By implementing a subset-selection strategy such as a stepwise procedure, we can reduce the effect of overfitting and improve interpretability by reducing the number of predictors; however, the regression estimates remain highly variable.

Ridge regression performed very well. Ridge regression reduces the variability of estimates by shrinking them, resulting in improved prediction accuracy at the cost of biased estimates. Highly correlated predictors are shrunk together, reducing the impact that any one of the estimates has on the model. Ridge regression, however, does not shrink predictors to 0 and thus does not provide more interpretable models than ordinary least squares. In fact, interpretability is arguably weakened, as the parameter estimates are no longer unbiased.

We speculate that ridge regression may have outperformed the LASSO due to the possible presence of "small" effects, that is, effects with true population parameters that are nonzero but sufficiently small that the LASSO might set the relevant coefficient estimate to 0 but the ridge regression would not. Fu (1998) showed that the LASSO shrinks small least-squares estimates to 0 and large estimates by a constant, while ridge regression shrinks all least-squares estimates proportionally.

The intention behind the LASSO is to improve prediction accuracy and model interpretability by combining subset selection with the shrinkage properties of ridge regression. However, the LASSO does have limitations. In the presence of highly correlated predictors, the LASSO tends to keep only one of them in the model, setting parameter estimates for the others to 0. While this strategy seems advantageous, it can hinder the construction of models that focus on explanation. This consideration is important if we would like to see what variables contribute or have a relationship to the response. Using the LASSO, we may overlook important variables that have had their parameter estimates set to 0, just because they are highly correlated with another variable that can provide sufficient predictive power. Also, in the $p > N$ case, the LASSO can at most select N predictors.

When the true model includes small but nonzero parameters, the LASSO will perform poorly, as it has a tendency to shrink these to 0. Ridge regression performs well in this situation but worse when true model contains 0 parameters along with large parameters, where the LASSO performs well. Many simulation studies have been done comparing the two techniques (e.g., see Tibshirani 1996; Fu 1998) and no technique appears to particularly dominate the other. While in this setting, ridge regression offered the best performance, the LASSO may still be preferred due to its ability to produce parsimonious models.

PLS behaves similarly to ridge regression (Hastie et al. 2001) but the resulting models are shrunk in steps, as opposed to continuously. Our simulation results suggest that for the scenarios in the scope of our study, the difference between PLS and the penalized regression techniques is minimal.

Only a fraction of the results for the simulation study were reported. Four scenarios out of 16 performed were selected to give a brief overview of the performance characteristics of the techniques over a wide range of regression situations. The performance trends were exhibited throughout the simulations.

A number of model selection strategies have also been considered. We recommend model selection via minimization of the 0.632+ bootstrap prediction error, but this approach can be time consuming and resource intensive. The simple to calculate model selection statistics seem to perform well but have varying performance dynamics given the nature of the data and the "true" model. It should be noted that nGCV selected ridge and LASSO models that generally performed well. Model selection via AIC and BIC were also considered but performed poorly compared with nGCV.

Shrinkage regression methods are useful for analyzing high-dimensional data such as high temporal or spectral satellite data. Remotely sensed time series can be analyzed to identify temporal zones (e.g., seasons within a year) that are most strongly related to specific forest health issues (e.g., defoliation) (Verbesselt et al. 2009). Regression techniques analyzed in this study can help to select specific temporal zones within time series to optimize the use of change metrics derived from this type of satellite data. We advocate the use of LASSO when the objective is to select specific pre-

dictor variables, although it is possible to embed ridge regression and PLS into stepwise variable-selection procedures.

Similarly, hyperspectral satellite data comprising spectral measurements in highly multicollinear wavelengths of the spectrum (e.g., 400–2500 nm) can be analyzed with shrinkage regression techniques to identify zones in the spectra that are most strongly correlated with the in situ measured biophysical variables. For example, Coops et al. (2003) predicted foliage nitrogen content of eucalypts from satellite-derived hyperspectral data using PLS. Shrinkage regression techniques can also be used for mapping and modeling forest biophysical variables (e.g., forest structure) at a regional scale using high-resolution satellite imagery. Wolter et al. (2009) used PLS regression of spatial neighborhood statistics (e.g., semivariogram analysis) to map forest structural properties and support modeling of the spruce budworm insect–host dynamics in Minnesota and Ontario. Our results showed that ridge and LASSO regression could also be suitable for such analysis.

