BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 24 2011, pages 3399–3406
doi:10.1093/bioinformatics/btr591

Gene expression

Optimized application of penalized regression methods to diverse genomic data

Levi Waldron[1,3], Melania Pintilie[2], Ming-Sound Tsao[3], Frances A. Shepherd[3], Curtis Huttenhower[1,*] and Igor Jurisica[3,*]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA, [2]Department of Biostatistics, Ontario Cancer Institute, University Health Network and [3]Ontario Cancer Institute, PMH/UHN, Campbell Family Institute for Cancer Research, Toronto, ON, Canada

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Penalized regression methods have been adopted widely for high-dimensional feature selection and prediction in many bioinformatic and biostatistical contexts. While their theoretical properties are well-understood, specific methodology for their optimal application to genomic data has not been determined.

Results: Through simulation of contrasting scenarios of correlated high-dimensional survival data, we compared the LASSO, Ridge and Elastic Net penalties for prediction and variable selection. We found that a 2D tuning of the Elastic Net penalties was necessary to avoid mimicking the performance of LASSO or Ridge regression. Furthermore, we found that in a simulated scenario favoring the LASSO penalty, a univariate pre-filter made the Elastic Net behave more like Ridge regression, which was detrimental to prediction performance. We demonstrate the real-life application of these methods to predicting the survival of cancer patients from microarray data, and to classification of obese and lean individuals from metagenomic data. Based on these results, we provide an optimized set of guidelines for the application of penalized regression for reproducible class comparison and prediction with genomic data.

Availability and Implementation: A parallelized implementation of the methods presented for regression and for simulation of synthetic data is provided as the *pensim* R package, available at http://cran.r-project.org/web/packages/pensim/index.html.

Contact: chuttenh@hsph.harvard.edu; juris@ai.utoronto.ca

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 3, 2011; revised on October 12, 2011; accepted on October 17, 2011

1 INTRODUCTION

Multivariate regression is a flexible machine learning method, suited to prediction of discrete, continuous and censored time-to-event (survival) outcomes from arbitrary combinations of predictor variable classes. In genomic settings, collinear predictors typically greatly outnumber available samples ($p > n$), a now-classic example being the prediction of cancer patient survival from tumor gene expression data (Beer *et al.*, 2002; Shedden *et al.*, 2008; Sørlie *et al.*, 2001; van de Vijver *et al.*, 2002; Wigle *et al.*, 2002). In this setting, ordinary regression is subject to overfitting and instability of coefficients (Harrell *et al.*, 1996), and stepwise variable selection methods do not scale well (Yuan and Lin, 2006). Regression has been successfully adapted to high-dimensional situations by penalization methods (review by Hesterberg, 2008), and penalized regression has been shown to outperform univariate and other multivariate regression methods in multiple genomic datasets (Bøvelstad *et al.*, 2007).

Two penalization methods, and a hybrid of these, are most commonly used. Ridge regression (Hoerl and Kennard, 1970) uses a penalty on the L_2 norm of the coefficients, which introduces bias in the prediction error in exchange for reduced variance. However, ridge regression keeps all variables in the model and thus cannot produce a parsimonious model from many variables. LASSO regression (Tibshirani, 1996; 1997) penalizes the L_1 norm, which tends to reduce many coefficients to exactly zero and thus performs variable selection in addition to prediction. However, the LASSO has been noted to be inferior to Ridge regression for prediction in lower dimensional situations, and tends to select only one of a group of collinear variables, which may not always be desirable (Zou and Hastie, 2005). Zou and Hastie (2005) thus proposed the Elastic Net, penalizing both the L_1 and L_2 norms with individual tuning parameters, as a way to achieve the best of both LASSO and Ridge. These three variants of penalized regression—LASSO, Ridge and Elastic Net—have since been applied to a variety of phenotype prediction tasks using genomic data (for example, Sharma *et al.*, 2008; Shedden *et al.*, 2008).

Several previous simulation studies have investigated properties of the Elastic Net (Zou and Hastie, 2005), the LASSO and Ridge regression (Bøvelstad *et al.*, 2007; Gui and Li, 2005; Yuan and Lin, 2006), but have not compared all these methods with alternative strategies for their application. We present a comprehensive assessment and optimization of these methods, using two contrasting configurations of simulated genomic data and two genome-scale experimental datasets. Comparative studies of this nature provide the most realistic and unbiased assessments of available machine learning methods, issues which have been identified as critical to researchers' selection of appropriate methodology (Boulesteix, 2006; Jelizarow *et al.*, 2010).

We introduce a 2D optimization of the Elastic Net penalty parameters, and show that it or a comparable procedure is necessary

*To whom correspondence should be addressed.

to distinguish the Elastic Net from LASSO and Ridge regression. In particular, we show that successive 1D tuning of the Elastic Net restricts the search for tuning parameters sufficiently that it can yield inferior prediction to a single-penalty counterpart. A univariate pre-filter is commonly used to reduce dimensionality and computation time, but we demonstrate a simulated situation in which the pre-filter can reduce predictive performance of the Elastic Net by reducing the importance of the L_1 penalty relative to the L_2 penalty.

