CISC 371 Class 27

Ridge Regression and The Lasso

Texts: [1] pp. 61–73

Main Concepts:

- *Ridge regression as regularization*
- *Ridge regression solved as constrained least squares*
- *The Lasso is constrained least squares with L_1 vector norm*

Sample Problem, Data Analytics: How can we compute a regression using fewer variables?

We now revisit the data analysis problem that we posed when we considered constrained least squares, which is how to improve an estimate of a linear relationship in a set of data. We will do this in three ways:

Standardized Data: Instead of using a given data vector, use the number of standard deviations from the vector's mean of each entry

Ridge Regression: Instead of using a maximum θ on the squared Euclidean norm of a solution vector $\vec{w}^*$, provide the Lagrange multiplier as a hyper-parameter in Tikhonov regularization

LASSO Regression: Instead of using a maximum θ on the squared Euclidean norm of a solution vector $\vec{w}^*$, use the sum of the absolute values of the entries of $\vec{w}^*$

27.1 Standardized Data or *Z* Score

In regression – especially for methods that are developed from a statistical point of view – it is usual practice to use *standard* or standardized data. These are data that have a mean of zero and a variance of one. Historically, the concept seems to have been suggested, at around 1900, by Karl Pearson [2] when he introduced the concept of a probability "ellipsoid", in which data were scored by their number of standard deviations from the mean in each relevant direction.

We can derive a standardization transformation for a single vector $\vec{a}$ and then generalize to a matrix of data. The mean of a vector $\vec{a}$ is defined as the sum of the entries divided by the number

of entries. This is a scalar value that is commonly written as the Greek letter μ; note that this is *not* a Lagrange multiplier but is a statistical value. We will define the mean of a vector $\vec{a}$ as

$$\mu_{\vec{a}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} a_i \tag{27.1}$$

For a matrix A, we will define the mean as a 1-form. The entries of the mean 1-form are the means of the columns. We will define the mean of a matrix A, which is a generalization of Definition 27.1, as

$$\underline{\mu}_A \stackrel{\text{def}}{=} \begin{bmatrix} \mu_{\vec{a}_1} & \mu_{\vec{a}_2} & \cdots & \mu_{\vec{a}_n} \end{bmatrix} \tag{27.2}$$

Because a vector is a matrix that has a single column, we could also use Definition 27.2 to write the mean of a vector $\vec{a}$ as $\underline{\mu}_{\vec{a}}$.

Statistically, the variance of a set of data has two definitions: the *sample* variance and the *population* variance. Here, we will use the population variance; we will be cautious because other writings and code may explicitly or implicitly use the sample variance.

For a vector $\vec{a}$, the population variance is the sum of the squares of the entries of the vector divided by the size of the vector. Using the conventional statistical symbol σ^2, we will define the population variance of a vector as

$$\sigma_{\vec{a}}^2 \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} (a_i - \mu_{\vec{a}})^2 \tag{27.3}$$

We will define the population variance of a matrix A as a diagonal matrix Σ. The j^{th} diagonal entry of Σ is the population variance of the j^{th} column of A. The definition of the population variance of a matrix A, using Definition 27.3, is

$$\Sigma_A \stackrel{\text{def}}{=} \begin{bmatrix} \sigma_{\vec{a}_1}^2 & & & \\ & \sigma_{\vec{a}_2}^2 & & \\ & & \ddots & \\ & & & \sigma_{\vec{a}_n}^2 \end{bmatrix} \tag{27.4}$$

We can use these definitions to concisely state how we will standardize data. Suppose that we are given a regression problem with a matrix of independent data M and a vector of dependent data $\vec{c}$, that is to be approximated as

$$M\vec{u} = \vec{c} \tag{27.5}$$

We want to convert Equation 27.5 to a standardized form, which implies that we seek:

- A design matrix X with columns that are zero-mean and unit variance
- A vector of dependent data $\vec{y}$ with zero mean and unit variance
- A linear model of the data that is

$$X\vec{w} = \vec{y} \tag{27.6}$$

For the independent data in M, we must subtract the mean of M from M and scale the difference to have a population variance of one. We must likewise modify the dependent data in C. These operations can be expressed as

$$\begin{aligned} X &= [M - [\vec{1}\underline{\mu}_M]]\Sigma_M^{-1/2} \\ \vec{y} &= [\vec{c} - [\vec{1}\mu_{\vec{c}}]]\Sigma_{\vec{c}}^{-1/2} \end{aligned} \tag{27.7}$$

