Systematic Trading from First Principles

Oden Petersen

2025-10-31

"$y = X\beta + \epsilon$, the rest is commentary."

About Me

- UNSW Maths/Compsci (graduated T1 2025)
- Data Engineering @ Arnott's Biscuits (2021)
- CPMSoc Maths Director 2021
- Data Scientist @ Daisee (2021-2022)
- Started QuantSoc 2022
- Quant Intern @ Autumn Compass (2022)
- Quant Intern @ Citadel Securities (Summer 2023)
- Trading Intern @ Vivcourt Trading (Summer 2024)
- Starting work next year
- Polymarket trading go brr

Point of This Talk

- Mathematise some intuitions about markets, trading systems, financial economics
- Give a precise picture of the many moving parts
- Build concepts axiomatically from the ground up
- Present a common basis for high-frequency and low-frequency trading (increasingly convergent)
- Starting point for exploring more resources. Like markets, a lot of financial literature is low-signal. *"People without dirty hands are wrong."*

Three approaches: systematic, semi-systematic, discretionary. Why systematic?

- Low-touch
- Consistency over time and across assets
- Low-latency opportunities

"We're mediocre traders, but our system never has rows with its girlfriends." Nick Patterson, Renaissance Technologies

Not the Point of This Talk

- How to get a job. In an efficient market the way to get a job is just to get good at the work
- How to do the job. The work itself is often extremely heuristic by necessity. The ability to appreciate which parts of the problem are most important represent things in a simple way while capturing the important qualitative parts is a key skill in both manual and systematic trading
- What the job is like. Communication at work focuses much more on concrete facts, relies on greater implicit shared context
- What you need to know for the job. Many of the best traders don't think in mathematical language
- The 'correct' model for anything. Not all uncertainty is quantifiable (Frank Knight, Nassim Taleb). "All models are wrong" (George Box).
- Politics of financialisation and market structure

Outline

1. Securities Markets
2. Trading
3. Market Microstructure
4. Portfolio Management
5. Factor Models
6. Dynamic Portfolio Selection
7. Appendix: Accounting
8. Appendix: Microprice
9. Appendix: PPCA Stabilisation
10. Appendix: Statistical Arbitrage
11. Appendix: Asset Pricing
12. Appendix: Research Process
13. Bibliography

Securities Markets

"Governing by the power of virtue can be compared to the pole star, which remains fixed in place while all the other stars orbit respectfully around it." Confucius

Spot Transactions

The point of trading is to obtain an asset by giving up money, or obtain money by giving up an asset.
If I give you $q > 0$ units of some asset A, and you give me $\$pq$, then:

- I have **sold** q units of A to you at $\$p$
- You have **bought** q units of A from me for $\$p$

Buying and selling are collectively called 'trading'.
Suppose I own some amount of A and some amount of money. If we let s be $+1$ for buying and -1 for selling, then the result of any trade is to add qs to the amount of A I own, and add $-\$qps$ to the amount of money I have.

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.
A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.
A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.
For example, $T = \ldots$

- t ('spot', e.g. blockchain)
- $t + 1, t + 2, \ldots$ ('clearing', e.g. equities)
- Last Thursday of month ('futures')

Securities Markets and Exchanges

The **market** is the collective activity of all traders. When we don't care who we trade with, we can just 'trade with the market'.
A **securities market** for some asset A, open at a time t, is any standardised way for traders to reach agreements to buy or sell A at a specified **settlement time** $T \geq t$.
For example, $T = \ldots$

- t ('spot', e.g. blockchain)
- $t + 1, t + 2, \ldots$ ('clearing', e.g. equities)
- Last Thursday of month ('futures')

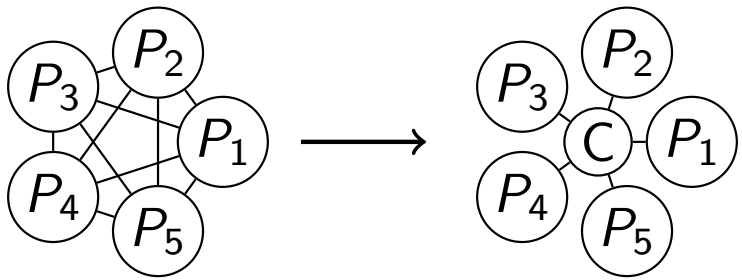
If you agree to give something to someone, you have an **obligation**. If someone agrees to give you something, you have a **right**.

Counterparty Risk

If I have an agreement with P_1 to buy 10 units for $\$p_1$ at T, and an agreement with P_2 to sell 10 units at $\$p_2$ at T, and no further rights/obligations, am I guaranteed to meet my obligations?

Centralisation

A **securities exchange** is a centralised venue serving a securities market for **exchange participants** (e.g. ASX, NYSE, TSE, HKEX, LME). Agreements not made through an exchange are often called OTC (over-the-counter).



Centralisation generally reduces **search costs** and **counterparty risk**.

Netting

Centralisation allows for **netting** of rights and obligations.
For any settlement time T, I only need to keep track of the difference between money owed to and by me, and units owed to and by me.
The quantity of A owned by me, plus the quantity owed to me, minus the quantity owed by me to others, is known as my **net position** in A.
If this is positive, I have a **long position**. If it is negative, I have a **short position**. If it is zero, I am **flat**.

Collateralisation

At certain intermediate times t' ($t \leq t' \leq T$), participants may be required to physically give ('post') something to the exchange to **collateralise** their obligations.

- Money ('margin')
- Assets ('locate'/'borrow')

If an agreement made on the exchange gives you rights to money or assets at T, this is typically as good as posting actual money or assets for an obligation at $T' \geq T$.
Some amount of interest may be charged to make up for the difference between the size of our obligations and the size of our collateral. For cash, this is according to an **interest rate**; for other assets, it is according to a **borrow rate**/**short rate**.

Summary

- **Trading** is swapping money and assets
- A **market** is whatever you use to trade
- A **securities market** is a standardised way to agree to trades
- Agreements consist of **rights** and **obligations**
- Finding a **counterparty** may involve **search cost**
- Agreements between two parties are subject to **counterparty risk**
- A **securities exchange** is a centralised trading venue
- After trades are agreed to on an exchange, they will be **settled** in some standardised way
- The net quantity of A that I have some claim to can either be positive (**long position**), negative (**short position**), or zero (**flat**).
- Traders may be obligated to post assets ('locate') or money ('margin')

Trading

"To gain an advantage from better knowledge of facilities of communication or transport is sometimes regarded as almost dishonest, although it is quite as important that society make use of the best opportunities in this respect as in using the latest scientific discoveries." Hayek

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.
Suppose that at each time t we have cash holdings of $\$c_t$ and net holdings of a_t units of some asset A.
Suppose also that trades $(s_t, q_t, \$p_t)$ take place at a finite set of distinct times

$$\tau = \{t_1, \ldots t_n\} \subset T = [t_-, t_+],$$

where $t_- < t_1 < \ldots < t_n < t_+$.

Setup

A sequence of trades that collectively increases the amount of money you have and leaves the amount of each asset you have unchanged is clearly favourable.
Suppose that at each time t we have cash holdings of $\$c_t$ and net holdings of a_t units of some asset A.
Suppose also that trades $(s_t, q_t, \$p_t)$ take place at a finite set of distinct times

$$\tau = \{t_1, \ldots t_n\} \subset T = [t_-, t_+],$$

where $t_- < t_1 < \ldots < t_n < t_+$.
Suppose further that p_t is a right-continuous function $\mathbb{R} \to \mathbb{R}$ with left-limits.
For instance, we could take $p_t = p_{\max(\tau \cap (-\infty, t])}$ for $t \geq \min \tau$ and $p_t = x$ otherwise for some arbitrary x. This is known as the last traded price.

