Columbia University
Statistical Analysis and Time Series

IEOR-4709

A. Capponi | **Problem Set #1** | Issued: January 27, 2025
Spring 2025 | Due: **BEFORE CLASS** February 5, 2025

Note: Please put the number of hours that you spent on this homework set on top of the first page of your homework. The TA in charge of this homework is Boxuan Li. The CA in charge of this homework is Kishore Kuppusamy.

Ex. 1.

Suppose $\{X_n\}_{n\geq 1}$ is a sequence of independent and identically distributed random variables with mean μ_X and variance σ_X^2. Additionally, suppose $\{Y_n\}_{n\geq 1}$ is a sequence of independent and identically distributed random variables with mean μ_Y and variance σ_Y^2. Let $\mu_{XY} := \mathbb{E}[XY]$ and $\sigma_{XY} = Cov(X,Y)$. Consider the estimator $s_{XY} := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)(Y_i - \bar{Y})$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

- Show that s_{XY} is a consistent estimator of σ_{XY}.
- Show that s_{XY} is an unbiased estimator of σ_{XY}.

Solution

- To show that s_{XY} is a consistent estimator of σ_{XY}, we need to prove that $s_{XY} \to \sigma_{XY}$ in probability as $n \to \infty$.

$$\begin{aligned}
s_{XY} :&= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \frac{1}{n-1}\sum_{i=1}^n (X_iY_i - \bar{X}Y_i - X_i\bar{Y} + \bar{X}\bar{Y}) \\
&= \frac{n}{n-1}\frac{\sum_{i=1}^n X_iY_i}{n} - \bar{X}\frac{n}{n-1}\frac{\sum_{i=1}^n Y_i}{n} - \bar{Y}\frac{n}{n-1}\frac{\sum_{i=1}^n X_i}{n} + \frac{n}{n-1}\bar{X}\bar{Y} \\
&= \frac{n}{n-1}\frac{\sum_{i=1}^n X_iY_i}{n} - \frac{n}{n-1}\bar{X}\bar{Y} \\
&\to \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sigma_{XY}
\end{aligned} \tag{1}$$

  in probability, as $n \to \infty$ by law of large numbers.

- To show that s_{XY} is an unbiased estimator of σ_{XY}, we need to prove that $E(s_{XY}) = \sigma_{XY}$. From the above derivations, we have

$$\begin{aligned}
s_{XY} =& \frac{n}{n-1}\frac{\sum_{i=1}^n X_iY_i}{n} - \frac{n}{n-1}\bar{X}\bar{Y} \\
=& \frac{n}{n-1}\frac{\sum_{i=1}^n X_iY_i}{n} - \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] + \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] - \frac{n}{n-1}\bar{X}\bar{Y}
\end{aligned} \tag{2}$$

Thus

$$\begin{aligned}
\mathbb{E}s_{XY} =& \frac{n}{n-1}\frac{\sum_{i=1}^n \mathbb{E}X_iY_i}{n} - \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] + \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] - \frac{n}{n-1}\mathbb{E}\bar{X}\bar{Y} \\
=& \frac{n}{n-1}\big(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y\big) + \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] - \frac{n}{n-1}\mathbb{E}\frac{(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n^2} \\
=& \frac{n}{n-1}\big(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y\big) + \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] - \frac{n}{n-1}\frac{\sum_{i,j=1}^n \mathbb{E}[X_iY_j]}{n^2} \\
=& \frac{n}{n-1}\big(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y\big) + \frac{n}{n-1}\mathbb{E}[X]\mathbb{E}[Y] - \frac{n}{n-1}\frac{n\mathbb{E}[XY] + n(n-1)\mathbb{E}[X]\mathbb{E}[Y]}{n^2} \\
=& \frac{n}{n-1}\big(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y\big) - \frac{1}{n-1}\big(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y\big) = \sigma_{XY}. \qquad (3)
\end{aligned}$$

Notice the equations follows from $\mathbb{E}[X_iY_i] = \mathbb{E}[XY]$ and $\mathbb{E}[X_iY_j] = \mathbb{E}[X_i]\mathbb{E}[Y_j]$ for $i \neq j$ because of independence of observations.

Ex. 2.
Let $X_1, X_2, \ldots, X_n$ be random variables with mean μ and variance σ^2. What are the method of moment estimators of the mean μ and variance σ^2?

