Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 642 – 658
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

An even faster algorithm for ridge regression of reduced rank data

Berwin A. Turlach*

School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

Received 13 January 2003; received in revised form 15 September 2004; accepted 21 September 2004
Available online 13 October 2004

Abstract

Hawkins and Yin (Comput. Statist. Data Anal. 40 (2002) 253) describe an algorithm for ridge regression of reduced rank data, i.e. data where p, the number of variables, is larger than n, the number of observations. Whereas a direct implementation of ridge regression in this setting requires calculations of order $\mathrm{O}(np^2 + p^3)$, their algorithm uses only calculations of order $\mathrm{O}(np^2)$. In this paper, we describe an alternative algorithm based on a factorization of the (transposed) design matrix. This approach is numerically more stable, further reduces the amount of calculations and needs less memory. In particular, we show that the factorization can be calculated in $\mathrm{O}(n^2p)$ operations. Once the factorization is obtained, for any value of the ridge parameter the ridge regression estimator can be calculated in $\mathrm{O}(np)$ operations and the generalized cross-validation score in $\mathrm{O}(n)$ operations.

Keywords: Bidiagonal factorization; Cross validation; Matrix factorization; Ridge regression

1. Introduction

To address problems with highly correlated predictors, Hoerl and Kennard (1970) proposed the use of ridge regression. Writing $\mathbf{X}$ for the $n \times p$ design matrix and $\boldsymbol{y}$ for the

* Tel.: +61 8 6488 3383; fax: +61 8 6488 1028.
E-mail address: berwin@maths.uwa.edu.au (B.A. Turlach).
URL: http://www.maths.uwa.edu.au/~berwin.

0167-9473/$ - see front matter
doi:10.1016/j.csda.2004.09.007

vector of responses, the ridge regression estimator for $\boldsymbol{\beta}$, for a given ridge parameter k, is given by

$$\hat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}^*\right)^{-1}\mathbf{X}'\boldsymbol{y}. \tag{1}$$

Here $\mathbf{I}^*$ is the $p \times p$ identity matrix if the regression has no intercept. Otherwise, assuming that the first component of $\boldsymbol{\beta}$ corresponds to the intercept term, it is an identity matrix with its top left element replaced by zero.

Hawkins and Yin (2002) point out that calculating $\hat{\boldsymbol{\beta}}(k)$ directly using (1) requires computations of order $\mathrm{O}\left(np^2 + p^3\right)$, namely $\mathrm{O}\left(np^2\right)$ to set up the normal equations and $\mathrm{O}\left(p^3\right)$ to perform the matrix inversion. However, in some applied sciences, e.g. chemometrics, it is not unusual that p, the number of variables, is larger than n the number of observations—sometimes substantially larger. In these situations such a direct approach would be prohibitively expensive. Hawkins and Yin (2002) propose an alternative algorithm that, essentially, calculates the inverse of the matrix $\mathbf{X}'\mathbf{X} + k\mathbf{I}^*$ via repeated applications of a rank-1 updating formula and allows to calculate $\hat{\boldsymbol{\beta}}(k)$ in $\mathrm{O}\left(np^2\right)$ operations.

Thus the algorithm of Hawkins and Yin (2002) offers some substantial savings in computing effort. Although we have no doubts that their algorithm is adequate for a wide variety of problems we have some concerns about its numerical stability and, hence, about its suitability as a "black-box" algorithm. The algorithm has several features that the numerical analysis literature considers problematic. In particular, the algorithm uses rank-1 matrix updates and it calculates explicitly a matrix inverse.

For calculating ridge regression estimates for the more familiar case where n is larger than p several numerically stable algorithms based on various factorization of the design matrix have been proposed. These algorithms have also the advantage that, once the factorization is calculated, it becomes relatively cheap to calculate the ridge regression estimator for different values of the ridge parameter k or to calculate the generalized cross-validation (GCV) score. For example, Golub and van Loan (1996) propose to calculate a QR factorization of the design matrix $\mathbf{X}$. For the case $n > p$, this factorization can be calculated in $\mathrm{O}\left(np^2\right)$ operations and, once the factorization is calculated, for any given k, one needs $\mathrm{O}\left(p^3\right)$ operations to update the factorization and $\hat{\boldsymbol{\beta}}(k)$ can be determined in $\mathrm{O}\left(p^2\right)$ operations. Lawson and Hanson (1995) suggest to first calculate the singular value decomposition (SVD) of the design matrix $\mathbf{X}$. Once the SVD is calculated, for any given k, one can update the SVD in $\mathrm{O}(p)$ operations and calculate $\hat{\boldsymbol{\beta}}(k)$ in $\mathrm{O}\left(p^2\right)$ operations. Unfortunately, calculating the SVD of a matrix is an expensive and non-trivial task.

An approach which uses a factorization that is easy to calculate was proposed by Eldén (1977); see also Björke (1996, Chapter 5.3.4). Eldén (1977) proposed to calculate a bidiagonal factorization of the design matrix $\mathbf{X}$. This factorization can be computed with the same effort as a QR factorization and its implementation is not more difficult than the implementation of a QR factorization. However, the bidiagonal factorization approach offers substantial computational savings for subsequent computations that equal the savings of the SVD approach. For a given value of the ridge parameter k, once the bidiagonal factorization is calculated, one can update the factorization in $\mathrm{O}(p)$ operations, calculate $\hat{\boldsymbol{\beta}}(k)$ in $\mathrm{O}\left(p^2\right)$ operations (Eldén, 1977) and the GCV score in $\mathrm{O}(p)$ operations (Eldén, 1984). Elden's algorithms have good numeric stability properties and can be adapted to quite