Conclusion

We have shown that tree mortality using MODIS satellite data is best predicted using shrinkage regression techniques such as ridge regression, LASSO, and PLS. However, based on the simulated data, no single modeling strategy worked best for all situations. Generally, shrinkage estimators resulted in better predictions than stepwise regression using least squares. The shrinkage regression techniques in combination with model assessment by bootstrapping can be used to select best-performing change metrics while analyzing time series of remotely sensed data.

Acknowledgements

This work was undertaken within the Cooperative Research Centre for Forestry Program 1.1: Monitoring and Measuring (www.crcforestry.com.au). Discussions with forest health experts Angus Carnegie and David and Ian Smith and CSIRO researcher Darius Culvenor contributed significantly to this study. The authors are grateful to the Associate Editors and two anonymous reviewers for detailed and generous commentary.

References

Akaike, H. 1974. New look at statistical-model identification. IEEE Trans. Automat. Contr. **19**(6): 716–723. doi:10.1109/TAC.1974.1100705.

Asner, G.P., and Martin, R.E. 2008. Spectral and chemical analysis of tropical forests: scaling from leaf to canopy levels. Remote Sens. Environ. **112**(10): 3958–3970. doi:10.1016/j.rse.2008.07.003.

Carrascal, L.M., Galván, I., and Gordo, O. 2009. Partial least squares regression as an alternative to current regression methods used in ecology. Oikos, **118**(5): 681–690. doi:10.1111/j.1600-0706.2008.16881.x.

Coops, N.C., Smith, M.-L., Martin, M.E., and Ollinger, S.V. 2003. Prediction of eucalypt foliage nitrogen content from satellite-derived hyperspectral data. IEEE Trans. Geosci. Rem. Sens. **41**(6): 1338–1346. doi:10.1109/TGRS.2003.813135.

Coops, N.C., Wulder, M.A., and White, J.C. 2006. Integrating remotely sensed and ancillary data sources to characterize a mount pine beetle infestation. Remote Sens. Environ. **105**(2): 83–97. doi:10.1016/j.rse.2006.06.007.

de Beurs, K.M., and Townsend, P.A. 2008. Estimating the effect of gypsy moth defoliation using MODIS. Remote Sens. Environ. **112**: 3938–3990.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1): 1–26. doi:10.1214/aos/1176344552.

Efron, B., and Tibshirani, R. 1997. Improvements on cross-validation. The 0.632+ bootstrap method. J. Am. Stat. Assoc. **92**(438): 548–560. doi:10.2307/2965703.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. Least angle regression. Ann. Stat. **32**(2): 407–451. doi:10.1214/009053604000000067.

Fraser, R.H., and Latifovic, R. 2005. Mapping insect-induced tree defoliation and mortality using coarse spatial resolution satellite imagery. Int. J. Remote Sens. **26**(1): 193–200. doi:10.1080/01431160410001716923.

Fu, W.J. 1998. Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Stat. **7**(3): 397–416. doi:10.2307/1390712.

Fu, W.J. 2005. Nonlinear GCV and quasi-GCV for shrinkage models. J. Statist. Plann. Inference, **131**(2): 333–347. doi:10.1016/j.jspi.2004.03.001.

Golub, G.H., Heath, M., and Wahba, G. 1979. Generalised cross-validation as a method for choosing a good ridge parameter. Technometrics, **21**(2): 215–223. doi:10.2307/1268518.

Goodwin, N.R., Coops, N.C., Wulder, M.A., Gillanders, S., Schroeder, T.A., and Nelson, T. 2008. Estimation of insect infestation dynamics using a temporal sequence of Landsat data. Remote Sens. Environ. **112**(9): 3680–3689. doi:10.1016/j.rse.2008.05.005.

Harrell, F.E. 2000. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J.H. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York.

Hoerl, A.E., and Kennard, R.W. 1970*a*. Ridge regression ? applications to non-orthogonal problems. Technometrics, **12**(1): 69–82. doi:10.2307/1267352.

Hoerl, A.E., and Kennard, R.W. 1970*b*. Ridge regression ? biased estimation for non-orthogonal problems. Technometrics, **12**(1): 55. doi:10.2307/1267351.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., and Ferreira, L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. **83**(1-2): 195–213. doi:10.1016/S0034-4257(02)00096-2.