Solution We match the sample's first and second moments with population's first and second moments:

$$\frac{1}{n}\sum_{i=1}^n X_i = \mathbb{E}(X) = \hat{\mu}$$

$$\frac{1}{n}\sum_{i=1}^n X_i^2 = \mathbb{E}(X^2) = \hat{\mu}^2 + \hat{\sigma}^2.$$

Therefore

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n^2}(\sum_{i=1}^n X_i)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Ex. 3.
Consider a mixture distribution with two components. Let X have the density function:

$$f(x;\theta) = \frac{1}{2}f_1(x;\theta_1) + \frac{1}{2}, f_2(x;\theta_2)$$

where $f_1(x;\theta_1)$ and $f_2(x;\theta_2)$ are two different probability density functions with parameters θ_1 and θ_2, respectively. Assume

$$f_1(x;\theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\theta_1)^2}{2\sigma_1^2}\right)$$

and

$$f_2(x;\theta_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\theta_2)^2}{2\sigma_2^2}\right).$$

Assume σ_1 and σ_2 are known. Find the method of moment estimator for θ_1 and θ_2.

Solution Consider a mixture distribution

$$f(x) = \sum_i p_i f_i(x)$$

As $Var(X) = E(X^2) - (EX)^2 = m_2 - m_1^2 = (\sigma^2 + \mu^2) - \mu^2$, where $m_1 = \mu, m_2 = \sigma^2 + \mu^2$ are the first and second moments, the variance of mixture distributions is given by

$$Var(f) = \sum_i p_i(\sigma_i^2 + \mu_i^2) - \left(\sum_i p_i\mu_i\right)^2 = \sum_i p_i\sigma_i^2 + \sum_i p_i(\mu_i)^2 - \left(\sum_i p_i\mu_i\right)^2$$

By plugging in $p_1 = p_2 = \frac{1}{2}$, we have

$$Var(f) = \frac{1}{2}(\sigma_1^2 + \sigma_2^2) + \frac{1}{2}(\theta_1^2 + \theta_2^2) - \left(\frac{1}{2}(\theta_1 + \theta_2)\right)^2$$

Denote by $\bar{x}$ the sample mean and by s^2 the sample variance. The system of equations is given by:

$$\bar{x} = E(X)$$
$$s^2 + \bar{x}^2 = E(X^2) = Var(f) + (E(X))^2$$

which is equivalent to

$$\bar{x} = \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)$$
$$s^2 = \frac{1}{2}(\hat{\theta}_1^2 + \hat{\theta}_2^2) + \frac{1}{2}(\sigma_1^2 + \sigma_2^2) - \left(\frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)\right)^2$$

Simplifying the equations, we get:

$$\hat{\theta}_1 + \hat{\theta}_2 = 2\bar{x}$$
$$s^2 = \frac{1}{2}(\hat{\theta}_1^2 + \hat{\theta}_2^2) - \frac{1}{4}\bar{x}^2 + \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$$

For θ_1, the estimators are:

$$\hat{\theta}_1 = \bar{x} - \frac{\sqrt{4s^2 - 2\sigma_1^2 - 2\sigma_2^2}}{2} = \bar{x} - \sqrt{s^2 - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$$

and

$$\hat{\theta}_1 = \bar{x} + \sqrt{s^2 - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$$

Corresponding to each $\hat{\theta}_1$ solution, the estimators for $\hat{\theta}_2$ are:

$$\hat{\theta}_2 = \bar{x} + \sqrt{s^2 - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$$

and

$$\hat{\theta}_2 = \bar{x} - \sqrt{s^2 - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)}$$

These expressions provide two potential solutions for θ_1 and θ_2. The choice between these solutions would typically depend on additional information on the problem such as restrictions on the parameter ranges.

Ex. 4.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Gaussian random variables with mean μ and standard deviation σ. Do the following

- Is the sample mean $M_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ a Gaussian random variable? Compute the mean and variance of the sample mean.
- Find the value of n such that $\mathbb{P}(79 \le M_n \le 81) = 0.99$. Assume $\mu = 80$ and $\sigma = 22$.