Accounting

For any time-varying quantity x_t, let x_t^+ and x_t^- denote the right- and left-limits respectively.
Furthermore, define a signed measure x_ω such that for any interval $T' = [t'_-, t'_+]$ we have

$$x_{T'} := x_{t_+}^+ - x_{t_-}^- .$$

Accounting

For any time-varying quantity x_t, let x_t^+ and x_t^- denote the right- and left-limits respectively.
Furthermore, define a signed measure x_ω such that for any interval $T' = [t'_-, t'_+]$ we have

$$x_{T'} := x^+_{t_+} - x^-_{t_-}.$$

Then we have

$$\begin{aligned}
a_{T'} &= \sum_{t \in \tau} s_t q_t && = \int_{t \in T} da, \\
c_{T'} &= \sum_{t \in \tau} -p_t(s_t q_t) && = \int_{t \in T} -p_t da, \\
p_{T'} &= p_{t'_+} - p^-_{t'_-}. && = \int_{t \in T} dp.
\end{aligned}$$

Cash Holdings

It can be shown (see appendix) that the cashflow over the entire interval $T = [t_-, t_+]$ is

$$\$c_T = \int_{t \in T} -\$p_t da = \$p_{t_-} a_{t_-} - \$p_{t_+} a_{t_+} + \$ \int_{t \in T} a_t^- dp.$$

This is similar in spirit to integration by parts:

$$\int_a^b f \frac{dg}{dx} dx = f(b)g(b) - f(a)g(a) - \int_a^b g \frac{df}{dx} dx.$$

Then we have

$$\$(c_{t_+} + p_{t_+} a_{t_+}) - \$(c_{t_-} + p_{t_-} a_{t_-}) = \$ \int_{t \in T} a_t^- dp.$$

The quantity $\$v_t = \$p_t a_t$ is known as the **dollar value** of our A holdings **marked** to the price $\$p_t$.

Portfolio Valuation

Suppose now that we trade multiple assets, such that p_t, a_t and v_t are vector-valued, with v_t the elementwise product of p_t and a_t.
A collection of assets held in quantities a_t is known as a **portfolio**.
We can write

$$\$(c_{t_+} + p_{t_+} \cdot a_{t_+}) - \$(c_{t_-} + p_{t_-} \cdot a_{t_-}) = \$ \int_{t \in T} a_t^- \cdot dp,$$

where p_ω is now a vector-valued measure.

Portfolio Valuation

Suppose now that we trade multiple assets, such that p_t, a_t and v_t are vector-valued, with v_t the elementwise product of p_t and a_t.
A collection of assets held in quantities a_t is known as a **portfolio**.
We can write

$$\$(c_{t_+} + p_{t_+} \cdot a_{t_+}) - \$(c_{t_-} + p_{t_-} \cdot a_{t_-}) = \$ \int_{t \in T} a_t^- \cdot dp,$$

where p_ω is now a vector-valued measure. Let

$$\$\Pi_t = \$c_t + \$p_t \cdot a_t = \$c_t + \$ \sum v_t.$$

We call $\$\Pi_t$ the **value** of our portfolio **marked** to p_t.

Profitability

The quantity $\$\Pi_{t_+} - \$\Pi_{t_-}$ is our **net P&L** (profit and loss) over the interval T, marked to p_t. Then we have

$$\Pi_T = \int_{t \in T} a_t^- \cdot dp.$$

we can write

$$\Pi_{[t_i, t_{i+1}]} = \int_{[t_i, t_{i+1}]} a_t^- \cdot dp = \int_{t \in [t_i, t_{i+1}]} v_t^- \cdot \frac{dp}{p_t^-},$$

where the quotient $\frac{dp}{p_t^-}$ is computed elementwise.

Leverage

Suppose we can always make any trade we like at time t with price $\$p_t$ (in practice, there are limits on the trades we can make at a particular price and time).
Then we can freely convert a portfolio with value $\$\Pi_t$ to that much in cash.
Conversely, we can convert $\$\Pi_t$ worth of cash into any portfolio with that value.

Leverage

Suppose we can always make any trade we like at time t with price $\$p_t$ (in practice, there are limits on the trades we can make at a particular price and time).
Then we can freely convert a portfolio with value $\$\Pi_t$ to that much in cash.
Conversely, we can convert $\$\Pi_t$ worth of cash into any portfolio with that value.
Typically, $\$\Pi_t$ can change in two ways: trading assets, or transferring cash into and out of the portfolio. We will generally ignore the possibility of transfers.
If we begin with a portfolio worth $\$\Pi_{t_1}$ and make a sequence of trades of the form (s_t, q_t, p_t) that result in a portfolio worth $\$\Pi_{t_n}$, then we could instead begin with a portfolio worth $L\$\Pi_{t_1}$ and make trades (s_t, Lq_t, p_t) to arrive at a portfolio worth $L\$\Pi_{t_n}$. The ratio L is known as the **leverage ratio**.

Return on Capital

Because of collateralisation requirements, portfolio management uses up cash.

Consider long-only spot-settled trading. If we were to turn our portfolio into cash, or convert cash into an identical portfolio, we would receive/require $\$\Pi_t$.[1]

If our initial portfolio value were $\$\Pi_{t_-} + \$N$ instead of Π_{t_-}, and we could simply scale up trade sizes at the same prices, then set

$$L = \frac{\Pi_{t_-} + N}{\Pi_{t_-}}.$$

The **return on capital** is defined as the increase in P&L per dollar added to initial portfolio value, i.e.

$$R_T = \frac{L \int_T d\Pi - \int_T d\Pi}{N} = \frac{\int_T d\Pi}{\Pi_{t_-}}.$$

[1] We generally want c_t to be small, but we may want some spare cash for trading, so it still forms part of the collateralisation requirement.

Log Returns

If we define

$$\ell_t = \log \Pi_t,$$

then for any interval $T' = [t'_-, t'_+]$, we have

$$\ell_{T'} = \log \left(\frac{\Pi^+_{t'_+}}{\Pi^-_{t'_-}} \right) = \log(R_{T'} + 1),$$

and so

$$R_{T'} = \exp(\ell_{T'}) - 1.$$

Then for any measurable set ω we can define

$$R_\omega = \exp(\ell_\omega) - 1 \approx \ell_\omega + O(\ell_\omega^2) \text{ (for small } \ell_\omega).$$

We call ℓ_T the **log-return** over the interval T.

Properties of Returns and Log-Returns

Let $w_t := \frac{1}{\Pi_t} v_t$ be the **weight vector**.
For an interval $T' = (t_i, t_{i+1}]$, we have a_t^- equal to a constant over T', and

$$R_{T'} = w_{t_i}^+ \cdot r_{T'},$$

where the elementwise quotient

$$r_{T'} = \frac{p_{t_{i+1}} - p_{t_i}}{p_{t_i}}$$

is known as the **asset returns** vector over T'. In contrast, $\ell_{T'}$ is not linear in $r_{T'}$.
For a disjoint collection of measurable sets $\omega_1, \ldots \omega_n$ whose union is Ω, we have

$$\ell_\Omega = \sum_{i=1}^{n} \ell_{\omega_i},$$

$$R_\Omega = \left(\prod_{i=1}^{n} (1 + R_{\omega_i}) \right) - 1 \approx \sum_{i=1}^{n} R_{\omega_i} + O\left(\sum_{i=1}^{n} \sum_{j=1}^{n} |R_{\omega_i} R_{\omega_j}| \right).$$

Summary

- A sequence of trades that increases c_t without changing a_t is almost always a good thing
- Trades $(s_t, q_t, \$p_t)$ will take place at discrete times τ
- P&L marked to p is $\$\Pi_T = \int_{t \in T} a_t^- \, dp$
- **Portfolio returns** are the percentage change in Π_t
- Portfolio returns are the dot product of **portfolio weights** and **asset returns**
- **Portfolio log-returns** are the change in $\log \Pi_t$
- Log-returns and returns are approximately equal when small
- Log-returns are additive over time

Market Microstructure

"Certainly, the modern compendium of mental illnesses (DSM-5) takes a dim view of people who think everyone is out to get them. Yet financial markets are different: people really are out to get you, after all." - Agustin Lebron, The Laws of Trading

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.
On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.
On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.
For each trade $(s, q, \$p)$, the exchange will typically charge a fee proportional to the **dollar volume** $\$pq$ of the trade. Fee rates may vary depending on trade type and between participants in accordance with exchange policy.

Trade Formation

In practice, the trades we can make at a time t and a price p_t are limited by our ability to find a willing counterparty.
On an electronic exchange, trades are formed by interacting with the exchange's **matching engine**.
For each trade $(s, q, \$p)$, the exchange will typically charge a fee proportional to the **dollar volume** $\$pq$ of the trade. Fee rates may vary depending on trade type and between participants in accordance with exchange policy.
The most common type of matching engine design is a **limit-order book** (sometimes called a double auction), which can operate in either a **continuous** or **batched** fashion.

Limit Order Book

At any point in time, market participants can create a request ('**limit order**') of the form $(s, q, \$p)$ to trade up to q units in direction $s = \pm 1$ at any price $\$(p - sm), m \geq 0$.
The value $\$m$ is known as the **price improvement**.
They are then said to be "**bid** for $\$p$" $(s = +1)$ or "**ask**ing/**offer**ing at $\$p$" $(s = -1)$.