Observation: **Equation 27.5 and Equation 27.6 solve different problems**

We can explore some effects of using zero-mean data by considering a problem where the population variances are one, that is, where $\Sigma_M = I$ and $\Sigma_{\vec{c}} = 1$. For any such unit-variance problem, we would estimate a solution of Equation 27.6 to be the vector $\hat{w}$. We would then estimate the dependent data, which are in the vector $\vec{c}$ of Equation 27.5, from the zero-mean independent data in the matrix X as

$$\hat{c} = X\hat{w} + [\vec{1}\mu_{\vec{c}}] \tag{27.8}$$

In general, the solution $\hat{w}$ of Equation 27.6 is *not* the same as a solution $\hat{y}$ of Equation 27.5.

Methods that are derived from the statistical point of view generally try to find solutions that are based on standard data, as stated in Equation 27.6. In particular, we generally want to avoid problems where the independent data in the matrix M are expressed as a Vandermonde matrix, or where a constant offset term is stated. We can see that, computationally, the transformation in Equation 27.7 will produce a division by zero when any column of M is a constant value, because we would be trying to invert a scaling matrix that refers to the variance of the zero vector $\vec{0}$.

Example: 9 data with 2 outliers

We can test linear regression and partial standardization on a simple data set. Suppose that we use the integers from 1 to 9 as independent data in a matrix M with a single column, so $M = \vec{m}$. For dependent data we will use a formula based on Euler's number e:

$$\begin{aligned} c_1 &= e\, m_1 - 5 \\ c_i &= e\, m_i \qquad\qquad \text{for } i = 2 \ldots 8 \\ c_9 &= e\, m_9 + 3 \end{aligned} \tag{27.9}$$

We will round the entries in $\vec{c}$ to two decimal places. If we solve Equation 27.6, using the data Equation 27.9 and the method of ordinary least squares, we find the estimated weight scalar to be

$$\hat{y} \approx 2.7954$$

We can find the zero-mean versions of the data as

$$X = M - \vec{1}\underline{\mu}_M \qquad \vec{y} = \vec{c} - \vec{1}\mu_{\vec{c}}$$

Using Equation 27.6, the estimated weight scalar is

$$\hat{w} \approx 3.2515$$

We can estimate the dependent data using M and $\hat{y}$ as $\hat{c}_M = M\hat{y}$, and using X, $\hat{w}$, and $\mu_{\vec{c}}$ as $\hat{c}_X = X\hat{w} + \vec{1}\mu_{\vec{c}}$. The data and the estimates are plotted in Figure 27.1.

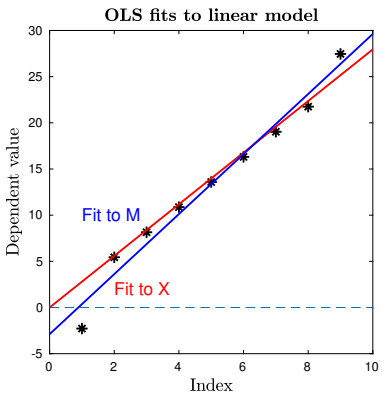

Figure 27.1: Plots of linear regressions on given data and on standardized data. The given data are shown as black asterisks. The OLS solution on the given data is shown as a red line. The OLS solution on the zero-mean version of the given data is shown as a blue line.

Computing the root-mean-square error, or RMSE, of each estimate gives

$$\text{RMSE}(\vec{c} - \hat{c}_M) \approx 1.89 \qquad \text{RMSE}(\vec{c} - \hat{c}_X) \approx 1.35$$

By using standardized data, we get a lower error and a visually better estimate of the linear relationship between the independent data and the dependent data.

27.2 Ridge Regression

There is a close mathematical relationship between constrained least squares, Tikhonov regularization, and a well known optimization method that is called *ridge regression*. This relationship has a curious history.

Working independently from Tikhonov, a scientist employed by the E. I. DuPont Company was trying to solve optimization problems that arose in the course of business. In 1959, Roger Hoerl observed "the frequent occurrence of nonsensical estimates from least squares multiple regression" [3], p. 190. He reasoned that these difficulties could be addressed in part by changing the optimization in the direction of minimal eigenvectors of the symmetric positive definite matrix $X^T X$ in Equation 26.5. In collaboration with Robert Kennard in 1970, a method for constrained optimization was derived [4]. Their derivation of *ridge regression* of was motivated by statistics and used statistical methods for analysis.

