Computational Statistics and Data Analysis 70 (2014) 198–211

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Stabilizing the lasso against cross-validation variability

S. Roberts *, G. Nowak

Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University, Canberra, 2600, Australia

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form 5 June 2013
Accepted 11 September 2013
Available online 30 September 2013

Keywords:
Model-selection
$p \gg n$
Penalized regression
Regularization
Shrinkage

ABSTRACT

An abundance of high-dimensional data has meant that L_1 penalized regression, known as the *lasso*, has become an indispensable tool of the practitioner. A feature of the lasso is a "tuning" parameter that controls the amount of shrinkage applied to the coefficients. In practice, a value for the tuning parameter is chosen using the method of cross-validation. It is shown that the model that is selected by the lasso can be extremely sensitive to the fold assignment used for cross-validation. A consequence of this sensitivity is that the results from a lasso analysis can lack interpretability. To overcome this model-selection instability of the lasso, a method called the *percentile-lasso* is introduced. The model selected by the percentile-lasso corresponds to the model selected by the lasso, when the lasso is fitted using an appropriate percentile of the possible "optimal" tuning parameter values. It is demonstrated that the percentile-lasso can achieve substantial improvements in both model-selection stability and model-selection error compared to the lasso. Importantly, when applied to real data the percentile-lasso, unlike the lasso, produces interpretable results, that is, results that are robust to the assignment of observations to folds for cross-validation. The percentile-lasso is easily applied to extensions of the lasso and in the context of penalized generalized linear models.

1. Introduction

An abundance of high-dimensional data has meant that methods for performing penalized regression have become indispensable tools of the practitioner. One of the most well-known methods of penalized regression is the lasso (Tibshirani, 1996). The lasso has the attractive property that it returns parsimonious results, due to the fact that it sets some parameter estimates to exactly zero. The importance of the lasso can be appreciated by the extensive amount of research interest it has generated. Examples of this research include extensions of the lasso such as the "adaptive lasso", the "relaxed-lasso" and the "elastic net" (Meinshausen, 2007; Zou, 2006; Zou and Hastie, 2005). One feature of the lasso, and its extensions, is the addition of "tuning" parameters that control some aspect of the fitted model such as the amount of shrinkage applied to the coefficients.

Cross-validation is the most commonly used method for choosing tuning parameter values in penalized regression. This method proceeds by splitting the data into "folds". Then over a sequence of tuning parameter values, penalized models are fitted to all but one of the folds and the predictive performance of each model gauged over the "left-out" fold. This process is repeated until each fold has been left out. The value of the tuning parameter is then, typically, chosen to be the value in the sequence that has the smallest aggregated prediction error (or cross-validation error). The widespread use of cross-validation is due to the fact that it is intuitively appealing, easy to implement and can provide a good estimate of the expected prediction error (Hastie et al., 2009, Chap. 7). A downside of cross-validation is the fact that the chosen tuning parameter value, and hence the model that is selected by the lasso, depends on the random assignment of observations to folds (the

* Corresponding author. Tel.: +61 261253470; fax: +61 2 612 50087.
E-mail addresses: steven.roberts@anu.edu.au (S. Roberts), gen.nowak@anu.edu.au (G. Nowak).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.008

"fold assignment"). This issue has been noted in previous studies (see, for example, Bovelstad et al. (2007)). In these studies the authors observe that the model that is selected by the lasso can differ substantially depending on the fold assignment. In order to allow for this variability, these studies investigated fitting the lasso using different fold assignments.

In this paper we provide a detailed analysis of the variability of the model that is selected by the lasso ("model-selection variability") arising from the assignment of observations to folds for cross-validation. We find that this variability is frequently large. In practice this means that it is a common occurrence for the fold assignment to have a demonstrable impact on the model that is selected by the lasso. This model-selection instability can cause the results from a lasso analysis to lack interpretability. We then show that the simple process of using an appropriate percentile of the possible "optimal" lasso tuning parameter values, obtained over different fold assignments, can result in markedly improved performance. In particular, when the true model is sparse we show that this method, termed the *percentile-lasso*, can result in large reductions in both model-selection instability and model-selection error, compared to the lasso. Consequently, this means that the percentile-lasso returns interpretable results in situations where the lasso does not. A further attractive feature of the percentile-lasso is that the method is easily applied to extensions of the lasso. For example, we show that the percentile-lasso applied in the context of the relaxed-lasso results in similarly large reductions in both model-selection instability and model-selection error.

2. The lasso

Here we describe the lasso and the method of cross-validation for choosing the value of the lasso tuning parameter, in more detail. For illustrative purposes, we will assume a linear regression setting. Suppose we have a sample of n observations, where for each observation we have measured a response y and p predictors or covariates $\left(x_1, \dots, x_p\right)$. The lasso estimate, denoted $\hat{\boldsymbol{\beta}}^{\text{lasso}}$, solves the following:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}. \tag{1}$$

Expression (1) contains the "usual" residual sum of squares plus a penalty on the L_1 norm of the parameter vector. This particular form of the penalty term results in the lasso shrinking some of the parameter estimates to zero. The degree of shrinkage is determined by the parameter λ. In order to implement the lasso a means of choosing this parameter is needed.

As mentioned in Section 1, cross-validation is the most commonly used method for choosing λ. Cross-validation provides a convenient means of overcoming limitations that arise from a lack of readily available data due to, for example, cost and/or time constraints. In a perfect setting, with plentiful data, independent test data could be used to estimate the prediction error of a number of candidate models. For the lasso, the candidate models are defined by a sequence of λ values. The optimal value of λ, denoted $\hat{\lambda}$, could then be taken to be the value corresponding to the candidate model with the smallest prediction error. However, due to a general scarcity of data, obtaining independent test data is generally not feasible. The method of cross-validation overcomes this issue by dividing the available data into groups and, in turn, treating each group of data "like" an independent test dataset. Specifically, the steps involved in K-fold cross-validation are as follows:

1. Make the fold assignment by randomly splitting the available data into K equal (or approximately equal) groups or "folds". The K folds will be used to estimate the prediction error for each candidate model (or alternatively, for each value of λ).
2. Withhold one of the K folds and fit each candidate model to the remaining $K - 1$ folds.
3. Compute the prediction error of each candidate model fitted in step 2 over the withheld fold.
4. Repeat steps 2–3 until each of the K folds of data have been withheld.
5. For each candidate model, aggregate the prediction errors obtained over the K folds.
6. The optimal value of λ, denoted $\hat{\lambda}$, is chosen to be the value that corresponds to the candidate model with the smallest aggregated prediction error.

From the above steps, it is clear that the cross-validation estimate of the prediction error is dependent on the fold assignment in step one. In particular, we will show that the fold assignment can lead to large variability in $\hat{\lambda}$ and, consequently, the model that is selected by the lasso.

We briefly mention that the R package `glmnet` (Friedman et al., 2010) is used for fitting the lasso in this paper. The `glmnet` package fits a range of penalized generalized linear models using an efficient algorithm. The package contains functions that can quickly fit the lasso over a sequence of λ values and that also perform cross-validation to find $\hat{\lambda}$. Additionally, the functions allow the fold assignments to be set by the user, making it an appropriate tool to implement the "repeated fold assignment" process that is part of our percentile-lasso method. We note that throughout this paper we implement cross-validation with $K = 10$, the default value in the `glmnet` package. When necessary to avoid confusion, the terminology "standard-lasso", rather than "lasso", will be used to refer to the lasso implemented using a single fold assignment.