10/25/2010

Linear Regression - continued

The Gauss-Markov theorem captures the inevitable tradeoff between the bias of an estimator and it's variance. In this lecture we stipulate restrictions on our estimator, so as to minimize its variance on the account of its bias.

1 Ridge Regression

Recall the linear regression model, were the set of observed data points is

$$\mathcal{D} = \{x_n, y_n\}$$

Ridge regression is a variant of ordinary linear regression, whose goal is to circumvent possible collinearity of the predictors, that is, when the design matrix is not invertible. This method can be viewed as artificially modifying the design matrix so as to make its determinant "sufficiently" different from 0 - This modification causes the estimator to be biased (as opposed to the RSS estimator), but significantly reduces the variance of the estimator.

Formally, the Ridge regression estimation problem is the following:

$$\min_{\beta}\{RSS(\mathcal{D}, \beta)\} \;\; s.t \;\; \sum_{i=1}^{n} \beta^2 \leq s$$

Therefore, the Ridge estimator is:

$$\hat{\beta}_{Ridge} = arg\min_{\beta}\{\sum_{n=1}^{N} \frac{1}{2}(y_n - \beta x_n)^2 + \lambda \sum_{i=1}^{n} \beta^2\}$$

Notes:

1) For a fixed choice of λ, the above expression is convex, and therefore an optimal solution w.r to β indeed exsists.

2) The exists a $1-1$ mapping between $\lambda \longleftrightarrow s$ (and thus the two equations above are equivalent).

3) λ is called "the **complexity** parameter", and has a large effect on the fit of the estimator. Intuitively, choosing lage values of λ incurrs a large "penalty" for overfitting estimators β with a large second norm (moment), and so a low-variance but highly biased solution will be preferred. The choice of λ can be viewed as the extent to which we modify the design matrix so as to decrease the correlation between the covariates.

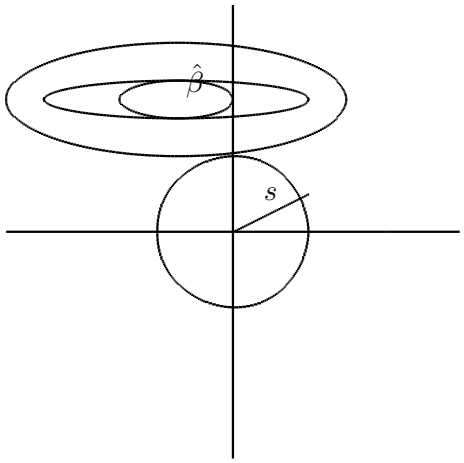


Image 1. An illustration of the Ridge regression estimation. The estimator is restricted to the s-ball of the origin. The unbiased RSS estimator is shown outside the ball, and the contours around it correspond to the value of the residual sum of squares at any given point (In general, the RSS is a quadratic function in x and y, and so it has the geometric shape of an ellipsoid). The value of the Ridge estimator is obtained in the tangent point of a contour and the ball.

In practice, the choice of λ is done by using the following **cross validation** technique:

1) Divide the data into k folds.

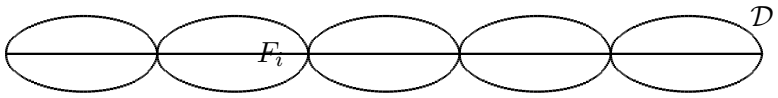
2) For each fold F_i and each candidate value for λ (we assume the candidates are a finite set of real values), define:

$$\epsilon_{n,\lambda} \stackrel{\text{def}}{=} (y_n - \hat{\beta}^{F_i,\lambda}_{Ridge} x_n)^2 \quad \forall n \in F_i$$

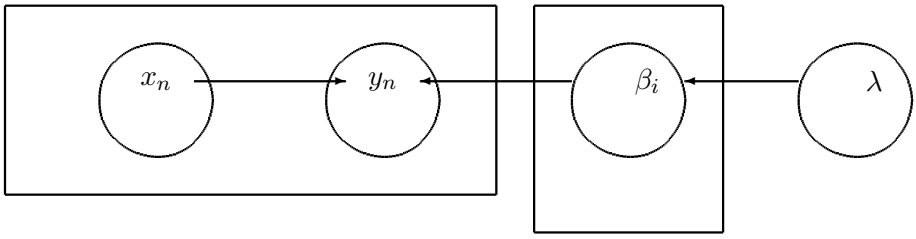
where $\hat{\beta}^{F_i,\lambda}_{Ridge}$ is estimated on $\mathcal{D}_{-F_i}$ (the data not including the ith fold). I.e, we use the ith fold as a "test" set, and the rest of the data as the "learning" set.