Ridge regression is mathematically equivalent to the elementary Tikhonov regularization of Equation 26.16. In developing constrained least squares (CLS), we imposed a constraint on the squared Euclidean norm, which was

$$\|\vec{w}^*\|^2 \leq \theta$$

From this constraint, we formulated the Lagrange function

$$\mathcal{L}(\vec{w}, \lambda) = f(\vec{w}) + \lambda(\|\vec{w}\|^2 - \theta) \tag{27.10}$$

Using Equation 26.29, we can compute the optimal Lagrange multiplier λ^* from the constraint hyper-parameter θ by finding the zero of a monotonically decreasing function.[1]

Ridge regression is a particular form of Tikhonov regularization, in which the hyper-parameter λ is provided by the user. In ridge regression, we seek the minimizer

$$\vec{w}^* = \underset{\vec{w} \in \mathbb{R}^n}{\operatorname{argmin}} f(\vec{w}) + \lambda \|\vec{w}\|^2 \tag{27.11}$$

The existence of a map from θ to λ^* implies that constrained least squares and ridge regression are effectively the same. The difference is whether the user supplies the hyper-parameter θ, from which we calculate λ^*, or supplies λ directly.

The optimization problem of Equation 27.11 has a simple closed-form solution. Differentiating and re-arranging terms produces the closed-form solution

$$\vec{w}^* = [X^T X + \lambda I]^{-1} X^T \vec{y} \tag{27.12}$$

Ridge regression, using Equation 27.12, is simple to implement.

[1] It is slightly tedious, but not difficult, to show that the function in Equation 26.29 is monotonically decreasing for the interval $\lambda \in [0, \infty]$; because the function is also continuous, it has a zero in the interval.

Ridge regression is still in active use for machine learning. One clear difficulty is the choice of the argument λ, which balances the ordinary least-squares solution and the constrain on the norm of the weight vector.

A major difficulty with ridge regression is that it does not, in general, change the individual weight values to zero: each w_j value may be small but we often compute $|w_j| > 0$. For data vectors $\vec{x} \in \mathbb{R}^n$ where n is large, an individual weight value $w_j = 0$ corresponds to a data "feature" that can be neglected for the purpose of linear regression. A *selection* operator is one that selects some number of weight values as non-zero and requires that the remaining weight values be zero. Selecting these non-zero weight values is not a trivial process.

In 1995, Leo Breiman proposed an algorithm called a "non-negative garotte" [5]. It was an iterative method that modified the ordinary least squares solution $\vec{w}^*_{LS}$ by selecting some weight values and "shrinking" these selected weights. This garotte was a computational improvement on both ridge regression and on subset selection.

27.3 The Lasso

In 1996, Robert Tibshirani, improved on the garotte with an algorithm [6] that is in active use in many academic disciplines. The name he proposed is clearly a play on words with "garotte"; however, for those who insist on an acronym, he obligingly invented one: Least Absolute Shrinkage and Selection Operator, or *LASSO*.

In our preferred terminology, the lasso is a linear least-squares problem that is constrained by the L_1 norm of the weight vector. Recall that one commonly used vector norm for a vector $\vec{w} \in \mathbb{R}^n$ is

$$L_1(\vec{w}) \text{ or } \|\vec{w}\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \|w_i\| \tag{27.13}$$

The lasso is simply the least-squares objective function with a threshold constraint θ:

$$\vec{w}^* = \underset{\vec{w} \in \mathbb{R}^n}{\operatorname{argmin}} [X\vec{w} - \vec{y}]^T [X\vec{w} - \vec{y}] \quad \text{such that} \quad \|\vec{w}^*\|_1 \leq \theta \tag{27.14}$$

The change from the L_2 norm in Equation 26.9 to the L_1 norm in Equation 27.14 is seemingly small but has a substantial effect on the minimization.

For both ridge regression and the lasso, if the ordinary least squares solution is not feasible, then the CLS or ridge regression produces a weight vector that is on the boundary of the level set

described by the constraint. Geometrically, the level set of the constraint in CLS or the equivalent ridge regression is a hypersphere of radius $\sqrt{\theta}$ around the origin. Also geometrically, the level set of the constraint in the lasso is a hypercube of width θ that is centered at the origin. Unit level sets for the constraints are illustrated, for a 2D problem, in Figure 27.2.

