Large p Small N Problems

When $p > N$ and in particular when $p \gg N$ new issues arise.

- We are never able to estimate all parameters without regularization. E.g. in a regression there are p parameters but the $\mathbf{X}$-matrix only has rank N.
- Signals can drown in noise.
- Big matrices, computational challenges.

As a rule of thumb; choose simple methods over complicated methods when $p \gg N$, regularize and bet on "sparsity".

Figure 18.1

Simulation study with $Y = \sum_{i=1}^{p} \beta_j X_j + \sigma\epsilon$.

Diagonal or Independence LDA

Recall that the estimated LDA classifier can be determined by

$$\delta_k(x) = \log \pi_k - \frac{1}{2}(x - \hat{\mu}_k)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_k)$$

and we classify to $\text{argmax}_k\{\delta_k(x)\}$.

If

$$\hat{\Sigma} = \text{diag}(s_1^2, \ldots, s_p^2)$$

this simplifies to

$$\delta_k(x) = -\sum_{j=1}^{p} \frac{(x_j - \bar{x}_{kj})^2}{s_j^2} + 2 \log \pi_k$$

where $x = (x_1, \ldots, x_p)^T$ and

$$\bar{x}_{kj} = \frac{1}{N_k} \sum_{i:y_i=k} x_{ij}$$

is the average of the j'th coordinate in the k'th group.

Shrunken Centroids

Note that the variance of $\bar{x}_{kj} - \bar{x}_j$ is

$$m_k^2 \sigma^2 \quad \text{with} \quad m_k^2 = \frac{1}{N_k} - \frac{1}{N}.$$

Introduce the general shrunken centroids

$$\bar{x}'_{kj} = \bar{x}_j + m_k(s_j + s_0) g\left(\frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0)}\right)$$

with s_0 a small, positive constant.

$$g_\Delta(d) = \text{sign}(d)(|d| - \Delta)_+$$

is known as soft thresholding.

$$g_\Delta(d) = d1(|d| \geq \Delta)_+$$

as hard thresholding.

Figure 18.4 – Train and Test Error

The parameter Δ is a tuning parameter for shrunken centroids. With 43 genes, $\Delta = 4.3$, we get a training error of 0 – but also a test error of 0.

Figure 18.4 – Centroid Profiles and Shrunken Centroids

Figure 18.3 – Heat Map

Elastic Net

The penalization function

$$\sum_{j=1}^{p} \alpha|\beta_j| + (1-\alpha)\beta_j^2$$

is known as the elastic net penalty.

For multinomial regression the penalized minus-log-likelihood function is

$$-\sum_{i=1}^{N} \log \Pr(Y = y_i | X = x_i) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \alpha|\beta_{kj}| + (1-\alpha)\beta_{kj}^2$$

There is an efficient implementation in the `glmnet` package for R.

Note that intercepts are not penalized and subject to the constraint that they sum to 0. All other redundancies in the parameterization are dealt with by the penalization.

Regularized Discriminant Analysis

Choosing the estimator

$$\hat{\Sigma}(\alpha) = \alpha\hat{\Sigma} + (1-\alpha)\text{diag}(\hat{\Sigma})$$

for $\alpha \in [0, 1]$ we get a regularized covariance estimator usable for LDA.

The `rda` function in the `rda` library does this in combination with nearest shrunken centroids with `regularization="R"`. With `regularization="S"` one gets

$$\hat{\Sigma}(\alpha) = \alpha\hat{\Sigma} + (1-\alpha)I_p$$

It is a little unclear which of three suggested centroid shrinkage methods from the paper Guo et al. (2006), see book, that is implemented in `rda`. ... but I have a qualified guess.

Computational Shortcuts

Suppose we consider the problem of minimizing

$$\sum_{i=1}^{N} L(y_i, \beta_0 + x_i^T \beta) + \lambda ||\beta||^2$$

over β_0 and $\beta \in \mathbb{R}^p$ with $p > N$. With $\mathbf{X} = UDV^T$ the singular value decomposition (when $p > N$, U is $N \times N$ orthogonal, D is $N \times N$ diagonal and V is $p \times N$ with $V^T V = I_N$), then $R = UD$ is an $N \times N$ matrix and

$$\mathbf{X}\beta = UDV^T \beta = RV^T \beta = R\theta$$

where $\theta = V^T \beta$ is N-dimensional. Writing $\beta = V\theta + \beta^\perp$ with $\beta^\perp$ orthogonal to the columns in V we see that

$$||\beta||^2 = ||V\theta||^2 + ||\beta^\perp||^2 \geq ||V\theta||^2 = \theta^T V^T V \theta = ||\theta||^2.$$

Since $\mathbf{X}\beta$ is unaffected by $\beta^\perp$ this term equals 0 and we need to minimize

$$\sum_{i=1}^{N} L(y_i, \theta_0 + r_i^T \theta) + \lambda ||\theta||^2, \quad \theta \in \mathbb{R}^N.$$

Support Vector Classifiers

Support vector machines are popular two class classifiers and have a reputation for being among the best performing.

With $y_i \in \{-1, 1\}$, $x_i \in E$ and $f : E \to \mathbb{R}$ we compute the predictor of y_i as $\text{sign}(f(x_i))$. With f in the reproducing kernel Hilbert space $\mathcal{H}$ estimation is done by minimization of

$$\sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda ||f||^2_{\mathcal{H}}$$

Thus the loss function $L : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$ is special and given as

$$L(y, z) = [1 - yz]_+$$