EC794: Financial Econometrics

Notes

Anming Gu
Boston University
agu2002@bu.edu

May 16, 2023

Abstract

This work covers the materials I learned from Boston University's PhD class in financial econometrics, EC794, taught by Professor Zhongjun Qu during Spring 2023. The class covers theoretical and empirical topics on the efficient market hypothesis, asset pricing (low- and high-frequency), capital asset pricing model, multifactor models (arbitrage pricing theory, Fama-French), analysis of the stochastic discount factor, and continuous-time models (Black-Scholes).

Contents

1 Asset Return Predictability

The readings for this chapter are

- Campbell J.Y. 2014. "Empirical Asset Pricing: Eugene Fama, Lars Peter Hansen, and Robert Shiller.' Scandinavian Journal of Economics. Sections 3, 4.
- Lewellen, J. 2004. "Predicting Returns with Financial Ratios." Journal of Financial Economics 74, 209-235.

In this chapter, we focus on aggregate stock returns in the low-frequency setting, *i.e.*, monthly and lower frequency. We examine return predictability in the context of the efficient market hypothesis (EMH). Our main econometric focus is valid inference for predictive regressions.

Let $\{P_t\}_1^T$ be a sample of prices. *Simple returns* are calculated as $r_{t+1} = \frac{P_{t+1}-P_t}{P_t}$, and *continuously compounded returns* are calculated as $r_{t+1} = \log P_{t+1} - \log P_t$.

1.1 The Efficient Market Hypothesis

A market in which asset prices always fully reflect available information is called *efficient*. This is an informal characterization of the efficient market hypothesis (EMH).

Let P_{t+1} be the vector of payoffs at time $t+1$ (prices plus dividends and interest payments) on assets available at time t, $\mathcal{F}_t$ be the information available at time t, $f(P_{t+1} \mid \mathcal{F}_t)$ be the conditional distribution of P_{t+1} at t, $\mathcal{F}_{mt}$ be the information used in the market to set price P_t,[1] $f(P_{t+1} \mid \mathcal{F}_{mt})$ be the distribution of P_{t+1} implied by $\mathcal{F}_{mt}$, and R_{mt} be the expected returns of $t+1$ implied by $f(P_{t+1} \mid \mathcal{F}_{mt})$ and P_t.

Note that $f(P_{t+1} \mid \mathcal{F}_m t)$ depends on the assumed model of market equilibrium. This distribution is undefined unless we specify a model.

The *efficient market hypothesis* that the prices at time t fully reflect available information is

$$f(P_{t+1} \mid \mathcal{F}_t) = f(P_{t+1} \mid \mathcal{F}_{mt}),$$

which implies $r_{t+1} = R_{mt} + u_{t+1}$ and $\mathbb{E}[u_{t+1} \mid \mathcal{F}_t] = 0$, where r_{t+1} and R_{mt} can be either simple or continuously compounded returns.

Remark 1.1. Efficiency means information efficiency. The EMH implies that it is impossible to make economic profits by trading on the basis of $\mathcal{F}_t$. This is potentially testable.

We want to test the hypothesis $\mathbb{E}[u_{t+1} \mid \mathcal{F}_t] = 0$ where $u_{t+1} = r_{t+1} - R_{mt}$. However, this isn't immediately testable because R_{mt} isn't observed. We need a model of market equilibrium.

Example 1.2. If the equilibrium model is such that $R_{mt} =$ constant, then the EMH implies $\mathbb{E}[r_{t+1} \mid \mathcal{F}_t] =$ constant. This is testable. However, if the test rejects, we don't know whether the problem is an inefficient market or a misspecified market equilibrium model.

In this example, the H_0 is $\mathbb{E}[r_{t+1} \mid \mathcal{F}_t] =$ constant. Let $\Omega_t \subsetneq \mathcal{F}_t$. Then H_0 implies $\mathbb{E}[r_{t+1} \mid \Omega_t] =$ constant due to the law of iterated expectations.

The literature has considered three types of tests for this hypothesis, *weak form test*, *semi-strong form test*, and *strong form test*. We will focus on the semi-strong form test. In this test, Ω_t includes information that is publicly available, *e.g.*, valuation ratios, announcements of annual earnings, stock splits, etc.

Remark 1.3. We can have three types of random walks.

I. $p_t = \mu + p_{t-1} + e_t$, where $e_t \sim \mathcal{N}(0, \sigma^2)$ *i.i.d.*

II. $p_t = \mu + p_{t-1} + e_t$, where $e_t \sim \mathcal{N}(0, \sigma_t^2)$ *n.i.d.*[2]

III. $\mathrm{Cov}(e_t, e_{t-k}) = 0$ for $k \neq 0$.

1.2 Predictive Regressions and Empirical Findings

We test the semi-strong EMH hypothesis under the assumption $R_{mt} =$ constant. This gives us the regression

$$\begin{aligned} r_{t+1} &= \alpha + \beta X_t + e_{t+1} \\ H_0 &: \beta = 0. \end{aligned}$$

Fama and French (1988) ran regressions using horizons between 1 month and 4 years. On nominal CRSP value-weighted NYSE portfolio returns on dividend yield from 1941 – 1986, they found that there appeared to be strong evidence of predictability.

However, their data had three nonstandard features:

1. The sample size was small
2. The regressor was persistent
3. The errors are negatively correlated between the two equations

[1] Note that $\mathcal{F}_{mt} \subset \mathcal{F}_t$.

[2] *n.i.d.* stands for independent but not necessarily identical increments.

Suppose that the return r_t and the predictor x_t satisfy

$$\begin{aligned} r_{t+1} &= \alpha + x_t\beta + u_{t+1} \\ x_{t+1} &= (1-\rho)\mu + \rho x_t + v_{t+1} \\ u_t &= \gamma v_t + e_t, \end{aligned}$$

where $0 < \rho \leq 1$, u_t and v_t are martingale differences, $\mathbb{E}[u_t^2] = \sigma_u^2$, $\mathbb{E}[v_t^2] = \sigma_v^2$, $\mathbb{E}[u_t^4] < \infty$, $\mathbb{E}[v_t^4] < \infty$, $\gamma < 0$, $\mathbb{E}[v_t e_s] = \mathbb{E}[e_t x_s] = 0 \ \forall t, s$.

Looking at the OLS estimate of β, we can derive $\mathbb{E}[\hat{\beta} - \beta] = \gamma\mathbb{E}[\hat{\rho} - \rho]$. In practice, $\hat{\rho}$ is downward biased because ρ is close to 1. Also note that $\gamma < 0$, so $\hat{\beta}$ is upward biased in finite samples. Hence, we need to account for this bias when conducting inference. Alternatively, we can look at the asymptotic distribution of the t-statistic.