We applied these optimized regression procedures in two very differing genomic settings: predicting survival of cancer patients from microarray data, and classification of lean and obese individuals from metagenomic sequence data. We use a cross-validation strategy for assessment of model prediction (Molinaro *et al.*, 2005; Simon *et al.*, 2011), with an additional inner level of cross-validation for model tuning in training data (Goeman, 2011). Both examples proved favorable to the L_2 penalty, and models trained by Ridge regression and Elastic Net showed independent predictive ability, whereas models trained by the LASSO did not. We emphasize evidence of overfitting in both simulated and real experimental data, and summarize methods for realistic assessment of prediction accuracy with limited sample size. Based on results from this study and best practices for high-dimensional model validation, we conclude with an end-to-end methodology for effective application of penalized regression to diverse genomic data for prediction and variable selection.

2 METHODS

We first consider the use of penalized regression to select features and predict outcome in simulated high-dimensional data, focusing specifically on the Cox proportional hazards model for potentially censored survival data (Cox, 1972). We further apply the guidelines for penalized regression developed from these synthetic data to real expression data and to penalized logistic regression for metagenomic data from the gut microbiomes and obesity status of subjects in the MetaHIT study (Qin *et al.*, 2010).

2.0.1 Creation of synthetic data Five hundred simulations were generated as summarized in Table 1, in one configuration favoring the LASSO penalty and one configuration favoring the Ridge penalty. Simulated variables were standard normal distributed, with covariance matrix specified by the within-group correlations in Table 1 and between-group covariance of zero. 'True' hazards were generated from a weighted sum of the predictor variables according to the proportional hazards assumption:

$$h_j = h_0 \exp\left(\sum_{i=1}^{i=p} \beta_i X_{ij}\right) \quad (1)$$

Where h_j is the hazard for each patient j ($j = 1, 2, \ldots, n$), h_o is an arbitrary baseline hazard (set to a constant 0.2), i is the feature index, each X_{ij} is the simulated expression value of feature i in patient j, and β_i are the associations of each predictor with outcome, given in Table 1. Survival times for 100 patients were then sampled from a random exponential distribution with decay rate h_j and censored on a uniform U(2,10) distribution as, for example, in Gui and Li (2005). Under these conditions, ~34% of events were censored. Results without censoring for 500:40 and 2000:40 noise:associated variables, and for 150 patients with a 500:40 variable ratio with censoring, in the LASSO-favoring scenario, are shown in the Supplemental Materials.

2.0.2 Methods for penalized regression Penalized regression was performed using the *pensim* R package, described in the Implementation, and the *penalized* R package (Goeman, 2010; Version 0.9-33). Five different penalization schemes were considered: LASSO regression (Tibshirani, 1996), Ridge regression (Hoerl and Kennard, 1970) and the Elastic Net with three methods of tuning the λ_1 and λ_2 penalties:

- optimization of λ_1 with λ_2 set to zero, followed by optimization of λ_2 ($\lambda_1 - \lambda_2$ method),
- optimization of λ_2 with λ_1 set to zero, followed by λ_1 ($\lambda_2 - \lambda_1$ method) and
- a 2D optimization of λ_1 and λ_2 simultaneously ($\lambda_1 + \lambda_2$ method).

Table 1. Simulated genomic predictor variables associated with a survival outcome

Variable	Within-group correlation		Association (β_i)		No. of variables	
	$+\lambda_1$	$+\lambda_2$	$+\lambda_1$	$+\lambda_2$	$+\lambda_1$	$+\lambda_2$
A	0.8	0.2	0.5	0.2	10	100
B	0	0.2	0.5	0.2	10	100
C	0.8	0.2	0.3	0.2	10	170
D	0	0	0.3	0	10	170
E	0.8	–	0	–	50	0
F	0	–	0	–	450	0

Simulated variables were standard normally distributed with population mean of within-group pair-wise Pearson's correlations indicated, and zero correlation between groups. Association indicates the coefficient of the variables when creating the risk for each sample according to Equation (1). $+\lambda_1$ and $+\lambda_2$ refer to the LASSO and Ridge-favoring scenarios, respectively.

In the Elastic Net (Zou and Hastie, 2005), the usual partial log-likelihood is penalized by the L_1 and L_2 norms of the regression coefficients with weights λ_1 and λ_2, respectively, i.e.:

$$l(\beta)_{\text{penalized}} = l(\beta) - \lambda_1 \sum_{i=1}^{i=p} |\beta_i| - \lambda_2 \sum_{i=1}^{i=p} (\beta_i)^2 \quad (2)$$

where λ_1 and λ_2 are tuned by maximizing $l(\beta)$ (Gui and Li, 2005; Verweij and Van Houwelingen, 1993; 1994), and $l(\beta)$ is the cross-validated partial log-likelihood. LASSO and Ridge regression are described by Equation (2) with λ_1 or λ_2 non-zero, respectively.

The $\lambda_1 + \lambda_2$ Elastic Net involves 2D optimization of the penalties. Reasonable initial guesses for the penalties were determined by computing 10-fold cross-validated partial log-likelihood (CVL) of Elastic Net models on a regular λ_1/λ_2 grid. The 2D optimization of CVL as a function of λ_1 and λ_2 was then performed by maximizing CVL, as visualized in the contour plots in Figure 1. To avoid spurious results due to local maxima, initial values of λ_1/λ_2 were selected randomly from the five best positions determined from the scan of the regular λ_1/λ_2 grid. 2D tuning of the penalties was performed by the quasi-Newton method described by Byrd *et al.* (1995), as implemented by the *optim* function in the R stats package (R Development Core Team, 2010). The penalty parameters were tuned 50 times using different folding of the data for calculating CVL, and the penalty parameters with maximum CVL were selected.

