1. A least squares problem can be viewed as minimization of the objective,

$$||Ax - b||^2 = \sum_{i=1}^{m} (\tilde{a}_i^T x - b_i)^2,$$

where $\tilde{a}_i^T$ are the rows of $A$ and the $n$-vector $x$ is to be chosen. A slightly generalized formulation is called the "weighted least squares problem" where we minimize the objective,

$$\sum_{i=1}^{m} w_i (\tilde{a}_i^T x - b_i)^2,$$

with $w_i$ given positive weights. The weights allow to reassign different levels of importance to different components of the residual vector.

   (a) Show that the weighted least squares objective can be expressed as $||D(Ax - b)||^2$ for an appropriate diagonal matrix $D$. This allows us to solve the weighted least squares problem as a standard least squares problem, by minimizing $||Bx - d||^2$, where $B = DA$ and $d = Db$.

   (b) Show that when $A$ has linearly independent columns, so does the matrix $B$.

   (c) The least squares approximate solution is given by $\hat{x} = (A^T A)^{-1} A^T b$. Give a similar formula for the solution of the weighted least squares problem. You might want to use the matrix $W = \text{diag}(w)$ in your formula.

2. Suppose $A$ is an $m \times n$ matrix with linearly independent columns and QR factorization $A = QR$ and $b$ is an $m$-vector. The vector $A\hat{x}$ is a linear combination of the columns of $A$ that is closest to the vector $b$. That is, it is the projection of $b$ onto the set of linear combinations of the columns of $A$.

   (a) Show that $A\hat{x} = QQ^T b$. (Note that the matrix $QQ^T$ is called the projection matrix).

   (b) Show that $||A\hat{x} - b||^2 = ||b||^2 - ||Q^T b||^2$.

3. See page 255 of [VMLS]. Now consider a quadratic (degree two) model with 2 features. I.e. $x$ is a 2-vector. This has the form,

$$\hat{f}(x) = a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2,$$

where the scalar $a$, the 2-vector $b$ and the 3-vector $c$ are the zeroth, first and second order coefficients in the model.

   (a) Put this model into the general linear in the parameters form by giving $p$ and the basis functions $f_1, \ldots, f_p$ (which map 2-vectors to scalar).

   (b) Experiment with a situation where you obtain observations from the following:

$$f(x) = (x-d)^T A(x-d) + 30\cos(10x_1)\cos(10x_2), \qquad \text{with} \qquad A = \begin{bmatrix} 8 & 3 \\ 3 & 11 \end{bmatrix}, d = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

   over the grid $x_1 = -5, -4.9, \ldots, 4.9, 5$ and $x_2 = -5, -4.9, \ldots, 4.9, 5$. Use your answer to (a) to numerically estimate $a$, $b$ and $c$.

   (c) Are you able to (approximately) reconstruct $A$ and $d$ from these observations? If so, explain/demonstrate how.

4. Determine the eigenvalues of the matrix $\mathbf{1}\mathbf{1}^T$ where $\mathbf{1}$ is the $n$-vector of ones.

5. By generating many random $n \times n$ matrices with independent uniform(0,1) entries, conjecture about the mean spectral radius. Demonstrate numerically.

6. Consider the rotation matrix,

$$Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

   (a) Calculate the eigenvalues and normalized eigenvectors.

   (b) Check that the trace equals the sum of the eigenvalues.

   (c) Check that the determinant equals the product of the eigenvalues.

7. Suppose the same $X$ diagonalizes both $A$ and $B$. They have the same eigenvectors in $A = X\Lambda_1 X^{-1}$ and $B = X\Lambda_2 X^{-1}$. Prove that $AB = BA$.

8. Consider the rank 1 matrices,

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 & -1 \\ 8 & -4 \end{bmatrix}.$$

   (a) Using hand calculations, find the SVD of $A$.

   (b) Using hand calculations, find the SVD of $B$.

   (c) Using hand calculations, find the SVD of $A + B$.

9. Consider the $n$-dimensional multivariate standard normal distribution of the random vector $X$, with density,

$$f(x) = (2\pi)^{-n/2} e^{-\frac{1}{2}x^T x}.$$

   Take now a new random variable $Y = AX + \mu$ where $A$ is a non-singular $n \times n$ matrix and $\mu$ is an $n$-vector. It can be shown that $Y$ has density,

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

   Here $\Sigma = AA^T$ is the covariance matrix and $|\Sigma|$ is its determinant.

   (a) Argue why the expression is valid (no division by 0 and existence of the inverse of $\Sigma$).

   (b) Consider the case of $n = 2$ where it is common to use,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix},$$

   with $\sigma_1, \sigma_2 > 0$ and $\rho \in (-1, 1)$. Derive an explicit expression for $f(x)$.

   (c) Set $\mu_1 = 1$, $\mu_2 = 1$, $\sigma_1 = 1.3$, $\sigma_2 = 0.8$ and $\rho = 0.7$. Plot an elegant surface plot and/or contour plot of the density on a sensible region.

   (d) For the parameters above, numerically integrate (using a crude Riemann sum) to evaluate the probability of both coordinates of $Y$ being negative.

10. Consider the MNIST digit classification problem as presented in the handout of lecture 1. Reproduce the example to create a classifier between even and odd digits. There are two possible classifiers: (I) use the classification of digits '0'-'9' and then decide if even or odd. (II) train on binary classification between evens and odds. What is the test accuracy of each of the classifiers (after training on 60,000 images) and testing on 10,000? Which is better?