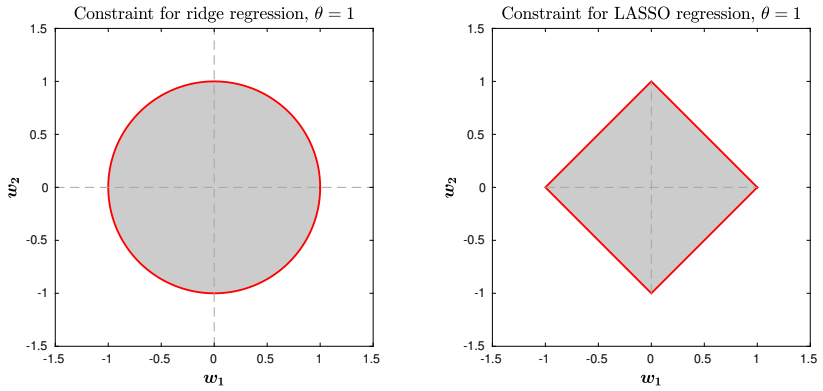

Figure 27.2: 2D level sets of the constraints for ridge regression and the lasso, using $\theta = 1$ for each problem. (A) Ridge regression is constrained to a unit disk centered at the origin. (B) The lasso is constrained to a unit square centered at the origin.

As shown by Tibshirani in 1996, with considerable work since then, the lasso works in practice by setting some weight values to exactly zero and ensuring that the sum of the absolute values of the other weight values meet the imposed constraint. We can explore the relationship between CLS (ridge regression) and the lasso by using a quadratic objective function

$$\begin{aligned} f(\vec{w}) &= [\vec{w} - \vec{w}_0]^T K [\vec{w} - \vec{w}_0] \\ \text{where} \quad K &= \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \\ \vec{w}_0 &= \begin{bmatrix} 0.95 \\ 2 \end{bmatrix} \end{aligned} \tag{27.15}$$

We can solve Equation 27.15 as a constrained optimization problem. Using a CLS, with $\theta = 1$, we compute $\vec{w}^*_{CLS}$; using the lasso, with $\theta = 1$, we compute $\vec{w}^*_{L1}$. The optimal weight vectors are

$$\vec{w}^*_{CLS} \approx \begin{bmatrix} 0.254 \\ 0.967 \end{bmatrix} \qquad \vec{w}^*_{L1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{27.16}$$

It is immediately apparent that there is a numerical difference between the CLS optimal weight vector $\vec{w}^*_{CLS}$ and the lasso optimal weight vector $\vec{w}^*_{L1}$. Using Equation 27.15, and the values computed in Equation 27.16, we can plot some contours of the objective function and the level sets of the constraint regions. These are shown in Figure 27.3.

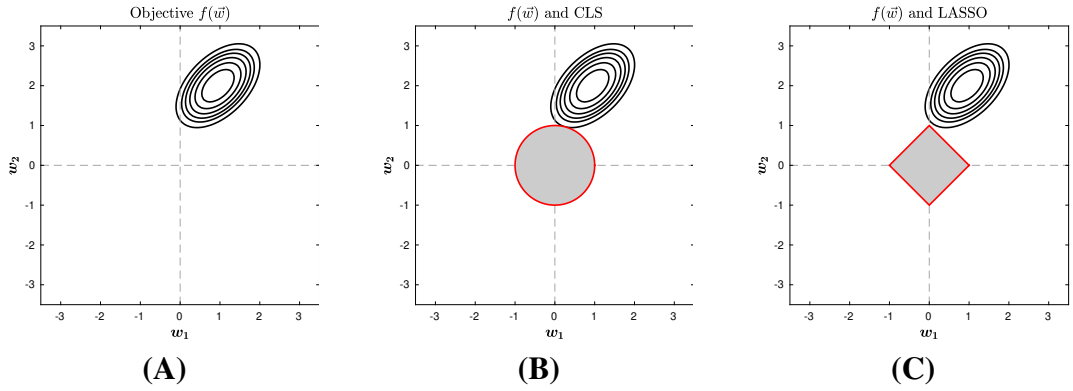

Figure 27.3: 2D contours of an objective function and level sets of the constraints for ridge regression and the lasso, using $\theta = 1$ for each problem. (A) Contours, or level curves, of a quadratic objective function are shown in blue. (B) Ridge regression, or constrained least squares, has a level set that is a unit disk centered at the origin. (C) The lasso has a level set that is a unit square centered at the origin.