2.0.3 Assessment of regression in simulated data Methods for regression and selection from multiple starts were assessed in terms of precision of variable selection and in terms of prediction of future events. Precision and prediction are presented as ranks within each dataset. Prediction was assessed in an independently generated test dataset simulated with the same properties as the training data, but with 200 samples. Three methods for assessing prediction accuracy were considered: (i) area under the survival ROC curve at the median survival time, using the *survivalROC* R package (Version 1.0.0); (ii) area under the Prediction Error Curve, using the *PEC* R package (Version 1.1.1); and (iii) Spearman's correlation between known true hazards and the hazards estimated by the penalized Cox regression model.

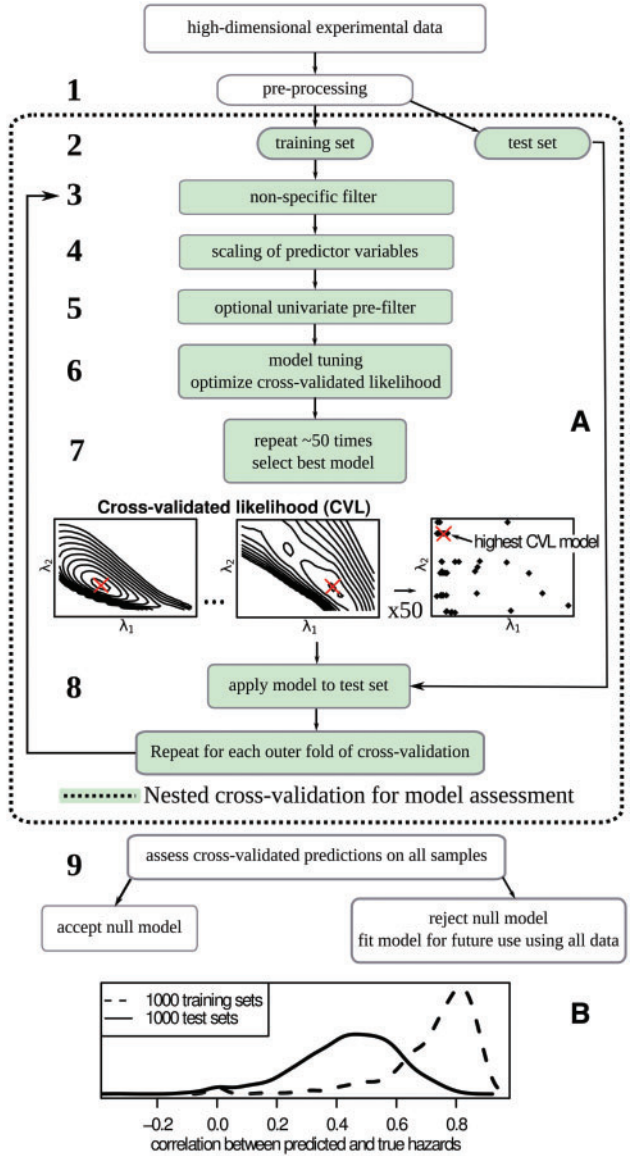

Fig. 1. (**A**) Methodology for model selection and validation of high-dimensional data. Objectives include both feature selection and outcome prediction, e.g. for patient survival given tumor gene expression data. A nearly unbiased assessment of prediction accuracy for small samples sizes is obtained by repeating all steps of model selection in each iteration of the cross-validation. Variable selection and model conditioning are achieved within the training sets by an optional, permissive univariate pre-filter followed by repeated cross-validation for parameter tuning. These steps are detailed in Section 4. **(B)** Over-fitting occurs in spite of tuning the models by cross-validation, as evidenced by reduced prediction accuracy in simulated test sets compared to resubstitution of training data.

These methods all generated similar rankings of the methods (Supplementary Fig. S2); however, we present area under the survival ROC curve for direct comparability to real datasets.

2.1 Application to genomic datasets

Methods for model selection and validation in the two genomic applications are summarized in Figure 1, and detailed in the Supplementary Material. In each case, a context-appropriate filter against low-variance features was applied. LASSO and Elastic Net without univariate pre-filter were performed for each dataset. Prediction accuracy was assessed by 10-fold cross-validation, with model tuning by a cross-validation nested within the training samples. Variable scalings were determined from the training samples, and applied to the test samples, as implemented in the *opt.nested.crossval* function of the *pensim* R package.

3 RESULTS

To establish optimized guidelines for reproducibly linking genomic features (such as gene expression) to outcome using penalized regression (Fig. 1), we compared the LASSO, Ridge regression and Elastic Net with three strategies for 2D tuning of its penalty parameters. We additionally varied the stringencies of an optional univariate pre-filter and evaluated two methods for selecting from repeated tunings, using simulated high-dimensional datasets with several configurations of signal-to-noise ratio and censoring of survival outcomes. Results of these simulations, along with current best practices for high-dimensional model assessment, were used to guide analyses of two publicly available experimental datasets: a microarray experiment investigating survival of cancer patients with lung adenocarcinomas (Beer *et al.*, 2002) and metagenomic data from the MetaHIT consortium (Qin *et al.*, 2010) describing the composition of gene function in the gut microbiotas of lean and obese individuals. Based on the results of these simulations and subsequent comparable analyses of real experimental data, we provide a set of guidelines for penalized regression analysis of diverse high-dimensional data.

