Stat 170 Spring 2023 Notes

Contents

Chapter 1

1/23/2023

Random walks, completing the square (for MGF expectations)

Chapter 2

1/25/2023

2.1 More on Random Walks and CLT

Consider the random walk that starts at 0 and at each step increases or decreases by 1, each with probability 1/2. More formally, we have $S_n = \sum_{i=1}^{n} x_i$ where each x_i is independently either 0 or 1 with probability 1/2 for $n \in [1, \cdots, N]$ for some $n \in \mathbb{Z}^+$. Note that $\mathrm{Var}(x_i) = 1$, so by CLT, for sufficiently large N, S_N approaches $\mathcal{N}(0, N)$. Thus, we can make an approximately 95% confidence interval of S_N using the two standard deviation estimation: $[-2\sqrt{N}, 2\sqrt{N}]$.

Theorem 2.1 Central Limit Theorem

If X_i are i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, then the distribution of the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

There's a bunch of nice things about this theorem: it's independent of the starting distribution of X_i, we can control the spread of the mean $\mathrm{Var}(\overline{X}) = \sigma^2/n$, the resulting distribution is normal. The normal distribution itself also has many nice properties: symmetry, maximizing entropy (among distributions with same mean and variance), etc.

2.2 Generalizing the Normal Distribution to Higher Dimensions

We can define two variables X and Y which are given by a bivariate distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \underbrace{\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}}_{\text{covariance matrix } \boldsymbol{\Sigma}} \right).$$

Since the covariance matrix is $\boldsymbol{\Sigma}$, its eigenvalues are all real. It is also positive definite – its eigenvalues are all positive. Also, $\mathrm{Cov}(X, Y) \leqslant \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$, and

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Note that correlation is unitless and is between -1 and 1 inclusive. The analog of the standard normal distribution in two dimensions is

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where the correlation $\rho \in [-1, 1]$. Note that since the variances are both 1, correlation and covariance are equal here. Note that $\rho = 0$ implies that X and Y are independent. This is not trivial – correlation of 0 does not always imply independence. Furthermore, if we generate X and Y using this bivariate distribution, then the eigenvectors of $\mathbf{\Sigma}$ are the axes of symmetry of the plot of the joint distribution of X and Y, and this is the key idea behind PCA. One other nice property is that if

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

then

$$\mathbf{\Sigma}^{1/2} \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma}\right).$$

Chapter 3

1/30/2023: Introduction to Market Making

Brian Yates – Trader at Clear Street Markets

3.1 Market Making

Market Making Model: Check out paper by Avellaneda and Stoikov: "High-frequency trading in a limit order book."

3.2 ETF Market Making

Just do sumproducts.

Chapter 4

2/1/2023

4.1 Linear Regression Review

Many packages by default do not include an intercept. Unless there's a compelling reason why you shouldn't include one, you should add it manually if necessary. Furthermore, no causalities can be drawn from a regression; it's a purely correlational thing. The generally important quantities to look at in the output of the regression:

- Betas: association between the predictor and response, holding all the other predictors constant
- p-values: the significance of the difference of the betas from 0
- R^2; the percent variance (unexplained by an intercept/mean-only model) explained by the model

A simple regression example is in CAPM, which regresses $r_i = \alpha + \beta Z_i + \varepsilon_i$, where r_i is the return of some stock on day i, and Z_i is the return of "the market" (generally some index or ETF such as SPY) on day i.

One common measure for risk-adjusted returns is the Sharpe ratio, calculated as $S_a = \mathbb{E}[R - R_F]/\sigma_R$, where R is the return of the asset, R_F is the returns of a risk-free asset, and σ_R is the standard deviation of the asset excess return.

Chapter 5

2/6/2023

Portfolio:

$$\mathbf{P}_t = \begin{pmatrix} P_{t,1} \\ P_{t,2} \\ \vdots \\ P_{t,q} \end{pmatrix}$$

and weights $\boldsymbol{\alpha}_t^\top = (\alpha_{t,1}, \alpha_{t,2}, \cdots, \alpha_{t,q})$. The initial value of the position is

$$\Pi_t = \boldsymbol{\alpha}_t^\top \mathbf{P}_t = \sum_{j=1}^{q} \alpha_{t,j} P_{t,j}.$$

A long position has $\alpha_{t,j} > 0$ while a short position has $\alpha_{t,j} < 0$. We define

$$w_{t,j} = \frac{\alpha_{t,j} P_{t,j}}{\Pi_t}.$$

Clearly, $\sum_{j=1}^{q} w_{t,j} = 1$ for all t. The total long position is

$$\text{long}_t = \sum_{j=1}^{q} w_{t,j} \mathbf{1}[w_{t,j} > 0]$$

and similarly the total short position is

$$\text{short}_t = -\sum_{j=1}^{q} w_{t,j} \mathbf{1}[w_{t,j} < 0].$$

By construction, $\text{long}_t - \text{short}_t = 1$. Assuming no change in weights, the portfolio returns are then

$$r_{t+1}^{\Pi} = \frac{\Pi_{t+1} - \Pi_t}{\Pi_t}.$$

We can verify that this equals

$$\sum_{j=1}^{q} w_{t,j} r_{t+1,j}$$

i.e. that the arithmetic returns of the portfolio is the weighted average of the returns of the individual substituents of the portfolio.

Chapter 6

2/8/2023: Bonds

6.1 Bonds

Suppose we have P_0 amount of capital and invest it at the risk free rate r^F. After one year, our capital will grow to $P_0(1+r^F)$. If we have a different risk free rate in each year, then our principal will compound to

$$P_T = P_0 \prod_{t=1}^{T} (1 + r_t^F).$$

For example, suppose that we have have an asset that will pay $Q_T = 100$ at time T. Then the price we would pay $P_0 = M_T Q_T$, where $M_T = \frac{1}{1+r^F}$ or whatever the interest rate compounds to over time. In this case, M_T is essentially the discount rate. The time T is called the time to maturity. Then the **yield** is the single interest rate that will give the same total compound interest the combined effects of the r_t^Fs:

$$(1 + \hat{i}_T)^T = \prod_{t=1}^{T} (1 + r_t^F) = \frac{1}{M_T}.$$

If we take the log of both sides, then we get

$$\begin{aligned} T \log(1 + \hat{i}_T)^T &= \sum_{t=1}^{T} \log(1 + r_t^F) \\ \hat{i}_T &= \exp\left\{ \frac{1}{T} \sum_{t=1}^{T} \log(1 + r_t^F) \right\} - 1 \\ &= \exp\left\{ \frac{1}{T} \log\left(\frac{Q_T}{P_0} \right) \right\} - 1. \end{aligned}$$

Then if we plot $\hat{i}_T$ vs. T, we get the **yield curve**.

6.2 PCA

We can transform our observed data by projecting the dataset onto the space defined by the top m PCA components, which are given by the eigenvectors of the covariance matrix of the data.

The key idea is so that each PC explains as much of the variance in the data as possible. Often times the first few principal components the vast majority of the variance in the data, and thus PCA can be a dimensionality reduction technique.

Notably with respect to bonds, we can look at the yields for the treasuries of various maturities to get a sense of what is happening in the treasuries market. The first principal component is mostly an average of the rates at all expiries, and is called the level. The second principal component is essentially the rates of the treasuries with long expiry minus the rates of those with short expiry; this check for an "inversion" of the yield curve has become an indicator of recession. This second principal component is called the slope. The third principal component roughly tells you the difference of the slope at high expiries and the slope at low expiries and is called the curvature. This third principal component explains much less of the variance than the first two PCs do. Note that this principal component analysis is totally unsupervised – we did not provide any labels for the data; the PCs simply give the vectors that explain the most variance in the data itself.