If we look at the asymptotic distribution of the t-statistic, we note that it is the weighted average of two distributions, where one is $\sim \mathcal{N}(0,1)$ and the other is skewed with negative mean. Robust inference takes these features of the t-statistic into account.

Empirical results show that predictability seems to be time varying. Data from 1940 – 1990 has higher t-values than that from 1927 – 2009.

1.3 Conclusion

The statistical evidence shows that aggregate stock returns are weakly predictable.

The finding of predictability is NOT interpreted as a rejection of the EMH in the literature. Instead, it is taken as evidence that the rationally expected returns (*i.e.*, the equity premiums) are time-varying.

The EMH is a joint hypothesis. If the test rejects, we do not know whether the problem is an inefficient market or a misspecifed model of market equilibrium.

2 Market Microstructure and High-Frequency Returns

The readings for this chapter are

- Roll, R. (1984). "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," Journal of Finance, 39, 1127-1139.
- Glosten, L.R. (1987). "Components of the Bid-Ask Spread and the Statistical Properties of Transaction Prices," Journal of Finance, 42, 1293-1307.
- Battalio and Schultz (2011). "Regulatory Uncertainty and Market Liquidity: The 2008 Short Sale Ban's Impact on Equity Option Markets," Journal of Finance, 66, 2013-2053.
- Andersen and Bollerslev (1997), "Intraday Periodicity and Volatility Persistence in Financial Markets," Journal of Empirical Finance, Vol.4, No.2-3, pp.115-158.

In the last chapter, we took returns as the principal objects of interest without any reference to the institutional structure in which there are determined. In particular, we have ignored that security prices are generally denominated in fixed increments (ticks) and that transactions occur over uneven time intervals.

An analogy is going from Newtonian mechanics to quantum mechanics. In the low-frequency setting, returns act in a continuous manner, but in our new setting, returns are no longer continuous.

In fact, the very process of trading can have important effects on the statistical properties of financial asset prices. In this chapter, we will consider two models for the bid-ask spread to examine these effects. The first model is a purely statistical model while the second model is an economic model. We will first analyze the theoretical results, then look at empirical analysis.

2.1 Rolle's Model

Market makers provide liquidity to the market. Each market maker offers two prices, a *bid price* P_a and an *ask price* P_a simultaneously, with $P_b < P_a$.

Remark 2.1. The NYSE fact book (1994) reports that the spread was ≤ 0.25 in 90.8% of the NYSE bid-ask quotes in 1994. As electronic trading proliferated, this spread has decreased but not gone to 0.

Denote P_t^* as the time t fundamental value of a security in a frictionless economy[3], and let s be the bid-ask spread, which we assume to be constant over time. Furthermore, assume there is only one transaction per time. Then the observed market price P_t is assumed to satisfy

$$P_t = P_t^* + I_t \frac{s}{2}, \text{where } I_t = \begin{cases} +1 & \text{w.p. } 1/2 \text{ (buyer-initiated)}, \\ -1 & \text{w.p. } 1/2 \text{ (seller-initiated)}, \end{cases}$$

which implies $\Delta P_t = \Delta P_t^* + (I_t - I_{t-1})\frac{s}{2}$. If the fundamental price P_t^* is constant, then $\Delta P_t = (I_t - I_{t-1})\frac{s}{2}$. Consequently, we have the following statistical properties:

- $\text{Var}[\Delta P_t] = s^2/2$, meaning the bid-ask spread generates volatility.
- $\text{Cov}(\Delta P_{t-1}, \Delta P_t) = -s^2/4$, so the covariance depends only on s.
- $\text{Cor}(\Delta P_{t-1}, \Delta P_t) = -1/2$, meaning there is constant negative serial correlation.
- $\text{Cov}(\Delta P_{t-k}, \Delta P_t) = 0$ for $k > 1$, which has an MA(1) structure.

If P_t^* changes over time and ΔP_{t+1}^* is stationary, serially uncorrelated, and independent of I_t, then the covariance doesn't change but the correlation becomes $\text{Cor}(\Delta P_{t-1}, \Delta P_t) = \frac{-s^2/4}{(s^2/2)+\sigma(\Delta P_t^*)}$. Inverting the covariance formula, we obtain

$$s = 2\sqrt{-\text{Cov}(\Delta P_{t-1}, \Delta P_t)},$$

which gives us a way to estimate the spread.

Rolle found that estimates from weekly returns differ significantly from daily returns. The fundamental price and the spread are linked due to asymmetrical information. Order arrivals convey information, which may impact the fundamental price and the spread simultaneously.

2.2 Glosten's Model

Let P^* be the *full information price*, *i.e.*, the price that would result if everyone had access to all information. Let P^c denote the *common information price*, *i.e.*, $P^c = \mathbb{E}[P^* \mid \Omega]$, where Ω denotes the common information. P^c is the best estimate of P^* using Ω.

Market makers have access to common information, while some investors may have private information. Market makers need to protect themselves from investors when providing liquidity to the market.

Example 2.2. For example, in Rolle's model, we assumed that the chance of buy and sell transactions are equally likely. If many people are selling, this gives a market maker information. Hence, the market maker can utilize this information, adjusting the price downward and increasing the spread.

We assume that all market makers use the following price updating rule:

$$\alpha(x) = \mathbb{E}[P^* \mid \Omega \cup \{\text{investor buys at price } x\}]$$
$$\beta(y) = \mathbb{E}[P^* \mid \Omega \cup \{\text{investor sells at price } x\}],$$

where α, β are functions of x, y, respectively. These functions describe how common knowledge expectations are updated in response at various possible ask and bid prices.[4]

Recall that P_a, P_b are the bid and ask prices the market maker offers. Then the market maker's expected profits in the above two situations equal

$$P_a - \alpha(P_a) - C_a\beta(P_b) - P_b - C_b,$$

where C_a, C_b include order-processing and inventory components. Glosten takes them as exogenous.

Under risk neutrality, in equilibrium, the expected profit equals zero in each case due to competition between market makers. Then these equations determine the equilibrium ask and bid prices.

Let $Z_a, Z_b \geq 0$ be the belief updating relative to P^c:

$$Z_a = \alpha(P_a) - P^c, \text{ investor buys, revise price upward;}$$
$$-Z_b = \beta(P_b) - P^c, \text{ investor sells, revise price downward.}$$

[3]This means there are no transaction costs.

[4]Note that we can extend these functions to utilize other information as well.

Then we have

$$P_a = P^c + Z_a + C_a,$$
$$P_b = P^c - Z_b - C_b,$$

where the bid-ask spread is greater than the case without asymmetric information (Rolle's model only has C_a, C_b). Note that we can also estimate Z_a, Z_b using data on P_a, P_b, P^c, C_a, C_b.