3.1 Establishing optimal penalized regression guidelines using simulated genomic data

Using simulated data (see Section 2, Table 1), we compared LASSO, Ridge and Elastic Net regression for variable selection and for prediction accuracy in independent test sets. The LASSO performed variable selection as expected, by preferentially selecting variables associated with outcome over noise variables, and by preferring the strongest individually representative feature from within each correlated block, identifying strongly associated features more frequently than weakly associated features (Supplementary Fig. S1). We found that simultaneous tuning of the Elastic Net parameters was necessary to distinguish it from LASSO and Ridge regression, and assessed the value of repeating tuning of the penalty parameters on different groupings of the data for cross-validation. Finally, we examined the effects of several stringencies of pre-filter for univariate association with outcome.

3.1.1 Selection from multiple starts Repeated tuning of the penalty parameters on randomized partitions of the samples for cross-validation produced small improvements in prediction accuracy, at least for the LASSO (Supplementary Table S1). This benefit was realized by selecting the model with maximum CVL, but not when selecting the median penalty. The benefit of multiple tunings was reduced with increasing stringencies of the univariate pre-filter.

3.1.2 Effect of a univariate pre-filter The use of a pre-filter for univariate association with response is a standard approach in multivariate analysis (Hosmer and Lemeshow, 1999), chapter 5; Simon *et al.*, 2003, chapter 8). It is commonly used as a first step in

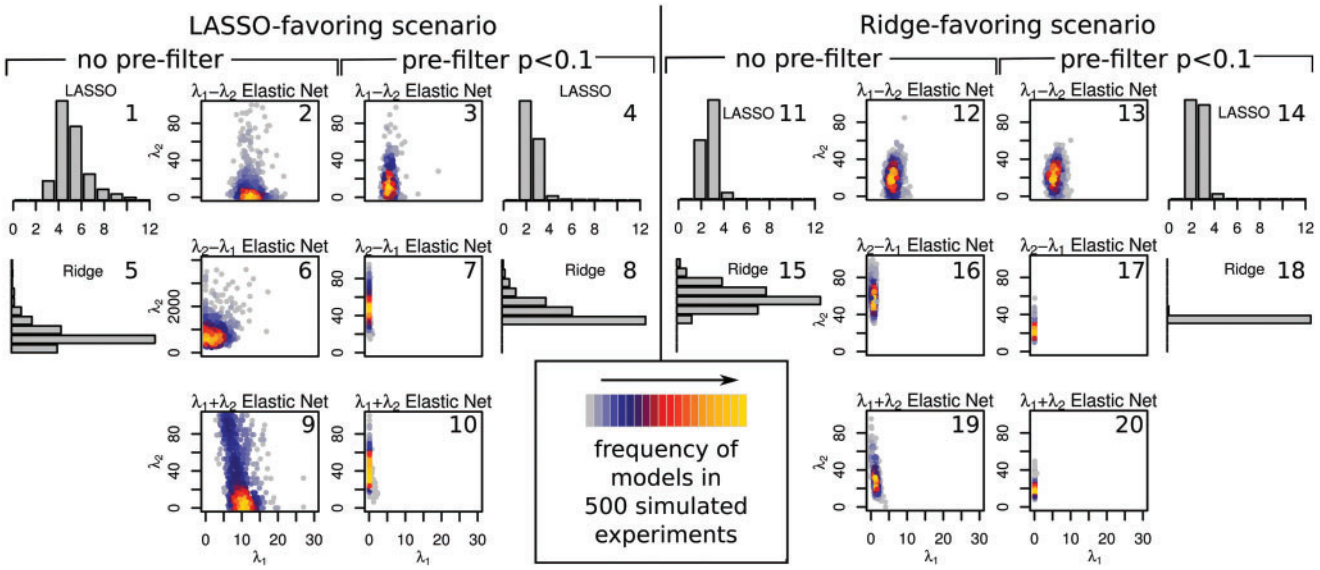

Fig. 2. Optimized values of the Elastic Net tuning parameters in simulated scenarios favoring the LASSO and the Ridge penalties, with comparison to LASSO and Ridge regression. Selected values of the Elastic Net tuning parameters depend on the nature of the problem at hand (left half versus right half), whether pre-filtering precedes the tuning (inner left versus right), and the tuning strategy (represented in each row). In both scenarios, with and without pre-filtering, sequential tuning of the Elastic Net penalties ($\lambda_1-\lambda_2$ and $\lambda_2-\lambda_1$ methods) was dominated by the first penalty tuned, as evidenced by the similarity of values of that penalty to the Ridge or LASSO penalty shown in the adjacent histogram, and smaller values of the second penalty tuned compared with the other single-penalty regression or sequential-tuning regression. Assessment of model prediction and precision of variable selection are correspondingly similar for these methods (Fig. 3 and Supplementary Fig. S3). Note the different *y*-axis scale for $\lambda_2-\lambda_1$ Elastic Net and Ridge regression in the LASSO-favoring scenario. Application of a univariate pre-filter reduced the relative influence of the λ_1 penalty, particularly in the LASSO-favoring scenario (for example, 9 versus 10). Univariate pre-filtering ($P<0.1$) reduced the tuned values of all penalty parameters and, in particular, reduced the influence of λ_1 relative to λ_2 in the $\lambda_1+\lambda_2$ Elastic Net (panels 9 versus 10 and 19 versus 20). These results show that sequential tuning of the λ_1 penalty Elastic Net ($\lambda_2-\lambda_1$ and $\lambda_1-\lambda_2$ methods) is not adequate to enjoy any benefit over LASSO and Ridge regression, and that even in a problem where the λ_1 penalty is preferred, application of a univariate pre-filter causes the λ_2 penalty to dominate the $\lambda_1+\lambda_2$ Elastic Net.