Chapter 7

2/13/2023: Martingales

7.1 Some Background

First recall Adam's Law, or the Law of Total Expectation, which states that $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$. Martingales are based on this concept in a slightly different language.

We will be thinking about conditional expectation using the framework of *information*. Define $\mathcal{F}_Y$ to be the information contained in Y. Then we can write $\mathbb{E}[X|\mathcal{F}_X] = X$; this is essentially the more formal ("mathematically correct") way to write $\mathbb{E}[X|X] = X$. Then we can rewrite Adam's Law as

$$\mathbb{E}[\mathbb{E}[X|\mathcal{F}_Y]] = \mathbb{E}[X].$$

Similar to the logic behind Adam's Law, we can also have $\mathbb{E}[\mathbb{E}[X|Y,Z]|Z] = \mathbb{E}[X|Z]$. Written in the information framework, we have

$$\mathbb{E}[\mathbb{E}[X|\mathcal{F}_Y]|\mathcal{F}_Z] = \mathbb{E}[X|\mathcal{F}_Z]$$

given that $\mathcal{F}_Y \supseteq \mathcal{F}_Z$, meaning that Y contains at least as much information as Z (for example, if $Y = X^2$, then $\mathcal{F}_X \supseteq \mathcal{F}_Y$).

Also recall Markov Chains, which have the Markov property

$$\mathbb{P}[X_{t+1}|X_t, X_{t-1}, \cdots, X_1] = \mathbb{P}[X_{t+1}|X_t].$$

7.2 Martingale Definition

Definition 7.1: Martingale

$(M_t, \mathcal{F}_t)$ is a **martingale** if

- $\mathcal{F}_t \supseteq \mathcal{F}_{t-1}$ where $\mathcal{F}_t = \text{Inf}(M_1, \cdots, M_t)$ is the information contained in the previous M_i values; this means that we have at least as much information at time t as we do at time $t-1$
- $\mathbb{E}[M_t|\mathcal{F}_{t-1}] = M_{t-1}$ for all t
- $\mathbb{E}[|M_T|] < \infty$

Concerning the last point, variance can be infinite, but infinite expected values would break everything. Also note that the key difference from Markov Chains is that Martingales make no claims about distributions and instead only care about means.

Example 7.1

Suppose that X_i are i.i.d. with $\mathbb{E}[X_i] = 0$, and define

$$\begin{aligned} M_0 &= 0 \\ M_n &= \sum_{k=1}^{n} X_k \\ \mathcal{F}_n &= \text{Inf}(X_1, \cdots, X_n). \end{aligned}$$

Then $(M_n, \mathcal{F}_n)$ is a martingale since we have

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = \mathbb{E}[X_n + M_{n-1}|\mathcal{F}_{n-1}] = \mathbb{E}[X_n|\mathcal{F}_{n-1}] + \mathbb{E}[M_{n-1}|F_{n-1}] = \mathbb{E}[X_n] + M_{n-1} = M_{n-1}.$$

Theorem 7.1

If $m < n$, then $\mathbb{E}[M_n|\mathcal{F}_m] = M_m$.

Proof. By the generalization of Adam's law from before,

$$\begin{aligned} \mathbb{E}[M_n|\mathcal{F}_{n-2}] &= \mathbb{E}[\mathbb{E}[M_n|\mathcal{F}_{n-1}]|\mathcal{F}_{n-2}] \\ &= \mathbb{E}[M_{n-1}|\mathcal{F}_{n-2}] \\ &= M_{n-2}. \end{aligned}$$

Then it's pretty clear we can formalize this iterative argument using induction. □

Corollary 7.1

$\mathbb{E}[M_n] = \mathbb{E}[M_m]$.

Proof. Take unconditional expectations on both sides of the previous theorem and use Adam's Law.

Consider another example of a security that pays off M_T at end time T. Then at times $t < T$, our expected end payoff is $P_t = \mathbb{E}[M_T|\mathcal{F}_t]$. We claim that $(P_t, \mathcal{F}_t)$ is a Martingale. The proof is by definition and generalized Adam's law:

$$P_t = \mathbb{E}[M_T|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[M_T|\mathcal{F}_{t+1}]|\mathcal{F}_t] = \mathbb{E}[P_{t+1}|\mathcal{F}_t].$$

Notably, for some Martingale $(P_t, \mathcal{F}_t)$, then the expectation of the returns is 0:

$$\begin{aligned} \mathbb{E}\left[\frac{P_{t+1} - P_t}{P_t}\middle|\mathcal{F}_t\right] &= \mathbb{E}\left[\frac{P_{t+1}}{P_t}\middle|\mathcal{F}_t\right] - 1 \\ &= \frac{1}{P_t}\mathbb{E}[P_{t+1}|\mathcal{F}_t] - 1 \\ &= \frac{1}{P_t} \cdot P_t - 1 \\ &= 0. \end{aligned}$$

Definition 7.2: Martingale Difference

Let $(Y_t, \mathcal{F}_t)$ be a Martingale. Then $\varepsilon_t = Y_t - Y_{t-1}$ is a **Martingale Difference**, sometimes denoted as $\varepsilon_t = \Delta Y_t$. This is essentially the discrete first derivative of the sequence.

We can also rewrite this definition to get

$$\begin{aligned} Y_{t+1} &= Y_t + \varepsilon_{t+1} \\ &= Y_{t-1} + \varepsilon_t + \varepsilon_{t+1} \\ &= \cdots \end{aligned}$$

The core properties of a Martingale and its differences are:

1. $\mathbb{E}[Y_{t+s}|\mathcal{F}_t] = Y_t \qquad \forall s \geqslant 0$
2. $\mathbb{E}[\varepsilon_{t+s}|\mathcal{F}_t] = 0 \qquad \forall s \geqslant 0$
3. $\text{Corr}(\varepsilon_{t+i}, \varepsilon_{t+j}|\mathcal{F}_t) = 0 \qquad i, j \geqslant 0, i \neq j$
4. $\text{Var}(Y_{t+s} - Y_t|\mathcal{F}_t) = \sum_{j=1}^{s} \text{Var}(\varepsilon_{t+j}|\mathcal{F}_t)$.

Note that conditions 2 and 3 are useful when doing a regression (probably some sort of autoregressive model predicting $Y_{t+1} = \beta Y_t + \varepsilon_{t+1}$. All four of these properties are also true unconditionally:

1. $\mathbb{E}[Y_{t+s}] = \mathbb{E}[Y_t]$
2. $\mathbb{E}[\varepsilon_{t+s}] = 0$
3. $\text{Corr}(\varepsilon_{t+i}, \varepsilon_{t+j}) = 0$.
4. $\text{Var}(Y_{t+s} - Y_t) = \sum_{j=1}^{s} \text{Var}(\varepsilon_{t+j})$.

Also to see the relationship between some of these properties, we can use property 3 to see that

$$\begin{aligned} \text{Var}(Y_{t+s} - Y_t|\mathcal{F}_t) &= \sum_{j=1}^{s}\sum_{i=1}^{s} \text{Cov}(\varepsilon_{t+j}, \varepsilon_{t+i}|\mathcal{F}_t) \\ &= \sum_{i=1}^{s} \text{Var}(\varepsilon_{t+j}|\mathcal{F}_t) + 2\sum_{i\neq j} \text{Cov}(\varepsilon_{t+i}, \varepsilon_{t+j}|\mathcal{F}_t) \\ &= \sum_{j=1}^{s} \text{Var}(\varepsilon_{t+j}|\mathcal{F}_t), \end{aligned}$$

which gives property 4.