Solution

- The sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is still Gaussian because the sum of independent Gaussian r.v's is a Gaussian r.v. This can be shown using MGF: Consider $X_1, ..X_n$ are i.i.d. Gaussian r.v. Let

$$S_n = X_1 + X_2 + \ldots + X_n$$

The MGF of S_n is:

$$\begin{aligned} M_{S_n}(t) &= \mathbb{E}\left[e^{tS_\mathrm{n}}\right] = \mathbb{E}\left[e^{t(X_1+X_2+\ldots+X_\mathrm{n})}\right] &(4)\\ &= \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\mu_1 t + l_i^2 e^2} = e^{\left(\sum_{i=1}^{n}\mu_i\right)t + \frac{1}{2}\left(\sum_{n=1}^{n}\sigma_i\right)t^2} &(5)\end{aligned}$$

Using the independence of $X_1, X_2, \ldots, X_\mathrm{n}$, we can factorize:

$$M_{S_n}(t) = \mathbb{E}\left[e^{tX_1}\right] \cdot \mathbb{E}\left[e^{tX_2}\right] \cdots \mathbb{E}\left[e^{tX_n}\right]$$

$$\mathbb{E}(M_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}n\mathbb{E}(X_1) = 80$$

$$Var(M_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}nVar(X_1) = \frac{22^2}{n} = \frac{484}{n}$$

- Since M_n is Gaussian with mean 80 and variance $\frac{484}{n}$, $Z := (M_n - 80)\frac{\sqrt{n}}{22}$ has standard normal distribution.

$$\begin{aligned} \mathbb{P}(79 \le M_n \le 81) &= 0.99\\ \mathbb{P}\Big((79-80)\frac{\sqrt{n}}{22} \le Z \le (81-80)\frac{\sqrt{n}}{22}\Big) &= 0.99\\ \mathbb{P}\Big(|Z| \le \frac{\sqrt{n}}{22}\Big) &= 0.99\end{aligned}$$

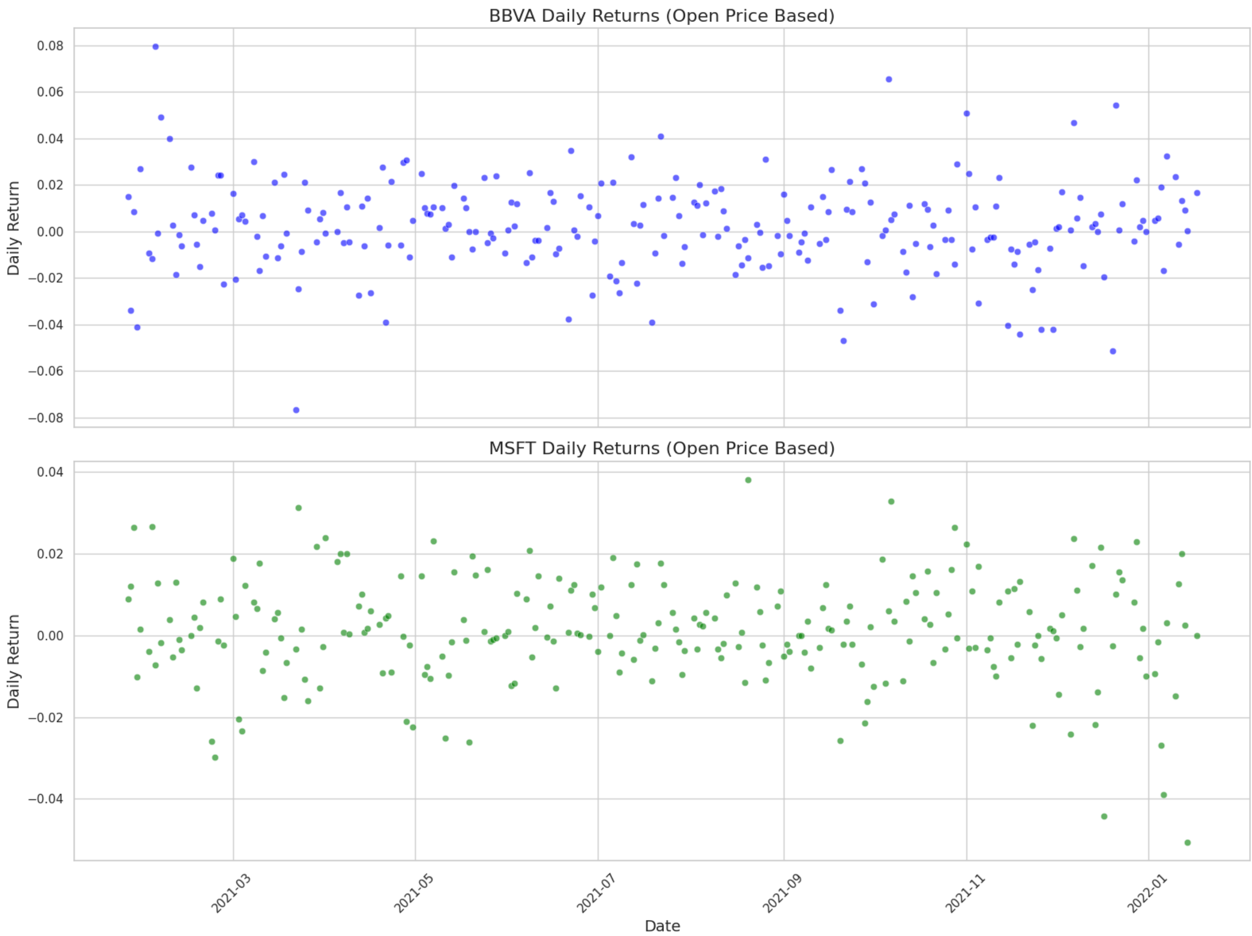
Therefore $\sqrt{n}/22 = 2.576$ and $n = 3,212$.