Let P_t be the price of the ith transaction, Ω_t be the information immediately before the tth transaction, and Ω_t^+ be the information immediately after the tth transaction, *i.e.*, $\Omega_t^+ = \Omega_t \cup \{\text{investor buys/sells at price } P_t\}$.

We start with the identity $P_t = P_{t,a} I_a + P_{t,b} I_b$, where I_a, I_b are indicators for the transaction direction. Then using math, we can show that $P_t = P_t^c + C_t Q_t$, where P_t^c is the common information price immediately after the trade, $C_t = \begin{cases} C_a & \text{if investor buys,} \\ C_b & \text{if investor sells} \end{cases}$, $Q_t = \begin{cases} +1 & \text{if investor buys,} \\ -1 & \text{if investor sells} \end{cases}$.

Note that the expected fundamental price P_t^c is correlated with the transaction direction Q_t. This is the most important difference from Rolle's model.

Then the return is

$$\Delta P_{t+1} = P_{t+1} - P_t = A_{t+1} Q_{t+1} + e_{t+1} + \{C_{t+1} Q_{t+1} - C_t Q_t\},$$

where

- A_{t+1} is Z_a or Z_b.
- $A_{t+1} Q_{t+1}$ is due to asymmetric information. It has a permanent effect on the price level, and it is serially uncorrelated.
- The price jumps if a transaction reveals important information.
- e_{t+1} is due to new information arrival between trades at t and $t+1$. It is common to all models of price dynamics, and it is serially uncorrelated.
- $C_{t+1} Q_{t+1} - C_t Q_t$ is due to transaction costs, as in Rolle's model. It introduces a negative serial correlation to returns, with only transitory effects on price levels.

Assume that $Z_a = Z_b = \alpha \frac{s}{2}$ and $C_a = C_b = (1-\alpha)\frac{s}{2}$, where α is the percentage of the half-spread attributable to adverse selection. Then we can show that $\Delta P_{t+1} = \alpha \frac{s}{2} Q_t + \frac{s}{2} \Delta Q_{t+1} + e_{t+1}$. We can use the restricted OLS to estimate the model if transaction-level data are available.

2.3 Theoretical Conclusion

The current literature tends to separate bid-ask spreads into three components:

- The adverse selection component: compensates market makers for the losses they incur from trading against better informed traders.
- The order processing costs component: the portion of the spreads that provide compensation to market makers for the bookkeeping, exchange fees, overhead, and other direct costs of making markets.
- The inventory holding cost component: compensate market makers for holding a non-zero inventory position in the assets they trade (overnight).

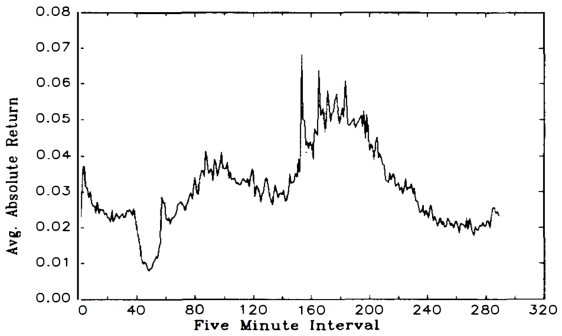
Market microstructure is a crucial aspect to consider when analyzing high-frequency financial data, *e.g.*, estimating volatility (spread) or estimating contemporaneous correlations between security prices (non-synchronous trading).

2.4 An Interlude on Data Processing

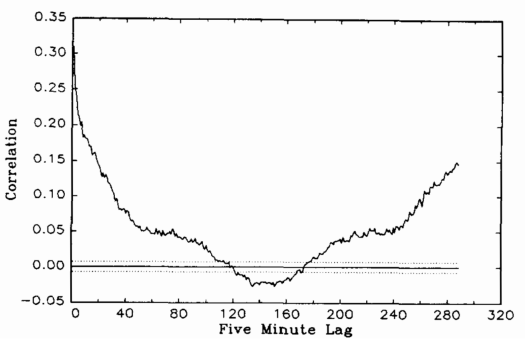
The data is based on the Deutschemark-USD (DM-$) spot exchange rate, traded 24 hours a day, 7 days a week. We will focus on mean, volatility, and autocorrelation.

The dataset has a short time stamp (Oct 1992 – Oct 1993). We will consider 5 minute intervals, where exchange rate levels for each interval is determined by the weighted average between the preceding and immediately following quotes.

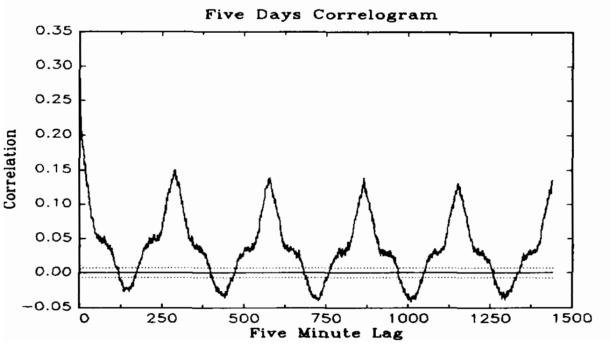
We have a sample of 74,880 observations, *i.e.*, $r_{t,n}$ with $n \in [288], t \in [260]$.



(a) Absolute returns for DM-$.



(b) Autocorrelation for absolute returns for DM-$.



(c) 5-day autocorrelation of the absolute returns for DM-$. The size of the autocorrelations at the daily frequencies decay slowly over the first four days, only to increase slightly at the fifth, or weekly, frequency. This signals the presence of a minor day-of-the-week effect.

2.5 Intraday Periodicity (Seasonality)

Sample mean: 0.000175%, sample standard deviation: 0.047%, sample skewness: 0.367, sample kurtosis: 21.5. Both skewness and kurtosis are highly statistically significant. The first order autocorrelation coefficient is -0.04, which is highly statistically significant.

We use absolute returns as proxies for return volatility, shown in Figure 1a. The strong drop in volatility between intervals 20 - 40 corresponds to lunch hours in Asian markets. Activity picks up when European markets open (interval 84). Volatility declines slowly until European lunch hour (interval 138) before it increases sharply when US markets open (interval 156).

Autocorrelation in the average returns resembles white noise after the first few lags. On the other hand, autocorrelations for the absolute returns has a distorted U-shape, induced by the strong intraday pattern, shown in Figure 1b. We also have the weekly autocorrelations in Figure 1c.