Limit Order Book

At any point in time, market participants can create a request ('**limit order**') of the form $(s, q, \$p)$ to trade up to q units in direction $s = \pm 1$ at any price $\$(p - sm), m \geq 0$.
The value $\$m$ is known as the **price improvement**.
They are then said to be "**bid** for $\$p$" $(s = +1)$ or "**ask**ing/**offer**ing at $\$p$" $(s = -1)$.
All limit orders active at time t are collected into a **limit-order book** $\mathcal{L}_t$.
By convention, $\mathcal{L}_t$ is right-continuous with left limits.
Users can add, cancel and modify orders, subject to exchange-specific rules.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\text{max}} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

Order Matching

Whenever $(+1, q_1, \$p_1), (-1, q_2, \$p_2) \in \mathcal{L}_t$ with $p_2 \leq p_1$, both orders could be at least partly satisfied by trading up to $q_{\max} = \min(q_1, q_2)$ units with one another at a price $\$p \in [\$p_2, \$p_1]$. If such a pair exists the book is said to be **in cross**.

If an order $(s, q, \$p)$ is in cross with another, it may be **matched** for q' units. The total matched quantity for all buy orders must equal the total matched quantity for all sell orders.

In this case, q' units will trade, and the order will become $(s, q - q', \$p)$. If $q = q'$ the order is said to be **fully filled** and will be removed from the book. Otherwise, it is said to be **partially filled**.

The ability to quickly find matches for a large number of units at a reasonable price is known as **liquidity**, and is another major benefit of centralisation.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.
Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1, q', \$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1, q', \$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.
Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1, q', \$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1, q', \$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

The functions $Q_t(-1, \$p)$ and $Q_t(+1, \$p)$ are known as the **supply curve** and **demand curve** respectively. The function $M_t(\$p)$ represents the **matchable quantity** at $\$p$.

Supply and Demand Curves

We can partition $\mathcal{L}_t$ into $\mathcal{L}_t = \mathcal{B}_t \cup \mathcal{A}_t$, with $\mathcal{B}_t$ the bid orders and $\mathcal{A}_t$ the ask orders.
Now define the functions

$$Q_t(+1, \$p) = \sum_{\substack{(+1, q', \$p') \in \mathcal{B}_t \\ \$p \leq \$p'}} q'$$

$$Q_t(-1, \$p) = \sum_{\substack{(-1, q', \$p') \in \mathcal{A}_t \\ \$p \geq \$p'}} q'$$

$$M_t(\$p) = \min(Q_t(+1, \$p), Q_t(-1, \$p))$$

The functions $Q_t(-1, \$p)$ and $Q_t(+1, \$p)$ are known as the **supply curve** and **demand curve** respectively. The function $M_t(\$p)$ represents the **matchable quantity** at $\$p$. The book $\mathcal{L}_t$ is in cross if and only if there exists some $\$p$ with $M_t(\$p) > 0$.

Batch Matching

In **batch** or **auction** style matching, orders are matched with one another only at particular discrete times.

1. Prior to the **match time** t^*, users can typically add, modify and cancel limit orders.
2. At each time $t \leq t^*$, an **indicative price** $\$p_t^*$ will be selected such that $M_t(\$p_t^*)$ is maximal. Tiebreaking will depend on exchange rules.
3. Finally, at the match time t^*, some subset of the crossed limit orders will be matched at $\$p^*$ for a total quantity $M_{t^*}(\$p_{t^*}^*)$. After the match, the book will no longer be crossed.

Batch Matching

In **batch** or **auction** style matching, orders are matched with one another only at particular discrete times.

1. Prior to the **match time** t^*, users can typically add, modify and cancel limit orders.
2. At each time $t \leq t^*$, an **indicative price** $\$p_t^*$ will be selected such that $M_t(\$p_t^*)$ is maximal. Tiebreaking will depend on exchange rules.
3. Finally, at the match time t^*, some subset of the crossed limit orders will be matched at $\$p^*$ for a total quantity $M_{t^*}(\$p_{t^*}^*)$. After the match, the book will no longer be crossed.

Maximising $M_t(\$p_t^*)$ is equivalent to maximising the sum of $\$qm$ across all orders, where q is the quantity filled and m is the price improvement.
It is common to use this matching style at the beginning or end of a trading day or lunch break, or when there is some kind of market instability such as following a large price move or company announcement.
Sometimes t^* is referred to as a **liquidity event** because of the large volume traded, and the relative insensitivity of $\$p_{t^*}^*$ to individual orders.

Batch Matching Properties

The following monotonicity properties typically hold:

- $\$p_t^*$ nondecreasing in $\mathcal{B}_t$ and nonincreasing in $\mathcal{A}_t$
- For each $\$p$, $M_t(\$p)$ nondecreasing in $\mathcal{L}_t$
- For individual orders $(s, q, \$p)$, we will have $\$p_t^*$ nondecreasing in $\$p$ and sq.
- For individual orders $(s, q, \$p)$ and each $\$p'$, we will have $M_t(\$p')$ nondecreasing in q and nondecreasing in $\$sp$.

Batch Matching Properties

The following monotonicity properties typically hold:

- $\$p_t^*$ nondecreasing in $\mathcal{B}_t$ and nonincreasing in $\mathcal{A}_t$
- For each $\$p$, $M_t(\$p)$ nondecreasing in $\mathcal{L}_t$
- For individual orders $(s, q, \$p)$, we will have $\$p_t^*$ nondecreasing in $\$p$ and sq.
- For individual orders $(s, q, \$p)$ and each $\$p'$, we will have $M_t(\$p')$ nondecreasing in q and nondecreasing in $\$sp$.

Price Priority

Because $M_t(\$p')$ is nondecreasing in sp, the matching will be designed to obey **price priority**.
If we have two orders $(s_1, q_1, \$p_1), (s_2, q_2, \$p_2)$ with $\$s_1 p_1 > \$s_2 p_2$, then the second order cannot be matched unless the first is completely filled.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.
The match time is typically chosen at random in some short interval in order to disincentivise this behaviour.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.
The match time is typically chosen at random in some short interval in order to disincentivise this behaviour.
To further encourage early submission of orders, many exchanges also implement a time priority rule.

Time Priority

If two orders exist at the same price $\$p$, the one that reached the matching engine later cannot be matched unless the earlier order is completely filled.

Order Timing

The order book is often visible to all participants. Traders may be incentivised to wait until immediately before the match time to post orders. If everyone does this, the matching engine may be overloaded, and p_t^* will change very rapidly leading up to t^*.
The match time is typically chosen at random in some short interval in order to disincentivise this behaviour.
To further encourage early submission of orders, many exchanges also implement a time priority rule.

Time Priority

If two orders exist at the same price $\$p$, the one that reached the matching engine later cannot be matched unless the earlier order is completely filled.

Tick Size

Time priority would not have much effect if we could just insert the later order at a price $\$p + s\epsilon$ for some very small $\epsilon > 0$.
To avoid this, prices must be integer multiples of some small increment $\$\delta$,

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.
Price priority is still used, and time priority is usually used.
This is mainly used for intraday trading, night session trading, and 24/7 markets like crypto.
If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.

Continuous Matching

In **continuous matching**, a match time is triggered every time a new limit order causes the book to become crossed.
Price priority is still used, and time priority is usually used.
This is mainly used for intraday trading, night session trading, and 24/7 markets like crypto.
If the matching engine receives an order at t, then immediately before and after t the book will be uncrossed, with $M_t(\$p) = 0$ at all $\$p$.
The only orders involved in the match will be the arriving order and some set of orders $\mathcal{M}_t$ in the opposite direction.
Typically we are not allowed to match with ourselves. Often the exchange will implement **self-trade protection** so that the quantity of our active and passive orders is simply cancelled out without recording a trade.

Order Types

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.

Order Types

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.
Depending on the price, the active order may not be fully filled. The rest will be converted to a passive order. We can avoid this by specifying a **fill-or-kill** or **fill-and-kill** order type.

Order Types

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.
Depending on the price, the active order may not be fully filled. The rest will be converted to a passive order. We can avoid this by specifying a **fill-or-kill** or **fill-and-kill** order type.
If we don't care about price, we can insert a market order $(s, q, s\infty)$. The traded price can be arbitrarily bad for us, so this is rarely a good idea.

Order Types

The arriving order is known as the **active** or **aggressive** order, and the pre-existing orders are known as **passive**.
Depending on the price, the active order may not be fully filled. The rest will be converted to a passive order. We can avoid this by specifying a **fill-or-kill** or **fill-and-kill** order type.
If we don't care about price, we can insert a market order $(s, q, s\infty)$. The traded price can be arbitrarily bad for us, so this is rarely a good idea.
Passive orders give all participants the option to trade at a particular price in the opposite direction.
However, they disappear when traded against, so it is important to be fast if we think other traders will behave similarly to us.

Latency

If we are too slow, we will only make a trade when faster participants have decided it is unfavourable. This is known as **adverse selection**.

