1 Shrinkage Methods

Shrinkage methods are continuous method for retaining a subset of the predictors and discarding the rest. The shrinkage methods don't suffer much from high variability.

1.1 Ridge Regression

The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{1}$$

An equivalent way to write the ridge problem is

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ &\text{subject to} \sum_{j=1}^{p} \beta_j^2 \le t, \end{aligned} \tag{2}$$

which makes explicit the size constraint on the parameters. To find the solution of the ridge regression, we rewrite the equality in (1) in matrix form,

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta. \tag{3}$$

Taking the gradient of (3), we are able to see that the ridge regression solutions are

$$\hat{\beta}^{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{4}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix.

Ridge Regression as a Posterior Distribution Suppose $y_i \sim \mathcal{N}\left(\beta_0 + x_i^\top \beta, \sigma^2\right)$, and the parameters β_j are each distributed as $\mathcal{N}\left(0, \tau^2\right)$, independently of one another. Thenm, the neagative log-posterior density of β, with τ^2 and σ^2 assumed known, is equal to the expression in (3), with $\lambda = \sigma^2/\tau^2$. Thus, in other words, the ridge estimate is the mode of the posterior distribution; since the distribution is Gaussian, it is also the posterior mean.

1.2 The Lasso

The lasso is a shrinkage method like ridge, with subtle but imporant differences. The lasso estimate is defined by

$$\begin{aligned} \hat{\beta}^{\text{lasso}} &= \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \\ &\text{subject to} \sum_{j=1}^{p} |\beta_j| \le t. \end{aligned} \tag{5}$$

We can also write the lasso problem in the equivalent *Lagrangian form*

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{6}$$

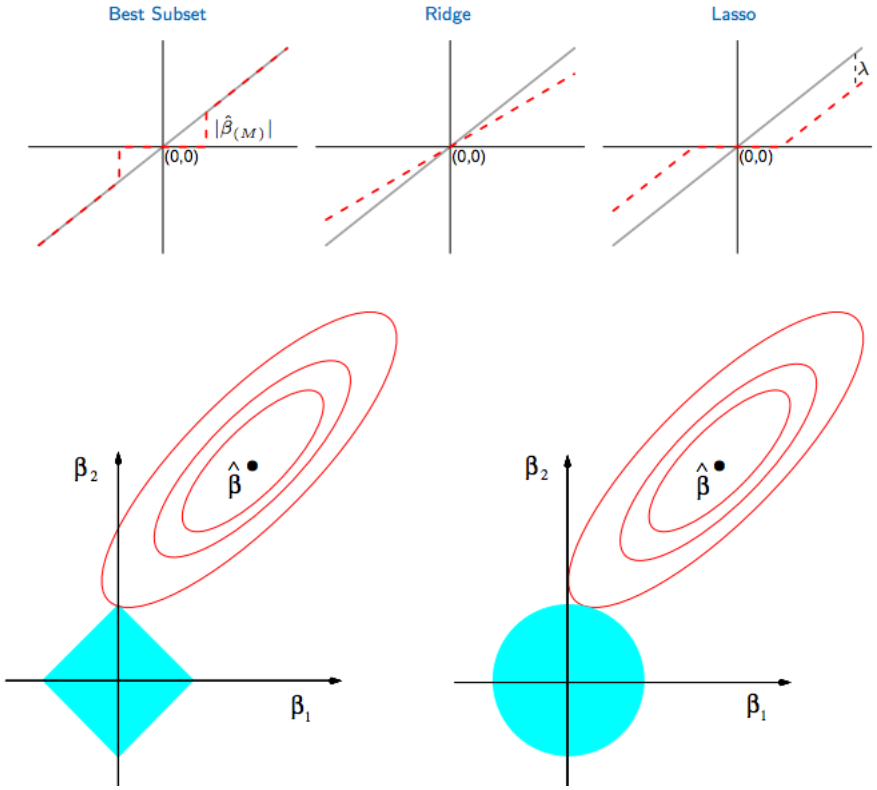
Computing the lasso solution is a quadratic programming problem. Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero. Thus, the lasso does a kind of continuous subset selection. If t is chosen larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ (where $\hat{\beta}_j = \hat{\beta}_j^{\text{lasso}}$, the least squares estimates), then the lasso estimates are the $\hat{\beta}_j$'s.

1.3 Discussion: Subset Selection, Ridge Regrssion and the Lasso

In this subsection, we will discuss and compare three approaches for restricting linear regression model. In the case of an orthonormal input matrix $\mathbf{X}$ the three procedures have explicit solutions. Ridge regression does a proportional shrinkage. Lasso translates each coefficient by a constant factor λ, truncating at zero. This is called "soft thresholding," and is used in the context of wavelet-based smoothing. Best-subset selection drops all variables with coefficient smaller than the Mth largest; this is a form of "hard thresholding."

