Ridge Regression and Lagrange Multipliers

Ben Wing
Bakhtiyar Uddin

April 30, 2007

1 Intro

The objective function under ridge regression is

$$\beta^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j_1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

We want to show that this is equivalent to a standard least-squares objective

$$\beta^{ols} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j_1}^{p} x_{ij}\beta_j)^2$$

subject to the constraint

$$\sum_{j=1}^{p} \beta_j^2 \leq c$$

for some constant c.

2 Lagrange Multipliers

The standard theory of constraint optimization involves Lagrange multipliers. It is normally phrased as optimizing $f(\beta)$ (where β is a vector) subject to a constraint $g(\beta) = c$.

To relate this to ridge regression, consider for the moment a case where we have only two predictor variables, so that

$$f(\beta) = f(\beta_1, \beta_2) = \arg\min_{\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2)^2$$
$$g(\beta) = g(\beta_1, \beta_2) = \beta_1^2 + \beta_2^2$$

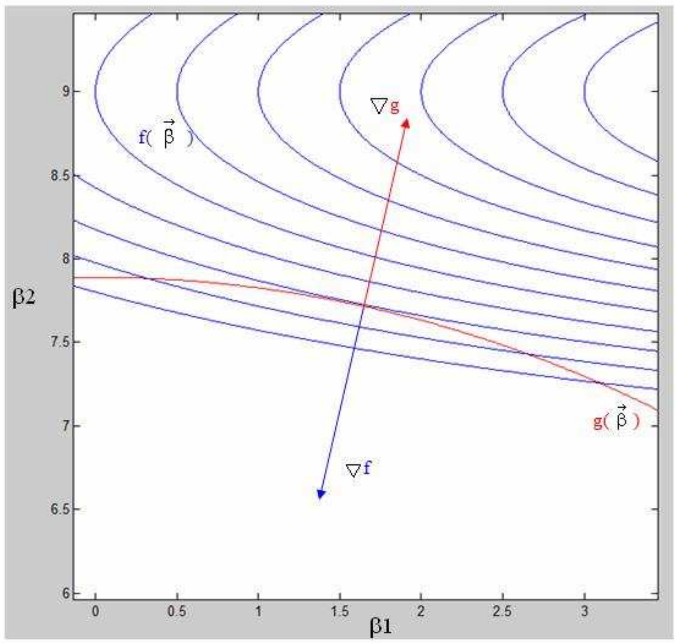

Figure 1: Zoom-in showing optimal point on constraint line

The situation is represented graphically in Figure 1, where the constraint $g(\beta_1, \beta_2)$ is represented as a red circle around the origin, and contour lines for $f(\beta_1, \beta_2)$ are shown. Each contour line represents all the (β_1, β_2) points with a particular value of $f(\beta_1, \beta_2)$. (These are ellipses surrounding the minimal value $(\hat{\beta}_1, \hat{\beta}_2)$.)

Figure 2 is a zoomed-out equivalent of Figure 1, showing the full constraint and contour lines.

Imagine moving along the constraint line. As can be seen, the value of $f(\beta)$ steadily decreases until the point where the constraint line $g(\beta)$ is tangent to $f(\beta)$, and then increases again. The minimum occurs at the point of tangency, where

$$\nabla f(\beta) = -\lambda \nabla g(\beta)$$

λ is called a *Lagrange multiplier*.

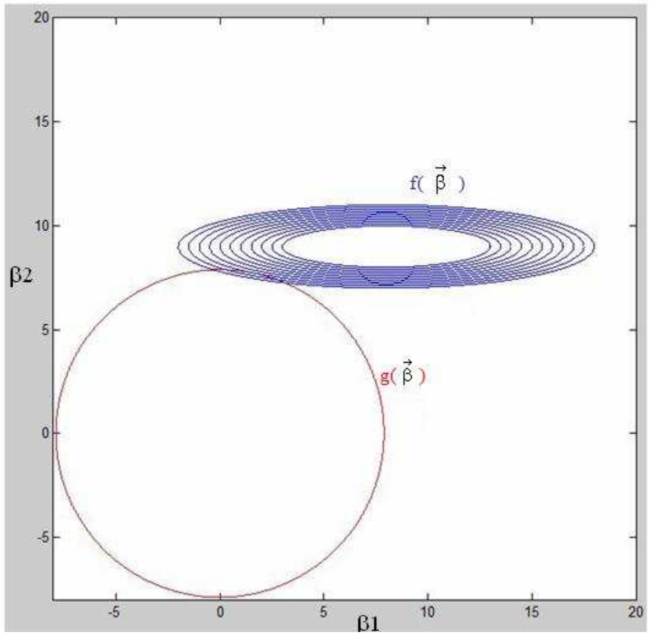

Figure 2: Unzoomed equivalent

The sign of λ is important. As shown in the figure, the minimum occurs when the gradient vectors point in opposite directions, and hence λ is positive. There is another tangent point, however, on the opposite side of the circle, where the gradients point in the same direction. In this case, λ is negative, and the objective is *maximized*, which is not what we want.

The normal method of Lagrange multipliers uses an objective function

$$L(x, \lambda) = f(\beta) + \lambda(g(\beta) - c)$$

By taking partial derivatives with respect to β and λ, and setting them to zero, it can easily be seen that the correct equations

$$\nabla f(\beta) = -\lambda \nabla g(\beta)$$
$$g(\beta) = c$$

result.

3 Lagrange Multipliers, Ridge Variation

Normally, these equations are solved for β and λ. In our case, however, λ is given.

Consider again the ridge objective, which we can write as

$$\beta^{ridge} = f(\beta) + \lambda g(\beta)$$

Note that this is nearly identical to the Lagrange objective. In fact, take the gradient with respect to β, and set to zero, to yield

$$\nabla f(\beta) = -\lambda \nabla g(\beta)$$

This is identical to the equation for constrained optimization, and shows why the ridge objective is equivalent to a constrained optimization problem.

The value of c is not particularly important, but it can be found by finding the optimum β values and plugging them into the equation $g(\beta) = c$. Note that this will depend on both λ and x. However, for a specific x, varying λ will cause c to vary accordingly. Note that when λ is 0, the ridge objective will be the same as the normal OLS objective, and the value of β will be the same as the value of β that minimizes the normal OLS objective. Since the constraint still applies, the value of c will be the one that causes the constraint circle to pass through the minimal value of $f(\beta)$, at the center of the contour ellipses in the graph. When λ goes towards infinity, β, and hence c, will shrink towards 0.

4 Equal or Less-Than-or-Equal?

Note that the above discussion concerned an equality constraint of the form $g(\beta) = c$. A constraint of the form $g(\beta) \leq c$ is a bit trickier. In this situation, there are two cases: Either the unconstrained optimum occurs within the constraint region, or it occurs outside. In the former case, the constraint is inactive, and the constrained and unconstrained optima are identical. In the latter case, the constraint is active, the constrained optimum occurs on the constraint line, and the constraint becomes an equality rather than an inequality. Hence, the normal procedure for dealing with an inequality constraint is to first compute the unconstrained optimum and check whether it satisfies the constraint, and if not to compute the connstrained optimum.

For ridge regression, however, it turns out that the former case (inactive constraint) never applies. This follows from the above discussion of the relation between c and λ. When λ is 0, the value of c is such that the unconstrained optimum falls on the constraint line. As λ increases, β decreases – in fact, all values in β decrease – and hence c decreases. Thus, the unconstrained optimum for β always lies on or outside the constraint line, and the equality constraint always applies. This means that we can equally well write the ridge constraint as $g(\beta) \leq c$.