Latency

If we are too slow, we will only make a trade when faster participants have decided it is unfavourable. This is known as **adverse selection**.
If we are too slow to cancel our passive orders, they will be traded against even when we have already decided this is no longer favourable.

Latency

If we are too slow, we will only make a trade when faster participants have decided it is unfavourable. This is known as **adverse selection**.
If we are too slow to cancel our passive orders, they will be traded against even when we have already decided this is no longer favourable.
If we are slow to cancel, we will need to make the decision to cancel based on less information. This will increase the frequency with which we regret cancelling an order due to loss of queue priority.

Latency

If we are too slow, we will only make a trade when faster participants have decided it is unfavourable. This is known as **adverse selection**.
If we are too slow to cancel our passive orders, they will be traded against even when we have already decided this is no longer favourable.
If we are slow to cancel, we will need to make the decision to cancel based on less information. This will increase the frequency with which we regret cancelling an order due to loss of queue priority.
We can respond more quickly to events on the exchange by paying for **colocation**, writing faster code, using specialised hardware (fibre optic, FPGAs), or exploiting aspects of exchange design.
Latency can sometimes be nontrivial to quantify due to **clock drift**, which may need to be smoothed out.

Trades Resulting from Continuous Matching

For each q' matched against a passive order $(s, q, \$p)$, the active order will trade q' units with the passive order at $\$p$.
The per-unit price achieved by the active trader will be

$$\$p_t^* = \frac{\sum_{(s,q,\$p)\in\mathcal{M}_t} \$pq}{\sum_{(s,q,\$p)\in\mathcal{M}_t} q}.$$

Instantaneous Price Impact

If we aggressively trade a very large quantity, we will exhaust all passive orders we would most prefer to trade with and $\mathcal{M}_t$ will need to include orders at worse price levels. This is sometimes known as **walking the book**.

Assume continuous matching, and consider a market order of $q > 0$ units in direction s.

The least favourable price in $\mathcal{M}_t$ will be given by

$$\$P_t(sq) = s \min_{\{\$p : Q_t(-s,p) \geq q\}} \$sp.$$

The unit price of the match will be given by

$$\$p_t^*(sq) = \frac{1}{q} \int_0^q \$P_t(sq')dq'.$$

We call the sensitivity of $\$p_t^*$ to sq the **instantaneous price impact**.

Bid-Ask Spread

We call the prices

$$\begin{aligned} \$b_t &= \lim_{q \to 0^+} \$p_t^*(-q) &&= \max_{(+1, q, \$p) \in \mathcal{B}_t} \$p \\ \$a_t &= \lim_{q \to 0^+} \$p_t^*(q) &&= \min_{(-1, q, \$p) \in \mathcal{A}_t} \$p \end{aligned}$$

the **bid price** and **ask price** respectively. All bid orders have price at most $\$b_t$ and all ask orders have price at least $\$a_t$.
The interval $[\$b_t, \$a_t]$ is known as the **spread**, and $\$a_t - \$b_t$ is the **width** of the spread. If $\$a_t - \$b_t = \$\delta$, we say that the market for the asset is **large-tick** or **tick-constrained**.

Arbitrage

Suppose the same asset (or extremely similar assets) can be traded in two separate order books $\mathcal{L}_t^1, \mathcal{L}_t^2$ with spreads $\delta_1 = [\$b_1, \$a_1]$ and $\delta_2 = [\$b_2, \$a_2]$.

If $\delta_1 \cap \delta_2 = \emptyset$ (i.e. if $a_1 < b_2$ or $a_2 < b_1$), then $\mathcal{L}_t^1 \cup \mathcal{L}_t^2$ is in cross. We can buy from one book and immediately sell to the other book for a higher price. This is known as **cross-listing arbitrage**.

Our positions will probably not be netted by the exchange. So this may require collateral. If the order books get close together we may be able to get out of the position. If they get further apart we may require even more collateral, or need to exit the position at a loss.

We will also need to take into account fees. For a fee rate f we can adjust the spreads to be $[(1-f)\$b_t, (1+f)\$a_t]$.

If $\mathcal{L}_t^1$ or $\mathcal{L}_t^2$ changes before our order reaches the matching engine, we may not get the fill, or the price may no longer be favourable.

Even if $\delta_1 \cap \delta_2 \neq \emptyset$, it may be possible to make one of the trades passively, and make the other immediately after the first is filled.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.
This is variously called the **theo**retical price, **microprice** or **price proxy** depending on context.

Price Proxies

Note that $\$p_t^*(0)$ is not yet defined. So long as we choose some price $\$m_t$ satisfying $\$m_t \in [\$b_t, \$a_t]$, setting $\$p_t^*(0) := \$m_t$ will make $\$p_t^*(\cdot)$ nondecreasing.

This is variously called the **theo**retical price, **microprice** or **price proxy** depending on context.

It is well known that the last traded price $\$p_t$ exhibits **negative serial autocorrelation** of returns $\frac{dp}{p_t}$ (sometimes called **bid-ask bounce**).

Most price proxies exhibit some degree of statistical predictability that cannot be exploited for profit. This is known as **microstructural noise**.

Common Microprices

Some simple choices for m_t include:

- $\frac{1}{2}b_t + \$\frac{1}{2}a_t$ (arithmetic **midprice**)
- $\$\sqrt{b_t a_t}$ (**geometric midprice**)
- $\$(1 - I_t)b_t + \$I_t a_t$ (depth-$\$d$ **weighted midprice**)
- $\$b_t^{1-I_t} a_t^{I_t}$ (depth-$\$d$ **geometrically weighted midprice**)

We call I_t the **book imbalance**.

Common Microprices

Some simple choices for m_t include:

- $\frac{1}{2}b_t + \$\frac{1}{2}a_t$ (arithmetic **midprice**)
- $\$\sqrt{b_t a_t}$ (**geometric midprice**)
- $\$(1 - I_t)b_t + \$I_t a_t$ (depth-$\$d$ **weighted midprice**)
- $\$b_t^{1-I_t} a_t^{I_t}$ (depth-$\$d$ **geometrically weighted midprice**)

We call I_t the **book imbalance**.
A popular choice for this is

$$I_t = \frac{Q_t(+1, \$b_t)}{Q_t(+1, \$b_t) + Q_t(-1, \$a_t)}.$$

Alternative price proxies are described in the appendix.

Persistent Price Impact

We call the difference $\$\lambda_t(sq) = \$p_t^*(sq) - \$m_t$ the **instantaneous price impact curve** of trading q units in direction s.
Buy orders will have nonnegative instantaneous price impact, while sell orders will have nonpositive instantaneous price impact.
Because aggressive trades remove liquidity from one side of the book, there is also a persistent effect on $\mathcal{L}_t$ and consequently $\$m_t$. This is known as **persistent price impact**.
The realised instantaneous price impact is given by

$$\$\lambda_t = \$p_t - \$m_t,$$

while the realised persistent price impact is given by

$$\$\nu_t = \$m_t - \$m_t^{\emptyset},$$

where $\$m_t^{\emptyset}$ is the path the microprice process would have taken had we not interacted at all with the matching engine.

P&L with transaction costs

We can write $\$p_t = \$m_t^{\emptyset} + \$\nu_t + \$\lambda_t$.

$$\begin{aligned}\$\Pi_T &= \$\int_{t\in T} a_t^-(dm^{\emptyset} + dn + d\lambda) \\ &= \underbrace{\$\int_{t\in T} a_t^- dm^{\emptyset}}_{\text{Midprice P\&L}} - \underbrace{\$\int_{t\in T} \nu_t da}_{\text{PPI Penalty}} - \underbrace{\$\int_{t\in T} \lambda_t da}_{\text{IPI Penalty}} \\ &+ \underbrace{\$((\nu_{t_+} + \lambda_{t_+})a_{t_+} - (\nu_{t_-} + \lambda_{t_-})a_{t_-})}_{\$0 \text{ if } a_{t_+}=a_{t_-}=0}.\end{aligned}$$

Attempts to make ν_t and da covary negatively are very hard to pull off and usually considered manipulative.
But if we only use passive execution, it is guaranteed that λ_t and da will covary negatively, and this term will change from a loss to a profit. Trying to make money solely from this term is known as **market making** or **liquidity provision**.

