Elements of Statistical Learning
Homework Sheet 4 (Rev 2: Apr 28 2021)

Optional: you can hand in solutions to questions 1 and/or 3 for marking if you like. These questions are marked below in blue and with $*$. This is for formative assessment only and not official coursework and does not count towards your final course grade. Hand in solutions via Blackboard on Tuesday Mar 21st. Ensure your name is CLEARLY written on them. If you submit more than one page, ensure that the pages uploaded as a single document and in the right order, please!

1. (*) Suppose X is a centred and sphered $n \times p$ data matrix. Let $y = Xa$, where a is a unit norm projection vector, where $y = (y_1, \ldots, y_n)$. Show that the mean and variance of y are zero and one respectively. Suppose b is another unit norm vector and $b^T a = 0$. Show that the data $z = Xb$ is uncorrelated with y. Why is it pointless to apply principal components analysis to X?

2. Let X be a p-dimensional random vector with mean zero and identity covariance matrix. Let θ be a unit norm vector and write the projection of X onto θ as $Y_\theta = X^T\theta$. Let $f_\theta(y)$ be the density of Y_θ. Then, the $L2$ distance between $f_\theta(y)$ and $\phi(y)$ is
$$J(\theta) = \int_{-\infty}^{\infty} \{f_\theta(y) - \phi(y)\}^2 \, dy, \tag{1}$$
where $\phi(y)$ is the standard normal density. Let $H_0, H_1, \ldots$ be Hermite polynomials, orthogonal on $\mathbb{R}$ with respect to the weight function $\phi^2(x)$ and standardised by
$$\int H_j^2(y)\phi^2(y) \, dy = j!\pi^{-1/2}2^{j-1}, \tag{2}$$
and that the term of the highest degree in H_i has positive coefficient. Note: $H_0(x) = 1$.

   Show that the Hermite functions
$$h_j(y) = (j!)^{-1/2}\pi^{1/4}2^{-(j-1)/2}H_j(y)\phi(y), \quad -\infty < y < \infty, \tag{3}$$
   are orthonormal.

   The Hermite function representation of $f_a(x)$ is given by
$$f_\theta(y) = \sum_{i=0}^{\infty} a_i(\theta)h_i(y), \tag{4}$$
   where
$$a_i(\theta) = \int f_\theta(y)h_i(y) \, dy = \mathbb{E}\{h_i(Y_\theta)\}. \tag{5}$$
   Show that
$$J(\theta) = \sum_{i=0}^{\infty} a_i(\theta)^2 - (2^{1/2}/\pi^{1/4})a_0(\theta) + (2\pi^{1/2})^{-1}. \tag{6}$$

[Interesting info: $J(\theta)$ can be approximated by $\hat{J}_m(\theta) = \sum_{i=0}^{m} \hat{a}_i(\theta)^2 - (2^{1/2}/\pi^{1/4})\hat{a}_0(\theta) + (2\pi^{1/2})^{-1}$, i.e. truncating the infinite sum (which cannot be computed directly) to m terms, which can be, and $\hat{a}_i(\theta) = n^{-1}\sum_{j=1}^{n} h_i(Y_{\theta,j})$, where $Y_{\theta,j}$ is the jth individual in the projected data Y_θ. For projection pursuit, we want to find vectors θ that maximise $J(\theta)$, i.e. look for densities that are very non-normal. Let θ_1 be a local maximum of $J(\theta)$. It can be shown that if m increases sufficiently quickly as a function of n, yet more slowly than $n^{2/3}$, then there exists a $\hat{\theta}_1$, which gives at least a local maximum of $\hat{J}_m(\theta)$ and is $\sqrt{n}$-consistent for θ_1. This means that $\hat{J}_m(\theta)$ is a computable statistic that ought to find interesting projections (and it does). See Hall, P. (1989) On polynomial-based projection indices for exploratory projection pursuit. *The Annals of Statistics*, **17**, 589–605. Peter Hall was one of the most outstanding and prolific statisticians of the 20th and 21st Centuries.]

3. (*) Let Y_j be the jth component of the random p-vector Y. Show that $\sum_{j=1}^{p} H(Y_j)$ is the entropy of the 'independence version' of Y, i.e. $\prod_{j=1}^{p} g_j(y_j)$, where g_j is the marginal density of Y_j.

4. From slide 11 in Lecture 16 we assumed X is a random vector with covariance I, $Y = A^T X$, where A is orthogonal. Show that

$$I(Y) = \sum_{j=1}^{p} H(Y_j) - H(X), \tag{7}$$

   where $H(Y)$ is the entropy of random variable Y.

5. Apply independent components analysis to the `iris` data within `R`. E.g. use the `fastICA` package. ICA here should produce projections/solutions that are non-Gaussian (and assumed independent) and so should do a job similar to exploratory projection pursuit.

6. Visit the site:

   `https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises`

   and play with some of the nice interactive graphics to build your own neural networks and assess performance.

7. (This is question 11.7 from ELS) Fit a neural network to the `spam` data of Section 9.1.2, and compare the results to those for the additive model given in that chapter. Compare both the classification performance and interpretability of the final model.