Lokhorst, J., Venables, B., and Turlach, B. 2009. Lasso2: L1 constrained estimation aka 'lasso'. R package version 1.2-10. Available from http://www.maths.uwa.edu.au/ berwin/software/lasso.html.

Miller, A.J. 2002. Subset selection in regression. 2nd ed. Chapman and Hall/CRC, Boca Raton, Fla.

Osborne, M.R., Presnell, B., and Turlach, B.A. 2000*a*. A new approach to variable selection in least squares problems. IMA J. Numer. Anal. **20**(3): 389–403. doi:10.1093/imanum/20.3.389.

Osborne, M.R., Presnell, B., and Turlach, B.A. 2000*b*. On the LASSO and its dual. J. Comput. Graph. Statist. **9**(2): 319–337. doi:10.2307/1390657.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN3-900051-07-0. Available from http://www.R-project.org.

Schwarz, G. 1978. Estimating dimension of a model. Ann. Stat. **6**(2): 461–464. doi:10.1214/aos/1176344136.

Smith, M.-L., Martin, M.E., Plourde, L., and Ollinger, S.V. 2003. Analysis of hyperspectral data for estimation of temperate forest canopy nitrogen concentration: comparison between an airborne (AVIRIS) and a spaceborne (Hyperion) sensor. IEEE Trans. Geosci. Rem. Sens. **41**(6): 1332–1337. doi:10.1109/TGRS.2003.813128.

Stone, C., Turner, R., and Verbesselt, J. 2008. Integrating plantation health surveillance and wood resource inventory systems using remote sensing. Aust. For. **71**(3): 245–253.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B, **58**(1): 267–288.

Tikhonov, A.N. 1943. On the stability of inverse problems. C.R. Acad. Sci. URSS, **39**: 176–170.

Venables, W.N., and Ripley, B.D. 2002. Modern applied statistics with S. 4th ed. Springer, New York. ISBN 0-387-95457-0. Available from http://www.stats.ox.ac.uk/pub/MASS4.

Verbesselt, J., Robinson, A., Stone, C., and Culvenor, D. 2009. Forecasting tree mortality using change metrics derived from MODIS satellite data. For. Ecol. Manag. **258**(7): 1166–1173. doi:10.1016/j.foreco.2009.06.011.

Wehrens, R., and Mevik, B.-H. 2007. pls: partial least squares regression (PLSR) and principal component regression (PCR), R package version 2.1-0. Available from http://mevik.net/work/software/pls.html.

White, J.C., Coops, N.C., Hilker, T., Wulder, M.A., and Carroll, A.L. 2007. Detecting mountain pine beetle red attack damage with EO-1 Hyperion moisture indices. Int. J. Remote Sens. **28**(10): 2111–2121. doi:10.1080/01431160600944028.

Wold, S., Sjöström, M., and Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**(2): 109–130. doi:10.1016/S0169-7439(01)00155-1.

Wolter, P.T., Townsend, P.A., Sturtevant, B.R., and Kingdon, C.C. 2008. Remote sensing of the distribution and abundance of host species for spruce budworm in northern Minnesota and Ontario. Remote Sens. Environ. **112**(10): 3971–3982. doi:10.1016/j.rse.2008.07.005.

Wolter, P.T., Townsend, P.A., and Sturtevant, B.R. 2009. Estimation of forest structural parameters using 5 and 10-meter SPOT-5 satellite data. Remote Sens. Environ. **113**(9): 2019–2036. doi:10.1016/j.rse.2009.05.009.

Wulder, M.A., Dymond, C.C., White, J.C., Leckie, D.G., and Carrol, A.L. 2006. A review of remote sensing opportunities. For. Ecol. Manag. **221**: 27–41.

Wulder, M.A., White, J.C., Coops, N.C., and Butson, C.R. 2008. Multi-temporal analysis of high spatial resolution imagery for disturbance monitoring. Remote Sens. Environ. **112**(6): 2729–2740. doi:10.1016/j.rse.2008.01.010.

Appendix A. The 0.632+ bootstrap

The following exposition borrows heavily from Hastie et al. (2001).

The bootstrap (Efron 1979) is a useful, but computationally intensive method that can be used for assessing statistical accuracy. The basic idea is that we randomly select B subsets from the original data set with replacement, with each sample being the same size as the original data set, so we have B bootstrap samples. We then refit the model to each of the bootstrap data sets and examine the behaviour of the fits over the B replications. The technique was originally developed with a focus on classification problems.