Downsides of Market Making

Assuming $a_{t_-} = a_{t_+} = 0$,

$$\$\Pi_T = \underbrace{\$ \int_{t \in T} a_t^- \, dm^{\emptyset}}_{\text{Midprice P\&L}} - \underbrace{\$ \int_{t \in T} \nu_t da}_{\text{PPI Penalty}} - \underbrace{\$ \int_{t \in T} \lambda_t da}_{\text{IPI Penalty}}.$$

With a market-making strategy, we lose a lot of control over a_t^-.
If market participants in general is making money on this term we will tend to lose money. This tendency is another kind of **adverse selection**.
In particular, if the priority of our orders are quite low, we will only trade against the aggressive orders with the largest quantity, which tend to be most predictive of midprice changes over a short time horizon. High order priority is therefore extremely valuable for a market making strategy.
However, it is still possible to make money on both terms. This is particularly true if a_t changes relatively slowly (**low-frequency trading**).
Whether this results in better performance overall is a different question.

Market Making Strategy

A highly simplified model of optimal market making is given by Avellaneda and Stoikov (2008). The market maker maintains two limit orders at any time in opposite directions, of the form

$$\left(s, 1, \$\left(m_t - \gamma q - \frac{1}{2} s\varsigma)\right)\right),$$

where γ is a risk-aversion parameter[2] and ς is the difference between the two prices.
In general, to avoid large a_t, (which would make our P&L very sensitive to price changes), we want the amount we buy to match the amount we sell. It can also be useful to **hedge** with an opposing position in a correlated asset.

[2] Defined differently in the paper

Market Impact Modeling

A very simple model for ν_t takes the form

$$\nu_t = \int_0^t f(s_{t'} q_{t'}) G(t - t') dt',$$

for some $f, G : \mathbb{R} \to \mathbb{R}$. This is known as a **propagator model**.
Persistent impact is empirically seen to obey a **square-root law** (due to Grinold and Kahn, 1994),

$$|\nu| \propto \sigma \sqrt{\frac{|\int da|}{V}},$$

where σ is the typical variance of returns and V is the typical volume over the period.
Gatheral (2016) derives this from a variety of different models of market impact. Execution strategies are discussed here and in Johnson (2010).
Most reasonable market impact models are said not to admit price manipulation, in the sense that if $a_{t_-} = a_{t_+} = 0$, our P&L from price impact cannot be positive.

Summary

- Trades are usually formed by a matching engine maintaining a **limit order book**
- Matching can be **continuous** or **batched**
- The ability to find counterparties at a favourable price is called **liquidity**
- Orders in continuous matching can be **passive** or **active**
- Passive orders and slow active orders are subject to **adverse selection**
- The state of the order book lets us estimate the fair value of an asset using the **microprice**
- Actions we take have **price impact** that affecting the traded price and microprice

Portfolio Management

"Cars have brakes so you can drive faster." Ben Rady

Utility

We now consider probabilistic models of P&L with respect to probability measure $\mathbb{P}$, where processes are adapted to some filtration $\mathcal{F}_t$.
Suppose for simplicity that there are a finite number of possible P&L outcomes $\Pi^1, \Pi^2, \ldots \Pi^n$ between now and some later time t.
We can construct a $\mathcal{F}_t$-measurable function X that equals Π^i on a set of probability $\mathbb{P}(X = \Pi^i)$.
We might decide that for some pair of outcome distributions X, X' we prefer our P&L to follow X rather than X'.
A theorem of Von Neumann and Morgenstern (1947) shows that under quite reasonable assumptions about our preferences, there exists a function $u : \{\Pi^1, \Pi^2, \ldots \Pi^n\} \to \mathbb{R}$ such that we prefer X to X' if and only if

$$\mathbb{E}_{\mathbb{P}}[u(X)] > \mathbb{E}_{\mathbb{P}}[u(X')].$$

If different trading strategies produce different outcome distributions of P&L, we might therefore seek a **utility function** to compare them.

Compounding

Recall that

$$\ell_{t_+}^+ = \ell_{t_-}^- + \ell_T.$$

If we have a series of intervals $T_1, T_2, \ldots$, we can define $T^n = \bigcup_{i=1}^n T_i$.
Then we have $\ell_{T^n} = \sum_{i=1}^n \ell_{T_i}$.
Suppose we have two strategies, one with log-returns measure ℓ and another with log-returns measure ℓ'.
If we assume for simplicity that $\ell_{T_i} - \ell'_{T_i}$ are independent and identically distributed, it follows from the law of large numbers that

$$\frac{\ell_{T^n} - \ell'_{T^n}}{n} \to \mathbb{E}[\ell_{T_i}] - \mathbb{E}[\ell'_{T_n}].$$

This means that in the long-run, ℓ will exceed ℓ' with probability approaching one if and only if

$$\mathbb{E}[\ell_{T_i}] > \mathbb{E}[\ell_{T_i}].$$

We call the stochastic process ℓ_t the **Kelly utility**, due to Kelly (1956).

Kelly Criterion

The Kelly utility of our P&L over some interval T is given by

$$u_T = \ell_T = \log(1 + R_T)$$

In practice, it may make sense to use the more conservative **fractional Kelly criterion**,

$$u_T = \log\left(1 + \frac{R_T}{L}\right),$$

where $L \in [0, 1]$ is a leverage parameter.
For small values of $\frac{R_T}{L}$, we can approximate this using a Taylor expansion:

$$\mathbb{E}[\log(1 + \frac{R_T}{L})] \approx \frac{1}{L}\mathbb{E}[R_T] - \frac{1}{2L^2}\mathbb{E}[R_T^2] + \mathbb{E}[O(\frac{R_T^3}{L^3})].$$

The quadratic approximation is known as the **quadratic utility** of the return on capital.

Quadratic Utility

If the central limit theorem applies to the u_{T_i}, then in the limit the distribution of u_{T^n} will be governed only by the first two moments. Recall that over certain intervals T', $R_{T'}$ is given by $w \cdot r_{T'}$, where w is the vector of portfolio weights. If w is deterministic, the quadratic utility over T' is given by

$$\frac{1}{L} w \cdot \mathbb{E}[r_{T'}] - \frac{1}{2L^2} w^\top \mathbb{E}[r_{T'} r_{T'}^\top] w,$$

and the optimal value of w is given by

$$w^* = L\mathbb{E}[r_{T'} r_{T'}^\top]^{-1} \mathbb{E}[r_{T'}] = L(\Sigma + \mu\mu^\top)^{-1}\mu,$$

where μ and Σ are the expectation and covariance of $r_{T'}$.

Markowitz Portfolio Optimisation

In practice, μ is relatively small compared to Σ, and so we have

$$w* = L(\Sigma + \mu\mu^\top)^{-1} \approx L\Sigma^{-1}\mu.$$

We can get this by maximising the quadratic loss function

$$Lw \cdot \mu - \frac{1}{2} w^\top \Sigma w.$$

This is equivalent to maximising the ratio

$$\frac{\mathbb{E}[R_{T'}]}{\sqrt{\mathrm{Var}[R_{T'}]}},$$

known as the **Sharpe ratio**, subject to $\mathrm{Var}[R_{T'}]$ having some desired positive value.
The problem of maximising expected returns subject to an upper bound on returns variance is known as **Markowitz portfolio optimisation**.

Estimation

If we have a number of returns vectors $r_1, r_2, \ldots r_n$, we might try to invoke the law of large numbers and write

$$\mu \approx \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} r_n,$$

$$\Sigma \approx \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} r_n r_n^\top.$$

Unfortunately, n might be smaller than the number of assets p. In this case, the matrix $[r_1, r_2, \ldots r_n]^\top$ will have a nullspace containing portfolios whose estimated volatility is 0. Portfolios close to this nullspace will have arbitrarily large Sharpe ratios, and no optimal portfolio will exist, since we cannot solve $\Sigma w = L\mu$.
Furthermore, even if $n > p$, errors in the estimation may have a significant effect on the expected utility out-of-sample.

Regularisation

We can augment the loss function slightly by adding a **regularisation penalty**,

$$Lw \cdot \mu - \frac{1}{2} w^\top \Sigma w - \frac{\lambda}{2} \|w\|_2^2,$$

which is guaranteed to admit an optimum $w*(\lambda) = L(\Sigma + \lambda I)^{-1}\mu$.
If we compute the SVD $\Sigma = UDU^\top$, we will have

$$\lim_{\lambda \to 0^+} w*(\lambda) = UD^+U^\top \mu,$$

where D^+ replaces the nonzero elements of D with their inverse. It may be good to z-score the returns first.
Even though these produce a solution, the out-of-sample utility may still be unsatisfactory due to estimation error.
We need to make use of known qualitative facts about the distribution of asset returns to better estimate the covariance matrix.

