Homework 8

Due Apr 2 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Random vector) A random vector $\tilde{x}$ with zero mean has a covariance matrix $\Sigma_{\tilde{x}}$ with the following eigendecomposition

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}. \tag{1}$$

   (a) What is the variance of each of the entries of the random vector $\tilde{x}[1]$, $\tilde{x}[2]$ and $\tilde{x}[3]$?

   - we know that the variance of the enteries of a random vector x are the diagnoal enteries of it's covariance matrix
   - so in this case as we are given an eigendecomposition of the form $Q\lambda Q^{-1} = \Sigma_{\tilde{x}}$
   - we can compute $\Sigma_{\tilde{x}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$
   - so thus we can see $var(\tilde{x}[1]) = 1$ and $var(\tilde{x}[2]) = var(\tilde{x}[3]) = \frac{1}{4}$

   (b) Is it possible to find a unit-norm vector u such that the inner product between $\tilde{x}$ and u (i.e. the amplitude of the projection of $\tilde{x}$ onto that direction) has variance greater than 1?

   - we can write the variance of any linear combination of $\tilde{x}$ and a vector $a \in \mathbb{R}^d$ as $var(a^t\tilde{x}) = a^t\Sigma_{\tilde{x}}a$
   - doing this out for an arbitrary vector $a = \begin{pmatrix} a_1 \\ 1_2 \\ 1_3 \end{pmatrix}$ we can see that $j(a) = a^t\Sigma_{\tilde{x}}a$
   - we know we are looking for the max such that a has unit norm so we can write this as a lagrangian
   - $max_{a\in\mathbb{R}^3} j(a)$ such that $||a|| = 1 \iff max(\mathcal{L}(a,\lambda)) = max_{a,\lambda} j(a) - \lambda(||a|| - 1)$

- so now to maximize we can find $\frac{\partial \mathcal{L}}{\partial a} = 2a^t\Sigma_{\tilde{x}} - 2\lambda a \Rightarrow a^t\Sigma_{\tilde{x}} = \lambda a$ so in other words our function max must be achieved when a is an eigen vector of $\Sigma_{\tilde{x}}$
- then we can find $\frac{\partial \mathcal{L}}{\partial \lambda} = ||a|| - 1 = 0 \rightarrow ||a|| = 1$ie at the max a must have unit norm
- this gives us three canaanite points (the three eigenvectors ie columns of Q)
- doing this out we can see $j(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}) = 1, j(\begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}) = .5, j(\begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}) = 0$
- thus no we can never have a linear combination $\tilde{x}$ and some vector a with variance greater than 1.

(c) Find three constants a_1, a_2 and a_3, such that at least one of them is nonzero and $\mathrm{P}(a_1\tilde{x}[1]+a_2\tilde{x}[2]+a_3\tilde{x}[3] = 0) = 1$. Justify your answer mathematically, and interpret it geometrically.

- we know a random variable with 0 variance will always equal it's mean
- we know that for any vector $a \in \mathbb{R}^3$ by the properties of gaussian random vectors $E[a^t\tilde{x}] = a^tE[\tilde{x}] = 0$
- so in other words for our linear combination $a^t\tilde{x}$ to always equal zero we just need to find any vector a such that $var(a^t\tilde{x}) = 0$
- we know that $var(a^t\tilde{x}) = a^t\Sigma_{\tilde{x}}a$ we showed in the last problem that
- $a^t\Sigma_{\tilde{x}}a = 0$ when $a = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$
- geometrically this makes sense as $a = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$ is an eigenvecotr of $\Sigma_{\tilde{x}}$ associated to eigen value 0 so in other words, it is an eigenvalue in the null space of our covariance matrix

2. (Basketball team) The coach of a basketball needs to choose two out of three players to play as guards during the last quarter of a game. The covariance matrix of the points scored by the players in a quarter is the following

	Player A	Player B	Player C
Player A	100	-80	10
Player B	-80	81	50
Player C	10	50	100

.

Compute the variance of the total number of points if the coach plays the three possible combinations of two players. Assuming all three players score the same number of points on average, which combination would you recommend to the coach if the team is winning by a lot? Which would you recommend if they are losing by a lot?

