Homework 6

Solutions

1. (Correlation coefficient)
   If $\rho_{\tilde{a},\tilde{b}} = 1$, then $E\left(\tilde{a}\tilde{b}\right) = \sigma_{\tilde{a}}\sigma_{\tilde{b}}$, so the covariance matrix of $\tilde{a}$ and $\tilde{b}$ equals

$$\Sigma_{\tilde{a},\tilde{b}} := \begin{bmatrix} \sigma_{\tilde{a}}^2 & \sigma_{\tilde{a}}\sigma_{\tilde{b}} \\ \sigma_{\tilde{a}}\sigma_{\tilde{b}} & \sigma_{\tilde{b}}^2 \end{bmatrix}. \tag{1}$$

   For this matrix we have

$$\begin{aligned} d &:= \sqrt{a^2 + 4b^2 + c^2 - 2ac} && (2)\\ &= \sqrt{\sigma_{\tilde{a}}^4 + 4\sigma_{\tilde{a}}^2\sigma_{\tilde{b}}^2 + \sigma_{\tilde{b}}^4 - 2\sigma_{\tilde{a}}^2\sigma_{\tilde{b}}^2} && (3)\\ &= \sigma_{\tilde{a}}^2 + \sigma_{\tilde{b}}^2. && (4) \end{aligned}$$

   By the provided formula, the smallest eigenvalue equals

$$\lambda_2 = \frac{a + c - d}{2} = 0. \tag{5}$$

   This means that the variance in the direction of the corresponding eigenvector

$$\begin{aligned} \mathrm{Var}\left(\left(\frac{u_2}{||u_2||_2}\right)^T \begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix}\right) &= \left(\frac{u_2}{||u_2||_2}\right)^T \Sigma_{\tilde{a},\tilde{b}} \left(\frac{u_2}{||u_2||_2}\right) && (6)\\ &= \lambda_2 && (7) \end{aligned}$$

   equals zero. By Chebyshev's inequality this means that

$$\mathrm{P}\left(u_2^T \begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix} = 0\right) = 1. \tag{8}$$

   By (4) and the formula for u_2 we have

$$\begin{aligned} u_2^T \begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix} &= \frac{\sigma_{\tilde{a}}^2 - \sigma_{\tilde{b}}^2 - \sigma_{\tilde{a}}^2 - \sigma_{\tilde{b}}^2}{2\sigma_{\tilde{a}}\sigma_{\tilde{b}}}\tilde{a} + \tilde{b} && (9)\\ &= -\frac{\sigma_{\tilde{b}}}{\sigma_{\tilde{a}}}\tilde{a} + \tilde{b}. && (10) \end{aligned}$$

   Eq. (8) is consequently equivalent to

$$\mathrm{P}\left(\tilde{b} = \frac{\sigma_{\tilde{b}}}{\sigma_{\tilde{a}}}\tilde{a}\right) = 1. \tag{11}$$

2. (Financial data)

   (a) The first princpal direction is shown below.

AAPL	AMZN	MSFT	GOOG	XOM	APC
0.0546	0.8679	0.0367	0.4827	0.0079	0.0098
CVX	C	GS	JPM	AET	JNJ
0.0139	0.0124	0.0534	0.0207	0.0086	0.0133
DGX	SPY	XLF	SSO	SDS	USO
0.0120	0.0544	0.0050	0.0442	-0.0169	0.0015

The second principal direction is shown below.

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.0419	0.4950	-0.0268	-0.8511	-0.0280	-0.0092
CVX	C	GS	JPM	AET	JNJ
-0.0288	-0.0259	-0.1150	-0.0371	-0.0280	-0.0411
DGX	SPY	XLF	SSO	SDS	USO
-0.0113	-0.0695	-0.0091	-0.0567	0.0214	-0.0004

As can be seen above, in both vectors the largest two are amzn and goog. This is due to the fact that they have the largest variance. To see this, we show the sample standard deviations below.