Summary

- We want to maximise reward per unit risk
- In theory we hold a portfolio $L\Sigma^{-1}\mu$
- In practice Σ is hard to estimate

Factor Models

"Every researcher has their own research process. This is part of their competitive advantage; it's indeed part of what they are, of thoughts and learned lessons accumulated over a lifetime of experiences and studying."
Giuseppe Paleologo

Capital Asset Pricing Model

There are certain common factors that jointly influence all stocks.
In the Capital Asset Pricing Model (CAPM), due initially to Treynor (1961), the returns[3] on the jth asset over T_i are given by

$$r_i^j = \underbrace{\beta^j}_{\text{Asset Sensitivity}} \underbrace{f_i - \mathbb{E}[f_i]}_{\text{Market Return}} + \underbrace{\alpha^j + \sigma^j \epsilon_i^j}_{\text{Idiosyncratic Return}},$$

where ϵ_i^j have mean zero and unit variance, and are uncorrelated with the f_i and with each other.
We have

$$\Sigma = \beta \Sigma_f \beta^\top + \Sigma_\epsilon,$$
$$\mu = \alpha,$$

where α, β are populated by α^j, β^j, Σ_ϵ is a diagonal matrix with entries σ^{j2}, and Σ^f is the variance of f_i.

[3] Traditionally, these are excess returns relative to some benchmark interest rate.

Spanned and Orthogonal α

Let $H_\beta = \beta(\beta^\top\beta)^{-1}\beta^\top$ be the projection matrix onto the span of β. Then

$$\alpha = \underbrace{H_\beta\alpha}_{\text{Spanned}} + \underbrace{(I - H_\beta)\alpha}_{\text{Orthogonal}} = \beta\alpha^\beta + \alpha^\perp,$$

where $\alpha^\perp$ is orthogonal to β.

Pervasiveness of $\alpha^\top$

If we set $w = \alpha^\perp$, we will have $w \cdot \mu = \|\alpha^\perp\|_2^2$ and $w^\top \Sigma w = \|\Sigma_\epsilon^{\frac{1}{2}} \alpha^\perp\|_2^2$.
The Sharpe ratio of this portfolio is

$$\frac{\|\alpha^\perp\|_2^2}{\|\Sigma_\epsilon^{\frac{1}{2}} \alpha^\perp\|_2}.$$

As $p \to \infty$, this grows as

$$\frac{p^2 \overline{(\alpha^{\perp j})^2}}{p\sqrt{\overline{(\sigma^j \alpha^{\perp j})^2}}} = p \frac{\overline{(\alpha^{\perp j})^2}}{\sqrt{\overline{(\sigma^j \alpha^{\perp j})^2}}}$$

Since an infinite Sharpe ratio would allow riskless profits (arbitrage), it must be the case that $\frac{\overline{(\alpha^{\top j})^2}}{\sqrt{\overline{(\sigma^j \alpha^{\top j})^2}}}$ decays to zero faster than p^{-1} as we add more assets.
Assuming $\sigma^j < \sigma^{\max}$ for all j, this means that $\overline{(\alpha^{\top j})^2} \to 0$ faster than p^{-2}.
Then $r_i \approx \beta((f_i - \mathbb{E}[f_i]) + \alpha^\beta) + \sigma^j \epsilon_i^j$. We call $(\beta^\top \beta)^{-1} \beta^\top \alpha^\beta$ the **risk premium** of the factor.

Factor Models

In general, we replace β with a matrix B and f_i becomes a vector. Then we have

$$r_i = B(f_i - \mathbb{E}[f_i] + \alpha^B) + \alpha^\perp + \Sigma_\epsilon^{\frac{1}{2}} \epsilon_i.$$

Centering the factor returns f_i will not affect the model, so we can assume $\mathbb{E}[f_i] = 0$.

$$r_i = B(f_i + \alpha^B) + \alpha^\perp + \Sigma_\epsilon^{\frac{1}{2}} \epsilon_i.$$

Then we have

$$\Sigma = B\Sigma_f B^\top + \Sigma_\epsilon,$$
$$\mu = \alpha.$$

We say that the covariance matrix of r_i is **spiked**, meaning that it has a low-rank approximation.

"Rotational" Indeterminacy

If we construct matrices $R, F, \alpha^{\beta}, \alpha^{\perp}, \varepsilon$ whose ith columns are $r_i, f_i, \alpha^{\beta}, \alpha^{\perp}, \epsilon_i$ respectively, then we have

$$R = B(F + \alpha^{B}) + \alpha^{\perp} + \Sigma_{\epsilon}^{\frac{1}{2}}\varepsilon = BF + \mathbb{E}[R] + \Sigma_{\epsilon}^{\frac{1}{2}}\varepsilon = BCC^{-1}F + \mathbb{E}[R] + \Sigma_{\epsilon}^{\frac{1}{2}}\varepsilon,$$

for any invertible square matrix C. The invariance to C is sometimes called **rotational indeterminacy**.

Estimation

Assuming the ε are jointly normal and B has linearly independent columns, the following parameter estimates are maximal likelihood:

$$
\begin{aligned}
\mathbb{E}[r_i] \approx \overline{R} &= \frac{1}{n}\sum_{i=1}^{n} r_i, \\
F \approx \hat{F} &= (B^\top \Sigma_\epsilon^{-1} B)^{-1} B^\top \Sigma_\epsilon^{-1}(R - \overline{R}), \\
\varepsilon\hat{\varepsilon} &= \Sigma_\epsilon^{-1}(R - \overline{R} - BF), \\
\Sigma_\epsilon \approx \hat{\Sigma}_\epsilon &= \frac{1}{n}((R - \overline{R} - BF)(R - \overline{R} - BF)^\top), \\
B \approx \hat{B}_{\mathrm{FM}} &= (R - \overline{R})F^\top (FF^\top)^{-1}, \text{ if } F \text{ is known, or} \\
B \approx \hat{B}_{\mathrm{PPCA}} &= \Sigma_\epsilon^{\frac{1}{2}} U_m,
\end{aligned}
$$

where U_m consists of the first m principal components of $\Sigma_\epsilon^{-\frac{1}{2}}(R - \overline{R})$. This is Probabilistic Principal Component Analysis, due to Tipping and Bishop (1999).

Fundamental Factor Models

We don't necessarily need to estimate B. We can use **characteristics** of the various assets. Some common choices:

- 1 if the asset belongs to a particular country or industry, 0 otherwise
- All 1s (equal weighting)

It is also possible to extend the model to include characteristics that vary day-to-day, such as

- Realised volatility
- Illiquidity (e.g. $\left(\frac{|\text{Daily Returns}|}{\text{Daily Volume}}\right)$, due to Amihud (2002))
- Crowding
- Market capitalisation
- Recent returns (momentum)
- Numbers derived from financial reports, e.g. net asset value
- Options Greeks $(\Delta, \theta, \nu, \Gamma, \ldots)$

It is usually best to ensure the characteristics are linearly independent.

If B and Σ_ϵ^{-1} are known but F are estimated, an unbiased estimator for the factor covariance Σ_f is given by $\frac{1}{n}FF^\top - (B^\top \Sigma_\epsilon^{-1} B)^{-1}$.

Summary

- The cross-section of asset returns is partly described by various **risk factors**
- Identifying these factors can help us find a low-rank approximation to the covariance matrix, plus a diagonal term
- This greatly simplifies covariance estimation

Dynamic Portfolio Selection

"Values which do not 'yet' exist, except as probabilistic estimations, or risk structures, acquire a power of command over economic (and therefore social) processes, necessarily devalorizing the actual." Nick Land, Teleoplexy

Market Efficiency

If w is fixed, da will be very small.[4]
If all investors hold a fixed optimal portfolio $w*$, the total dollar value invested in each asset will be proportional to w^*.
However, most assets give their holder the right to potential future cashflows ('dividends'). These dividends are affected by things in the real world, which we can improve our forecast of over time.
If market participants are generally good at forecasting dividends, prices will reflect dividend expectations.[5]
Forecasting is expensive. We could just look at total dollar value invested in each asset to infer a dynamic w^*. This is known as **passive investing**.
If everybody did this, w^* would be constant and stop being a good forecast of dividends. This is known as the **Grossman-Stiglitz paradox**.

[4] More precisely, w_t is fixed, but $a_t = \Pi_t \frac{w_t}{p_t}$ may need to change a little as p_t changes.

[5] The result of price impact is to make such forecasts a public good, in the microeconomic sense.

Structural Edge

Price impact and fees make trading expensive. We should only do it if we think we can do better than passive investing.

- Access to multiple markets simultaneously, or ability to move information between markets faster than others
- Access to better market data
- Access to more historical data for model estimation purposes
- Ability to provide liquidity without losing to adverse selection, e.g. PFOF
- Lower taxes
- More favourable terms with exchange/broker for fees, margin requirements, funding rates, borrow, etc.
- Market-maker rebates from the exchange
- Economies of scale
- ...