Chapter 8

2/15/2023: Martingales Continued

8.1 Computations

We will first prove the third property of Martingales from last class:

Claim 8.1
For $i \neq j$, we have $\text{Cov}(\varepsilon_{t+j}, \varepsilon_{t+i}|\mathcal{F}_t) = 0$.

Proof: Utilizing the definition of covariance, we can write

$$\text{Cov}(\varepsilon_{t+i}, \varepsilon_{t+j}|\mathcal{F}_t) = \mathbb{E}[\varepsilon_{t+i}\varepsilon_{t+j}|\mathcal{F}_t] - \mathbb{E}[\varepsilon_{t+i}|\mathcal{F}_t]\mathbb{E}[\varepsilon_{t+j}|\mathcal{F}_t].$$

By property 2 from last class, we have $\mathbb{E}[\varepsilon_{t+i}|\mathcal{F}_t]\mathbb{E}[\varepsilon_{t+j}|\mathcal{F}_t] = 0$. Then without loss of generality assume that $i < j$. By Generalized Adam's Law, we then get

$$\mathbb{E}[\varepsilon_{t+i}\varepsilon_{t+j}|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[\varepsilon_{t+i}\varepsilon_{t+j}|\mathcal{F}_{t+i}]\mathcal{F}_t] = \mathbb{E}[\varepsilon_{t+i}\mathbb{E}[\varepsilon_{t+j}|\mathcal{F}_{t+i}]|\mathcal{F}_t].$$

Again by property 2, we get $\mathbb{E}[\varepsilon_{t+j}|\mathcal{F}_{t+i}] = 0$, which means that our original covariance is 0. □

Next, suppose we have a self-financing portfolio of stocks over time with prices $\{\mathbf{P}_t\}_{t=1}^T$ and histories $\{\mathcal{F}_t\}_{t=1}^T$, and suppose at time t we have weights $\boldsymbol{\alpha}_t$ in each of the stocks. Because this portfolio is self-financing, we do not want any cash to flow into or out of the portfolio and, assuming that we can instantaneously reweight our portfolio at each time step, this means that $\boldsymbol{\alpha}_{t-1}^\top \mathbf{P}_t = \boldsymbol{\alpha}_t^\top \mathbf{P}_t$.

Let $\Pi_t = \boldsymbol{\alpha}_t^\top \mathbf{P}_t$ be the total value of the portfolio at time t. Then the profit gained between days t and $t+1$ is

$$\begin{aligned}\Pi_{t+1} - \Pi_t &= \boldsymbol{\alpha}_t^\top \mathbf{P}_{t+1} - \boldsymbol{\alpha}_{t-1}^\top \mathbf{P}_t \\ &= \boldsymbol{\alpha}_t^\top \mathbf{P}_{t+1} - \boldsymbol{\alpha}_t^\top \mathbf{P}_t \\ &= \boldsymbol{\alpha}_t^\top (\mathbf{P}_{t+1} - \mathbf{P}_t).\end{aligned}$$

This is relatively obvious, it just says your total profit is the weighted average of the profits from your portfolio.

Claim 8.2

If $(\mathbf{P}_t, \mathcal{F}_t)$ is a Martingale, then $(\Pi_t, \mathcal{F}_t)$ is also a Martingale.

Proof: This follows fairly directly from above and from the properties of Martingale differences: we first see that the expected difference between consecutive Π_t given the relevant information is 0:

$$\mathbb{E}[\Pi_{t+1} - \Pi_t | \mathcal{F}_t] = \mathbb{E}[\boldsymbol{\alpha}_t^\top (\mathbf{P}_{t+1} - \mathbf{P}_t) | \mathcal{F}_t] = \boldsymbol{\alpha}_t^\top \mathbb{E}[\mathbf{P}_{t+1} - \mathbf{P}_t | \mathcal{F}_t] = 0.$$

This then implies that $(\Pi_t, \mathcal{F}_t)$ satisfies the key condition of being a Martingale:

$$\mathbb{E}[\Pi_{t+1} | \mathcal{F}_t] = \mathbb{E}[\Pi_t | \mathcal{F}_t] = \Pi_t.$$

□

This shows us that we can construct new Martingales from old ones; in particular, in this example we are creating the new Martingale $(\Pi_t, \mathcal{F}_t)$ from the Martingale differences of the original Martingale $(\mathbf{P}_t, \mathcal{F}_t)$. More generally, if we have Martingale $(M_t, \mathcal{F}_t)$ and random variables $\{A_t\}_{t=1}^n$ with $\mathbb{E}[A_t | \mathcal{F}_{t-1}] = A_t$ (i.e. A_t is completely determined by the information that we have from previous days $\mathcal{F}_{t-1}$), then if we define

$$Z_n = \sum_{t=1}^{n} A_t (M_t - M_{t-1}),$$

then $(Z_t, \mathcal{F}_t)$ is a Martingale. If we have continuous time periods, then this becomes the integral $\int A_t \mathrm{d}M_t$, which is the basis of stochastic calculus.

8.2 Doob's Optional Stopping Theorem

8.2.1 Definitions and Theorem

Definition 8.1: Martingale Stopping Times

The **stopping time** τ of martingale $(M_n, \mathcal{F}_n)$ is a random variable whose value is interpreted as the time at which a given stochastic process exhibits a specific behavior of interest. In particular, we always need to be able to answer whether or not time τ has already occurred by time t given information $\mathcal{F}_t$.

Example 8.1 (Stopping Times and Not Stopping Time)

Here are a few examples to emphasize what defines a stopping time and examples of random variables that do not satisfy this condition.

- If τ is the first time that our portfolio is down \$5, then it is a stopping time.
- If τ is the time at which a stock reaches its daily minimum, then it is not a stopping time, because we can never know if it happened or not without looking into the future (or until the day ends).

Theorem 8.1 Doob's Optional Stopping Theorem

If we have stopping time τ with $\mathbb{E}[\tau] < \infty$ and some bound B such that $|M_n - M_{n-1}| \leqslant B$, then $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.

8.2.2 Examples

To show how powerful this theorem is, we'll look at a few nice problems.

Question 8.1: Gambler's Ruin

Let $X_i \sim \text{Bern}(1/2)$ be a sequence of random variables and let $M_n = \sum_{i=1}^{n} X_i$ so that M_n when paired with the relevant information is a Martingale. Given some $A, B \in \mathbb{Z}^+$, what is the probability that the sequence $\{M_t\}_{t=1}^{\infty}$ reaches the value A before it reaches $-B$?

Solution

Let τ be the first time the sequence hits either A or $-B$. Note that this is a stopping time since we can always verify whether this has happened yet and since the bounding conditions holds. Then by Doob (it's straightforward to verify that the necessary conditions are satisfied), $\mathbb{E}[M_\tau] = 0$, which tells us that the probability of M_τ being A (i.e. that the sequence hits A first) is $\frac{B}{A+B}$.

Note that if our stopping condition was only the first time we hit A (and not bounding the other side), then Doob's Theorem would not hold because $\mathbb{E}[\tau]$ would not be finite.

Question 8.2: oh no chris is drunk again

Suppose we have a drunk monkey (named Chris) with a typewriter jamming away at the keys. What's the expected time (number of key presses) it will take him to type *ABRACADABRA*? (Assume the monkey only types capital letters.)