AAPL	AMZN	MSFT	GOOG	XOM	APC
2.1289	19.4385	1.0525	13.2065	0.7770	1.0380
CVX	C	GS	JPM	AET	JNJ
1.2916	0.8195	3.0984	1.1657	1.7488	1.2298
DGX	SPY	XLF	SSO	SDS	USO
1.1167	1.7223	0.2518	1.3955	0.5511	0.1823

This is somewhat expected since amzn and goog have the largest share prices in the group. To see this, we show the prices from the last day in the data set below.

AAPL	AMZN	MSFT	GOOG	XOM	APC
219.2650	1944.3000	113.0815	1186.8700	83.9688	63.4848
CVX	C	GS	JPM	AET	JNJ
118.2641	73.7602	236.4426	116.8560	204.6858	141.0816
DGX	SPY	XLF	SSO	SDS	USO
106.7525	290.5603	28.6738	128.6087	32.5735	14.8000

(b) Below we give the first principal direction.

AAPL	AMZN	MSFT	GOOG	XOM	APC
0.1952	0.1913	0.2539	0.2512	0.2020	0.1510
CVX	C	GS	JPM	AET	JNJ
0.2034	0.2533	0.2631	0.2734	0.1117	0.1792
DGX	SPY	XLF	SSO	SDS	USO
0.1450	0.3275	0.2953	0.3266	-0.3240	0.1136

Next we give the second principal direction.

AAPL	AMZN	MSFT	GOOG	XOM	APC
-0.1949	-0.2193	-0.2164	-0.1860	0.3803	0.4622
CVX	C	GS	JPM	AET	JNJ
0.4167	0.0270	0.0202	0.0086	-0.0830	-0.0700
DGX	SPY	XLF	SSO	SDS	USO
-0.1958	-0.0656	-0.0083	-0.0644	0.0745	0.4850

The first principal direction is an average of all the stocks (except sds). This represents the trend in the market (as a whole). The reason that sds has a negative coefficient is that it is designed to move in the opposite direction of the market. The second principal component appears to group financial and oil stocks together, and computes the difference in their returns against the technology and health care stocks. Note that the PCA algorithm did not know about the meanings of the stocks, so these relationships were extracted from the data.

(c) The portfolio standard deviation is computed by

$$\sqrt{\alpha^T \Sigma \alpha} \approx 6962.07$$

(d) Since $\tilde{y}$ is normally distributed with mean 879.782454 and standard deviation 6962.07 we obtain

$$\Pr(\tilde{y} \leq -1000) = 0.3936$$

This isn't as startling as it may appear, since the value of our portfolio (as of the last day in the dataset) is about 856755 dollars.

3. (Streaks)

(a) To compute the pmf we consider the $2^5 = 32$ possible sequences:

00000 00001 00010 00011 00100 00101 00110 00111 01000 01001 01010 01011 01100 01101 01110 01111 10000 10001 10010 10011 10100 10101 10110 10111 11000 11001 11010 11011 11100 11101 11110 11111

The pmf is

$$p_{\tilde{s}}(0) = \frac{1}{32}, \qquad p_{\tilde{s}}(1) = \frac{12}{32}, \qquad p_{\tilde{s}}(2) = \frac{11}{32}, \tag{12}$$

$$p_{\tilde{s}}(3) = \frac{5}{32}, \qquad p_{\tilde{s}}(4) = \frac{2}{32}, \qquad p_{\tilde{s}}(5) = \frac{1}{32}. \tag{13}$$

(b) The code is

```
def pmf_longest_streak(n, tries):
    pmf_longest = np.zeros(n+1)
    for i in range(tries):
        current_streak = 0
        longest_streak = 0
        for j in range(n):
            if np.random.rand() > 0.5:
                current_streak = current_streak + 1
```

```
            else:
                if current_streak > longest_streak:
                    longest_streak = current_streak
                current_streak = 0
        if current_streak > longest_streak:
            longest_streak = current_streak
        pmf_longest[longest_streak] = pmf_longest[longest_streak] + 1./tries
    return pmf_longest
```

The images are shown in Figure 1.