Skill Edge

- Responding to publicly available information faster than prices can move to reflect it
- Better understanding of exchange mechanics
- Knowledge of proprietary datasets other than market data ('alternative data')
- Knowledge of typical market dynamics (from experience)
- More intelligent pricing and/or execution (e.g. intelligent placement of orders to get good queue position)
- Being able to know (or guess) who is making particular trades
- Better models (to see things humans don't)
- Better human traders (to see things models don't)
- Better software for humans to interface with models
- Better risk management
- ...

Linear Regression

Suppose that

$$R = BF + \alpha + \Sigma_\epsilon^{\frac{1}{2}} \epsilon,$$

and that f_i are known in advance of r_i.
If we view this as a factor model, we can recall the Fama-MacBeth estimate for B,

$$B \approx \hat{B}_{\mathrm{FM}} = (R - \overline{R})F^\top (FF^\top)^{-1}.$$

Then we can forecast returns according to

$$\mathbb{E}[r_i] = Bf_i + \alpha \approx (R - \overline{R})F^\top (FF^\top)^{-1} + \overline{R}.$$

The covariance of r_i conditional on f_i is Σ_ϵ.
It follows that the optimal portfolio conditional on f_i has the form

$$w = L\Sigma_\epsilon^{-1}(Bf_i + \alpha) \approx L\Sigma_\epsilon^{-1}((R - \overline{R})F^\top (FF^\top)^{-1} f_i + \overline{R}).$$

In the most general case, Σ_ϵ need not be diagonal.

Feature Selection

Variables that let us forecast the value of a **target variable** are called **features**.
As a general heuristic, we want to choose features such that the target variable will have a different mean for different values of the feature.
Especially useful are features that have nonnegligible correlation ρ with the target, since we can then use a simple linear model for forecasting.
However, we are not always so lucky.
This is a bit harder than factor selection. It's easier to model how markets go right than how they go wrong. *"All happy families are alike", Tolstoy.*
Even when we do find a useful feature, exploiting it may make it easier for other people to identify the anomaly. As more people begin to use a signal, it will experience **alpha decay**.

Feature Engineering

Once we have raw numbers, we can create new features by applying various transformations.

- Applying a fixed nonlinear function like $\tan^{-1}$ or $x \mapsto x^2$
- Setting large values of the feature to some constant (winsorisation)
- Multiplying features together (interaction terms, used in Friedman's MARS)
- Returning 1 if the feature exceeds some value, 0 otherwise (as in decision trees)
- Computing a linear combination of features then applying a nonlinear transform (as in neural networks)
- Computing a weighted average of variables over time (e.g. EWMA)

Model Constraints

Interpretability

It may be desirable to restrict ourselves to features that are **interpretable** in some way.
This can help us direct our research in a more fruitful direction.
Constraining our portfolio in accordance with an interpretable risk model is especially helpful for peace of mind out-of-sample.
However, for the purposes of α forecasting, it is not quite as important that our final model be fully interpretable.

Regularisation

It can be useful to penalise or constrain the model in some way to improve out-of-sample generalisation, such as with a ridge penalty or a monotonicity constraint.
This will sometimes require hyperparameter optimisation, e.g. with Leave-One-Out Cross-Validation (LOOCV) and a gradient-free optimiser (gaussian process, CMAES, genetic algorithm, etc.).

Signal Aggregation

In general, the best estimates for Σ and α should vary with time based on data relevant to the assets we are trading.
Consider a single asset A with returns $r_1^A, r_2^A, \ldots r_n^A$.
Suppose we have a vector x_i of many different model forecasts for r_i^A.
Construct a **feature matrix** $X = [x_1, x_2, \ldots x_n]^\top$.
We can hold asset j in proportion to some linear combination of x_i given by $x_i \cdot \beta$.
Let $F = [r_1^A x_1, r_2^A x_2, r_n^A x_n]^\top$ and $\overline{F} = \frac{1}{n} \sum_{i=1}^{n} r_i^A x_i$.
Assuming no transaction costs, our mean quadratic loss over the entire dataset will be given by

$$L\overline{F} \cdot \beta - \frac{1}{2n} \beta^\top F^\top F \beta,$$

and the optimal β is given by

$$L \left(\frac{1}{n} F^\top F \right)^{-1} \overline{F}.$$

Sharpe Ratio Hypothesis Testing (In-Sample)

The portfolio return for the ith time period will be $R_i^A = r_i^A x_i \cdot \beta$.

We can compute the Sharpe ratio $\lambda = \frac{\overline{R_i^A}}{\sqrt{\overline{(R_i^A - \overline{R_i^A})^2}}}$.

If we assume that $r_1^A x_1, r_2^A x_2, \ldots$ are approximately i.i.d. normal with mean zero, we will have

$$\frac{n^2(n-p)}{p(n-1)^2}\lambda^2 \sim F_{p,n-p}.$$

The test statistic will be smaller when p is larger, and larger when n is larger.

Sharpe Ratio Hypothesis Testing (Out-Of-Sample)

We can also compute the Sharpe ratio out-of-sample. Making the weaker assumption that $R_1^A, R_2^A, \ldots$ are approximately i.i.d. normal with mean zero, we have

$$\frac{n}{\sqrt{n-1}}\lambda \sim t_{n-1}.$$

If we regularise the covariance matrix, the in-sample distribution of λ does not have an analytic form, but p-values can be obtained by parametric bootstrap. Bounds on the test statistic distribution are described in Issouani, Bertail and Gautherat (2024).
If we test multiple feature sets and take the best, our results will be inflated. This is known as the problem of multiple testing. One remedy is the **Rademacher Anti-Serum** (Paleologo, 2025).

Transaction Cost Analysis

Recall that

$$\$\Pi_T = \underbrace{\$ \int_{t\in T} a_t^- \, dm^{\emptyset}}_{\text{Midprice P\&L}} - \underbrace{\$ \int_{t\in T} \nu_t da}_{\text{PPI Penalty}} - \underbrace{\$ \int_{t\in T} \lambda_t da}_{\text{IPI Penalty}}.$$

We can use a modified propagator model for $\nu_t + \lambda_t$,

$$\nu_t + \lambda_t = p_t \int_{t_-}^{t} \lambda p_{t'} G(t - t') da(t'),$$

with $G(0) = 1$.
The P&L lost to price impact is

$$\int_{t\in T} p_t \int_{t_-}^{t} \lambda p_{t'} G(t-t') da(t') da(t) = \int_{t\in T} \Pi_t \int_{t_-}^{t} \lambda \Pi_{t'} G(t-t') dw(t') dw(t).$$

Transaction Cost Optimisation

Let $dX = [0, x_2 - x_1, x_3 - x_2, \ldots x_n - x_{n-1}]^\top$. Furthermore, define an upper-triangular matrix Λ whose i, j entry is $\lambda G(t_j - t_i)$.
If we assume portfolio value $\$\Pi_t$ is roughly equal to a constant $\$\pi$, and ignore the contribution of $\nu_t + \lambda_t$ to P&L variance, our quadratic utility becomes

$$L\overline{F} \cdot \beta - \frac{L\pi^2}{n}\beta^\top dX^\top \Lambda dX \beta - \frac{1}{2n}\beta^\top F^\top F \beta,$$

and the optimal β is given by

$$\beta = L\left(\frac{2L\pi^2}{n} dX^\top \Lambda dX + \frac{1}{n}F^\top F\right)^{-1} \overline{F}.$$

Note that as $\pi \to \infty$ we have $\beta \to 0$.
More sophisticated execution can be achieved with techniques like reinforcement learning.

Statistical Arbitrage

Consider the modified factor model

$$r_i = Bf_i + \Theta_0\epsilon_i + \Theta_1\epsilon_{i-1} = \begin{bmatrix} B & \Theta_0 & \Theta_1 \end{bmatrix} \begin{bmatrix} f_i \\ \epsilon_i \\ \epsilon_{i-1} \end{bmatrix},$$

where B is known, f_i are i.i.d. multivariate normal with mean zero and covariance Σ_f, Θ_0, Θ_1 are known diagonal matrices, and ϵ_i are i.i.d. standard multivariate normal.

We can apply a Kalman filter to forecast r_i (see appendix). Over the interval $(t_i, t_{i+1}]$, we will hold a portfolio

$$w = L(\Theta_1 P_i^\epsilon \Theta_1 + B\Sigma_f B^\top + \Theta_0^2)^{-1}\Theta_1\hat{\epsilon}_i,$$

where $\hat{\epsilon}_i$ is the filtered value of ϵ_i and P_i^ϵ is the uncertainty matrix of the filtering.

As the number of assets $p \to \infty$, we will be able to infer f_i exactly.