Solution

Suppose a casino takes fair bets on the letters the monkey types such that if someone bet \$1 that the next letter the monkey will type is an A, then they will get \$26 if the monkey does indeed type an A with its next letter and nothing if not. The casino will stop taking bets once the monkey has successfully typed *ABRACADABRA*.

Suppose further that at first there is one bettor and that with each key the monkey presses, another bettor shows up. Each of these bettors have the exact same behavior:

- The bettor shows up to the casino with \$1.
- The bettor leaves the casino once either they lose all their money or the
- The bettor bets all their money on the first letter that they have not yet bet on.

Since each bettor bets exactly \$1 and never has any other cash flows (since they will subsequently re-bet their entire amount on the next letter) unless the monkey has typed the magic word, the casino will have $\$\tau$ after the monkey has typed τ letters before having

to pay off any bets. Furthermore, since each of the bets is fair (i.e. has expected value 0), the amount of money M_t the casino has at each time step t (before the bettors re-bet their earnings from the previous key press) is a Martingale with respect to the information in the previous time steps, and we can easily check that the conditions of Doob's OST hold. Thus, if we have stopping time τ, then we have $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$. Since the casino starts with no money and only takes fair-valued bets, $\mathbb{E}[M_0] = 0$. Furthermore, at time τ the casino will have garnered $\$\tau$ worth of payments, so if the casino has to pay off P to the bettors at the end (once the monkey has typed *ABRACADABRA*), then

$$\mathbb{E}[M_\tau] = \mathbb{E}[\tau - P] = \mathbb{E}[\tau] - P = 0 \implies \mathbb{E}[\tau] = P.$$

Thus, the expected stopping time is exactly the amount that the casino will have to pay the bettors in the end. When the monkey has typed the magic word, the casino will have to pay exactly four bettors:

- One bettor will win $\$26^{11}$ since they started betting 11 time steps previously and won all 11 bets.
- One bettor will win $\$26^4$ since they started betting 4 time steps previously and won all 4 bets due to the fact that the first and last four letters of *ABRACADABRA* are the same.
- One bettor will win \$26 since they won the bet that the monkey would type an A, the last letter of the magic phrase.

Thus, the expected stopping time is

$$\mathbb{E}[\tau] = P = 26^{11} + 26^4 + 26.$$

Chapter 9

2/22/2023: Even More Martingales

9.1 More Martingale Computations

Recall the drunk monkey problem from the end of last class. We can solve similar problems such as the number of (fair) coin flips expected to flip the following series

1. HHH ($2^3 + 2^2 + 2^1 = 14$)
2. HTH ($2^3 + 2^1 = 10$)
3. HTT ($2^3 = 8$)

Note:

This isn't important but I thought it was interesting; feel free to skip this note.

Upon first glance, it may seem fairly unintuitive that *more* repetition in the desired sequence is associated with a *later* expected stopping time. After all, the sequence HHH has a lot of symmetry, and any individual heads toss can show up in the stopping sequence in three different places. Shouldn't this mean that the sequence HHH is "easiest" to obtain and thus have the lowest expected stopping time?

I think this logic fails because of the sequential nature with which these games are played. In particular, when we report an individual flip, that result is not in isolation; it just becomes the last result in a sequence of flips that have already been decided. Thus, if the previous flip was tails and the new flip is heads, then this heads can only be in the first position of the HHH goal sequence (if it is part of that sequence at all). Similarly, if the previous two flips were tails then heads, then flipping a heads could only put it in the second position of the goal HHH sequence (again, *if* it ends up being part of that sequence at all). Thus, the symmetry that seems like it should make sequences such as HHH easier to obtain actually does not have a helpful effect at all.

This explains why the expected stopping time for the "symmetric" sequences shouldn't be lower than those of the "assymmetric" sequences, but as we have calculated, the "symmetric" sequences stopping times are in fact *higher* – why is this the case? For this, consider again the scenario of the monkey typing letters A-Z, and suppose we want to find the expected number of key presses until it types a certain phrase. We will think about what happens if this phrase is ABCXYZ compared to if it is ABCABC.

Using the same methods as before, we find that the expected number of key presses it takes the monkey to type ABCXYZ and ABCABC are 26^6 and $26^6 + 26^3$, respectively. If we are waiting until the monkey types ABCXYZ, then at any given point, if the monkey has not recently typed a useful sequence, then the monkey has a 1/26 chance of typing an X and starting off the sequence. Once the monkey has typed an X, then there are three scenarios for the next letter:

- The monkey types another A (with probability 1/26), restarting the phrase ABCXYZ
- The monkey types a B (with probability 1/26) so that it has now typed two of the characters in the phrase
- The monkey types a different letter (with probability 24/26), essentially taking it back to the beginning in terms of progress

These three cases are the same no matter where the monkey is in typing the word. However, if we instead consider the case when we are waiting for the monkey to type ABCABC, then these three cases are the same *unless* the last three letters it has typed are ABC. In this case, the first two cases are the same, and the monkey instead has a 25/26 chance of typing a letter that destroys all its progress. Thus, since this progress-destroying probability has increased, the monkey will have to restart more often (in expectation) from this juncture, and thus the expected number of key presses it takes to type the phrase should be higher compared to the ABCXYZ case. Notably, when length of the substring repeated at the beginning and end of the key phrase is longer, the more consequential this increase in the probability of the bad scenario, and thus the greater the expected value of number of key presses.

9.2 Importance of the Amount of Information Available

Suppose we have $P_{t+1} = P_t + X_t + \varepsilon_{t+1}$ where $\mathbb{E}[X_t] = 0$, $\mathbb{E}[\varepsilon_{t+1}] = 0$, X_t is independent of P_t, and ε_{t+1} is some random noise variable independent of the other variables. Furthermore let $\mathcal{F}_t$ and $\mathcal{G}_t$ be the information generated by the sequences $\{P_k\}_{k=1}^t$ and $\{(P_k, X_k)\}_{k=1}^t$, respectively. Then, we can see that

$$\begin{aligned}\mathbb{E}[P_{t+1}|\mathcal{F}_t] &= \mathbb{E}[P_t + X_t + \varepsilon_{t+1}|\mathcal{F}_t] \\ &= P_t + \mathbb{E}[X_t|\mathcal{F}_t] + 0 \\ &= P_t.\end{aligned}$$

Thus, $(P_t, \mathcal{F}_t)$ is a Martingale. However, note that

$$\mathbb{E}[P_{t+1}|\mathcal{G}_t] = P_t + \mathbb{E}[X_t|\mathcal{G}_t] + 0 = P_t + X_t,$$

and thus $(P_t, \mathcal{G}_t)$ is not a Martingale. This is essentially insider trading – if we have access to $\mathcal{G}_t \supset \mathcal{F}_t$ which has all the information about X_t, then we can predict price movement better than anyone else.

9.3 More Properties of Stopping Times

Before proving Doob's OST, we explore another important theorem about stopping times.

Theorem 9.1 Wald's Theorem

If X_i are i.i.d. with finite expectation $\mathbb{E}[|X_i|] < \infty$ for all i, and if N is the stopping time also with $\mathbb{E}[N] < \infty$, then

$$\mathbb{E}\left[\sum_{n=1}^{N} X_n\right] = \mathbb{E}[N]\mathbb{E}[X_i].$$

This theorem is really nice because it can apply to such a wide variety of stopping times with the only limitation being that the expectation of this stopping time must be finite.