For the nonorthogonal case,

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



From the diagram above, it suggests that lasso tends to give a sparser solution.

2 Linear Methods for Classification

In this section, we'll discuss about the classification problem using linear methods. We require some monotone transformation of δ_k or $\Pr(G = k|X = x)$ be linear for the decision boundaries to be linear. For example, if htere are two cases, a popular model for the posterior probabilities is

$$\begin{aligned}\Pr(G = 1|X = x) &= \frac{\exp\left(\beta_0 + \beta^\top x\right)}{1 + \exp\left(\beta_0 + \beta^\top x\right)},\\ \Pr(G = 2|X = x) &= \frac{1}{1 + \exp\left(\beta_0 + \beta^\top x\right)}.\end{aligned} \tag{7}$$

Here the monotone transformation is the *logit* transformation: $\log[p/(1-p)]$, and in fact we see that

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^\top x. \tag{8}$$

The decision boundary is the set of points for which the log-odds are zero, and this is a hyperplane defined by $\{x|\beta_0 + \beta^\top x\}$. We provide two popular methods that result in linear log-odds or logits: linear discriminant analysis and linear logistic regrssion.

2.1 Linear Discriminant Analysis

Suppose $f_k(x)$ is the class-conditional density of X in class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^K \pi_k = 1$. A simple application of Bayes theorem gives us

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}. \tag{9}$$

Suppose that we model each class density as multivariate Gaussian:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x - \mu_k\right)^\top \mathbf{\Sigma}_k^{-1}\left(x - \mu_k\right)\right). \tag{10}$$

Linear discriminant analysis (LDA) arises in the special case when we assume that the classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ for all k. In comparing two classes k and l, it is sufficient to look at the log-ratio, and we see that

$$\begin{aligned}\log \frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}\\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left(\mu_k + \mu_l\right)^\top \mathbf{\Sigma}^{-1}\left(\mu_k - \mu_l\right) + x^\top \mathbf{\Sigma}^{-1}\left(\mu_k - \mu_l\right)\end{aligned} \tag{11}$$

3 Principal Component Analysis

Consider a dataset $\mathcal{D}$. We want to find the subspace so we can project $\mathcal{D}$ onto a lower dimensional dataset that accounts for the variance, $\tilde{\mathcal{D}}$. The process is as follows:

(a) Calculate the mean of the dataset $\mathcal{D}$.

$$\mu = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i.$$

(b) Standardize the data.

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mu.$$

(c) Compute the covariance matrix.

$$\mathbf{\Sigma} = \sum_i \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top = \frac{1}{N} \tilde{X}^\top \tilde{X}.$$

(d) Compute the eigendecomposition of $\mathbf{\Sigma}$.

$$\mathbf{\Sigma} = P \Lambda P^\top$$

where Λ is a diagonal matrix of eigenvalues sorted in a decreasing way.

(e) Project data onto the eigenvectors corresponding to the top k eigenvectors to get $\mathcal{D}$.

4 Gaussian Models

In this chapter, we discuss the multivariate Gaussian which is very popular for calculating the joint probability density function.

The pdf for an multivariate normal (MVN) in D dimensions is defined by

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}\right) = \frac{1}{\left(2\pi\right)^{D/2} \left|\mathbf{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right). \tag{12}$$

Consider the Mahalanobis distance beteween a data vector $\mathbf{x}$ and the mean vector $\boldsymbol{\mu}$. Using an eigendecomposition of $\mathbf{\Sigma}$, we could write $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ is an orthonormal matrix of eigenvalues satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. Then, we can find the inverse of covariance matrix as follows:

$$\mathbf{\Sigma}^{-1} = \left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\right)^{-1} = \mathbf{U}^{-\top}\Lambda^{-1}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top \tag{13}$$

where $\mathbf{u}_i$ is the ith column of $\mathbf{U}$, containing the ith eigenvector. Now, we can rewrite the Mahalanobis distance as

$$\begin{aligned}
\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) &= \left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \left(\sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top\right)\left(\mathbf{x}-\boldsymbol{\mu}\right) \\
&= \sum_{i=1}^{D} \frac{1}{\lambda_i}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \mathbf{u}_i\mathbf{u}_i^\top\left(\mathbf{x}-\boldsymbol{\mu}\right) \\
&= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}
\end{aligned} \tag{14}$$

where $y_i = \mathbf{u}_i^\top\left(\mathbf{x}-\boldsymbol{\mu}\right)$. Notice that the last equality of (14) is an equaltion of an ellipse in D dimension. In general, we can interpret the Mahalanobis distance as the distance corresponding to Euclidean distance in the transformed coordinate system, where we shift by $\boldsymbol{\mu}$ and rotate by $\mathbf{U}$.