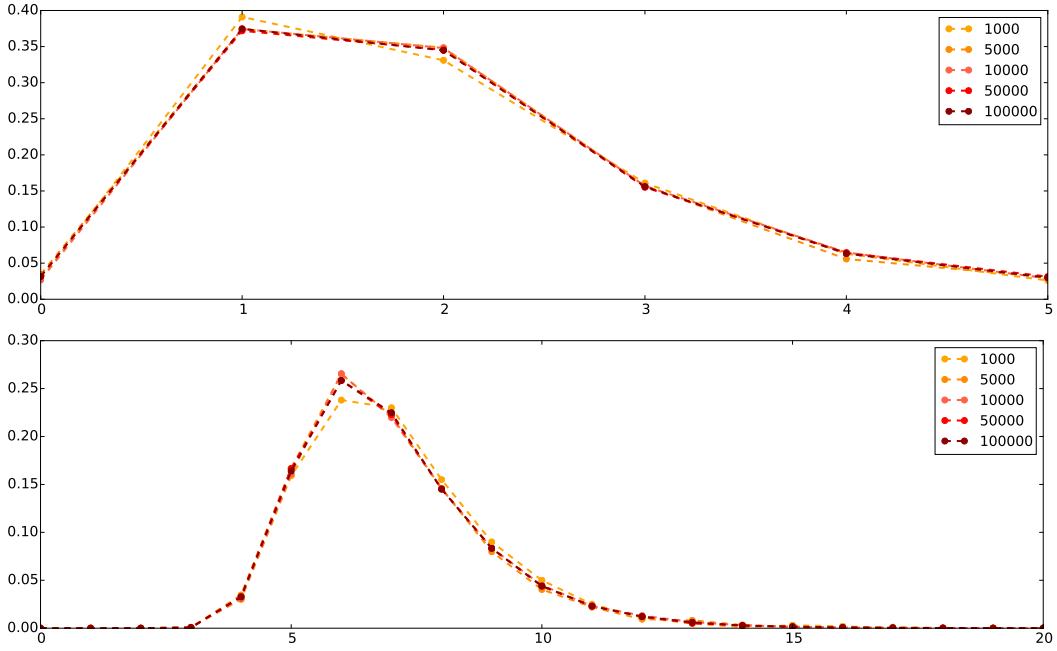

Figure 1: Pmf of $\tilde{s}$ for sequences of length 5 (above) and 200 (below).

(c) The probability is 0.319. It is therefore not unlikely to find a streak of 8 or more ones in a sequence of 200 iid Bernoulli random variables, so it is very possible that the random generator is fine.

4. (Radioactive particle)

(a) The moving average is plotted in Figure 2. It converges to 3.5. If we assume that the expected value of $\tilde{x}$ is finite then it is equal to zero, since $f_{\tilde{x}}$ is an even function and consequently $x f_{\tilde{x}}(x)$ is odd so that

$$\mathrm{E}(\tilde{x}) = \int x f_{\tilde{x}}(x) \mathrm{d}x = 0.$$

By the Law of Large Numbers we should estimate c to equal 3.5 since $\tilde{a} \to c$ in mean square and in probability. This works under the assumption that the $\tilde{x}$ has finite mean and variance.

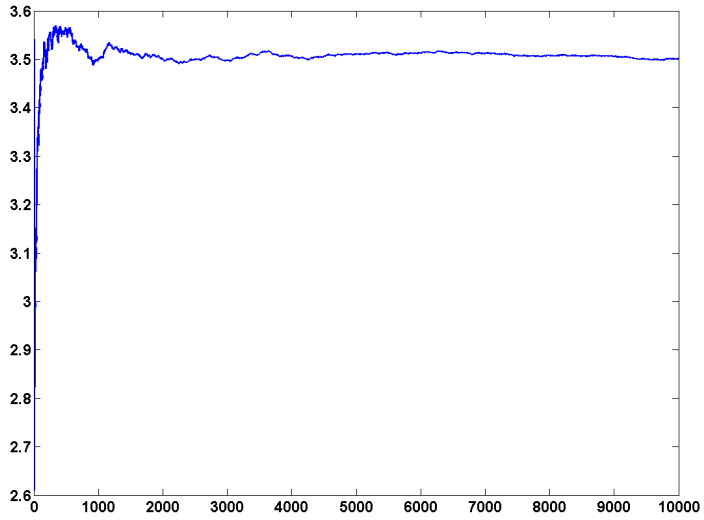

Figure 2: Sample running average of the data in *radioactive_sample_1.txt*.