2.6 Model for Intraday Returns

Andersen and Bollerslev (1997) introduces the model

$$r_t = \sum_{n=1}^{N} r_{t,n} = \sigma_t \frac{1}{N^{1/2}} \sum_{n=1}^{N} s_n z_{t,n},$$

where r_t is the daily continuously compounded return, N is the number of return intervals in a day, σ_t is the stochastic component of day t's volatility, s_n is a deterministic intraday periodic component with normalization $N^{-1} \sum s_n = 1$, and $z_{t,n} \sim \mathcal{N}(0,1)$ are *i.i.d.*

We can examine whether formal time series modeling of volatility is affected by the intraday seasonality by fitting MA(1)-GARCH(1, 1) models to high-frequency data using the aggregation of k time intervals ($k = 1, \ldots 144$). We find that $\alpha + \beta$ varies a lot between the different time horizons. In particular, there is high persistence ($\alpha + \beta$ close to 1) in short intervals (< 30 minutes) and long intervals (> 2 hours).

2.7 Estimating the Seasonality Component in Volatility

Now, we consider a generalized model, where $s_{t,n}$ can depend on σ_t:

$$r_{t,n} = \mathbb{E}[r_{t,n}] + \frac{\sigma_t s_{t,n}}{N^{1/2}} z_{t,n},$$

where the first term is small and $s_{t,n}$ represents seasonality.

Our goal is to estimate $s_{t,n}$ and obtain filtered results $(r_{t,n} - \mathbb{E}[r_{t,n}])/s_{t,n}$. There are two approaches available: series approximation and kernel estimation. We will focus on the first approach. The idea is that we approximate the model using a Fourier series after applying a logarithmic transformation. This gives us a better GARCH estimate where $\alpha + \beta$ stays more consistent over different horizons.

2.8 Conclusion

The return distribution varies systematically within a trading day. Hence, we should not apply the GARCH model to intraday data without filtering out the intraday periodicity. If intraday periodicity is constant over time, then it has no effect on the formal econometric modeling of volatility at daily and lower frequencies, otherwise it can complicate the analysis.

3 The Capital Asset Pricing Model

The readings for this chapter are

- Gibbons, M.R., Ross, S.A, and Shanken, J. 1989. A Test of the Efficiency of a Given Portfolio. Econometrica 57, 1121-52.
- Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, Review of Economics and Statistics, 47, 13-37.

Markowitz (1959) formulated the investor's portfolio selection problem in terms of the returns' means and covariances. Sharpe (1964) and Lintner (1965) studied the equilibrium implications.

3.1 Portfolio Optimization

There are N risky assets of price 1, with a mean vector μ and covariance matrix Ω, where Ω is positive definite. We assume that μ and Ω are known, *i.e.*, with no estimation uncertainty.[5] First, we consider a setting with no riskfree asset.

We call a portfolio p the *minimum-variance portfolio of all portfolios with mean return* μ_p if its portfolio weight vector solves

$$\min_{\omega} \omega'\Omega\omega \quad \text{s.t.} \quad \omega'\mu = \mu_p, \omega'\mathbf{1} = 1.$$

Theorem 3.1. *The solution to the above constrained optimization problem is given by*

$$\omega = g + h\mu_p, \tag{1}$$

where g, h depend only on μ, Ω: $g = \frac{1}{D}[B\Omega^{-1}\mathbf{1} - A\Omega^{-1}\mu], h = \frac{1}{D}[C\Omega^{-1}\mu - A\Omega^{-1}\mathbf{1}]$, with $A = \mathbf{1}'\Omega^{-1}\mu, B = \mu'\Omega^{-1}\mu, C = \mathbf{1}'\Omega^{-1}\mathbf{1}, D = BC - A^2$.

Proposition 3.2. *The efficient frontier has the following properties.*

1. *g is a minimum-variance portfolio with zero expected return.*

   Proof. Set μ_p in Equation (1) to zero. □

2. *$g+h$ is a minimum-variance portfolio with expected return equal to 1. Also, h is an arbitrage portfolio, i.e., its elements sum up to zero.*

   Proof. This is because the elements of g sum to 1. □

3. *The minimum-variance frontier can be generated from any two distinct minimum-variance portfolios.*

   Proof. Suppose $\omega_1 = g + h\mu_{p_1}, \omega_2 = g + h\mu_{p_2}$ are two such portfolios. Let $\omega^* = g + h\mu^*$ be an arbitrary minimum-variance portfolio. If we choose λ s.t. it solves $\mu^* = \lambda\mu_{p_1} + (1-\lambda)\mu_{p_2}$, then $\lambda\omega_1 + (1-\lambda)\omega_2$ generates ω^*. □

[5] Note that this assumption cannot hold because μ and Ω change over time.

4. *Any portfolio of minimum-variance portfolios is also a minimum-variance portfolio.*

   Proof. This is because the resulting weight vector satisfies the optimal solution formula in Equation (1). □

5. *Let p, r be two minimum-variance portfolios. Then the covariance of their returns is*

$$\mathrm{Cov}(R_p, R_r) = \frac{B - A\mu_p - A\mu_r + C\mu_p\mu_r}{D}. \tag{2}$$

   Proof. Let ω_p, ω_r be their respective weight vectors. Their returns' covariance equals $\omega_p'\Omega\omega_r$. Next, apply the expression Equation (1) to ω_p, ω_r. □

6. *There exists a global minimum-variance portfolio.*

   Proof. Let $p = r$ and compute the minimum of Equation (2). □

7. *For each minimum-variance portfolio p, except the global minimum-variance portfolio, there exists a minimum portfolio that has zero covariance with it. This portfolio is called the zero beta portfolio w.r.t. p.*

   Proof. Set Equation (2) to zero and solve for the expected return μ_r. □

8. *Consider a regression of the return of any arbitrary portfolio (not necessarily efficient) on any minimum-variance portfolio R_p (except the global minimum variance portfolio) and its zero beta portfolio R_{op}:*

$$R_a = \beta_0 + \beta_1 R_{op} + \beta_2 R_p + e_p.$$

   Then the population regression coefficients satisfy:

$$\begin{aligned}
\beta_0 &= 0 \\
\beta_1 &= \frac{\mathrm{Cov}(R_p, R_a)}{\sigma_p^2} \\
\beta_2 &= \frac{\mathrm{Cov}(R_{op}, R_a)}{\sigma_{op}^2} = 1 - \beta_2.
\end{aligned}$$

9. *We have*

$$\mathbb{E}[R_a] = (1 - \beta_2)\mathbb{E}[R_{op}] + \beta_2\mathbb{E}[R_p].$$

   Proof. Consider the previous property and compute the expectation. □

Remark 3.3. In the population regression coefficients of 8, $\beta_0 = 0$ represents an arbitrage condition, *i.e.*, a free lunch doesn't exist.

Now we assume there is an asset with a fixed return R_f. Draw a ray that passes R_f and an arbitrary point Y on the efficient frontier. Any point on this ray corresponds to a portfolio consisting of Y and some amount of the riskfree asset. All assets that consist of a nonnegative amount of Y and some riskfree asset are on the ray that passes through R_f and Y. This is shown in Figure 2.