Computation of the lasso is not straightforward. Considering the 2D case, the constraint of Equation 27.14 must be written as

$$\begin{aligned} w_1 + w_2 &\leq \theta \\ -w_1 + w_2 &\leq \theta \\ w_1 - w_2 &\leq \theta \\ -w_1 - w_2 &\leq \theta \end{aligned}$$

When we describe the constraint as a set of linear inequalities, we discover that we need 2^n inequalities to constrain a weight vector $\vec{w} \in \mathbb{R}^n$. Current methods use solutions that are linear in the number of dimensions n of the data space.

The lasso can be effectively computed using the inequalities sequentially, which is a form of coordinate descent. Current solutions use an interior-point method or an alternating-direction method.

The Elastic Net

In 2005, Zou and Hastie [7] provided an elegant method that combined ridge regression and the lasso. The concept was stated as being "like a stretchable fishing net that retains 'all the big fish'." [*op cit*, p. 302]

The concept is simple, although the statistical derivation is somewhat complicated. Let us begin by recalling that ridge regression is regularized by the squared L_2 norm and the lasso is regularized by the L_1 norm.

Suppose that, in addition to a hyper-parameter for ridge regression and a hyper-parameter for the lasso, we add a hyper-parameter that blends these regularization terms. If we want to linearly interpolate between these regularizers, we can use a hyper-parameter α and regularize with

$$(1 - \alpha)\|\vec{w}\|_2^2 + \alpha\|\vec{w}\|_1 \tag{27.17}$$

A subsequent statistical improvement to Equation 27.17 used the regularization term

$$\frac{(1 - \alpha)}{2}\|\vec{w}\|_2^2 + \alpha\|\vec{w}\|_1 \tag{27.18}$$

We can visualize the effects of the elastic net in two dimensions. Figure 27.4 shows the boundary of the elastic-net regularizer for a few values of α. For $\alpha = 0$, the elastic net is the L_2 regularizer of ridge regression; for $\alpha = 1$, the elastic net is the L_1 regularizer of the lasso; and for intermediate values of α, the elastic net appears like the holes of a fishing net that have been expanded by viscous drag in the water.

The elastic net provides a way for us to avoid a hard decision between ridge regression and the lasso. The lasso works poorly on under-determined problems and when sets of variables are closely correlated. Rather than switching to ridge regression for such problems, we can blend these methods to better fit our problem.

For further reading on the elastic net, the original paper by Zou and Hastie [7] is clear and approachable. Later writings by either of these authors are also worth exploring.

Extra Notes

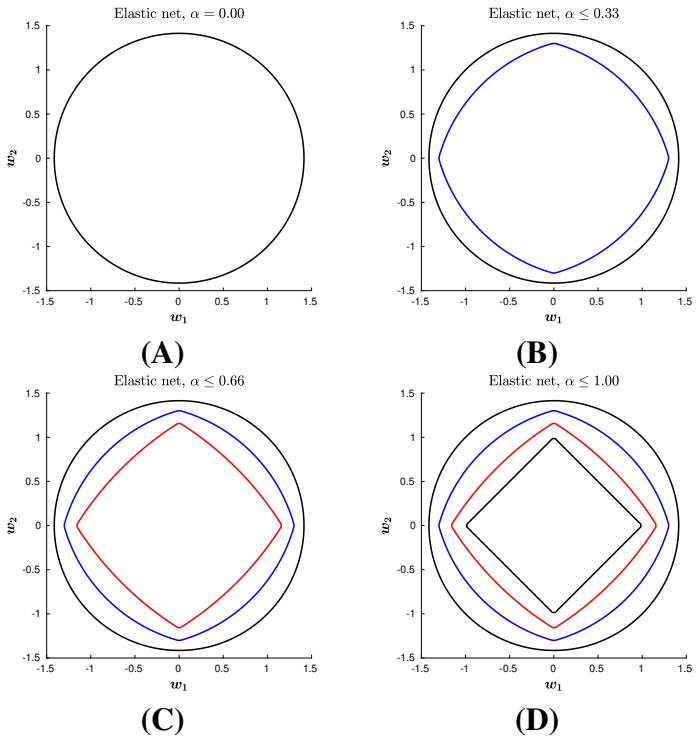

Figure 27.4: Regularizer of the elastic net, in two dimensions, for select values of $0 \leq \alpha \leq 1$. The boundary of the regularizer is interpolated between a circle and a square.