- assume that the random vector for number of points scored by the players is given by $\tilde{x} = \begin{pmatrix} \text{number of points made by player 1} \\ \text{number of points made by player 2} \\ \text{number of points made by player 3} \end{pmatrix}$ given this we can define our different combinations of players as $a_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, a_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, a_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$
- so we can compute the variance with combination i as $a_i^t \Sigma_{\tilde{x}} a_i$
- doing this out we find
  (a) $a_1^t \Sigma_{\tilde{x}} a_1 = 21$
  (b) $a_2^t \Sigma_{\tilde{x}} a_2 = 220$
  (c) $a_3^t \Sigma_{\tilde{x}} a_3 = 280$
- so given you were winning by a lot it would stand to reason you want to take a safe bet, and thus choosing a_1 that is players 1 and 2 is optimal
- if on the other hand you were badly loosing, you may want to pick a combination of players that has the potential to score a lot of points (regardless of if they could potentially mess up badly as well) and thus would chose combination a_3 that is players 2 and 3

3. (Estimating a direction) We consider a dataset of d-dimensional vectors that is modeled as samples from a random vector

$$\tilde{y} := \tilde{x}v + \tilde{z}, \tag{2}$$

where $v \in \mathbb{R}^d$, $\tilde{x} \in R$ is a random variable with mean 0 and variance σ^2_{signal}, v is a fixed deterministic unit-norm vector, and $\tilde{z} \in R^d$ is a Gaussian random vector with independent entries, each of which has mean zero and variance σ^2_{noise}. Assume that $\tilde{x}$ and $\tilde{z}$ are independent.

(a) Sketch some samples of $\tilde{y}$ for $d = 2$ when σ_{signal} is much larger than σ_{noise}. You can assume any v for the diagram.

(b) For the v you picked in part (a), sketch some samples of $\tilde{y}$ for $d = 2$ when σ_{signal} is much smaller than σ_{noise}.

(c) Is averaging the dataset a good algorithm for estimating v?

- no averaging the dataset will not get us a good estimate of v
- we can examine the expected value of our variables
- $E[v\tilde{x} + \tilde{z}] = E[v\tilde{x}] + E[\tilde{z}] = vE[\tilde{x}] + E[\tilde{z}] = v * 0 + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- which may not be v

(d) Compute the covariance matrix of $\tilde{y}$.

- we can write the variance as $var(\tilde{y}) = E[(\tilde{y} - E[\tilde{y}])^2] = E[(\tilde{x}v + \tilde{z} - E[\tilde{x}v + \tilde{z}])^2] = E[(\tilde{x}v + \tilde{z} - E[\tilde{x}]v + E[\tilde{z}])^2] = E[(\tilde{x}v + \tilde{z})^2] = E[(\tilde{x}v)^t(\tilde{x}v) + (\tilde{x}v)^t\tilde{z} + \tilde{z}^t(\tilde{x}v) + \tilde{z}^t\tilde{z}] = v^tE[\tilde{x}^t\tilde{x}]v + v^tE[\tilde{x}^t\tilde{z}] + E[\tilde{z}^t\tilde{x}]v + E[\tilde{z}^t\tilde{z}] = v^t\sigma^2_{signal}v + 2v^tcov(\tilde{x}, \tilde{z}) + \sigma^2_{noise}I = v^t\sigma^2_{signal}v + \sigma^2_{noise}I$

(e) Express the eigendecomposition of the covariance matrix in terms of σ_{signal}, σ_{noise}, v, $u_2, \ldots, u_d$. Here $u_2, \ldots, u_d$ are unit ℓ_2-norm vectors that are orthogonal to v and each other.

- given the covariance matrix $v^t\sigma^2_{signal}v + \sigma^2_{noise}I = \Sigma_{\tilde{x}}$
- to find the part that is diagonalizable given our eigenvecotrs wich we can write in matrix form as $P = \begin{pmatrix} u_1 \\ ... \\ u_n \end{pmatrix}$ we first need to think about the eigenvectors of $cov(y) = v^tv\sigma^2_{singal} + \sigma^2_{noise}I$
- we know that the eigenmatrix of $\Lambda = cov(y)$ should eb able to be written as $P\Lambda P^t = cov(y) \Rightarrow \Lambda = P^tcov(y)P$
- so doing this out we have $P^t(v^t\sigma^2_{signal}v + \sigma^2_{noise}I)P = \sigma^2_{noise}P^tv^tvP + \sigma^2_{noise}P^tIP = \sigma^2_{noise}I + \sigma^2_{signal}$
- so we can thus write $cov(y) = P(\sigma^2_{signal} + \sigma^2_{noise}I)$

(f) Suggest an algorithm to estimate the direction of v from the data.

- we showed above that $\Sigma_{\tilde{x}} = v^t\sigma^2_{signal}v + \sigma^2_{noise}I = cov(y) = P(\sigma^2_{signal} + \sigma^2_{noise}I)$

- so this tells us that we can think of the eigenvalues of our covariance matrix as representing the noise in our data
- so thus eigen value i can be expressed as $\lambda_i = \sigma^2_{signal} + \sigma^2_{noise}[i] \Rightarrow \sigma^2_{signal} = \lambda[i] - \sigma^2_{noise}[i]$
- so taking the largest eigenvalue of $cov(y)$ will give us our best estimate of v

4. (Financial data) In this exercise you will use the code in the findata folder. For the data loading code to work properly, make sure you have the pandas Python package installed on your system.