Proof: We prove this essentially by direct computation. Letting $\mathcal{F}_n$ be the information contained in $\{X_n\}_{n=1}^{N}$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N} X_n\right] &= \mathbb{E}\left[\sum_{n=1}^{\infty} X_n \cdot \mathbf{1}_{N \geqslant n}\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[X_n \cdot \mathbf{1}_{N \geqslant n} \middle| \mathcal{F}_{n-1}\right]\right] \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{N \geqslant n} \mathbb{E}[X_n | \mathcal{F}_{n-1}]\right].
\end{aligned}$$

Note that the last step is valid because given $\mathcal{F}_{n-1}$, $\mathbf{1}_{N \geqslant n}$ is simply a constant (since we know if the stopping condition has happened before n). Then, this sum becomes

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{N \geqslant n} \mathbb{E}[X_n | \mathcal{F}_{n-1}]\right] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbf{1}_{N \geqslant n}] \cdot \mathbb{E}[X_i] = \left(\sum_{n=1}^{\infty} \mathbb{P}[N \geqslant n]\right) \mathbb{E}[X_i] = \mathbb{E}[N]\mathbb{E}[X_i]$$

□

Note:

To see that $\sum_{n=1}^{\infty} \mathbb{P}[N \geqslant n] = \mathbb{E}[N]$ in the last equality of the proof above:

$$\begin{array}{rcccccccc}
\mathbb{P}[N \geqslant 1] & = & \mathbb{P}[N = 1] & + & \mathbb{P}[N = 2] & + & \mathbb{P}[N = 3] & + & \cdots \\
\mathbb{P}[N \geqslant 2] & = & & & \mathbb{P}[N = 2] & + & \mathbb{P}[N = 3] & + & \cdots \\
\mathbb{P}[N \geqslant 3] & = & & & & & \mathbb{P}[N = 3] & + & \cdots \\
+ \quad \vdots & & \vdots & & \vdots & & \vdots & & \\
\hline
\sum_{n=1}^{\infty} \mathbb{P}[N \geqslant n] & = & \mathbb{P}[N = 1] & + & 2 \cdot \mathbb{P}[N = 2] & + & 3 \cdot \mathbb{P}[N = 3] & + & \cdots \\
& = & \mathbb{E}[N] & & & & & &
\end{array}$$

Proving Doob's Theorem is pretty tricky, so we will go part of the way there by proving what is essentially a precursor to it.

Theorem 9.2 Doob's OST Precursor

For any stopping time τ, let $\tau' = \min(\tau, n) = \tau \wedge n$. Then, for the Martingale $\{M_n, \mathcal{F}_n\}$,

$$\mathbb{E}[M_{\tau \wedge n}] = \mathbb{E}[M_0].$$

Note that Doob's OST essentially comes by taking the limit as n goes to ∞. This isn't super straightforward (can't always switch limit and expectation), but it works out nicely in this case.

To prove this theorem, we will make use of the following Lemma:

Lemma 9.1

Without loss of generality, let $M_0 = 0$. Then, if

$$Z_n = \sum_{k=1}^{n} A_k(M_k - M_{k-1}),$$

then $(Z_n, \mathcal{F}_n)$ is a Martingale.

This lemma will be proved in the homework (presumably PSet 3). We now use this lemma to prove the precursor to Doob's OST.

Proof: Note that

$$\begin{aligned} M_{\tau \wedge n} &= M_\tau \cdot \mathbf{1}_{\tau \leqslant n-1} + M_n \cdot \mathbf{1}_{\tau \geqslant n} \\ &= M_0 + \sum_{k=1}^{n} A_k(M_k - M_{k-1}) \end{aligned}$$

where $A_k = \mathbf{1}_{\tau \geqslant k}$. The last equality can be seen through some simple expansion and telescoping. Then the sum in the last expression is a Martingale that initially has an unconditional expected value of 0, and thus taking unconditional expectations of both sides completes the proof. (Note that the M_0 was not on the board when Natesh wrote it, but I'm pretty sure we need it; I think it was just assumed to be 0 as in the lemma.) □

Chapter 10

2/27/2023: Portfolio Theory

10.1 Portfolio Optimization Basics

Suppose we have a portfolio of two stocks with expected returns $\mu_1 = \mathbb{E}[r_1]$ and $\mu_2 = \mathbb{E}[r_2]$. For this example, first let $\mu_1 = 1$ and $\mu_2 = 0$. We want to allocate weights (either positive or negative) to each of the stocks that sum to 1. If we allocate $1 + \alpha$ to asset 1 and $-\alpha$ to asset 2, then our expected return is $1 + \alpha$, which goes to ∞ as α goes to ∞. There's a few problems with just concluding that we should assert that an infinite α is optimal:

- Larger α values exposes us to larger variance, which makes this strategy arbitrarily **risky** as α becomes arbitrarily large.
- (Less important for this theory but a real consideration in practice:) Since the weight on the second asset is $-\alpha$, we are essentially borrowing capital via shorting asset 2 in order to invest in asset 1. This has issues such as not taking into account the cost of borrowing.

Thus, expected value is not enough to evaluate a portfolio – we also need to think about some metric of risk such as variance.

10.1.1 Diversification ("Vignettes")

One common way to reduce risk is through diversification. Suppose we have two assets A, B both with variances σ^2. Then, the variance of our portfolio given weights w_1 and w_2 on A and B respectively is

$$\begin{aligned}\operatorname{Var}(w_1 r^A + w_2 r^B) &= w_1^2\sigma^2 + w_2^2\sigma^2 + 2w_1 w_2 \operatorname{Cov}(r^A r^B) \\ &\leqslant \sigma^2(w_1^2 + w_2^2 + 2w_1 w_2) \\ &= \sigma^2(w_1 + w_2)^2 \\ &= \sigma^2.\end{aligned}$$

The inequality is simply the Cauchy-Schwarz Inequality, $\operatorname{Cov}(w_1, w_2) \leqslant \sqrt{\operatorname{Var}(w_1)\operatorname{Var}(w_2)} = \sigma^2$, and the last equality comes from the fact that the weights must sum to 1.

Instead suppose there are n assets with variances $\sigma_1^2, \cdots, \sigma_n^2$ with weights $w_i = \frac{1}{n}$. Then the variance of the returns of our portfolio is

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} w_i r^i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{n^2}\sum_{i\neq j}^{n}\operatorname{Cov}(r^i, r^j) \\
&= \frac{1}{n}\left(\frac{\sum_{i=1}^{n}\sigma_i^2}{n}\right) + \frac{n(n-1)}{n^2}\left(\frac{\sum_{i\neq j}\operatorname{Cov}(r^i, r^j)}{n(n-1)}\right) \\
&= \frac{1}{n}\underbrace{\left(\frac{\sum_{i=1}^{n}\sigma_i^2}{n}\right)}_{\text{“average variance”}} + \left(1-\frac{1}{n}\right)\underbrace{\left(\frac{\sum_{i\neq j}\operatorname{Cov}(r^i, r^j)}{n(n-1)}\right)}_{\text{“average covariance”}}.
\end{aligned}$$

Then as n becomes large, essentially all of the weight will be on the “average covariance” term, so the variance of our portfolio will be almost entirely determined by the covariances between the assets.