Now consider the problem

$$\min_{\omega} \omega'\Omega\omega \quad \text{s.t.} \quad \omega'\mu + (1 - \omega'\mathbf{1})R_f = \mu_p.$$

Theorem 3.4. *The solution to the above minimization problem is*

$$\omega = \frac{\mu_p - R_f}{(\mu - R_f\mathbf{1})'\Omega^{-1}(\mu - R_f\mathbf{1})}\Omega^{-1}(\mu - R_f\mathbf{1}) = c_p\overline{\omega},$$

where $c_p = \frac{\mu_p - R_f}{(\mu - R_f\mathbf{1})'\Omega^{-1}(\mu - R_f\mathbf{1})}$ is a scalar, and $\overline{\omega} = \Omega^{-1}(\mu - R_f\mathbf{1})$ is a portfolio weight vector that does not depend on p.

Therefore, all minimum-variance portfolios are a combination of the riskfree asset and a particular risky asset portfolio with weights proportional to $\overline{\omega}$. The latter portfolio is called the *tangent portfolio*.

Recall that the *Sharpe ratio* of a portfolio p, with expected return μ_p and standard deviation σ_p, is defined as

$$\frac{\mu_p - R_f}{\sigma_p}.$$

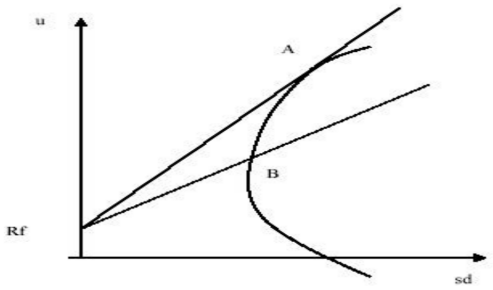


Figure 2: The efficient frontier is the curved line. Note that all points to the right of the line are feasible, but not efficient while all points the left of the line are infeasible (in the context of no riskfree asset).

The Sharpe ratio can be interpreted as the price for a single unit of risk.

The tangent portfolio X has the highest Sharpe ratio among all portfolios of risky assets. This maximum Sharpe ratio is equal to

$$[(\mu - R_f\mathbf{1})'\Omega^{-1}(\mu - R_f\mathbf{1})]^{1/2}.$$

For any asset or portfolio that is included in X, we must have

$$\mathbb{E}[R_a - R_f] = \beta\mathbb{E}[R_X - R_f],$$

where $\beta = \frac{\text{Cov}(R_X, R_a)}{\sigma_X^2}$.

This result implies that if any asset in X is found to have a positive α w.r.t. X, then X cannot be the tangent portfolio. This property is very useful for testing portfolio efficiency:

Suppose we want to test whether a risky asset portfolio p is efficient, where p consists of N assets and the riskfree return is R_f. Let R_{pt} be the excess return on p and R_t be the vector of excess returns on the N assets. Consider the following system of regressions

$$R_t = \alpha + \beta R_{pt} + e_t.$$

Under the null hypothesis, R_{pt} is mean-variance efficient, and thus $H_0 : \alpha = 0$.

All results so far are mathematical facts. In other words, there is no theory that can be tested using data. To derive a testable theory about the market equilibrium, we need to make assumptions about investors' risk preferences and the characteristics of the market.

The CAPM model makes such assumptions. Its main result shows that the tangent and market portfolios are equivalent. This result also shows that a mean-variance optimizer can't do better than simply "holding the market," so there is no need to compute any optimal portfolio.

3.2 CAPM and Its Testable Implications

Market Assumptions:

- Each individual investor can invest any part of his capital in certain riskfree assets, all of which pay interest at a common positive rate, exogenously determined.
- She can invest any fraction of her capital in any or all of a given finite set of risky securities.
- Risky securities are traded in a single purely competitive market, free of transaction costs and taxes.
- Any investor may borrow funds to invest at the riskfree rate. There is no limit on the amount she can borrow at this rate.
- She makes all transactions at discrete points in time.[6]

Investor Assumptions:

- Each investor has a fixed amount of capital for investment in riskless and risky assets after optimal cash holdings have been deducted.

[6]This condition is critical.

- She will have assigned a joint probability distribution on all individual stocks. All expected values of returns are finite, all variances are non-zero and finite, and the covariance matrix is positive-definite.[7]
- If any two mixtures of assets have the same expected return, the investor will prefer the one having the smaller variance of return, and if any two mixtures of assets have the same variance of returns, she will, prefer the one having the greater expected value.

Theorem 3.5. *Under the above assumptions, for any investor, the optimal composition of the risky assets is independent of the division of capital between risky and riskless assets.*

Thus, under the stated assumptions, all investors will make identical decisions regarding the proportionate composition of his stock portfolio. Only a single point on the Markowitz efficient frontier is relevant to the investor's decision regarding his investments in risky assets.

Remark 3.6. This leads to a question: why would we utilize hedge funds or investment banks? Alternatively, we can view that our models are inadequate.

The *idealized uncertainty* assumption states that for any given set of market prices, all investors assign identical sets of means, variances, and covariances to the joint distribution of the returns.

We will use (r_m, σ_m) to denote the mean and the standard deviation of the market portfolio. Recall that the market portfolio weights are the capitalization weights, *i.e.*, w_k is the kth asset's total capital value divided by the total capital value of all assets (for a stock this would be the total number of outstanding shares multiplied by the share price).

Theorem 3.7. *Under the above assumptions, including idealized uncertainty and with $R_f < A/C$:*

1. *The tangent portfolio coincides with the market portfolio.*
2. *For any given portfolio a, we must have $\mathbb{E}[R_a - R_f] = \beta \mathbb{E}[R_m - R_f]$, where $\beta = \mathrm{Cov}(R_m, R_a)/\sigma_m^2$. This implies that for any portfolio, $\mathbb{E}[R_a - R_f]/\mathrm{Cov}(R_m, R_a) =$ constant.*

CAPM Implications:

1. The market portfolio of invested wealth is mean-variance efficient. The market portfolio has the highest Sharpe ratio.
2. All investors make identical decisions regarding the proportionate composition of his stock portfolio.
3. In a cross-section regression of excess asset returns on their market betas, the intercept should be zero and the slope should be positive and equal to the expected excess returns on the market portfolio.
4. Market betas are sufficient to describe the cross section of expected returns.

3.3 Time Series Approach to Testing Implication 1

Let R_{mt} represent the excess returns on the market portfolio. Suppose there are $N = 10$ test assets with excess returns summarized by the vector R_t. Consider the following regression system $R_t = \alpha + \beta R_{mt} + e_t$. Under the CAPM, the market portfolio is efficient, therefore $H_0 : \alpha = 0$ and the alternative hypothesis is $H_1 : \alpha \neq 0$. Assume e_t are stationary and uncorrelated over time, with $\mathbb{E}[e_t e_t'] = \Sigma$.