4.1 Gaussian Discriminant Analysis

One application of Multivariate Normal is to define the class conditional densities in a generative classfier, i.e.,

$$\Pr\left(\mathbf{x}|y=c, \boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{\Sigma}_c\right). \tag{15}$$

Applying Bayes Rules, one could obtain

$$\Pr\left(y=c|\mathbf{x}, \boldsymbol{\theta}\right) = \frac{\Pr\left(\mathbf{x}|y=c, \boldsymbol{\theta}\right)\Pr\left(y=c|\boldsymbol{\theta}\right)}{\sum_k \Pr\left(\mathbf{x}|y=k, \boldsymbol{\theta}\right)\Pr\left(y=k|\boldsymbol{\theta}\right)}. \tag{16}$$

Then, we can classify a feature vector by following decision rule:

$$\hat{y}(\mathbf{x}) = \arg\max_c \left(\log\Pr\left(y=c|\boldsymbol{\pi}\right) + \log\Pr\left(\mathbf{x}|\boldsymbol{\theta}_c\right)\right). \tag{17}$$

4.2 Quadratic Discriminant Analysis

Plugging in the definition of Multivariate Gaussian, we can obtain the followin equality:

$$\Pr\left(y=c|\mathbf{x},\boldsymbol{\theta}\right)=\frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-1/2}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_c\right)^\top\boldsymbol{\Sigma}_c^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_c\right)\right)}{\sum_k\pi_k|2\pi\boldsymbol{\Sigma}_k|^{-1/2}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)^\top\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)\right)}. \tag{18}$$

4.3 Linear Discriminant Analysis

Consider a special case when the covariance matrices are shared across all classes, e.g. $\boldsymbol{\Sigma}_c=\boldsymbol{\Sigma}$ for all c. In this case, we can further simplify (18) to

$$\begin{aligned}\Pr(y=c|\mathbf{x},\boldsymbol{\theta})&=\frac{\pi_c\exp\left(\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right)}{\sum_k\pi_k\exp\left(\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right)}\\&=\frac{\exp\left(\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c+\log\pi_c\right)}{\sum_k\exp\left(\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\log\pi_k\right)}.\end{aligned} \tag{19}$$

If we define $\gamma_c=-\frac{1}{2}\boldsymbol{\mu}_c^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c+\log\pi_c$ and $\boldsymbol{\beta}_c=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c$, we can further simplify the above equality to

$$\Pr\left(y=c|\mathbf{x},\boldsymbol{\theta}\right)=\frac{\exp\left(\boldsymbol{\beta}_c^\top\mathbf{x}+\gamma_c\right)}{\sum_k\exp\left(\boldsymbol{\beta}_k^\top\mathbf{x}+\gamma_k\right)}. \tag{20}$$

Note that we can further regularize the covariance matrices as follows:

$$\hat{\boldsymbol{\Sigma}}=\lambda\text{diag}\left(\hat{\boldsymbol{\Sigma}}_{\text{MLE}}\right)+\left(1-\lambda\right)\hat{\boldsymbol{\Sigma}}_{\text{MLE}} \tag{21}$$

where λ controls the amount of regularization.

4.4 Inference in jointly Gaussian distributions

Suppose $\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2)$ is jointly Gaussian with parameters $\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\\boldsymbol{\mu}_2\end{bmatrix}$, $\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11}&\boldsymbol{\Sigma}_{12}\\\boldsymbol{\Sigma}_{21}&\boldsymbol{\Sigma}_{22}\end{bmatrix}$, and $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}\boldsymbol{\Lambda}_{11}&\boldsymbol{\Lambda}_{12}\\\boldsymbol{\Lambda}_{21}&\boldsymbol{\Lambda}_{22}\end{bmatrix}$. Then, the marginal probabilities are given by

$$\begin{aligned}p(\mathbf{x}_1)&=\mathcal{N}\left(\mathbf{x}_1|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}\right)\\p(\mathbf{x}_2)&=\mathcal{N}\left(\mathbf{x}_2|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_{22}\right)\end{aligned} \tag{22}$$

and the posterior conditional is given by

$$\begin{aligned}p\left(\mathbf{x}_1|\mathbf{x}_2\right)&=\mathcal{N}\left(\mathbf{x}_1|\boldsymbol{\mu}_{1|2},\boldsymbol{\Sigma}_{1|2}\right)\\\boldsymbol{\mu}_{1|2}&=\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2-\boldsymbol{\mu}_2\right)\\\boldsymbol{\Sigma}_{1|2}&=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{aligned} \tag{23}$$

The result of (**??**) follows from the Schur's Complement.