10.1.2 General Portfolio Setup

Now suppose we have a vector of asset returns at time t $\mathbf{r}_t$ with expectations that are conditional on information $\boldsymbol{\mu}_t = \mathbb{E}[\mathbf{r}_{t+1}|\mathcal{F}_t]$ and similarly conditional variance $\boldsymbol{\Sigma}_t = \operatorname{Var}(\mathbf{r}_{t+1}|\mathcal{F}_t)$. We wish to find $\boldsymbol{\alpha}_t$, which is the optimal portfolio. The value of our portfolio is $V_t = \boldsymbol{\alpha}_t^\top \mathbf{r}_t$, and the conditional expectation and variance become (by properties proved in PSet 2)

$$\mathbb{E}[V_{t+1}|\mathcal{F}_t] = \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t$$

and

$$\operatorname{Var}(V_{t+1}|\mathcal{F}_t) = \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t.$$

One simple formulation is to simply find $\operatorname{argmax}_{\boldsymbol{\alpha}_t} \boldsymbol{\alpha}^\top \boldsymbol{\mu}_t$, but we saw earlier that this is too simplistic as it does not consider risk (among other things). Some other possibilities for our optimization problem formulation:

1. $\operatorname{argmax}_{\boldsymbol{\alpha}_t} \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t - \frac{\Gamma_t}{2}\boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t$
2. $\operatorname{argmax}_{\boldsymbol{\alpha}_t} \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t$ such that $\boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}$ is under some limit
3. Divide the returns by standard deviation, giving the Sharpe ratio

The third optimization problem is very hard to solve, but the first and second are actually the same and are much more tractable using Lagrange multipliers.

Theorem 10.1

Let $\hat{\boldsymbol{\alpha}}_t$ to be the solution to the optimization problem in (1) from above. Then,

$$\hat{\boldsymbol{\alpha}}_t = \frac{1}{\Gamma_t}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t,$$

and the expected returns and variance of the portfolio are

$$\mathbb{E}[V_{t+1}|\mathcal{F}_t] = \frac{1}{\Gamma_t}\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t$$

and

$$\mathrm{Var}(V_{t+1}|\mathcal{F}_t) = \frac{1}{\Gamma_t^2}\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t.$$

Proof: The proof is fairly straightforward optimization:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\alpha}_t}\left(\boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t - \frac{\Gamma_t}{2}\boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t\right) &= \boldsymbol{\mu}_t - \Gamma_t \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t \\ \implies \Gamma_t \boldsymbol{\Sigma}_t \hat{\boldsymbol{\alpha}}_t &= \boldsymbol{\mu}_t \\ \implies \hat{\boldsymbol{\alpha}}_t &= \frac{1}{\Gamma_t}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t.\end{aligned}$$

Then finding the expected values and variances are fairly straightforward computation. □

Note:

The term denoted as the partial derivative with respect to a vector $\mathbf{v} \in \mathbb{R}^n$ is the same as the gradient with respect to that vector, i.e.

$$\frac{\partial f}{\partial \mathbf{v}} = \nabla_{\mathbf{v}} f = \begin{bmatrix} \frac{\partial f}{\partial v_1} \\ \vdots \\ \frac{\partial f}{\partial v_n} \end{bmatrix}.$$

In this scenario, we can also calculate the Sharpe ratio:

$$S_a = \frac{\mathbb{E}[V_{t+1}|\mathcal{F}_t]}{\sqrt{\mathrm{Var}(V_{t+1}|\mathcal{F}_t)}} = \sqrt{\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}.$$

Notably, this Sharpe ratio does not depend on our choice of Γ_t.

10.2 Lagrange Multipliers Review

If we want to maximize $f(x)$ subject to some constraint $g(x) = c$, we usually rely on a technique called Lagrange Multipliers. This method relies on the observation that at the optimum, the curve $g(x) = c$ is tangent to the relevant level set of $f(x)$, with the intuition being that if this were not the case, we could move in at least one of the directions to increase the objective function while still satisfying the constraint.

Example 10.1 (Shannon Entropy)

For state space $\Omega = \{w_i\}_{i=1}^n$ with probability distribution $\mathbb{P}$ that has probabilities P_i of being in each state w_i, then the **Shannon Entropy** is

$$\text{Ent}(\mathbb{P}) = -\sum_{i=1}^{n} P_i \log P_i \geqslant 0.$$

Note that in this discrete space, the discrete uniform distribution maximizes entropy, and among continuous unbounded distributions with mean 0 and variance 1, the Gaussian distribution maximizes entropy. We will show that entropy is maximized when $P_i = \frac{1}{n}$ for all $i \in \{1, \cdots, n\}$.
Our objective is to solve $\max_{\mathbb{P}} -\sum_{i=1}^n P_i \log P_i$ subject to the constraint that $\sum_{i=1}^n P_i = 1$. Using Lagrange multipliers, this is equivalent to maximizing

$$\mathcal{L}(\mathbb{P}, \lambda) = -\sum_{i=1}^{n} P_i \log P_i + \lambda \left(\sum_{i=1}^{n} P_i - 1 \right),$$

which we do by setting the partial derivatives in terms of both P_i and λ to be 0. The derivatives with respect to P_i are

$$\frac{\partial}{\partial P_i} \mathcal{L}(\mathbb{P}, \lambda) = -1 - \log P_i + \lambda,$$

which, after equating this to 0, tells us that all $\hat{P}_i$ must be equal. Taking the partial derivative with respect to λ and maximizing (or using the constraint of optimization) leads to the desired result.

Chapter 11

3/1/2023

Before getting to the core information for the day, first here is a brief note about a specific important Martingale.

Note:

If $\{M_n, \mathcal{F}_n\}$ is a Martingale and $\{A_k\}_{k=1}^N$ is some random variable with $\mathbb{E}[A_k|\mathcal{F}_{k-1}] = A_k$, then

$$Z_n = \sum_{k=1}^{N} A_k(M_k - M_{k-1})$$

is also a Martingale. We can interpret this in a financial context as the total returns of a stock over the time period where the Martingale differences $M_k - M_{k-1}$ are the stock price changes between days and A_k is the quantity that we invest into the stock on a given day. This is also important probabilistically because it tells us one fundamental way of constructing new Martingales from old ones. When we change the time scale from discrete to continuous, this becomes the basis of stochastic calculus.

11.1 Back to Portfolio Basics

Recall that our optimization problem from last class was

$$\max_{\boldsymbol{\alpha}_t^\top} \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t - \frac{\Gamma_t}{2} \boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t,$$

where $\boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t$ is the returns, $\boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t$ represents the risk (variance) of the portfolio, and Γ_t is a parameter representing our risk tolerance. We can also formulate the optimization problem as

$$\min_{\boldsymbol{\alpha}_t} \boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t = \mu^\star.$$

This formulation tells us that we want to minimize our risk given some fixed returns that we need to attain. Both formulations are the same when $\Gamma_t = \left[\frac{\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\mu}_t}{\mu^\star}\right]$. We will solve this second formulation using a Lagrangian (with the quantity to minimize multiplied by 1/2 for convenience

which does not have any effect on the optimal quantities we will attain):

$$\begin{aligned}
\mathcal{L}_t &= \frac{1}{2}\boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t - \lambda\left(\boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t - \mu^\star\right) \\
\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{\alpha}_t} &= \mathbf{0} \implies \boldsymbol{\Sigma}_t \boldsymbol{\alpha} - \lambda \boldsymbol{\mu}_t = \mathbf{0} \\
\frac{\partial \mathcal{L}_t}{\partial \lambda} &= 0 \implies \boldsymbol{\alpha}_t^\top \boldsymbol{\mu}_t - \mu^\star = 0.
\end{aligned}$$

Then it is relatively straightforward to solve for the optimal quantity

$$\hat{\boldsymbol{\alpha}}_t = \left(\frac{\mu^\star}{c_t}\right)\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t$$

where

$$c_t = \boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t.$$

We can also rewrite the optimizer as

$$\hat{\boldsymbol{\alpha}}_t = \frac{\mu^\star \cdot \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}{\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t} = \underbrace{\mu^\star\left(\frac{\mathbf{1}_q^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}{\boldsymbol{\mu}_t^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}\right)}_{C_{\mu^\star}} \cdot \underbrace{\frac{\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}{\mathbf{1}_q^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t}}_{\mathbf{w}_{\mu,t}}.$$

Notably, $C_{\mu^\star}$ is a constant and the weights $\mathbf{w}_{\mu,t}$ are independent of $\mu^\star$.