   Throughout, we will be using the data obtained by calling `load_data()` in `findata_tools.py`. This will give you the names, and closing prices for a set of 18 stocks over a period of 433 days ordered chronologically. For a fixed stock (such as msft), let $P_1, \ldots, P_{433}$ denote its sequence of closing prices ordered in time. For that stock, define the daily returns series $R_i := P_{i+1} - P_i$ for $i = 1, \ldots, 432$. Throughout we think of the daily stock returns as features, and each day (but the last) as a separate datapoint in $\mathbb{R}^{18}$. That is, we have 432 datapoints each having 18 features.

   (a) Looking at the first two principal directions of the centered data, give the two stocks with the largest coefficients (in absolute value) in each direction. Give a hypothesis why these two stocks have the largest coefficients, and confirm your hypothesis using the data. The file `findata_tools.py` has `pretty_print()` functions that can help you output your results. You are not required to include the principal directions in your submission.

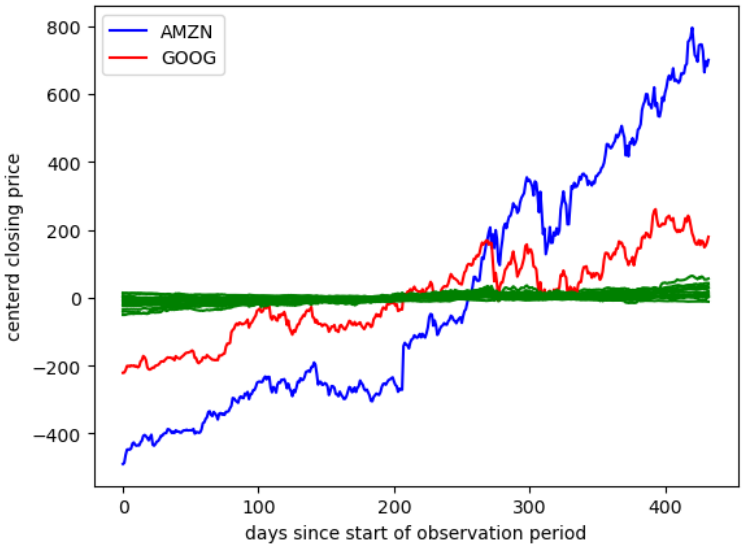
   - call out dataset X
   - so first of all we need to center our data set $ct(X) = X - m(x)$
   - then we can find the principal two directions of our data set as the eigenvectors associated to the two largest eigenvalues of our centered dataset
   - our first principal direction is

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.0017	-0.1677	-0.0881	0.0273	-0.1187	-0.6221
CVX	C	GS	JPM	AET	JNJ
-0.1469	0.3501	0.0950	-0.0074	0.1215	0.5089
DGX	SPY	XLF	SSO	SDS	USO
-0.0340	0.0194	0.0822	0.0343	-0.0513	-0.3511

   - our second principal direction is

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.2615	-0.2632	-0.2753	-0.2730	-0.1138	-0.1009
CVX	C	GS	JPM	AET	JNJ
-0.2414	-0.2276	-0.1361	-0.2734	-0.2721	-0.1669
DGX	SPY	XLF	SSO	SDS	USO
-0.1354	-0.2819	-0.2701	-0.2814	0.2802	-0.2350

   - this shows that in both case amazon and google have the largest variance



-
- in the above graph i plotted the stock price changes of all stocks in the centered dataset over the time period
- Amzn is plotted in blue and GOOG is plotted in red, all other stocks are plotted in green
- this confirms out hypothesis as these two stocks very clearly account for the majority of the variance in the dataset.

(b) Standardize the centered data so that each stock (feature) has variance 1 and compute the first 2 principal directions. This is equivalent to computing the principal directions of the correlation matrix (the previous part used the covariance matrix). Using the information in the comments of *generate_findata.py* as a guide to the stocks, give an English interpretation of the first 2 principal directions computed here. You are not required to include the principal directions in your submission.