Gibbons-Ross-Shanken (1989) proposed to test H_0 using "Hotelling's T^2 test," a multivariate generalization of the univariate t-test. The OLS is equivalent to the Gaussian MLE:

$$\hat{\beta} = \frac{\sum_{t=1}^T (R_{mt} - \overline{R}_m)(R_t - \overline{R})}{\sum_{t=1}^T (R_{mt} - \overline{R}_m)^2},$$
$$\hat{\alpha} = \overline{R} - \hat{\beta}\overline{R}_m,$$

where $\overline{R} = T^{-1}\sum R_t$ and $\overline{R}_m = T^{-1}\sum R_{mt}$. Then we can do the following decomposition:

$$\sqrt{T}\hat{\alpha} = \sqrt{T}(\overline{R} - \beta\overline{R}_m) - \sqrt{T}(\hat{\beta} - \beta)\overline{R}_m.$$

Then under H_0, $\sqrt{T}\hat{\alpha} \xrightarrow{d} \mathcal{N}(0, V)$, where $V = \left[1 + \frac{\mathbb{E}^2[R_{mt}]}{\mathrm{Var}[R_{mt}]}\right]\Sigma$. Then it is feasible to form a quadratic form in $\hat{\alpha}$ to test H_0. This leads to

$$GRS = \left(\frac{T - (N+1)}{N}\right)\left[1 + \frac{(\overline{R}_m)^2}{T^{-1}\sum_{t=1}^T (R_{mt} - \overline{R}_m)^2}\right]^{-1} \hat{\alpha}'\hat{\Sigma}^{-1}\hat{\alpha},$$

[7] This condition results in a unique portfolio.

where $\hat{\Sigma} = T^{-1}\sum_{t=1}^{T}\hat{e}_t\hat{e}_t'$, $\hat{e}_t$ is the OLS residuals, and $(N+1)$ is a finite sample correction because the variance is estimated. If the data are normally distributed, then $\hat{\alpha}$ and $T\hat{\Sigma}$ will have normal and Wishart distributions, respectively. As a result, $GRS \sim F_{N,T-(N+1)}$. Without normality, $GRS \xrightarrow{d} \chi^2_N/N$.

3.4 Cross Sectional Approach to Testing Implication 3

Suppose there are N portfolios, $i = 1, \ldots, N$, with excess returns $R_i - R_f$, and their market betas are unknown. Consider the following regression of R_i on β_i: $R_i - R_f = \mu + \gamma\beta_i + e_i$. Implication 3 implies that for any t, $\mu = 0$ and $\gamma = \mathbb{E}[R_m - R_f]$.

In practice, the betas are unknown, but they can be estimated using time series regressions. This approach leads to the two-pass regressions in Black-Jensen-Scholes (1972):

1. At the beginning of January of each year, estimate the β's of the NYSE stocks by $R_{j,t} - R_f = \alpha_j + \beta_j(R_{m,t} - R_f) + v_{j,t}$ using monthly returns for the past five years.
2. Form ten portfolios based on the estimated β's. The first portfolio has the 10% of the stocks with the lowest estimated β's, etc.
3. The return in each of the next 12 months for reach of the ten portfolios is calculated.
4. Repeat Steps 1 and 2 each year for the entire sample period.
5. Estimate the β's of the ten constructed portfolios using the time series regression in Step 1.
6. Estimate the cross section regression $R_i - R_f = \mu + \gamma\hat{\beta}_i + e_i$ $(i = 1, \ldots, 10)$ and test the restrictions $\mu = 0$ and $\gamma = \mathbb{E}[R_m - R_f]$ using the t-statistic.

Empirically, they reported t-values of 6.52 for the intercept and 6.53 for the slope coefficient relative to the observed excess return on the market portfolio, rejecting the CAPM. The standard errors for the t-statistic do not account for the estimation uncertainty in $\hat{\beta}_i$.

3.5 Asymptotic Distribution of the Two-Stage Cross Sectional Regression Estimator

Now we look at a more general setting that also can be used to study multifactor models. We have N portfolios returns observed for T periods: $R_{i,t},\ i = 1, \ldots, N, t = 1, \ldots, T$, where $T \ll N$, *e.g.*, $N = 25, T = 200$. $R_{i,t}$ is assumed to be a linear function of k factors:

$$\begin{aligned} R_{i,t} &= \alpha_i + \beta_{1i}f_{1t} + \cdots + \beta_{ki}f_{kt} + u_{i,t} \\ &= \alpha_i + B_i'F_t + u_{i,t}. \end{aligned}$$

Note the special case where $k = 1$ and f_t is the return on the market portfolio. Our goal is to test the factor model, with CAPM being a special case. Let Z_i be variables such as portfolio characteristics, *e.g.*, firm size, introduced to test the model.

1. Run N independent time series to estimate B_i: $R_{i,t} = \alpha_i + B_i'F_t + u_{i,t}$.
2. Estimate a single cross sectional regression $\overline{R}_i = \rho + \hat{B}_i'\lambda + Z_i'\gamma + e_i$, where $\overline{R}_i = \frac{1}{T}\sum_{t=1}^{T} R_{it}$. This can be estimated via OLS or GLS, and we will focus on OLS.

Under both the null and the alternative hypothesis, $\mathbb{E}[R_i] = \rho + B_i'\lambda + Z_i'\gamma$. Under the null hypothesis, B_i adequately describes asset returns, therefore $\gamma = 0$. Furthermore, in the CAPM case, λ should equal the expected return on the market portfolio.

We derive the asymptotic distributions of $\hat{\rho}, \hat{\lambda}, \hat{\gamma}$ under the null hypothesis. This will allow us to test, for example, $\gamma = 0$.

Let $\mathbf{1} = [1]_{N\times 1}, Z = [Z_i']_{N\times q}, \hat{B} = [\hat{B}_i']_{N\times k}, \overline{R} = [\overline{R}_i]_{N\times 1}$, and $\hat{X} = (\mathbf{1}, Z, \hat{B})$. Then $\hat{\theta} = (\hat{\rho}, \hat{\gamma}, \hat{\lambda})' = (\hat{X}'\hat{X})^{-1}\hat{X}'\overline{R}$. To relate this estimate to its true value, let $X = (\mathbf{1}, Z, B), \theta = (\rho, \gamma, \lambda)'$.