Note:

There are some practical challenges that arise when using these methods in practice to find estimates of all the relevant quantities. For example, if the number of dimensions (stocks) is greater than the number of data points, then the covariance matrix might have a lot of issues. We will explore this phenomenon in PSet 4.

One other possible optimization is to simply minimize the variance of the portfolio given that the weights sum to 1 as always,

$$\min_{\boldsymbol{\alpha}_t} \boldsymbol{\alpha}_t^\top \boldsymbol{\Sigma}_t \boldsymbol{\alpha}_t \quad \text{s.t.} \quad \boldsymbol{\alpha}_t^\top \mathbf{1}_q = 1,$$

where q is the number of stocks and where $\mathbf{1}_q$ is the column vectors of 1's. The solution is called the global minimum variance portfolio:

$$\hat{\mathbf{w}}_{1,t} = \frac{\boldsymbol{\Sigma}_t^{-1}\mathbf{1}_q}{\mathbf{1}_q^\top \boldsymbol{\Sigma}_t^\top \mathbf{1}_q}.$$

The next optimizer, which combines the ideas of the previous optimization problems, is so important that it has a name.

11.2 Markowitz

The Markowitz optimization problem is

$$\min_{\mathbf{w}_t} \mathbf{w}_t^\top \boldsymbol{\Sigma}_t \mathbf{w}_t \quad \text{s.t.} \quad \mathbf{w}_t^\top \mathbf{1}_q = 1 \quad \text{and} \quad \mathbf{w}_t^\top \boldsymbol{\mu}_t = \mu^\star.$$

The solution to this optimization problem is

$$\hat{\mathbf{w}}_t = \lambda_t \hat{\mathbf{w}}_{1,t} + (1 - \lambda_t)\hat{\mathbf{w}}_{\mu,t}$$

where

$$\hat{\mathbf{w}}_{1,t} = \frac{\boldsymbol{\Sigma}_t^{-1}\mathbf{1}_q}{\mathbf{1}_q^\top \boldsymbol{\Sigma}_t^{-1}\mathbf{1}_q},$$

$$\hat{\mathbf{w}}_{\mu,t} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_t}{\mathbf{1}^\top \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t},$$

and λ_t is kinda messy but can be found in the supplementary files on Canvas.

Note:

Note that $\boldsymbol{\alpha}_t$ and $\mathbf{w}_t$ in the past few sections both denote the same weights on the securities in the portfolio. This is because keeping consistent notation is difficult.

This optimization problem is solved simply using Lagrange multipliers with two constraints:

$$\mathcal{L}_t = \frac{1}{2}\mathbf{w}_t^\top \boldsymbol{\Sigma}_t \mathbf{w}_t - \lambda_1(\mathbf{w}_t^\top \mathbf{1}_q - 1) - \lambda_2(\mathbf{w}_t^\top \boldsymbol{\mu}_t - \mu^\star).$$

Then taking partial derivatives with respect to $\mathbf{w}$, λ_1, and λ_2, equating them to 0, and solving gives the desired optimal weights.

Note:

If we wanted to take into account inequalities such as a ban on short selling, we could instead solve this optimization using a (convex) linear program.

Chapter 12

3/6/2023: Tails and Sampling

12.1 Heavy Tailedness

The **Cauchy distribution**, which is the t-distribution with one degree of freedom, is essentially the poster child of empty tailedness. The PDF of the Cauchy distribution is

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{x^2 + 1}$$

for all $x \in \mathbb{R}$. This distribution has a nonexistent mean, variance – this implies that all moments do not exist (not trivial). We can also generate the Cauchy distribution by taking $\tan\theta$ if $\theta \sim \text{Unif}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. The geometric interpretation is that if we have a sphere of diameter 1 and a plane that it is tangent to, then a line that makes an angle of θ with the normal to the plane will intersect the plane at a distance of $\tan\theta$ from the point of tangency. Issues like this can arise especially when taking ratios, for example the ratio of the returns of a stock to the returns of an index. The fact that the mean is undefined means that we cannot rely on LLN or CLT to ensure that our estimates for mean will converge to some value.

Note that the t-distribution with $\nu > 1$ degrees of freedom has defined moments and is less heavy tailed than a Cauchy but is heavier tailed than a Gaussian.

12.1.1 Extreme Values

One other way to think about heavy tailedness is through a running maximum. Suppose we have the standard normal variables $Z_1, Z_2, \cdots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and let $\psi_n = \max_{1 \leqslant i \leqslant n} Z_i$. As $n \to \infty$, $\psi_n \to \infty$. Note that this is true for any distribution that has support that has an infinite supremum. In particular, we have something along the lines of $\psi_n \in O(\sqrt{2\log n})$ which comes from the $\exp\left\{-\frac{x^2}{2}\right\}$ in the PDF of the standard normal.

Now if we instead consider $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim}$ Cauchy and let $\mu_n = \max_{1 \leqslant i \leqslant n} Y_i$, then we get something along the lines of $\mu_n \sim O(n^\alpha)$ for $\alpha = 1/2$ (verify this), which is significantly faster than the rate of growth of ψ_n.

12.1.2 Estimators

If we have $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, then $\frac{1}{n}\sum_{i=1}^n X_i = \overline{X} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$. The real life implication is that as our sample size increases, the variance of our estimate of our mean decreases and

eventually converges to the true value. However, this nice consequence is not nearly as nice for the Cauchy distribution.

Claim 12.1 If $Z_1, \cdots, Z_n \overset{i.i.d.}{\sim}$ Cauchy, then $\frac{1}{n}\sum_{i=1}^{n} Z_i = \overline{Z} \sim$ Cauchy. This means that when we have undefined moments, taking the average does not reduce our uncertainty at all; the tails of the estimate of the mean is just as heavy as the tails of the individual data points.

Proof: The moment generating function of the Cauchy distribution does not exist since the moments themselves do not exist, so instead we will analyze the characteristic function $\varphi_z(t)$ which has a very similar function:

$$\varphi_z(t) = \mathbb{E}\left[e^{itz}\right].$$

Notably, the characteristic function of the Cauchy distribution is $\varphi_z(t) = e^{-|t|}$. In particular, there is a bijection between the set of distributions and and the set of characteric functions. If we take the mean of our Cauchy variables, then

$$\begin{aligned}
\mathbb{E}\left[e^{it\overline{z}}\right] &= \mathbb{E}\left[\exp\left\{it \cdot \frac{1}{n}\sum_{i=1}^{n} Z_i\right\}\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left\{i \cdot \frac{t}{n} \cdot z_n\right\}\right] \\
&= \prod_{i=1}^{n} \exp\left\{-\left|\frac{t}{n}\right|\right\} \\
&= e^{-|t|}.
\end{aligned}$$

This is exactly the characteristic function of the Cauchy distribution, so we are done. □

Note:

This isn't even as bad as it gets – if $Y \sim \mathcal{N}(0,1)$, then estimating the mean of $\frac{1}{Y^2}$ by averaging the data as usual will actually result in an estimator whose distribution's tails are even heavier than those of the sampling distribution.

