Inference in High Dimensions and Regularization

Guy Lebanon

Traditional statistics has concentrated on the case of $d \ll n$ (the number of parameters to be estimated is much smaller than the train set size). In that case the maximum likelihood estimator (MLE) performs well and its asymptotic optimality "kicks in" (recall the asymptotic variance of the MLE is the best possible - the inverse Fisher information). However, as higher and higher dimensional parameter spaces are considered it was discovered that the MLE performs poorly. It overfits the training data and performs poorly on future test data. In other words, when $d \not\ll n$ the MLE picks up noise as signal, for example by assuming that irrelevant features are relevant due to small train set size.

From a mathematical perspective, in these cases the mean squared error (MSE) of the MLE $\mathsf{E}_{p^{\text{true}}} \|\hat{\theta} - \theta^{\text{true}}\|^2$ is too high. Recall that the MSE is the sum of the squared bias and variance and that the MLE is unbiased (as in the case of linear regression) or asymptotically unbiased (in other cases). It turns out that by introducing a bias we can substantially reduce the variance and get overall lower MSE. Specifically, regularization introduces a bias by reducing the absolute value of $\hat{\theta}^{\text{mle}}$ (bringing it closer to 0). The resulting estimation variance is much lower which means the overall estimation accuracy in terms of MSE (or otherwise) is better. Intuitively shrinking $\hat{\theta}^{\text{mle}}$ towards 0 depresses the tendency of the MLE to overfit by picking up training set noise for signal.

We follow with some specific regularization techniques. These techniques are most commonly described in the context of linear regression. However, most can be applied for other models as well (Fisher's LDA, naive Bayes, logistic regression, etc) though without the the nice closed forms of linear regression.

The earliest form of regularization is probably James-Stein shrinkage

$$\hat{\theta}^{\text{JS}} = \frac{1}{1+\tau} \hat{\theta}^{\text{mle}}, \qquad \tau > 0 \tag{1}$$

which shrinks all dimensions of $\hat{\theta}$ uniformly. Breiman's garrote shrinks the MLE as

$$\hat{\theta}_j^{\text{garrote}} = \hat{\theta}_j^{\text{mle}} \hat{\tau}_j, \quad j = 1, \ldots, d \quad \text{where} \quad \hat{\tau} = \arg\max_{\tau} \ell(\hat{\theta}_1^{\text{mle}} \tau_1, \ldots, \hat{\theta}_d^{\text{mle}} \tau_d), \quad \text{subject to} \quad \sum_{j=1}^{d} \tau_j \leq c \tag{2}$$

(if τ_j are forced to be non-negative this is called the non-negative garrote). When $c < d$ some or all of the components of $\hat{\theta}^{\text{mle}}$ are reduced in absolute value toward 0. The ridge estimator is

$$\hat{\theta}^{\text{ridge}} = \arg\max_{\theta} \ell(\theta) \quad \text{subject to} \quad \|\theta\|_2^2 \leq c \tag{3}$$

where $\|v\|_p = (\sum_{j=1}^{d} |v_j|^p)^{1/p}$. Forming the Lagrangian $\mathcal{L}(\theta, \lambda) = \ell(\theta) - \lambda(\|\theta\|_2^2 - c)$ and maximizing with respect to θ by setting $\nabla_\theta \mathcal{L} = 0$ we see that (3) is equivalent to

$$\hat{\theta}^{\text{ridge}} = \arg\max_{\theta} \ell(\theta) - \lambda \|\theta\|_2^2, \qquad \lambda \geq 0 \tag{4}$$

where $\lambda(c)$ may be found by setting $\nabla_\lambda \mathcal{L} = 0$. However since c in (3) is usually chosen arbitrarily in most cases (4) is solved for several different values of λ and the best performing estimator is chosen based on cross validation (there is no need to find what λ corresopnds to c is c is chosen arbitrarily and has no special

meaning). Closely related to ridge is the lasso estimator where $\|\theta\|_2^2$ is replaced with $\|\theta\|_1 = \sum |\theta_j|$

$$\hat{\theta}^{\text{lasso}} = \arg\max_\theta \ell(\theta) \quad \text{subject to} \quad \|\theta\|_1 \leq c \tag{5}$$

$$= \arg\max_\theta \ell(\theta) - \lambda\|\theta\|_1, \qquad \lambda \geq 0. \tag{6}$$