One approach to use this method to measure prediction error is to make predictions on the original data using the bootstrap samples. Let $\widehat{f}^{*b}(x_i)$ be the predicted value at x_i using a model built on bootstrap sample b; then the bootstrapped error estimate is given by

$$[A1] \qquad \widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N}(y_i - \widehat{f}^{*b}(x_i))^2$$

In general, this estimate provides a poor estimate of prediction accuracy because the bootstrap samples used to fit the model can contain the same observations as those used to test its accuracy. This overlap can make predictions appear unrealistically good. For this reason, cross-validation explicitly uses nonoverlapping data for training and test sets.

The 0.632+ bootstrap can be thought of as mimicking cross-validation. For each observation i, we only keep track of bootstrap samples that do not contain it. We use these samples to predict the value of observation i and measure the error. This gives us the leave-one-out bootstrap prediction of error:

$$[A2] \qquad \widehat{\mathrm{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}(y_i - \widehat{f}^{*b}(x_i))^2$$

Here, C^{-i} is the set of bootstrap samples b that do not contain observation i and $|C^{-i}|$ is the number of such samples. This leave-one-out bootstrap method solves the overfitting problem suffered by $\widehat{\mathrm{Err}}_{\mathrm{boot}}$ but its estimate is still expected to be biased upward of the true error (Hastie et al. 2001). Efron and Tibshirani (1997) introduced the 0.632 estimator, which is designed to alleviate this bias. It is defined by

$$[A3] \qquad \widehat{\mathrm{Err}}^{(0.632)} = 0.368 \times \overline{\mathrm{err}} + 0.632 \times \widehat{\mathrm{Err}}^{(1)}$$

Here, $\overline{\mathrm{err}}$ is the training error rate given by

$$[A4] \qquad \overline{\mathrm{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f}(x_i))^2$$

The derivation of the 0.632 estimator is complex. The idea is that it pulls the leave-one-out bootstrap error estimate down towards the training error rate, reducing the estimator's upwards bias.

Efron and Tibshirani (1997) stated that the estimator can be improved by taking into account the amount of overfitting. For observations $i, j \in 1, ..., N$, define

$$[A5] \qquad \widehat{\gamma} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - \widehat{f}(x_j))^2$$

Then the "relative overfitting rate" is defined as

$$[A6] \qquad \widehat{R} = \frac{\widehat{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}}{\widehat{\gamma} - \overline{\mathrm{err}}}$$

Using this, we can introduce new weights for an improved 0.632+ estimator.

$$[A7] \qquad \widehat{\mathrm{Err}}^{(0.632+)} = (1 - \widehat{w} \times \overline{\mathrm{err}} + \widehat{w} \times \widehat{\mathrm{Err}}^{(1)})$$

$$[A8] \qquad \widehat{w} = \frac{0.632}{1 - 0.368\widehat{R}}$$

If it happens that $\widehat{\gamma} \leq \overline{\text{err}}$ or $\overline{\text{err}} < \widehat{\gamma} \leq \widehat{\text{Err}}^{(1)}$, then $\widehat{R}$ will lie outside of the range of [0,1]. In this case, the definitions of $\widehat{\text{Err}}^{(1)}$ and $\widehat{R}$ must be modified:

$$[A9] \qquad \widehat{\text{Err}}^{(1)\prime} = \min(\widehat{\text{Err}}^{(1)}, \widehat{\gamma})$$

$$[A10] \qquad \widehat{R}' = \begin{cases} (\widehat{\text{Err}}^{(1)} - \overline{\text{err}})/(\widehat{\gamma} - \overline{\text{err}}) & \text{if } \widehat{\text{Err}}^{(1)},\ \widehat{\gamma} > \overline{\text{err}} \\ 0 & \text{if otherwise} \end{cases}$$

References

Efron, B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1): 1–26. doi:10.1214/aos/1176344552.

Efron, B., and Tibshirani, R. 1997. Improvements on cross-validation. The 0.632+ bootstrap method. J. Am. Stat. Assoc. **92**(438): 548–560. doi:10.2307/2965703.

Hastie, T., Tibshirani, R., and Friedman, J.H. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York.