Corollary 12.1

If $\alpha_1, \alpha_2 \geqslant 0$ with $\alpha_1 + \alpha_2 = 1$, then $\alpha_1 Z_1 + \alpha_2 Z_2 \sim$ Cauchy.

Markowitz relates directly to this corollary. The Markowitz portfolio tries to minimize the variance of $\alpha_1 r_1 + \alpha_2 r_2$, but if the returns are Cauchy, then this is a useless minimization attempt. That by Natesh, this is also true when Z_1 and Z_2 are not independent.

12.2 Bootstrapping

Given samples $X_1, \cdots, X_n$, suppose we are trying to analyze the average value. We can take the average directly of these samples to get $\overline{X}$. If we want more information about the distribution of the sample means, we can resample. We take n samples from replacement $X_1^{(1)}, \cdots, X_m^{(1)}$

which we can then use to estimate resampled mean $\overline{X}^{(1)}$. The most common choice is $m = n$. Clearly we can repeat this procedure as many times as we want, and this is nice because we can essentially do more analysis without actually having to gather more data, which can be costly, inconvenient, or impossible. However, note that the bootstrap only estimates the true distribution and variance of the data since it is essentially assuming that the entire population looks like our sample.

Suppose we are tracking the price of a stock P_t with $P_{t+1} = P_t + r_t$. Next class we will explore various models for this. If we have P_t and past data $\{P_i\}_{i=0}^{t-1}$, then we could bootstrap the previous returns in order to estimate a confidence interval for future price $P_{t+\delta}$. In particular, we sample $r_t, r_{t+1}, \cdots, r_{t+\delta-1}$ from the returns $r_0, r_1, \cdots, r_{t-1}$. This procedure is nice because it makes no parametric assumptions about the data. However, this method does have some weaknesses; probably most importantly, we are assuming that future returns look like previous returns. It also has issues with dependencies between consecutive returns. One way to deal with this is by looking at serial returns, for example by looking at every three day stretch within the time period, and then sampling from that set of three day stretches.

Chapter 13

3/20/2023: Brownian Motion

13.1 Definition

Recall the random walk which we have talked about multiple times already in this class; these were all defined on discrete time scales. However, real-life events generally happen on continuous time scales – how can we generalize this idea to a continuous time interval?

First we think about our discrete random walk with times $t = \{0, 1, \cdots, n\}$ and which moves up or down with probability $1/2$ at each time step. Then suppose we compress the time dimension by a factor of n and the spatial dimension by a factor of $\sqrt{n}$ so that time will vary in $[0, 1]$ and the spatial dimension will have a variance of 1 at the terminal time point. As $n \to \infty$, the limiting object that we get will be Brownian Motion.

Note:

The physical origins of Brownian Motion came from some dude in the 1800s who was studying pollen and thought their motion was kinda cool. Then in 1905 Einstein came along and did his smart stuff to write down the equation for Brownian Motion. According to Natesh, this was "absolute bonkers".

With $N \to \infty$, we can treat the time scale as continuous.

Definition 13.1: Brownian Motion

$\{B_t, t_{\geqslant 0}\}$ is **Brownian motion** if:

1. $B_0 = 0$
2. B_t has *stationary* and *independent* increments
3. $B_{t+s} - B_s \sim \mathcal{N}(0, t)$ for any $t, s > 0$
4. B_t has continuous sample paths

Sometimes Brownian motion is also denoted as W_t after Norbert Wiener.

A couple of notes about the definition:

- If we have times $0 = t_0 \leqslant t_1 \leqslant t_2 \leqslant t_3 \leqslant \cdots \leqslant 1$, then the increments are $B_{t_1}, B_{t_2} - B_{t_1}, B_{t_3} - B_{t_2}, \cdots$.

- Independence of increments means that $B_{t_2} - B_{t_1}$ is independent of $B_{t_3} - B_{t_2}$ (and same for any increments). However, this does not imply that B_{t_2} and B_{t_1} are independent; we will see later exactly what the correlation is.
- The increments being stationary means that $B_{t_1+s} - B_{t_1}$ and $B_{t_2+s} - B_{t_2}$ have the same distribution. Note that this follows from property (3) of Brownian motion.
- The Gaussian distribution of the increments follows from the random walk foundation and from the central limit theorem – we are adding some large ($n \to \infty$) number of i.i.d. r.v.s, so the limiting distribution will be Gaussian by CLT.

13.2 Properties

We have that $(B_{t_1}, \cdots, B_{t_n}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We will now calculate the covariance matrix $\mathbf{\Sigma}$. We have $B_t = B_t - B_0 \sim \mathcal{N}(0, t)$.

Lemma 13.1

$$\mathrm{Cov}(B_t, B_s) = \min(s, t) = s \wedge t.$$

Proof: Assume without loss of generality that $s \leqslant t$. We can prove this lemma via direct calculation:

$$\begin{aligned}
\mathrm{Cov}(B_t, B - s) &= \mathbb{E}[B_t B_s] - \underbrace{\mathbb{E}[B_t]}_{0}\underbrace{\mathbb{E}[B_s]}_{0} \\
&= \mathbb{E}[(B_t - B_s + B_s)B_s] \\
&= \mathbb{E}[(B_t - B_s)B_s] + \mathbb{E}[B_s^2] \\
&= \mathbb{E}[B_t - B_s]\mathbb{E}[B_s] + \mathrm{Var}(B_s) \\
&= s.
\end{aligned}$$

□

We can then write the covariance matrix $\mathbf{\Sigma}$ as

$$\begin{bmatrix} t_1 & t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & t_2 & \cdots & t_2 \\ t_1 & t_2 & t_3 & \cdots & t_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & t_3 & \cdots & t_n \end{bmatrix}.$$

Notably, this covariance matrix is positive definite.

Next, consider if we "magnify" a small portion of the interval $[t_1, t_2] \subseteq [0, 1]$. Then this gives us another object of Brownian motion, which tells us that Brownian motion is **scale invariant** and self-similar.

Theorem 13.1 Scale Invariance

For any $c > 0$ let $Z_t = \frac{B_{ct}}{\sqrt{c}}$. Then, Z_t is Brownian motion.

Proof: PSet fun. □

13.3 Some Linear Algebra

If $Z \sim \mathcal{N}(0,1)$, then $\sigma Z \sim \mathcal{N}(0,\sigma^2)$. In the multivariate case, we have the analog that if $\mathbf{Z}_{d\times 1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, then if we have some $d \times d$ covariance matrix $\mathbf{\Sigma}$, then $\mathbf{\Sigma}^{1/2}\mathbf{Z}_{d\times 1} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. One common way to calculate $\mathbf{\Sigma}^{1/2}$ is through **Cholesky decomposition** which finds $\mathbf{\Sigma}^{1/2} = L$ such that $\mathbf{\Sigma} = LL^\top$ and such that L is a lower triangular matrix with positive diagonal entries. This lower triangular structure can also be useful for solving equations such as $A\mathbf{x} = \mathbf{b}$; by decomposing $A = LL^\top$, we sequentially solve two systems which both come from lower diagonal matrices of coefficients, which are relatively trivial to solve.

Definition 13.2: Random Fourier Series

If we have some function $f : [0,1] \mapsto \mathbb{R}$, then we can rewrite the function using the **Fourier series**

$$f(t) = \sum_{n=0}^{\infty} \left(a_n \cos(nt) + b_n \sin(nt)\right).$$

13.4 Simulations

Use ChatGPT!

13.5 PDEs