We can show that

$$\begin{aligned} \sqrt{T}(\hat{\theta} - \theta) &= (\hat{X}'\hat{X})^{-1}\hat{X}'\sqrt{T}[\overline{R} - \mathbb{E}[\overline{R}]] - (\hat{X}'\hat{X})^{-1}\hat{X}'\sqrt{T}[\hat{B} - B]\lambda \\ &= (I) - (II), \end{aligned}$$

where term (I) accounts for the effect of estimating the expected return and term (II) accounts for the effect of estimating B. Both of these terms are normally distributed by the CLT.

3.5.1 Asymptotic Distribution with Conditionally Uncorrelated Homoskedastic Errors

This follows the work of Shanken 1992. We have the two assumptions:

1. Assume $\mathbb{E}[u_t \mid \mathcal{F}] = 0$, $\mathbb{E}[u_t u_t' \mid \mathcal{F}] = \Sigma_u$, and $\mathbb{E}[u_t u_{t+s}' \mid \mathcal{F}] = 0$ for $s \neq 0$, where $\mathcal{F}$ is the information set generated by the entire factor sequence and Σ_m is an $N \times N$ PSD matrix.
2. The factors F_t are stationary with finite fourth moments, satisfying

$$\frac{1}{T}\sum_{t=1}^{T}(F_t - \mathbb{E}[F_t])'(F_t - \mathbb{E}[F_t]) \xrightarrow{p} \Sigma_{F_0},$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(F_t - \mathbb{E}[F_t]) \xrightarrow{d} \mathcal{N}(0, \Sigma_F),$$

where $\Sigma_F = \sum_{k=-\infty}^{\infty} \mathbb{E}[F_t - \mathbb{E}[F_t]](F_{t+k} - \mathbb{E}[F_{t+k}])'$.

Theorem 3.8. *Under Assumptions 1 and 2, as $T \to \infty$ and N fixed,*

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma_c),$$

where $\Sigma_c = \Sigma_F^* + (1 + \lambda' \Sigma_{F_0}^{-1} \lambda) D \Sigma_u D', \Sigma_F^* = \begin{bmatrix} 0 & 0 \\ 0 & \Sigma_F \end{bmatrix}, D = (X'X)^{-1}X'$.

3.5.2 Asymptotic Distribution Allowing for Serial Correlation and Heteroskedasticity

This follows the work of Jagannathan et al., 2009. We have the assumption:

1. Let $h_t = ((h_t^{(1)})', (h_t^{(2)})')'$. Assume that the CLT applies to the random sequence h_t, that is

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t \xrightarrow{d} \mathcal{N}(0, \Sigma_h)$$

with

$$\Sigma_h = \begin{bmatrix} \Psi & \Gamma \\ \Gamma' & \Pi \end{bmatrix},$$

where $\Psi = \sum_{k=-\infty}^{\infty} \mathbb{E}[(h_t^{(1)})(h_{t+k}^{(1)})'], \Pi = \sum_{k=\infty}^{\infty} \mathbb{E}[(h_t^{(2)})(h_{t+k}^{(2)})'], \Gamma = \sum_{k=-\infty}^{\infty} \mathbb{E}[(h_t^{(1)})(h_{t+k}^{(2)})']$.

Theorem 3.9. *Suppose X has full rank. Under Assumption 1, as $T \to \infty$ and N fixed,*

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma_c),$$

where $\Sigma_c = D[\Psi + \Pi - (\Gamma + \Gamma')D'], D = (X'X)^{-1}X'$.

For a given set of testing assets, one can always come up with a one factor model, using the mean-variance efficient portfolio as the factor, to fit the data. So the real question is whether the CAPM restrictions hold *ex ante* as well as *ex post*.

For testing, using more assets can reveal larger deviations from the model. However, this also increases the size of the deviations required to reject the model. Some ways of grouping of assets into portfolios are often desirable.

4 Multifactor Pricing Models

Ross (1976) argued that the apparent empirical success of the CAPM (at that time) is due to three assumptions, which are more plausible than the assumptions needed to derive the CAPM. These assumptions are:

(i) there are many assets;

(ii) the market permits no arbitrage opportunities;

(iii) asset returns have a factor structure with a small number of factors

Ross showed that under these three assumptions, expected asset returns are (approximately) a linear function of the factor loadings.

4.1 Arbitrage Pricing Theory (APT)

Assume there are N assets, with prices all normalized to 1. Assume their returns are linear functions of a common factor (we will allow for multiple factors later). Then we have $R_i = \alpha_i + \beta_i F + e_i$, where F is a common factor with $\mathbb{E}[|F|] < \infty, \mathbb{E}[e_i] = 0$, the e_i are sufficiently independent s.t. $1/N \sum e_i \xrightarrow{p} 0, \mathbb{E}[e_i F] = 0$. In matrix form, we have

$$R = \alpha + \beta F + e.$$

We construct the model as follows:

1. Construct an arbitrage portfolio using a weight vector η. By construction, the portfolio return is $\eta' R$ and $\eta' \mathbf{1} = 0$. We require this portfolio to be well-diversified, *i.e.*, $\eta_i = O(1/N)\ \forall i$.
2. By the LLN,
$$\eta' R = \eta' \alpha + (\eta' \beta) F + \eta' e \xrightarrow{p} \eta' \alpha + (\eta') \beta) F \text{ as } N \to \infty,$$
i.e., the influence on the well-diversified portfolio of the independent noise terms becomes negligible.
3. Further, we require that this portfolio has no systematic risk, *i.e.*, $\eta' \beta = 0$. Then $\eta' R \approx \eta' \alpha$.
4. Because η is an arbitrage portfolio and it is approximately riskfree, to prevent arbitrarily large disequilibrium positions, we must have $\eta' \alpha \approx 0$.
5. Since the above equation must hold for any portfolio η s.t. $\eta' \beta = \eta' \mathbf{1} = 0$, α must belong to the vector space spanned by $\mathbf{1}$ and β, that is, $\exists \rho, \delta \in \mathbb{R}$ s.t.
$$\alpha \approx \rho \mathbf{1} + \beta \delta.$$
Note that this means α is in a 2-dimensional vector space.
6. Let $\mu = \mathbb{E}[R] = \alpha + \beta \mathbb{E}[F]$, then $\mu \approx \rho \mathbf{1} + \beta[\delta + \mathbb{E}[F]] := \rho \mathbf{1} + \beta \lambda$. Therefore, the expected returns are approximately a linear function of the factor loading; explicitly,
$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} \approx \rho \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \lambda \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}.$$
In practice, the pricing errors are often ignored: $\mu = \rho \mathbf{1} + \beta \lambda$.

If the market includes a riskfree asset, then ρ equals the riskfree rate. If the factor is a portfolio of traded assets with expected return μ_F, then $\mu_F = \rho + \lambda$, which implies $\mu - \rho = \beta(\mu_F - \rho)$. This is the same prediction as the CAPM, where μ_F is the expected return on the market portfolio.