We note that the l_1 penalty of the lasso encourages sparse solutions i.e., for a given λ some of the components of $\hat{\theta}^{\text{lasso}}$ will be zero (more so as λ increases). The l_2 penalty of ridge does not encourage sparsity, that is the coefficients of $\hat{\theta}^{\text{ridge}}$ will be close to zero (perhaps very close) but not precisely zero. Since a sparse estimator has computational and interpretability advantages lasso is often preferred over ridge. The elastic net estimator combines both norms

$$\hat{\theta}^{\text{elnet}} = \arg\max_\theta \ell(\theta) \quad \text{subject to} \quad \|\theta\|_1 \leq c_1, \|\theta\|_2^2 \leq c_2 \tag{7}$$

$$= \arg\max_\theta \ell(\theta) - \lambda_1\|\theta\|_1 - \lambda_2\|\theta\|_2^2, \qquad \lambda \geq 0. \tag{8}$$

The fused lasso is useful when it is known apriori that neighboring dimensions should be similar (for example when $X_1, \ldots, X_d$ represent sequential or temporal measurements)

$$\hat{\theta}^{\text{fus}} = \arg\max_\theta \ell(\theta) \quad \text{subject to} \quad \|\theta\|_1 \leq c_1, \sum_i \|\theta_i - \theta_{i+1}\|_1 \leq c_2 \tag{9}$$

$$= \arg\max_\theta \ell(\theta) - \lambda_1\|\theta\|_1 - \lambda_2 \sum_i \|\theta_i - \theta_{i+1}\|_1, \qquad \lambda \geq 0. \tag{10}$$

Linear Regression

Although the above methods apply generally to maximum likelihood problems, the linear regression setting enables derivation of closed forms which provides insight into the differences between different regularizers.

Assuming that $\mathbf{X}$ is a matrix whose rows are the training data $X^{(1)}, \ldots, X^{(n)}$ and $\mathbf{Y}$ is a vector whose entries are the training labels $Y^{(1)}, \ldots, Y^{(n)}$, the conditional loglikelihood is $-(\mathbf{Y} - \mathbf{X}\theta)^\top(\mathbf{Y} - \mathbf{X}\theta) + c$ (see note on linear regression). Setting its gradient to zero gives the vector equation $-\mathbf{X}^\top\mathbf{X}\theta + \mathbf{X}^\top\mathbf{Y} = 0$ which imply $\hat{\theta}^{\text{mle}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. Additional insight can be gained by examining the special case of orthonormal training examples $\mathbf{X}^\top\mathbf{X} = I$: $\hat{\theta}_j^{\text{mle}} = \mathbf{X}^\top\mathbf{Y}$ (the orthogonal projection of the columns of $\mathbf{X}$ on $\mathbf{Y}$).

In the case of ridge, setting the penlized loglikelihood gradient to zero gives the vector equation $-\mathbf{X}^\top\mathbf{X}\theta + \mathbf{X}^\top\mathbf{Y} - \lambda\theta = 0 \Rightarrow \hat{\theta}^{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{Y}$. Intuitively, adding a small value to the diagonal of $\mathbf{X}^\top\mathbf{X}$ makes it non-singular and "larger", and consequentially $\hat{\theta}^{\text{ridge}}$ is "smaller" than $\hat{\theta}^{\text{mle}}$. In the orthonormal case $\hat{\theta}^{\text{ridge}} = \frac{1}{1+\lambda}\hat{\theta}^{\text{mle}}$ and each component is shrunk by the same constant factor $\hat{\theta}^{\text{ridge}} = \hat{\theta}^{\text{JS}}$.

In the case of lasso regression we have $0 = \nabla(\ell(\theta) - \lambda\|\theta\|_1) \Rightarrow -\mathbf{X}^\top\mathbf{X}\theta + \mathbf{X}^\top\mathbf{Y} - \lambda s(\theta) = 0$ where $s_j(\theta) = \text{sign}(\theta_j)$. A closed form solution is not possible since the gradient changes with the sign of the parameter vector θ. However, in the orthonormal case we do get a closed form as follows. The penalized loglikelihood is $-(\mathbf{Y} - \mathbf{X}\theta)^\top(\mathbf{Y} - \mathbf{X}\theta)/2 - \lambda\|\theta\|_1 = (\mathbf{Y}^\top\mathbf{X}\theta + \theta^\top\mathbf{X}^\top\mathbf{Y} - \|\theta\|_2^2)/2 - \lambda\|\theta\|_1 = \theta^\top\hat{\theta}^{\text{mle}} - \|\theta\|_2^2/2 - \lambda\|\theta\|_1$ which decomposes into a sum of d maximization problems $\theta_j\hat{\theta}_j^{\text{mle}} - \theta_j^2/2 - \lambda|\theta|_j$ that can be solved independently. For each j, if $|\hat{\theta}_j^{\text{mle}}| < \lambda_j$ the objective function will be negative unless $\theta_j = 0$ in which case the objective function is 0. Therefore if $|\hat{\theta}_j^{\text{mle}}| < \lambda_j$, $\hat{\theta}_j^{\text{lasso}} = 0$. If $|\hat{\theta}_j^{\text{mle}}| > \lambda_j$, having $\hat{\theta}_j^{\text{lasso}} = 0$ results in a zero objective function which may be improved upon by a non-negative solution (see later). We thus determine that if $|\hat{\theta}_j^{\text{mle}}| > \lambda_j$, $\hat{\theta}_j^{\text{lasso}} \neq 0$ in which case the objective function is differentiable and we can proceed with setting its derivative to 0: $0 = \hat{\theta}_j^{\text{mle}} - \theta_j - \lambda\text{sign}(\theta_j)$ to obtain $\hat{\theta}_j^{\text{lasso}} = \text{sign}(\hat{\theta}_j^{\text{mle}})(|\hat{\theta}_j^{\text{mle}}| - \lambda)$. Combining the two cases we get that in the orthonormal case

