Bayesian Lasso

Anonylise Wagner

Department of Statistics, University of Washington Seattle, WA, 98195, USA

Abstract

The Bayesian Lasso, building on the interpretation of Tibshirani, places Laplace priors on linear regression coefficients to allow for Bayesian approaches to parameter and error estimation. An efficient Gibbs sampler allows for quick computation and may be exanded to other forms of penalized regression.

1 Introduction

(VM's comment: **Your introduction looks more like beginning of the Methods section to me. Ideally, intro should introduce the problem without formulae. Need to motivate sparsity and give a literature review of sparse regression, explaining why it is useful and what people have done before Park and Casella paper.**) Linear regression is a broad problem with a myriad of proposed techniques for solving. At its heart, linear regression assumes that we have some vector of responses, $\boldsymbol{y}$, which depend linearly on some covariates, $\boldsymbol{X}$. Formally, accounting for an error term, we wish to fit the model

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector, μ is the mean of $\boldsymbol{y}$, $\boldsymbol{X}$ is a matrix of regressors (typically standardized), and $\boldsymbol{\epsilon}$ is a vector of indepedent and identically distributed zero mean normal variables with an unknown variance. For the sake of this discussion we will assume that the $\boldsymbol{y}$ has zero mean ($\mu = 0$), noting that in practice this can be easily achieved by subtracting the sample mean estimate.

What is really of interest is the regression coefficients, $\boldsymbol{\beta}$. These can be used for prediction when new values of regressors are given, or they may be of interest for their interpretation on the effects of certain regressor values on the dependent $\boldsymbol{y}$. Irrespective of the use, there are

a number of approaches to finding the regressor coefficients. One familiar method, ordinary least squares, seeks to minimize least squares error

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

While OLS estimates are a straight-forward and simple approach, Section 2 details some of the shortcomings of this method.

Penalized regression takes a similar approach, trying to minimize the square error of predicted values, but with further constraints on regression cofficients. Of particular interest is least absolute shrinkage and selection operator (Or Lasso), which penalizes the sum of the absolute value of the coefficients. For some penalty weight $\lambda \geq 0$, the Lasso has the form

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

In the Lasso original derivation it was noted that these estimates can be equivalently viewed as posterior modes of OLS estimates, when independent Laplace priors are placed on the $\boldsymbol{\beta}$s.Tibshirani [1996]

¡¡¡¡¡¡¡ HEAD This idea can be expanded to other penalized regressions of the form

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

for some $q \geq 0$.Park and Casella [2008] These are collectively known as Bridge Regression, though $q = 1$ (Lasso Regression) and $q = 2$ (Ridge Regression) have appealing and better understood properties.

2 Background and Motivation

Ordinary least squares estimates are well understood, allowing for estimation of fit parameters and error suitable to many applications.(I want to find something to cite for this) This method is not without its setbacks, and alternative approaches are required in many situations.

Correlated regressor variables, for example, cause problems in estimating the statistical properties of OLS $\boldsymbol{\beta}$ estimates. OLS will also tend to fit all β_j with at least non-zero value,

as this will usually produce minor decreases in the squared error.Seeger [2008] This last problem helps to illustrate the desire for sparsity, a property that can be assumed a priori or may be required due to underdetermined-ness of data.

While the Lasso does improve on OLS point estimates, until recently error estimates were difficult to derive. The more desirous property of the Lasso is its ability to produce sparce estimates of parameters. It accomplishes this task by restricting solutions onto an L_1 'ball' around the origin. This is more obvious from the dual formulation,

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq r$$

for some r determined by λ. Figure 2 shows how this penalization compares to a Ridge Regression, Bridge Regression with $q = 2$ which restricts solutions onto an L_2 ball.Park and Casella [2008]

While sparcity may be an a priori assumption made about the data, underdetermined data requires this assumption to create useful models. One common field often presented with this problem is the study of gene expression, where there may be thousands to hundreds of thousands of predictors but only a few hundred data points.Seeger [2008]

The Bayesian Lasso, in keeping with Tibshirani's original interpretation, places independent Laplace (or double exponential) distributions on the parameters. Specifically,

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}$$

is placed on each distribution, which results in a unimodal posterior. Other prior distributions have been considred, but unimodality is key for Lasso estimates as non-unimodal posterior distributions cause convergence problems.

In the paper by Park and Casella, they produce a heirarchical models along with a Gibbs sampling technique to sample from posterior distributions. Alongside this formulation they also present an approach to a data-driven estimation for λ hyperparameters.Park and Casella [2008] =======

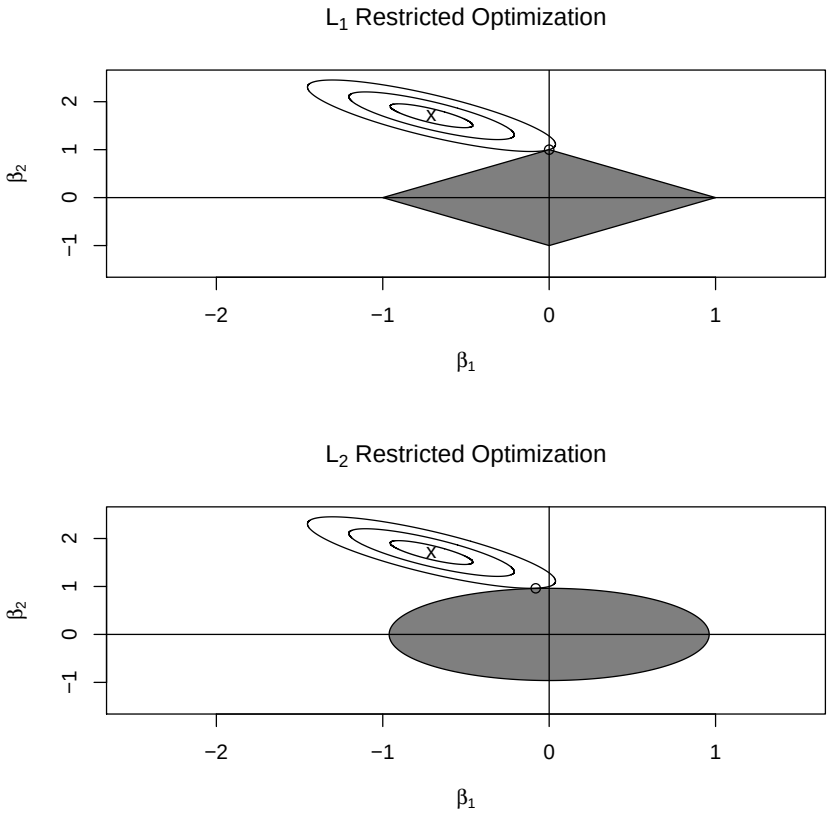

Figure 1: This example illustrates the tendancy for Lasso estimates to be driven to 0, compared with Ridge Regression estimates which are close to but not exactly 0 for β_1. (I want to make this side by side, and 1:1)

3 Background and Motivation

(VM's comment: **I don't think you need a separate Background section; make it a Methods subsection if needed**) ¿¿¿¿¿¿¿ 454fc9dcb7215da08f8dca241a67a31bd26acfc4

4 Methods

5 Results

6 Discussion

References

Trevor Park and George Casella. The bayesian lasso. Journal of the American Statistical Association, 103(482):681–686, 2008.

Matthias W Seeger. Bayesian inference and optimal design for the sparse linear model. The Journal of Machine Learning Research, 9:759–813, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.