3) Finally, choose

$$\lambda^{x-val} = arg\min_{\lambda} \frac{1}{N} \sum_{n=1} N\epsilon_{F_i,\lambda}$$



1.1 Bayesian Connection



Assume the linear regression model as in image 3, where:

$$\beta_i \sim \mathcal{N}(0, 1/\lambda) \quad y_n|x_n, \beta_{1:p} \sim \mathcal{N}(\beta^T x_n, \sigma^2)$$

Now, consider a MAP estimation of β:

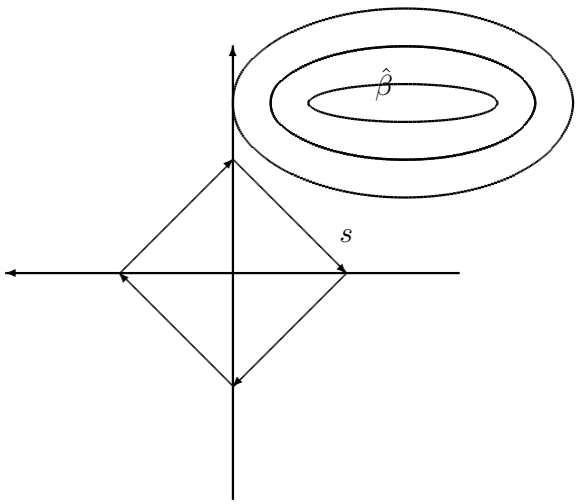
$$\begin{aligned}
\hat{\beta} &= arg\max_{\beta} \log p(\beta|x_{1:N}, y_{1:N}, \lambda) \\
&= arg\max_{\beta} \log p(y_{1:N}|x_{1:N}, \beta) \prod_{i=1}^{p} p(\beta_i|\lambda) \\
&= arg\max_{\beta} \sum_{n=1}^{N} \log p(y_n|x_n, \beta) + \sum_{i=1}^{p} \log p(\beta_i|\lambda) \qquad (1)
\end{aligned}$$

But $p(\beta_i|\lambda) = \frac{1}{\sqrt{2\pi/\lambda}} e^{-\lambda\beta_i^2} - \sum_{i=1}^{p} \lambda\beta_i^2$, where the last term is the "Ridge penalty" incurred by λ (a large choice of λ yields a lower variance of the β_i's). Note also that the first term in (1) above is minimized by the RSS estimator.

1.2 The Lasso method

In this part of the lecture we consider a different restriction on the estimator β, namely, that its L_1 norm will be bounded (as opposed to the L_2 norm restriction in the Ridge estimator). More formally,

$$\hat{\beta}^{Lasso} \stackrel{\text{def}}{=} arg\min_{\beta} RSS(\mathcal{D}, \beta) + \lambda \sum_{i=1}^{p} |\beta_i| \qquad (2)$$



Typically (when $p >> n$), the contours "hit" the restricted region at a vertx, which means $\beta_i = 0$ for some covariates i. This means our solution is **sparse** . The Lasso method therefore determines the importance of the covariates in a form of "feature selection".

Notes:
1) There are algorithms for exploring the whole regularization path (GLMnet in **R**).
2) The function in (2) is convex $\Longrightarrow$ an optimal solution exists.
3) Sparsistent.

Important remark: The Lasso method produces a sparse solution only for MAP estimation.

2 Exponential Families

-A very "flexible" family of distributions:

Multinomial - categorical
Gaussian - Real
Dirichlet - simplex
Gamma - R^+
Poisson - Naturals
Beta - [0,1]
Bernoulli - 0,1

The probability (density) function of a distribution in the Exponential family has the following general form:

$$p(x|\eta) = h(x)e^{\eta^T t(x) - a(\eta)}$$

where:

η - a natural parameter
$t(x)$ - sufficient statistics
$h(x)$ - the underlying measure under which $p(x, \eta)$ is defined (e.g the Lebesgue measure).
$a(\eta)$ - the log normalizer (ensures that $p(x, \eta)$ sums/integrates to 1): $a(\eta) = \log \int h(x)e^{\eta^T t(x)}dx$ (integrating out the un-normalized density over the sample space x).

Example: Bernoulli r.v:

$$p(x|\pi) = \pi^x(1-\pi)^{1-x} = e^{\log(\pi^x(1-\pi)^{1-x})} \quad = \quad e^{x\log\pi + (1-x)\log(1-\pi)} = e^{x\log(\pi/(1-\pi)) + \log(1-\pi)} \tag{3}$$

where $\eta = \log(\pi/(1-\pi))$, $t(x) = x$, $a(\eta) = -\log(1-\pi)$.
Now, invert the relationship between π and η: $\pi = \frac{1}{1+e^{-\eta}}$
$\Longrightarrow a(\eta) = -\log(1-\pi) = \log(1+e^{\eta})$.