

1. (Correlation coefficient) The entries of a two-dimensional random vector have a correlation coefficient equal to one. What is the variance of the second principal component? Provide both a proof and an intuitive justification.

Let X and Y the two components of a random vector. If $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = 1$ then $\text{Cov}(X, Y) = \sigma_X \sigma_Y$. The covariance matrix is then $\Sigma = \begin{bmatrix} \sigma_X^2 & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \\ \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix}$. $\det(\Sigma - \lambda I) = \lambda(\lambda - (\sigma_X^2 + \sigma_Y^2))$, $\lambda \in \mathbf{R} \Rightarrow$ the eigenvalues are 0 and $\sigma_X^2 + \sigma_Y^2$. The sample variance of the second principal component is the smaller eigenvalue 0. We can expect such result since there X and Y are positively correlated with a coefficient of one thus the variance of the data is completely expressed by the variance of the first principal component, there is no information on the second principal component.

2. (Not centering) To analyze what happens if we apply PCA without centering, let \tilde{x} be a d -dimensional vector with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\tilde{x}}$ equal to the identity matrix. If we compute the eigendecomposition of the matrix $E(\tilde{x}\tilde{x}^T)$ what is the value of the largest eigenvalue? What is the direction of the corresponding eigenvector?

3. (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, \dots, 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 $i = 1, \dots, 432$. 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.

The two stocks corresponding to the two principal directions of the centered data with the largest coefficients (in absolute value) in each direction are: "amzn" and "goog". It can be explained by computing the absolute return for each stock over the period of 433 days:

```

      aapl      amzn      msft      goog      xom      apc \
454.14342  3423.779352  208.844523  2799.940424  258.501684  388.899131
      cvx      c      gs      jpm      aet      jnj \
381.849154  247.140956  901.311051  277.9143  505.556593  275.823333
      dgx      spy      xlf      sso      sds      uso
257.409551  424.583189  64.368439  277.841774  276.939997  77.16

```

Figure 1: Stocks returns over a period of 433 days (output of `pretty_print`).

In term of return `goog` and `amzn` stocks returned about 4 and 3 times more than the next stock after `amzn` and `goog` stocks with the highest return: `gs`, 53 and 43 times more than the last stock in term of return among the 18 stocks: `xlf`.

amzn/gs	amzn/xlf	goog/gs	goog/xlf
3.8	53.2	3.1	43.5

So most of the variance in the data will be explained by these two stocks: `goog` and `amzn`.

- (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 think of each of the entries of the principal directions as a weighting on the corresponding stock.

- `SPY` - A security that roughly tracks the S&P 500, a weighted average of the stock prices of 500 top US companies. We have only 18 stocks and excluding the ETF

(Exchange traded products), we are left with 13 stocks. Computing the weighted average using the entries in the first principal direction, we obtain a weighted average price in the range of the SPY price reported for the same day. It is less true using the entries of the second principal direction.

```

-----
Weights coefficients for the first 13 stock prices using PD0
aapl  amzn  msft  goog  xom  apc  cvx  \
-0.184648 -0.168895 -0.227645 -0.197887 -0.288861 -0.194808 -0.223865

c  gs  jpm  aet  jnj  dgx  spy  \
-0.280147 -0.260443 -0.27294 -0.108555 -0.150783 -0.196612 -0.336632

xlf  sso  sds
-0.198364 -0.334826 0.327259
-----
Day0 prices
aapl  amzn  msft  goog  xom  apc  \
101.790649 636.98999 52.433533 741.840027 72.740799 48.801582

cvx  c  gs  jpm  aet  jnj  \
82.577927 50.149845 172.000156 61.18005 107.291946 95.826782

dgx  spy  xlf  sso  sds
68.369064 194.027725 14.100877 60.724442 82.0
-----
Day0, computed SPY weighted average prices:155.63987057982837
-----
Day0 SPY price
spy
194.027725

```

Figure 2: Weighted average of the 13 stock prices and SPY price using PD0.

```

-----
Weights coefficients for the first 13 stock prices using PD1
aapl  amzn  msft  goog  xom  apc  cvx  \
0.285821 0.3302 0.382005 0.389346 -0.293444 -0.329116 -0.308272

c  gs  jpm  aet  jnj  dgx  spy  \
-0.133288 -0.137664 -0.155599 0.071286 0.108239 0.165412 0.084872

xlf  sso  sds
-0.124933 0.091568 -0.056002
-----
Day0 prices
aapl  amzn  msft  goog  xom  apc  \
101.790649 636.98999 52.433533 741.840027 72.740799 48.801582

cvx  c  gs  jpm  aet  jnj  \
82.577927 50.149845 172.000156 61.18005 107.291946 95.826782

dgx  spy  xlf  sso  sds
68.369064 194.027725 14.100877 60.724442 82.0
-----
Day0, computed SPY weighted average prices:1674.2139367259144
-----
Day0 SPY price
spy
194.027725

```

Figure 3: Weighted average of the 13 stock prices and SPY price using PD1.

- XLF - A security that tracks a weighted average of top US financial companies. We find the same weighting scheme but to a less extent as we have only 3 financial stocks among the 18 stocks.

```

-----
Weights coefficients for the first 3 financial stock prices using PD0
c  gs  jpm
-0.280147 -0.260443 -0.27294
-----
Day100 prices
c  gs  jpm
45.329781 155.994034 62.998978
-----
Day100, computed XLF weighted average price:86.68575537909774
-----
Day100 XLF return
xlf
14.347868

```

Figure 4: Weighted average of the 3 financial stock prices and XLF price using PD0.

- SSO - ProShares levered ETF that roughly corresponds to twice the daily performance of the S&P 500. We have similar entries for SSO and SPY which indicate a strong correlation between these two stocks

	SSO	SPY
First PD	-0.3348	-0.3366
Second PD	0.09156	0.0848

- SDS - ProShares inverse levered ETF that roughly corresponds to twice the negative daily performance of the S&P 500. The entries for SDS and SPY are roughly opposite confirming the opposite trend of SDS compared to SPY.

	SDS	SPY
First PD	0.3272	-0.3366
Second PD	-0.0560	0.0848

- USO - Exchange traded product that tracks the price of oil in the US Taking the mean of the entries related to oil company (xom, apc, cvx) from the principal directions and comparing to the entry for USO, they are close, confirming the correlation between uso and (xom, apc, cvx):

	USO	Mean(XOM, APC, CVX)
First PD	-0.1592	-0.2091
Second PD	-0.3709	-0.3102

- (c) Assume the stock returns each day are drawn independently from a multivariate distribution \tilde{x} where $\tilde{x}[i]$ corresponds to the i th 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 .

Using the sample covariance matrix and taking the root square of $[\text{shares}^T \times \text{covariance} \times \text{shares}]$, we find that for such portfolio the standard deviation of 1 day is: 4309.94952.

- (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

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

For each day of the 432 days, we compute the daily return $\tilde{y} := \sum_{i=1}^{18} \alpha[i] \tilde{x}[i]$ and count the number of times over the 432 days: $\Pr(\tilde{y} \leq -1000)$, then divide the result by the number of days (432), we obtain: 0.3425.

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