Ex. 5.
Go on Google or Yahoo Finance and download the Banco Bilbao Vizcaya Argentaria SA, and Microsoft time series from January 18, 2024, back to January 22, 2021. Do the following:

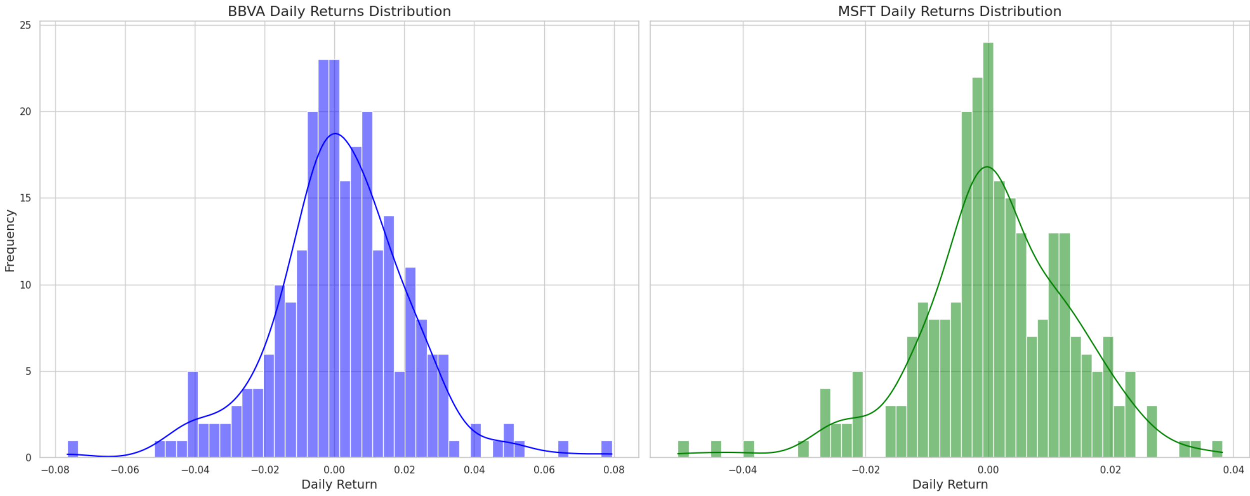
- Produce a scatterplot of each time return time series (compute the returns using open prices). What do you observe? Comment on where they are centered, and how spread out the graphs are.
- Produce a histogram of each time return time series. Does the histogram confirm your findings from the scatterplot?
- Compute the mean and variance of each return time series. Do these statistics reflect your perception from the graphs?
- Compute the covariance of the two return time series. What do you observe? Can you explain your findings using common arguments? (for instance the fact that they belong to different sectors, the fact that the two companies are located in different countries, etc...)

Solution

- Both BBVA returns and MSFT returns are centered around zero, BBVA presents a few extreme returns occurring in both sides of the distribution, but MSFT returns are more concentrated than BBVA returns.



- Yes, the histograms also show that both returns are centered around zero but BBVA returns has a heavier tail on both sides.



- The mean of BBVA returns is 0.001770, while the mean of MSFT returns is 0.001266. The variance of BBVA returns is 0.000385, and the variance of MSFT returns is 0.000161. The positive mean return of MSFT and BBVAL can be explained by the heavier positive tail of its distribution, pushing the mean up. The variance of BBVA is greater than the variance of MSFT, which again reflects the heavier tails that compensate for the more concentrated returns around zero.

- (It is important to consider common dates for both time series.) The covariance of the two time series equals 1.16e-5, and a correlation of 0.0465. Hence, the two returns series have a weak correlation. This may be explained by the fact that the two firms belong to different sectors (Finance vs Tech).