27.4 Extra Notes on Regression: Standardized Data and the Offset Term

A linear transformation of one vector to another vector is, strictly, a matrix multiplication. However, in common usage, a *linear regression* can include an "offset" term that we can write as

$$M\vec{u} + \alpha\vec{1} = \vec{c} \tag{27.19}$$

Alternatively, we can write Equation 27.19 as a partitioned matrix:

$$[M\ \vec{1}] \begin{bmatrix} \vec{u} \\ \alpha \end{bmatrix} = \vec{c} \tag{27.20}$$

The zero-mean form of Equation 27.20, where we do not correct for variances of the given data, is

$$[X\ \vec{0}] \begin{bmatrix} \vec{w} \\ \alpha \end{bmatrix} = \vec{y} \tag{27.21}$$

We can accomplish this task in two steps:

1. Solve $X\vec{w} = \vec{y}$ as $\hat{w}$
2. Solve for α in Equation 27.20

The second step can be derived by approximating $\hat{z} = \hat{w}$ and substituting $X = M - \vec{1}\underline{\mu}_M$ into Equation 27.20. This is

$$\begin{aligned}
& M\hat{w} + \vec{1}\alpha &=& \hat{c} \\
\equiv \quad & M\hat{w} + \vec{1}\alpha &=& X\hat{w} + \vec{1}\mu_{\vec{c}} \\
\equiv \quad & \left[X + \vec{1}\underline{\mu}_M\right]\hat{w} + \vec{1}\alpha &=& X\hat{w} + \vec{1}\mu_{\vec{c}} \\
\equiv \quad & X\hat{w} + \vec{1}\underline{\mu}_M\hat{w} + \vec{1}\alpha &=& X\hat{w} + \vec{1}\mu_{\vec{c}} \\
\equiv \quad & \vec{1}\underline{\mu}_M\hat{w} + \vec{1}\alpha &=& \vec{1}\mu_{\vec{c}} \\
\equiv \quad & \vec{1}\alpha &=& \vec{1}\mu_{\vec{c}} - \vec{1}\underline{\mu}_M\hat{w} \\
\equiv \quad & \vec{1}\alpha &=& \vec{1}\left[\mu_{\vec{c}} - \underline{\mu}_M\hat{w}\right] \\
\equiv \quad & \left[\vec{1}^T\vec{1}\right]\alpha &=& \left[\vec{1}^T\vec{1}\right]\left[\mu_{\vec{c}} - \underline{\mu}_M\hat{w}\right] \\
\equiv \quad & \alpha &=& \mu_{\vec{c}} - \underline{\mu}_M\hat{w}
\end{aligned} \tag{27.22}$$

To manage a requirement that the design matrix X and the dependent values $\vec{y}$ have a variance of 1, Equation 27.22 becomes

$$\alpha = \mu_{\vec{c}} - \underline{\mu}_M\Sigma_{\vec{c}}^{1/2}\left[\Sigma_M^{-1/2}\hat{w}\right] \tag{27.23}$$

We can reconcile our derivation – which uses linear algebra – with Tibshirani's statistical reasoning. He supposes that we are given a zero-mean matrix M, which implies that $\underline{\mu}_M = \underline{0}$, which reduces Equation 27.23 to $\alpha = \mu_{\vec{c}}$. Tibshirani's statement is that the offset term α is optimally the mean of the dependent data in $\vec{c}$; this is equivalent to our derivation.

Because the offset term α can be expressed in terms of the means and variances of the given data, it is both customary and acceptable to omit the offset term in optimization.

End of Extra Notes

References

[1] Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2009

[2] Pearson K: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag J Sci* 50(302):157–175, 1900

[3] Hoerl RW: Ridge analysis 25 years later. *Am Stat* 39(3):186–192, 1985

[4] Hoerl AE, Kennard RW: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67, 1970

[5] Breiman L: Better subset regression using the nonnegative garrote. *Technometrics* 37(4):373–384, 1995

[6] Tibshirani R: Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58(1):267–288, 1996

[7] Zou H, Hastie T: Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 67(2):301–320, 2005