high-dimensional model development (Guyon *et al.*, 2010) or even for development of the model itself (Beer *et al.*, 2002; Bøvelstad *et al.*, 2007; Chen *et al.*, 2007). The use of univariate selection for model development is developed formally for Cox regression by Tibshirani (2009), and implemented by the R package *uniCox*. Unsupervised dimension reduction can be used in addition to or instead of univariate screening, as described for example by Harrell (2001, Section 4.7).

We computed a nominal *P*-value for each predictor in the simulated training set by logrank test, and considered several stringencies of pre-filter: $P<0.1$, $P<0.3$, $P<0.5$ and no pre-filter. Normally, the correct stringency of pre-screening is unknown, and Fan and Lv (2008, R package *SIS*, Version 0.6) suggest a non-parametric method to screen variables with weak association to outcome while ensuring maintenance of the important variables.

In the situation of many weakly correlated variables of equal association with outcome, the pre-filter had little effect on prediction accuracy. However, with few true positives (a situation favoring the λ_1 penalty), Ridge regression improved greatly from the univariate pre-filter, but LASSO regression was barely affected. For the $\lambda_1+\lambda_2$ Elastic Net, the univariate pre-filter reduced the influence of the λ_1 penalty relative to λ_2, causing a loss of prediction accuracy in the LASSO-favoring scenario (Fig. 3).

3.1.3 Choice of regression method We evaluated three methods of tuning the Elastic Net. Tuning one parameter followed by the other ($\lambda_1-\lambda_2$ and $\lambda_2-\lambda_1$) caused the Elastic net to mimic LASSO and Ridge regression, respectively, in terms of the penalty parameters selected (Fig. 2) as well as prediction accuracy and variable selection (Fig. 3). All regression methods outperformed the null model in most cases, but in 1000 datasets simulated from the same population, each method performed near-best in at least one instance and near-worst in at least one, particularly with respect to prediction. Large overall performance differences existed between simulated datasets, but as relative performance of available methods for a particular dataset is typically of interest, we ranked the competing methods within each dataset and aggregated ranks across datasets to compare methods.

The relative prediction performance of LASSO and Ridge regression depends on the sparsity of true positive variables, as the LASSO penalty selects no more variables than there are samples, and tends to select a single representative from a group of correlated variables (Zou and Hastie, 2005). We simulated situations in which the LASSO and Ridge penalties each produced superior prediction, and showed that in each case, prediction performance of the Elastic Net was comparable to the better of the two, provided that a true 2D tuning was used ($\lambda_1+\lambda_2$ method) and no univariate pre-filter was applied. Sequential tuning of the Elastic Net produced behavior dominated by the first variable tuned.

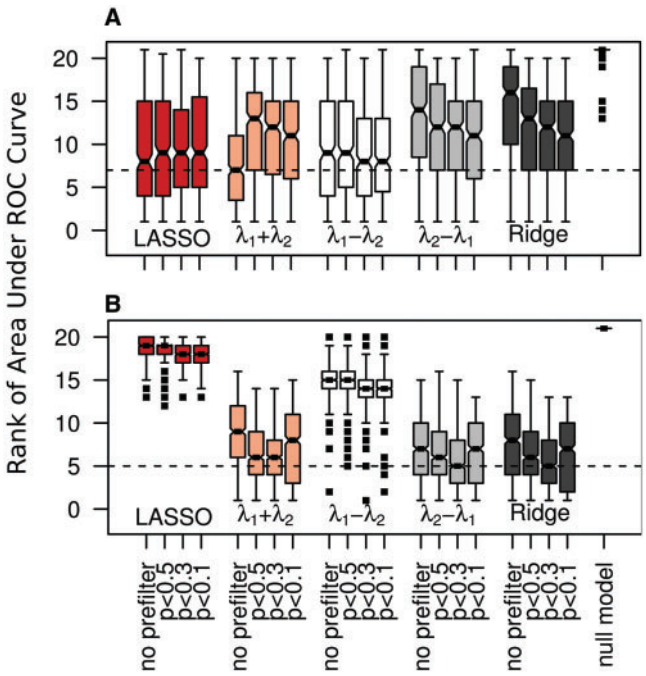

Fig. 3. Ranking of methods for prediction accuracy in two scenarios, simulated to favor (**A**) the LASSO penalty and (**B**) the Ridge penalty. In both scenarios and with all levels of pre-filtering, sequential tuning of the Elastic Net is dominated by the first penalty tuned ($\lambda_1-\lambda_2$ is similar to LASSO, and $\lambda_1-\lambda_2$ is similar to Ridge). Only with 2D tuning ($\lambda_1+\lambda_2$) does the Elastic Net perform comparably in both scenarios to the better single-penalty method. The pre-filter has little effect on prediction in most cases, except in the LASSO-favoring where it improves prediction by Ridge regression, and worsens prediction by $\lambda_1+\lambda_2$ Elastic Net by decreasing the relative importance of the λ_1 penalty.

