ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization

Peter D. Hoff

Department of Statistical Science, Duke University, Box 90251, Durham, NC, 27708-0251, USA

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form 12 June 2017
Accepted 14 June 2017
Available online xxxx

Keywords:
Cyclic coordinate descent
Generalized linear model
Linear regression
Optimization
Ridge regression
Sparsity
Spatial autocorrelation

ABSTRACT

Using a multiplicative reparametrization, it is shown that a subclass of L_q penalties with q less than or equal to one can be expressed as sums of L_2 penalties. It follows that the lasso and other norm-penalized regression estimates may be obtained using a very simple and intuitive alternating ridge regression algorithm. As compared to a similarly intuitive EM algorithm for L_q optimization, the proposed algorithm avoids some numerical instability issues and is also competitive in terms of speed. Furthermore, the proposed algorithm can be extended to accommodate sparse high-dimensional scenarios, generalized linear models, and can be used to create structured sparsity via penalties derived from covariance models for the parameters. Such model-based penalties may be useful for sparse estimation of spatially or temporally structured parameters.

1. Introduction

Consider estimation for the normal linear regression model $y \sim N_n(X\beta, \sigma^2 I)$, where $X \in \mathbb{R}^{n\times p}$ is a matrix of predictor variables and $\beta \in \mathbb{R}^p$ is a vector of regression coefficients to be estimated. A least squares estimate is a minimizer of the residual sum of squares $\|y - X\beta\|^2$. A popular alternative estimate is the lasso estimate (Tibshirani, 1996), which minimizes $\|y - X\beta\|^2 + \lambda\|\beta\|_1$, a penalized residual sum of squares that balances fit to the data against the possibility that some or many of the elements of β are small or zero. Indeed, minimizers of this penalized sum of squares may have elements that are exactly zero.

There exist a large variety of optimization algorithms for finding lasso estimates (see Schmidt et al. (2007) for a review). However, the details of many of these algorithms are somewhat opaque to data analysts who are not well-versed in the theory of optimization. One exception is the local quadratic approximation (LQA) algorithm of Fan and Li (2001), which proceeds by iteratively computing a series of ridge regressions. Fan and Li (2001) also suggested using LQA for non-convex L_q penalization when $q < 1$, and this technique was used by Kabán and Durrant (2008) and Kabán (2013) in their studies of non-convex L_q-penalized logistic regression. However, LQA can be numerically unstable for some combinations of models and penalties. To remedy this, Hunter and Li (2005) suggested optimizing a surrogate "perturbed" objective function. This perturbation must be user-specified, and its value can affect the parameter estimate. As an alternative to using local quadratic approximations, Zou and Li (2008) suggest L_q-penalized optimization using local linear approximations (LLA). While this approach avoids the instability of LQA, the algorithm is implemented by iteratively solving a series of L_1 penalization problems for which an optimization algorithm must be chosen as well.

This article develops a simple alternative technique for obtaining L_q-penalized regression estimates for many values of $q \leq 1$. The technique is based on a non-identifiable Hadamard product parametrization (HPP) of β as $\beta = u \circ v$, where

E-mail address: peter.hoff@duke.edu.

http://dx.doi.org/10.1016/j.csda.2017.06.007
0167-9473/© 2017 Published by Elsevier B.V.

"$\circ$" denotes the Hadamard (element-wise) product of the vectors u and v. As shown in Section 2, if $\hat{u}$ and $\hat{v}$ are optimal L_2-penalized values of u and v, then $\hat{\beta} = \hat{u} \circ \hat{v}$ is an optimal L_1-penalized value of β. An alternating ridge regression algorithm for obtaining $\hat{u} \circ \hat{v}$ is easy to understand and implement, and is competitive with LQA in terms of speed. Furthermore, a modified version of HPP can be adapted to provide fast convergence in sparse, high-dimensional scenarios. In Section 3 we consider extensions of this algorithm for non-convex L_q-penalized regression with $q \leq 1$. As in the L_1 case, L_q-penalized linear regression estimates may be found using alternating ridge regression, whereas estimates in generalized linear models can be obtained with a modified version of an iteratively reweighted least squares algorithm. In Section 4 we show how the HPP can facilitate structured sparsity in parameter estimates: The L_2 penalty on the vectors u and v can be interpreted as independent Gaussian prior distributions on the elements of u and v. If instead we choose a penalty that mimics a dependent Gaussian prior, then we can achieve structured sparsity among the elements of $\hat{\beta} = \hat{u} \circ \hat{v}$. This technique is illustrated with an analysis of brain imaging data, for which a spatially structured HPP penalty is able to identify spatially contiguous regions of differential brain activity. A discussion follows in Section 5.