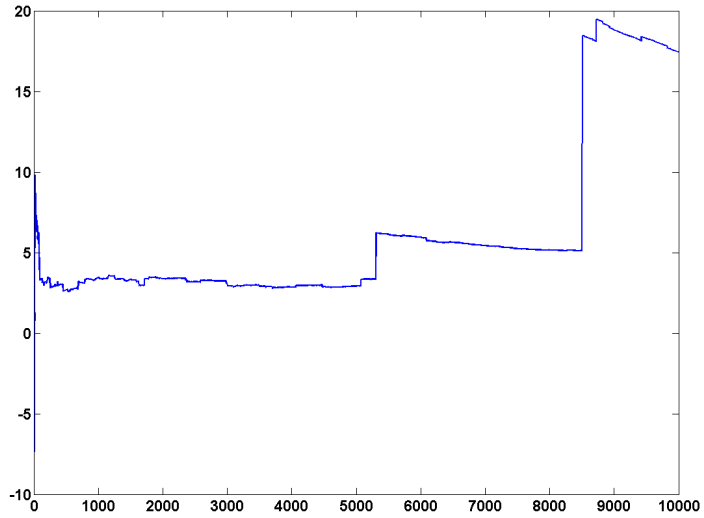

Figure 3: Moving average of the data in *radioactive_sample_2.txt*.

(b) The moving average is plotted in Figure 3. It does not seem to converge to any value, so the method does not work.

(c) The cdf of $\tilde{x}$ is equal to

$$\begin{aligned} F_{\tilde{x}(x)} &= \mathrm{P}(\tilde{x} \leq x) \\ &= \mathrm{P}(\tan \tilde{a} \leq x) \\ &= \mathrm{P}(\tilde{a} \leq \arctan x) \quad \text{by monotonicity of the tangent between } -\pi/2 \text{ and } \pi/2 \\ &= \frac{1}{\pi} \int_{-\pi/2}^{\arctan x} \mathrm{d}a \\ &= \frac{1}{2} + \frac{\arctan x}{\pi}, \end{aligned}$$

so the pdf is equal to

$$f_{\tilde{x}(x)} = \frac{1}{\pi(1+x^2)}.$$

$\tilde{x}$ is a Cauchy random value. As we saw in the notes, $\mathrm{E}(\tilde{x})$ does not exist, as it is the difference of two limits that tend to infinity. The condition in (a) does not hold for this distribution, which is the reason that we cannot estimate c from the sample average. As we can see in Figure 3 the probability of having samples that deviate very significantly from the mean is relatively high, so that the running average jumps up and down without converging.

(d) Using data from *radioactive_sample_2.txt*, we get sample median equal to 3.5. And, since $\tilde{x}$ is symmetric around the origin we know that $\mathrm{P}(\tilde{x} \leq 0) = \mathrm{P}(\tilde{x} \geq -0)$ and given that $\mathrm{P}(\tilde{x} \leq 0) + \mathrm{P}(\tilde{x} \geq -0) = 1$ we have $\mathrm{P}(\tilde{x} \leq 0) = \mathrm{P}(\tilde{x} \geq -0) = 0.5$. Thus using definition of median we get that $\text{median}(\tilde{x}) = 0$. Therefore, using the results from question 1b from homework 5 we have:

$$median(\tilde{m}) = c + median(\tilde{x}) \tag{14}$$

$$3.5 = c + 0 \tag{15}$$

$$c = 3.5. \tag{16}$$