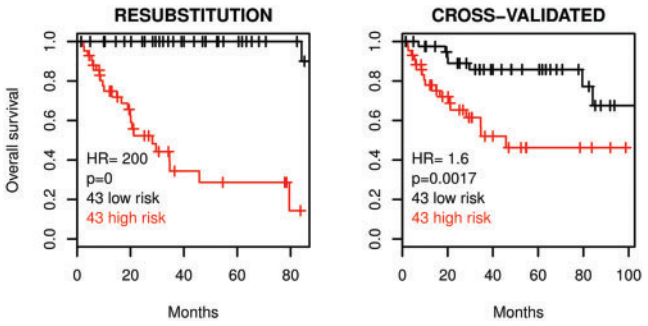

Fig. 4. Model selection guidelines allow reproducible outcome prediction from tumor gene expression data. Kaplan–Meier plots for cross-validated risk prediction for lung adenocarcinoma patients from Beer *et al.*, using the Elastic Net. A naive model is overfit to training data, as evidenced by reduced prediction accuracy in cross-validation compared with resubstitution in the training data.

3.2 Predicting survival of cancer patients from microarray data

The Beer *et al.* (2002) lung adenocarcinoma study was a seminal demonstration of the application of microarrays to predict overall survival of cancer patients based on tumor gene expression profiles. Subsequently, it has become a classic dataset for new analyses and method development (see for example Chen *et al.*, 2007; Michiels *et al.*, 2005; Subramanian *et al.*, 2005).

We applied a standard non-specific filter against unexpressed genes, without reference to patient outcome. Prediction accuracy was assessed by 10-fold cross-validation, with scaling of features and all steps of model training performed in training data only. Prediction models were trained and risk scores calculated for each fold of the cross-validation, and collected together, using the *opt.nested.crossval* function of the *pensim* R package. In separate instances, we performed model training by LASSO, Ridge and Elastic Net with the $\lambda_1+\lambda_2$ tuning method, in each case with no univariate pre-filter and selecting the penalties with highest cross-validated partial log likelihood in 50 starts. This prediction problem favored the L_2 penalty, as independently predictive models were obtained by Ridge regression ($P=0.003$) and Elastic Net ($P=0.002$), but not by LASSO ($P=0.72$, all tests of significance by Likelihood Ratio test). The models produced by Ridge regression and Elastic Net were very similar: the Elastic Net set only 43 of 1310 variables exactly to zero, and the continuous risk predictions of the two models were highly correlated (Pearson's correlation = 0.99, df = 84, $P<2e\text{-}16$). The assumption of sample independence is violated in cross-validated predictions, so following Simon *et al.* (2011), we also assessed significance for the Elastic Net model by permuting the outcome labels 500 times. For each permutation, we repeated the entire procedure of generating cross-validated risk predictions for all samples. The Likelihood Ratio test statistic did not exceed the observed value in 500 permutations ($P<0.002$). Superior performance of L_2-penalized regression, with associated complex prediction models, may not be uncommon in gene expression prediction problems, as Bøvelstad *et al.* (2007) also found Ridge regression to produce superior survival prediction compared with the LASSO in three other gene expression datasets.

We applied the same methods using all samples to determine coefficients for use in independent datasets. When this model was used to make predictions for these same samples (resubstitution), unrealistically accurate performance was observed due to overfitting (Fig. 4). This highlights the importance of validation in samples not used for any part of model training (see Section 4, Guideline 2).

3.3 Classification of obesity from metagenomics data

Metagenomic data, consisting of short DNA sequences sampled from uncultured microbial communities, are biologically very different from microarray gene expression data (Qin *et al.*, 2010) and represent a new biological application of high-dimensional regression. In a second example, we analyzed data from the combined genomes of microbes in stool samples from the MetaHIT Danish cohort of 85 lean and obese individuals. Whereas gene expression data represent expression of a single gene transcript, these sequence data were summarized by assigning open reading frames from the MetaHIT assembly to functionally characterized orthologous gene families (see Section 2) and calculating the relative abundances of 665 gene functions. We used the same three regression methods and the cross-validation procedure described in the first example, assessing model prediction by area under the ROC curve. This dataset also favored the L_2 penalty, with Ridge regression (AUC = 0.61, $P=0.04$) and Elastic Net (AUC = 0.59, $P=0.08$) outperforming the LASSO, which converged on the null

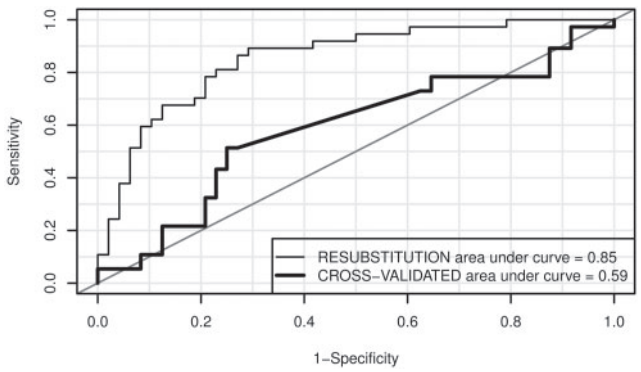

Fig. 5. Application to high-dimensional metagenomic data. ROC curves for classification of obese ($n=37$) and non-obese ($n=48$) individuals using metagenomic data describing the gut microbiota (Qin *et al.*, 2010), trained by Elastic Net. High-dimensional features are no longer gene expression but the relative abundance of specific microbial pathways in the stool microbiome. Overfitting to the training set is again observed in resubstitution predictions, but cross-validation shows marginal evidence of independent predictive ability (AUC = 0.59, $P=0.08$).

