Assignment 5

1. This problem will give you experience with the (sometimes counterintuitive) properties of high-dimensional spaces.
   a) Plot the average angle $\phi$ between two randomly chosen $N$-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ as a function of $N$ between 1 and 50. Assume that the elements of the vectors are chosen i.i.d. from a standard Gaussian distribution. Determine the angle using $\phi = \min(\theta', \pi - \theta')$, where $\cos \theta' = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$.
   b) Again as a function of $N$ from 1 to 50, plot the probability that a randomly chosen point in the $N$-dimensional unit cube (a vector $\mathbf{x}$ whose elements $x_i$, $i = 1 \cdots N$ are chosen i.i.d. from a uniform distribution between -1 and 1) lies within the $N$-dimensional unit sphere. A vector $\mathbf{x}$ lies within the unit sphere if $\sum x_i^2 < 1$.

2. Pick $P$ two-dimensional data points, say $P = 100$, by drawing the $x$ and $y$ coordinates from independent Gaussian distributions of standard deviation (s.d.) 1 for $x$ and 3 for $y$. Form the covariance matrix $C = \frac{1}{P} \sum \mathbf{x}_i \mathbf{x}_i^T$, where the $\mathbf{x}_i$ are the 2-dimensional data points. Find its eigenvectors and eigenvalues (you can do it by hand, since it’s a 2D matrix, or have software do it), and note which is the top principal component. How do these they relate to the distribution from which you drew the points, e.g. do they give a good description of your data? Now pick $P$ points from each of two Gaussians, one with a mean $(5, 1)^T$ and one with mean $(5, 1)^T$, both with the same s.d.’s as before. Form the covariance matrix of these data points and find its eigenvectors and eigenvalues. Now how do they relate to the underlying distribution?

3. Download the data file data.csv from lk.zi.columbia.edu/misc/data.zip and load its contents as a variable $X$ in your analysis software of choice. $X$ should be a $100 \times 1000$ matrix. Examine the data, for example by showing the elements of the matrix as a heatmap. You shouldn’t be able to discern any structure by eye (you don’t need to show anything for this part of the problem). The MATLAB function csvread('data.csv') or numpy function numpy.genfromtxt('data.csv',delimiter=',') may be helpful.
   a) Compute the SVD (svd in MATLAB or numpy.linalg.svd) of the data matrix, and plot the distribution of singular values $\sigma_i$ in order of descending magnitude. Optional: Also perform PCA and verify the eigenvalues $\lambda_i$ are related to the singular values in the way you expect. You can do this manually by computing the eigendecomposition of the covariance matrix and compare the results to built-in methods if you wish.
   b) Write code to generate a new matrix $X_{\text{shuf}}$ whose rows are shuffled versions of the rows of $X$. That is, the elements of the $i$th row of $X_{\text{shuf}}$ are a random permutation of the elements of the $i$th row of $X$. Make sure that each
row is shuffled using a different random permutation. The MATLAB function `randperm` and numpy function `numpy.random.shuffle` maybe be helpful.

c) Plot the distribution of the singular values of $X_{\text{shuf}}$ on the same axes as those of $X$. How many values stand out? **Optional:** To be more thorough, repeat the shuffling procedure 50 times, and plot the singular values or eigenvalues of the shuffled data along with error bars.

d) Plot, as a function of time, the factors/components that correspond to the values you identified as standing out in the previous part of the problem.

e) How much of the variance in the data is explained by the components you identified?

4. Optional (advanced): Computing the moments of a multi-dimensional Gaussian distribution: A Gaussian distribution of an $N$-dimensional random variable $\mathbf{s}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$ (meaning that $C_{ij} = \langle (s_i - m_i)(s_j - m_j) \rangle$) is given by

$$P(\mathbf{s}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{s} - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{s} - \mathbf{m})}$$  \hspace{1cm} (1)

To compute the moments, as for the 1D Gaussian, we’ll change to zero-mean variables, $\mathbf{x} = \mathbf{s} - \langle \mathbf{s} \rangle$, and find the moment-generating function $\langle e^{\mathbf{p}^T \mathbf{x}} \rangle$. To do this, work in the eigenvector basis. Let $\mathbf{O}$ be the matrix whose columns are the eigenvectors of $\mathbf{C}$. In the integral for computing $\langle e^{\mathbf{p}^T \mathbf{x}} \rangle$, insert $\mathbf{OO}^T = \mathbf{I}$ between the $\mathbf{p}$ and the $\mathbf{x}$ and between the $\mathbf{C}^{-1}$ and each of the $\mathbf{x}$’s alongside it. Accordingly, change variables in your integral to the variables in the eigenvector basis, e.g. to $\mathbf{y} = \mathbf{O}^T \mathbf{x}$, $\mathbf{q} = \mathbf{O}^T \mathbf{p}$ (note that the determinant of an orthogonal matrix is ±1, so the absolute value of the determinant of the Jacobian for this transformation is 1), and let $\Sigma$ be $\mathbf{O}^T \mathbf{C} \mathbf{O}$. Your problem should break up into a product of $N$ one-dimensional integrals, each giving $\langle e^{\mathbf{q_i} \mathbf{y_i}} \rangle$, $i = 1, \ldots, N$. Put these together in vector-matrix form, and, appropriately inserting $\mathbf{1} = \mathbf{O}^T \mathbf{O}$, transform back to the original basis. You should find $\langle e^{\mathbf{p}^T \mathbf{x}} \rangle = e^{\frac{1}{2} \mathbf{p}^T \mathbf{C} \mathbf{p}}$.

The moments can then be found as follows: write an arbitrary product of elements of $\mathbf{x}$ as $x_i x_j \ldots x_k$. Then $\langle x_i x_j \ldots x_k \rangle = \frac{d}{dp_i} \frac{d}{dp_j} \ldots \frac{d}{dp_k} \langle e^{\mathbf{p}^T \mathbf{x}} \rangle |_{\mathbf{p}=0} = (\frac{d}{dp_i} \frac{d}{dp_j} \ldots \frac{d}{dp_k} e^{\frac{1}{2} \mathbf{p}^T \mathbf{C} \mathbf{p}}) |_{\mathbf{p}=0}$. Use this to show that $\mathbf{C}$ is indeed the covariance of the distribution: $\langle x_i x_j \rangle = C_{ij}$.

If you can, go on to use this to prove, or at least get a feel for, Wick’s theorem. The average of a product of $N$ elements of the zero-mean variable $\mathbf{x}$ is called an $N$-point function. Wick’s theorem states that, for a Gaussian distribution, any $N$-point function for $N$ even is equal to a sum, over all distinct ways of grouping
the $N$ elements into pairs, of the products of the two-point functions of the pairs (and for $N$ odd, the $N$-point function is zero). For example: $\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle + \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle = C_{12} C_{34} + C_{13} C_{24} + C_{14} C_{23}$. 