$$\hat{\theta}_j^{\text{lasso}} = \text{sign}(\hat{\theta}_j^{\text{mle}})(|\hat{\theta}_j^{\text{mle}}| - \lambda)_+, \qquad \text{where} \qquad A_+ = \max(A, 0).$$

We see immediately the sparsity encouraging tresholding nature of lasso as it zeros out small coefficients (as opposed to ridge which will make them smaller but non-zero).

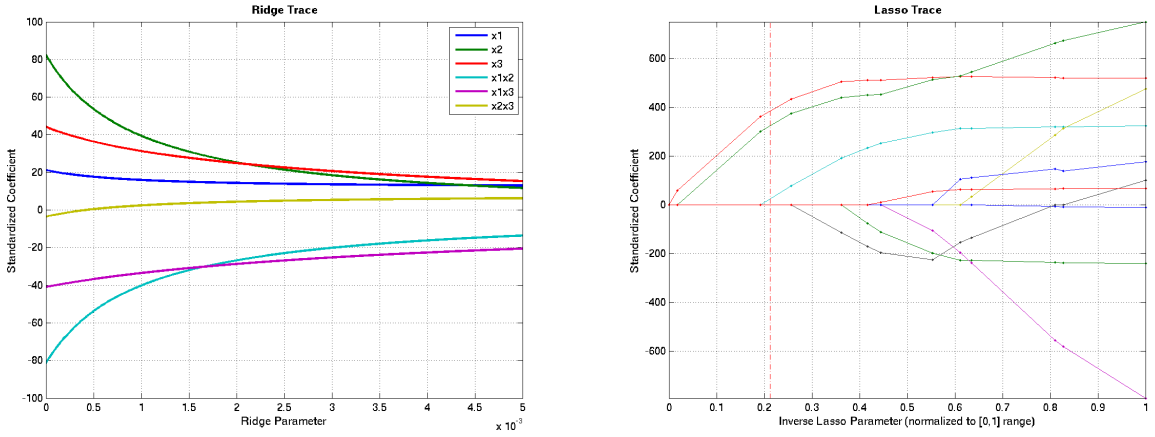

Figure 1: Left: A trace of the ridge regression coefficients as a function of the regularization parameter λ for the Matlab acetylene data. As the regularization parameter increases the parameter coefficients are shrunk towards zero (but the parameter is vector is never sparse). Right: A trace of the lasso regression coefficients as a function of the inverse regularization parameter $c\lambda^{-1}$ (c is chosen such that the range is $[0, 1]$) for the diabetes data. Note that as λ increases the parameter coefficients are shrunk towards zero with many components being identically zero.

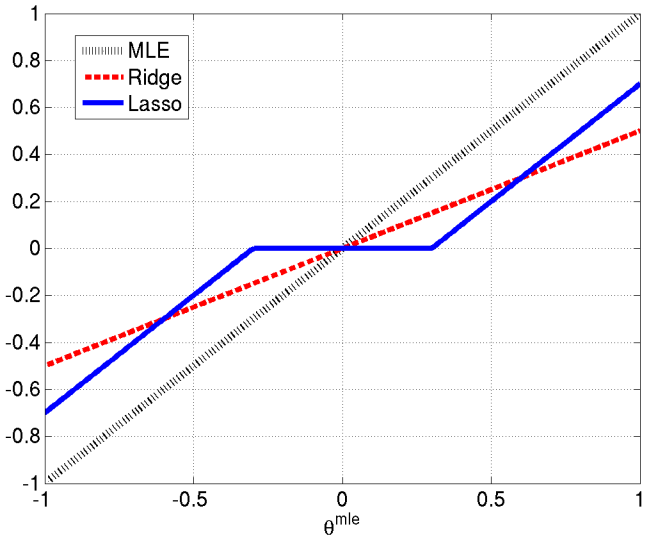

Figure 2: A comparison between the MLE, ridge ($\lambda = 1$), and lasso ($\lambda = 0.3$) for linear regression in the orthonormal case (where each dimension can be optimized separately and closed form expressions are available for all three regularization methods. The soft thresholding nature of the lasso encourages sparsity.