model in all folds of cross-validation. The permutation test for Elastic Net, with area under the ROC curve as the test statistic, produced a similar result ($P=0.03$). As with the first example, resubstitution predictions showed evidence of overfitting to the training set (Fig. 5, AUC = 0.85, $P=2\times10^{-9}$, results shown for Elastic Net).

Elastic Net and Ridge regression models were again similar, with the Elastic Net setting only 139 of 532 coefficients exactly to zero, with similar continuous predictions (Pearson's correlation = 0.91, df = 83, $P<$ 2e-16). In these models, protein localization, secretion and signaling were associated with leanness, and biosynthetic proteins determining cell wall phospholipid composition were positively associated with obesity. These are potentially indicative of the Gram-positive/negative shift, previously associated with obesity due to changes in the relative abundances of the *Bacteroidetes* and *Firmicutes* phyla (Ley *et al.*, 2006; Turnbaugh *et al.*, 2006).

4 GUIDELINES FOR THE APPLICATION OF PENALIZED REGRESSION TO DIVERSE GENOMIC DATA

As discussed above, the performance of different regression methods and training procedures varied widely in our simulated data, with each individual method performing both best and worst in individual instances of 500 simulations from the same population. This emphasizes the random variation in model development that can be expected, and the care that must be taken to avoid reporting on results solely due to overfitting. We summarize previously established and presently determined steps for development and assessment of predictive models from high-dimensional genomic data.

1. *Data quality control and normalization*: in addition to normalization and quality control appropriate for the data at hand, features should be transformed to the same scale, as the impact of penalization on coefficients is scale-dependent (Tibshirani, 1997).

2,8. *Assess prediction accuracy in independent test data or by cross-validation for small sample sizes*: overfitting in training data was evident in simulations and both experimental datasets analyzed, despite model tuning by optimizing cross-validated partial log-likelihood. Overfitting of high-dimensional data is a well-known issue (Simon *et al.*, 2011), that may be inevitable in high-dimensional model training. It is thus necessary to assess model prediction accuracy in samples not used in any way for model training. For small sample sizes, split training/test evaluation produces downward bias and instability in estimated prediction accuracy, and cross-validation or 0.632 bootstrap resampling is preferable (Molinaro *et al.*, 2005).

3. *Non-specific feature filter*: features with consistently low values or variance are likely to be affected primarily by noise, and should be removed without reference to the response variable prior to scaling this noise to the same variability as true signals in the next step. The specifics of this filter depend on knowledge of the data at hand, e.g. removing genes never significantly above background in a microarray dataset or metagenomic taxa never above a relative abundance cutoff.

4. *Scaling of predictor variables*: the effect of penalization depends on the magnitude of coefficients and therefore on the scale of the coefficients. Therefore, predictor variables should be on a comparable scale or scaled if they are not. Scaling is done in training data only, and the transformations determined in training data are then applied to the test data.

5. *Optional application of a univariate pre-filter*: univariate pre-filters have commonly been used for feature selection and to reduce computation time for model training. In an L_1-favoring simulated scenario, a pre-filter reduced the relative importance of the L_1 penalty in the Elastic Net, which worsened predictions, but it improved the otherwise inferior predictions of Ridge regression. It had little influence in an L_2-favoring simulated scenario. We therefore recommend caution when applying a univariate pre-filter with the Elastic Net.

6-7. *Model selection in training samples*: properties of a given dataset may favor either Ridge- or LASSO-penalized regression. The Elastic Net can provide the advantages of both penalties and perform no worse than the better of its single-penalty counterparts, provided an adequate 2D tuning strategy is used. In simulated high-dimensional data with few true predictors and many noise variables, LASSO provided the best prediction, with short computation time. In a contrasting scenario with many weakly correlated variables with equal predictive value, Ridge regression performed superior to the LASSO. In both instances, the Elastic Net performed comparably to the better of the single-penalty methods, provided that (i) a 2D tuning strategy was used ($\lambda_1+\lambda_2$ method) and no univariate pre-filter was applied. This flexibility of the Elastic Net comes at the cost of substantially greater computation time. Repeated tuning the penalty parameter(s) with different foldings of the samples for 10-fold cross-validation, and selecting the model associated with highest CVL, produced some improvement in prediction accuracy.

Other machine learning algorithms may be considered (for example, Breiman, 2001), as well as boosting to improve on a set of weak learners (Bühlmann, 2007).

9. *Final model selection*: if evidence of an independently predictive model is found in the cross-validation procedure, the same methods are used on all samples to select a model for studying selected variables, and for prediction on new, independent datasets.