The residual returns $r_i - Bf_i = \Theta_0\epsilon_i + \Theta_1\epsilon_{i-1}$ follow an independent $MA(1)$ process for each asset. We can filter ϵ_i[6] as

$$\sum_{n=1}$$

Summary

- Doing better than a fixed portfolio is nontrivial but valuable
- We need to identify predictive signals and combine them in an optimal way
- Optimisation should take into account the price impact of these signals

Appendix: Accounting

Proof Sketch for c_T Identity

$$
\begin{aligned}
c_T = \sum_{t \in \tau} -p_t(s_t q_t) \quad &= \sum_{i=1}^{n} -p_{t_i}(a^+_{t_i} - a^-_{t_i}) = -\sum_{i=1}^{n} p_{t_i} a^+_{t_i} + \sum_{i=1}^{n} p_{t_i} a^-_{t_i} \\
&= -\sum_{i=1}^{n-1} p_{t_i} a^-_{t_{i+1}} - p_{t_n} a^+_{t_n} + \sum_{i=1}^{n-1} p_{t_{i+1}} a^-_{t_{i+1}} + p_{t_1} a^-_{t_1} \\
&= p_{t_1} a^-_{t_1} + -p_{t_n} a^+_{t_n} + \sum_{i=2}^{n} (p_{t_i} - p_{t_{i-1}}) a^-_{t_i} \\
&= p_{t_1} a^-_{t_1} - p_{t_n} a^+_{t_n} + \int_{t \in [t_1, t_n]} a^-_t \, dp \\
&= p_{t_-} a_{t_-} - p_{t_+} a_{t_+} + \int_{t \in T} a^-_t \, dp.
\end{aligned}
$$

Annualised Returns

The **annualised log-return** over ω is $\ell_\omega \frac{1 \text{ year}}{\lambda_\omega}$, where λ_ω is the duration (lebesgue measure) of ω in units of time.
The **geometrically annualised return** over ω is
$(1 + R_\omega)^{\frac{1 \text{ year}}{\lambda_\omega}} - 1 = \exp\left(\ell_\omega \frac{1 \text{ year}}{\lambda_\omega}\right) - 1.$
The **arithmetically annualised return** over ω is $R_\omega \frac{1 \text{ year}}{\lambda_\omega}$.

Appendix: Microprice

Alternative Price Proxies

More generally, we can define the depth-$\$d$ imbalance,

$$I_t(\$d) = \frac{Q_t(+1, \$b_t - \$d)}{Q_t(+1, \$b_t - \$d) + Q_t(-1, \$a_t + \$d)}.$$

We can also define an **exponentially weighted imbalance**,

$$I_t(\zeta) = \frac{\sum_{(+1,q,p)\in\mathcal{B}_t} q \exp(-\zeta|b_t - p|)}{\sum_{(+1,q,p)\in\mathcal{L}_t} q \exp(-\zeta \min(|b_t - p|, |a_t - p|))}$$

Appendix: PPCA Stabilisation

Stabilising $\hat{B}_{\text{PPCA}}$

In practice, our estimate for B may change over time. If we have PPCA estimates at two different times, $\hat{B}^t_{\text{PPCA}}, \hat{B}^{t+1}_{\text{PPCA}}$, we may wish to rotate the latter to be close to the former.
To achieve this, we can use the estimate $\hat{B}^{t+1}_{\text{PPCA}} VU^\top$, where $UDV^\top$ is the SVD of ${B^t_{\text{PPCA}}}^\top B^{t+1}_{\text{PPCA}}$.

Appendix: Statistical Arbitrage

Kalman Filter

The state is

$$x_i = \begin{bmatrix} f_i \\ \epsilon_i \\ \epsilon_{i-1} \end{bmatrix}.$$

Maintain state estimates $\hat{x}_i = \begin{bmatrix} \hat{f}_i \\ \hat{\epsilon}_i \\ \hat{\epsilon}_{i-1} \end{bmatrix}$ and uncertainty matrix

$$P_i = \begin{bmatrix} P_i^f & \dots & \dots \\ \dots & P_i^\epsilon & \dots \\ \dots & \dots & P_{i-1}^\epsilon \end{bmatrix}.$$

Initialise $\hat{x}_i$ to zero and P_i to

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We have $r_i = Hx_i$, where

Appendix: Asset Pricing

Arrow-Debreu Securities

An **event** is a set of possible worlds distinguished from other possible worlds by some common feature.
For instance, there is a set of possible worlds where Australia wins its next game of cricket against the UK.
Suppose Ω is the set of all events.[7] At each time t', we can construct a set $\mathcal{F}_{t'} \subseteq \Omega$ containing all events whose outcome is known at time t'.[8]
An **Arrow-Debreu security** A_ω for $\omega \in \mathcal{F}_t$ is one which we can exchange at time t for \$1 if ω occurs and \$0 otherwise. Assume there exists t such that $\omega \in \mathcal{F}_t$.
Let $\$\mathbb{Q}_{t'}(\omega)$ be the market price of ω at time $t' \leq t$. Assume unlimited liquidity, no fees, and no collateral requirements or interest payments.
The market is **arbitrage-free** if there is no fixed set of trades that guarantees positive P&L in all possible worlds.
Assuming this is the case, we can deduce properties that $\mathbb{Q}_{t'}(\cdot)$ must satisfy.

[7] We require that Ω be a σ-algebra.
[8] $\mathcal{F}_t$ is a filtration, i.e. a nondecreasing σ-algebra

Dutch Book Theorems

$\mathbb{Q}_{t'}(\omega) \in [0, 1]$

If $\mathbb{Q}_{t'}(w) < 0$, we would actually receive money by buying A_ω, and we can sell it at t for at least \$0. This would guarantee positive P&L.
If $\mathbb{Q}_{t'}(w) > 1$, we could short A_ω, and buy it back at t for at most \$1. This would guarantee positive P&L.

If ω_1 and ω_2 are disjoint sets, we say they are **mutually exclusive events**.

$\mathbb{Q}_{t'}(\omega_1) + \mathbb{Q}_{t'}(\omega_2) = \mathbb{Q}_{t'}(\omega_1 \cup \omega_2)$

If the LHS exceeds the RHS, we buy $A_{\omega_1}, A_{\omega_2}$ and short $A_{\omega_1 \cup \omega_2}$; otherwise, we do the reverse.

An arbitrage strategy on Arrow-Debreu securities is sometimes called a **Dutch Book**, and properties derived from arbitrage-freeness are known as **Dutch Book Theorems**.

Fundamental Theorem of Asset Pricing

The market for Arrow-Debreu securities is arbitrage-free if and only if $\mathbb{Q}_{t'}(\cdot)$ obeys the axioms of Kolmogorov (1950), i.e. is a **probability measure** on $\mathcal{F}_t$.
In general, we can consider securities that are redeemable at time t for some amount given by a $\mathcal{F}_t$-measurable function (**random variable**) S.
A market is arbitrage-free at time t' if and only if the market price of each security S is given by

$$\mathbb{E}_{\mathbb{Q}_{t'}}[S] = \int_{\Omega} S d\mathbb{Q}_{t'},$$

where $\mathbb{Q}_{t'}$ is some probability measure on $\mathcal{F}_t$. This is known as the **fundamental theorem of asset pricing** (FTAP).

$\mathbb{Q}$-world vs $\mathbb{P}$-world

A set of securities M is said to form a **complete market** for Ω if and only if we can infer the arbitrage-free prices of every Arrow-Debreu security from the prices of M.
In practice, even if we erroneously assume identical bid and ask prices, markets are only ever complete for extremely trivial Ω. This means there are many different measures $\mathbb{Q}$ that are consistent with the prices of M.
The only way to see if asset prices are consistent with one another is by constraining them with some set of **model assumptions** to induce a particular probability measure $\mathbb{P}$.
This is one extremely general description of the family of strategies known as **statistical arbitrage**. Favourable outcomes are no longer guaranteed.

Appendix: Research Process

Data Cleaning

Prices may need to be adjusted to reflect corporate actions:

- Dividends
- Stock split, reverse stock split
- MA, spinoff
- Rights issues, tender offer, warrant issue
- Exchange offer
- Public offering, share buyback
- Liquidation
- Delisting
- Debt to equity conversion

Backtesting

Walkforward backtesting is best practice for avoiding bugs. Large sections of production code should ideally be identical to the code used for strategies in backtesting.
Market impact models should be calibrated against realised results.
Borrow costs, fees, taxes, etc. should be accounted for if non-negligible.
It is important to avoid survivorship bias and lookahead bias in backtests.
Particularly important for low-frequency equities trading is the possibility of delisted securities not showing up in old historical data.
Historical data may also have been corrected, redacted, or unredacted after its initial publication.

Bibliography

Bibliography