Solution proposal, exam in STK4030/STK9030, fall 2010.

Problem 1.

a) Since the volume of the p-dimensional ball with radius r is proportional to r^p, we have

$$F(z) = P(Z \leq z) = \frac{z^p}{1^p} = z^p.$$

Furthermore:

$$F_N(z) = P(Z_N \leq z) = 1 - P(Z_N > z) = 1 - P(\text{all points further from the origin than } z) = 1 - (1 - z^p)^N.$$

b) Let the median be m. Then

$$P(Z_N \leq m) = \frac{1}{2} = 1 - (1 - m^p)^N$$

$$m = (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{p}}.$$

As $p \to \infty$, $m \to 1$ for all finite N. Conclusion: For large dimension, the median distance is very close to 1, i.e., in at least 50% of all samples of size N all points will be very close to 1. This is an aspect of the curse of dimensionality.

Problem 2.

a) We assume that the order of the eigenvectors are chosen according to the size of the eigenvalues, the largest eigenvalues first. The method is called principal component repression. The predictor can be written

$$\widehat{\mathbf{y}}_{(M)} = \bar{y}\mathbf{1} + \mathbf{X}\sum_{m=1}^{M}\widehat{\theta}_m\mathbf{v}_m.$$

This shows that it is a linear method with regression coefficuet

$$\widehat{\boldsymbol{\beta}} = \sum_{m=1}^{M}\widehat{\theta}_m\mathbf{v}_m.$$

b) When $\mathbf{X}$ has full rank p, $\mathbf{X}^T\mathbf{X}$ has full rank p, and has p linearly independent eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_p$. Let $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_p)$ and $\mathbf{Z} = \mathbf{XV}$. Then $\mathbf{Z}$ has orthogonal coloumns $\mathbf{z}_i = \mathbf{X}\mathbf{v}_i$ from the properties of the eigenvectors. This gives:

$$\widehat{\mathbf{y}}_{(p)} = \bar{y}\mathbf{1} + \sum_{m=1}^{p}\mathbf{z}_m(\mathbf{z}_m^T\mathbf{z}_m)^{-1}\mathbf{z}_m^T\mathbf{y} = \bar{y}\mathbf{1} + \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$$

$$= \bar{y}\mathbf{1} + \mathbf{XV}(\mathbf{V}^T\mathbf{X}^T\mathbf{XV})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y} = \bar{y}\mathbf{1} + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \bar{y}\mathbf{1} + \mathbf{H}\mathbf{y}.$$

Problem 3.

a) Let G be the class of an object with observations X. A Bayes classifier classifies into the class with the highest value of $P(G = k|X = x)$.

Bayes rule gives

$$P(G = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)},$$

where π_l is the a priori probability for class l, and $f_l(x)$ is the probability density for observations in class l.

b) Linear discriminant analysis is based upon the assumption that X in class k is multi-normal with expectation μ_k and covariance matrix Σ common to all classes. Then

$$\log\frac{P(G=k|X=x)}{P(G=l|X=x)} = \log\frac{\pi_k f_k(x)}{\pi_l f_l(x)}$$

$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)$$

$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + x^T\Sigma^{-1}(\mu_k-\mu_l),$$

that is, linear in x. In practice, $\{\pi_k\}, \{\mu_k\}$ and Σ must be estimated from data.

Problem 4.

A kernel smoother tries to estimate a curve through a set of (x, y)-points. It is given

$$\widehat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i)y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)}.$$

[This is called the Nadaray-Watson kernel-weighted average.]

One may choose $K_\lambda(x_0, x) = D(\frac{|x_0-x|}{\lambda})$. Common choices of D are:

1) [The Epanechnikov kernel]

$$D(t) = \frac{3}{4}(1-t^2) \text{ if } |t| \le 1,\ 0 \text{ otherwise.}$$

2) [The tricube function]

$$D(t) = (1-|t|^3)^3 \text{ if } |t| \le 1,\ 0 \text{ otherwise.}$$

[3) The normal density.]