2. L_1 optimization using the HPP and ridge regression

2.1. The Hadamard product parametrization

The lasso or L_1-penalized regression estimate $\hat{\beta}$ of β for the model $y \sim N_p(X\beta, \sigma^2 I)$ is the minimizer of $\|y - X\beta\|^2 + \lambda\|\beta\|_1$, or equivalently of the objective function

$$f(\beta) = \beta^\top Q \beta - 2\beta^\top l + \lambda\|\beta\|_1, \tag{1}$$

where $Q = X^\top X$ and $l = X^\top y$. Now reparametrize the model so that $\beta = u \circ v$, where "$\circ$" is the Hadamard (element-wise) product. We refer to this parametrization as the Hadamard product parametrization (HPP). Estimation of u and v using L_2 penalties corresponds to the following objective function:

$$g(u, v) = (u \circ v)^\top Q (u \circ v) - 2(u \circ v)^\top l + \lambda(u^\top u + v^\top v)/2. \tag{2}$$

Consideration of this parametrization and objective function may seem odd, as the values of u and v beyond their element-wise product β are not identifiable from the data. However, g is differentiable and biconvex, and its local minimizers can be found using a very simple alternating ridge regression algorithm. Furthermore, there is a correspondence between minimizers of g and minimizers of f, which we state more generally as follows:

Lemma 1. *Let $f(\beta) = h(\beta) + \lambda\|\beta\|_1$ and $g(u, v) = h(u \circ v) + \lambda(u^\top u + v^\top v)/2$. Then*

1. *$\inf_\beta f(\beta) = \inf_{u,v} g(u, v)$;*
2. *if $(\hat{u}, \hat{v})$ is a local minimum of g, then $\hat{\beta} = \hat{u} \circ \hat{v}$ is a local minimum of f.*

Proof. To show item 1 we write $u = \beta/v$, where "/" denotes element-wise division, so that

$$\begin{aligned}
\inf_{u,v} g(u, v) &= \inf_{\beta,v} g(\beta/v, v) \\
&= \inf_\beta \inf_v \left\{h(\beta) + \lambda\left(\|\beta/v\|^2 + \|v\|^2\right)/2\right\} \\
&= \inf_\beta \left\{h(\beta) + \lambda \inf_v \left(\|\beta/v\|^2 + \|v\|^2\right)/2\right\}.
\end{aligned}$$

The inner infimum over v is attained, and a minimizer $\tilde{v}$ can be found element-wise. The jth element $\tilde{v}_j$ of a minimizer $\tilde{v}$ is simply a minimizer of $\beta_j^2/v_j^2 + v_j^2$. If β_j is zero then $\tilde{v}_j = 0$ is the unique global minimizer. Otherwise, this function is strictly convex in v_j^2 with a unique minimum at $\tilde{v}_j^2 = |\beta_j|$. The inner minimum is therefore

$$\begin{aligned}
\|\beta/\tilde{v}\|^2 + \|\tilde{v}\|^2 &= \sum_{j=1}^p \left(\beta_j^2/\tilde{v}_j^2 + \tilde{v}_j^2\right) \\
&= \sum_{j=1}^p \left(\beta_j^2/|\beta_j| + |\beta_j|\right) = 2\|\beta\|_1,
\end{aligned}$$

and so

$$\begin{aligned}
\inf_{u,v} g(u, v) &= \inf_{\beta,v} g(\beta/v, v) \\
&= \inf_\beta \left\{h(\beta) + \lambda \min_v \left(\|\beta/v\|^2 + \|v\|^2\right)/2\right\} \\
&= \inf_\beta \{h(\beta) + \lambda\|\beta\|_1\} = \inf_\beta f(\beta).
\end{aligned}$$

Please cite this article in press as: Hoff, P.D., Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.06.007.