- we can Standardize our data
- so our first principal direction is

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.0017	-0.1677	-0.0881	0.0273	-0.1187	-0.6221
CVX	C	GS	JPM	AET	JNJ
-0.1469	0.3501	0.0950	-0.0074	0.1215	0.5089
DGX	SPY	XLF	SSO	SDS	USO
-0.0340	0.0194	0.0822	0.0343	-0.0513	-0.3511

- this seems to Suggest that on the Standardized dataset C that is citigroup and JNJ (johnson and johnson) have the most variance
- our second principal direction is

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.2615	-0.2632	-0.2753	-0.2730	-0.1138	-0.1009
CVX	C	GS	JPM	AET	JNJ
-0.2414	-0.2276	-0.1361	-0.2734	-0.2721	-0.1669
DGX	SPY	XLF	SSO	SDS	USO
-0.1354	-0.2819	-0.2701	-0.2814	0.2802	-0.2350

- this seems to Suggest State Street's SPDR SP 500 ETF has the largest coefficient of negative variance on our Standardized data set
- and SDS ProShares inverse levered ETF has the largest positive coefficient of variance

(c) Assume the stock returns each day are drawn independently from a multivariate distribution $\tilde{x}$ where $\tilde{x}[i]$ corresponds to the ith stock. Assume further that you hold a portfolio with 200 shares of each of appl, amzn, msft, and goog, and 100 shares of each of the remaining 14 stocks in the dataset. Using the sample covariance matrix as an estimator for the true covariance of $\tilde{x}$, approximate the standard deviation of your 1 day portfolio returns $\tilde{y}$ (this is a measure of the risk of your portfolio). Here $\tilde{y}$ is given by

$$\tilde{y} := \sum_{i=1}^{18} \alpha[i]\tilde{x}[i],$$

where $\alpha[i]$ is the number of shares you hold of stock i.

- assume the random vector for stock return (in terms of dollars is) $\tilde{x} = \begin{pmatrix} 'aapl', \\ 'amzn', \\ 'msft', \\ 'goog', \\ 'xom', \\ 'apc', \\ 'cvx', \\ 'c', \\ 'gs', \\ 'jpm', \\ 'aet', \\ 'jnj', \\ 'dgx', \\ 'spy', \\ 'xlf', \\ 'sso', \\ 'sds', \\ 'uso', \end{pmatrix}$

- so our alpha vector is given by $\alpha = \begin{pmatrix} 200, \\ 200, \\ 200, \\ 200, \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 100 \end{pmatrix}$
  - we are told that $\tilde{y} = \Sigma_{i=1}^{n} \alpha[i]\tilde{x}[i]$
  - this is a linear combination of $\tilde{x}$ so we know that $var(\tilde{y}) = var(\Sigma_{i=1}^{18} \alpha^t \tilde{x}_i) = \Sigma_{i=1}^{18} \alpha^t \Sigma_X[i]\alpha = \alpha^t \Sigma_X \alpha = 12554439583.358356$ which is a lot of variance

(d) Assume further that $\tilde{x}$ from the previous part has a multivariate Gaussian distribution. Compute the probability of losing 1000 or more dollars in a single day. That is, compute

$$\Pr(\tilde{y} \leq -1000).$$

- we are told that $\tilde{x}$ is normally distributed, that is $\tilde{x} \sim \mathcal{N}(\mu, \Sigma_{\tilde{x}})$
- so we want to reason about $\tilde{y} = \alpha^t \tilde{x}$
- we know that $E[\tilde{y}] = E[\alpha^t \tilde{x}] = \alpha^t E[\tilde{x}] = \alpha^t \mu$
- and can see that $var(\tilde{y}) = E[(\tilde{y} - E[\tilde{y}])^2] = E[(\alpha^t \tilde{x} - \alpha^t \mu)^2] = E[(\alpha^t(\tilde{x} - \mu))((\alpha^t(\tilde{x} - \mu)))^t] = E[\alpha^t(\tilde{x} - \mu)((\tilde{x} - \mu)^t \alpha)] = \alpha^t E[(\tilde{x} - \mu)((\tilde{x} - \mu)^t)]\alpha = \alpha^t \Sigma_x \alpha$
- we can define $ct(\tilde{y}) := \tilde{y} - E[\tilde{y}]$
- we can see that $E[ct(\tilde{y})] = E[\tilde{y} - E[\tilde{y}]] = E[\tilde{y}] - E[\tilde{y}] = 0$
- and $var(ct(\tilde{y})) = E[ct(\tilde{y}) - E[(ct(\tilde{y}))]^2] = E[ct(\tilde{y})^2] = E[(\tilde{y} - E[\tilde{y}])^2] = \alpha^t \Sigma_{\tilde{x}} \alpha$
- so in other words we know $ct(\tilde{y}) \sim \mathcal{N}(0, \alpha^t \Sigma_{\tilde{x}} \alpha)$
- so thus we are dealing with gaussian random variable as opposed to a gaussian random vector
- we can see thus that $P(\tilde{y} \leq -1000) = 1 - F_{\tilde{y}}(-1000) = .5$
- so we have a 50% chance of losing more than 1000 dollars with this strategy

Note: The assumptions made in the previous parts are often invalid and can lead to inaccurate risk calculations in real financial situations.