5 CONCLUSIONS

In this study, we outline and implement guidelines to optimize the performance of penalized regression, as assessed by parameter tuning, variable selection and prediction in independent high-dimensional genomic data. We simulated two contrasting scenarios that favored the LASSO and Ridge penalties with high-dimensional collinear predictors and survival outcome. We found that a simultaneous tuning of the Elastic Net penalties, which was previously unavailable, was required to differentiate it from LASSO and Ridge regression. In the LASSO-favoring simulation scenario, application of a permissive univariate pre-filter reduced performance of the Elastic Net, by reducing the influence of the λ_1 penalty, but in other situations was neutral. Selecting the penalty associated with highest cross-validated partial log-likelihood from repeated tunings produced better predictions in independent data for nearly all methods.

Based on findings from the simulations, we demonstrated the application of penalized regression in two very different genomic contexts, differing in biological data type (gene expression and metagenomic structure) and in response variable type (censored survival time and binary obesity). In both scenarios, Ridge regression and Elastic Net produced similar models, which outperformed the LASSO.

We observed overfitting in both synthetic and real data, highlighting the importance of proper model validation with independent data. We summarize the methods employed here in a set of specific guidelines to guide the development of reproducibly predictive models from genomic data.

6 IMPLEMENTATION

To perform these studies, we introduced a 2D tuning of the Elastic Net and simplified the processes of repeated tuning of the penalty parameters and of cross-validated estimation of model accuracy. We parallelized the tasks of repeated tuning of the LASSO, Ridge and Elastic Net penalties, and of cross-validation for assessment of prediction accuracy. We also developed a function for generating synthetic high-dimensional data with time-to-event or binary outcome, and blocks of predictor variables defined by collinearity and association with outcome, with options for introducing labeling errors and for censoring of survival times.

These functions are available in the *pensim* R package from the CRAN repository. It utilizes the *penalized* R package (Goeman, 2010; Version 0.9–33) for regression, the *snow* R package Version 0.3-7 for parallelization of the multiple starts, the *MASS R* package (Venables and Ripley, 2002) for correlated random number generation and the *rlecuyer* R package Version 0.3-1 for parallelization of random number generation.

Funding: National Science Foundation grant (NSF DBI-1053486 to C.H.). Canada Foundation for Innovation (CFI #12301 and CFI #203383 to I.J.) and Ontario Research Fund (GL2-01-030 to I.J.) supporting computational analysis, partially; Canada Research Chair Program (to I.J. in part); Ontario Ministry of Health and Long Term Care, in part. The views expressed do not necessarily reflect those of the OMOHLTC.

Conflict of Interest: none declared.

REFERENCES

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Boulesteix,A.L. (2006) Reader's reaction to "Dimension reduction for classification with gene expression microarray data" by Dai et al (2006). *Stat. Appl. Genet. Mol. Biol.*, **5**, Article16.

Bøvelstad,H.M. *et al.* (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**, 2080–2087.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Bühlmann,P. (2007) Boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.*, **22**, 477–505.

Byrd,R.H. *et al.* (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Chen,H.-Y. *et al.* (2007) A five-gene signature and clinical outcome in non–small-cell lung cancer. *N. Engl. J. Med.*, **356**, 11–20.

Cox,D.R. (1972) Regression models and life-tables. *J. R. Stat. Soc. Ser. B*, **34**, 187–220.

Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849–911.

Goeman,J.J. (2010) L1 penalized estimation in the Cox proportional hazards model. *Biometr. J. Biometri. Zeitsch.*, **52**, 70–84.

Gui,J. and Li,H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Guyon,I. *et al.* (2010) Model selection: beyond the Bayesian/frequentist divide. *J. Mach. Learn. Res.*, **11**, 61–87.

Harrell,F.E. (2001) *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York.

Harrell,F.E. Jr *et al.* (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, **15**, 361–387.

Hesterberg,T. (2008) Least angle and $\ell 1$ penalized regression: a review. *Stat. Surv.*, **2**, 61–93.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hosmer,D.W. and Lemeshow,S. (1999) *Applied survival analysis: regression modeling of time to event data*. J. Wiley, New York.

Jelizarow,M. *et al.* (2010) Over-optimism in bioinformatics: an illustration. *Bioinformatics*, **26**, 1990–1998.

Ley,R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

Michiels,S. *et al.* (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Molinaro,A.M. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sharma,R. *et al.* (2008) Systemic inflammatory response predicts prognosis in patients with advanced-stage colorectal cancer. *Clin. Colorectal Cancer*, **7**, 331–337.

Shedden,K. *et al.* (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.*, **14**, 822–827.

Simon,R.M. (2003) Design and analysis of DNA microarray investigations. Springer, New York.

Simon,R.M. *et al.* (2011) Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief. Bioinformatics*, **12**, 203–214.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

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

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tibshirani,R.J. (2009) Univariate shrinkage in the Cox model for high dimensional data. *Stat. Appl. Genet. Mol. Biol.*, **8**, 21–21.

Turnbaugh,P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1131.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*. Springer, New York.

Verweij,P.J. and Van Houwelingen,H.C. (1993) Cross-validation in survival analysis. *Stat. Med.*, **12**, 2305–2314.

Verweij,P.J. and Van Houwelingen,H.C. (1994) Penalized likelihood in Cox regression. *Stat. Med.*, **13**, 2427–2436.

Wigle,D.A. *et al.* (2002) Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.*, **62**, 3005–3008.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, **68**, 49–67.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–301.