The assumptions of APT and CAPM are non-nested. The APT assumes: (i) asset returns have a factor structure, (ii) investors know the factor loadings, and (iii) $N \to \infty$. CAPM assumes: (i) investors engage in mean-variance optimization, (ii) investors know the first two moments of the assets, and (iii) N is finite.

Now suppose there are K factors with $K \ll N$. Assume their returns are linear functions of K common factors:

$$R_{it} = \alpha_i + B_i' F_t + e_{it},$$

where

$$B_i = \begin{bmatrix} \beta_{i1} \\ \vdots \\ \beta_{iK} \end{bmatrix}, F_t = \begin{bmatrix} f_{1t} \\ \vdots \\ f_{Kt} \end{bmatrix}.$$

Using the same argument as in the $K = 1$ case, we can obtain $\alpha_i = \rho + B_i' \delta_K$ for some δ_i, $\forall i$ and $\mathbb{E}[R_{it}] = \mu_i = \rho + B_i' \lambda_K\ \forall i$, with $\lambda_K = \delta_K + \mathbb{E}[F_t]$.

If the factors are portfolios of traded assets, then λ_K is a $K \times 1$ vector of factor risk premiums. Then we have

$$\mathbb{E}[F] = \mu_F = \rho + \lambda_K,$$

which implies $\lambda_K = \mu_F - \rho, \delta_K = -\rho \mathbf{1}$.

The APT is often viewed as a generalization of the CAPM, where K factors are used to model the cross sectional distribution of expected returns.

4.1.1 Estimation and Testing

A complete description of the APT consists of two equations:

$$R_{it} = \alpha_i + B_i' F_t + e_{it},$$
$$\mu_i = \rho + B_i' \lambda_K.$$

The first equation is an assumption for the DGP, *i.e.*, the returns have a linear factor structure with a small number of factors. The second equation is the theory's prediction, *i.e.*, the expected returns are a linear function of factor loadings. Any test of the APT is always a joint test of both. If the test rejects, we don't know whether it is because the theory's prediction is off or because we have misspecified the factors, either their number or their identity.

The testing of APT depends on how the factors are specified:

1. Factors are portfolios of traded assets and there exists a riskfree asset;
2. Factors are traded portfolios of traded assets and there is no riskfree asset;
3. Factors are not portfolios of traded assets, *e.g.*, inflation rate.

We focus on the first case because it is the most relevant case empirically.

Let $\rho = R_f$ (because there is a riskfree asset) and $\mu_F = R_f + \lambda_K$ (because factors are portfolios). These equations imply

$$\mu_i - R_f = B_i'(\mu_F - R_f).$$

Therefore, for the time series regression

$$R_{it} - R_f = \delta_i + B_i'(F_t - R_f) + e_{it},$$

we must have $\delta_i = 0 \ \forall i$. We can estimate the regressions for each i and then apply the t-test to test $\delta_i = 0$, but this suffers from the multiple testing problem.

Alternatively, we can estimate the regressions as a system for all i, and apply the GRS test to check $H_0 : \delta_1 = \delta_2 = \cdots = \delta_n = 0$. Under an *i.i.d.* normality assumption, $GRS \sim F_{N,T-K-N}$.

Without normality assumptions, we need to make two assumptions:

1. $\mathbb{E}[e_t e_s] = 0 \ \forall t \neq s, \mathbb{E}[e_t e_t' = \Sigma]$ where Σ is positive definite, $\mathbb{E}[\|e_t\|^4] < \infty$, and $\mathbb{E}[X_t e_s'] = 0 \ \forall t, s$.
2. The excess returns on factors are stationary with finite fourth moments, s.t. $1/T \sum (X_t - \overline{X})(X_t - \overline{X})'$ converges to a positive definite matrix in probability.

Theorem 4.1. *Under the null hypothesis and Assumptions 1 and 2, with N finite and* $T \to \infty$, $GRS \xleftarrow{d} \chi^2_N / N$.

Remark 4.2. Because the $F_{N,T-K-N}$ has a critical values than χ^2_N/N, using the former provides more conservative inference. Hence, if someone uses the F distribution, she is not necessarily assuming normality.

4.1.2 Empirical Findings: Fama and French (1993) and More

This section is based on the paper "Common Risk Factors in the Returns on Stocks and Bonds" (Fama and French, Journal of Financial Economics, 1993).

The market betas are insufficient to describe the cross section of average returns on U.S. common stocks. At the same time, variables that have no special standing in asset-pricing theory show reliable power to explain the cross-section of average returns. The list of empirically determined average-return variables including, size, leverage, earnings/price, book-to-market equity, etc.

The paper constructs two portfolios using U.S. common stocks based on the size and book-to-market equity ratio. It then uses these portfolios as proxies to risk factors to explain the cross section average returns. The main regression is

$$R_{it} - RF_t = \alpha_i + \beta_i (R_{mt} - RF_t) + s_i SMB_t + h_i HML_t + e_{it},$$

where R_{it} is the return on asset i at time t, RF_t is the riskfree rate, R_{mt} is the return on a proxy to the market portfolio, SMB_t is the return on the size factor, HML_t is the return on the book-to-market factor.

Remark 4.3. Using just the market portfolio, the R^2 of the regression is approximately 60-70%.

In the paper, the median NYSE size is used to split NYSE, Amex, and NASDAQ stocks into two groups: small and big (S, B). These stocks are then divided into three book-to-market equity groups based on the breakpoints for the bottom 30% (L), middle 40% (M), and top 30% (H) of the ranked values. Then six portfolios (S/L, S/M, ...) are constructed from the intersection of the two ME and the three BE/ME groups.

The SMB factor is meant to mimic the risk factor in returns relative to size. It is computed as the difference between the simple average of the returns on the three small-stock portfolios and the simple average of the returns on the big-stock portfolios. There are two ways to interpret the return on the SMB factor: the return on an arbitrage portfolio that goes long the small and goes short the large firms or the difference in returns of two portfolios, one holding small firms and the other big firms.

The factor HML is meant to mimic the risk factor in returns related to book-to-market equity. It is the difference between the simple average of the returns on the two high BE/ME portfolios and the average of the returns on the two low BE/ME portfolios.

The average value of market factor is 0.43% per month. This is large from an investment perspective (about 5% a year). The average SMB return is 0.27% per month ($t = 1.73$), and the average HML return is 0.43% per month ($t = 2.91$). Empirically, the three factor model captures strong common variations in stock returns. The R^2 becomes 90-99%.

4.1.3 Recent Developments

Carhart (1997) proposes a momentum factor to explain mutual fund returns. Using a sample free of survivor bias, the paper demonstrates that common factors in stock returns and investment expenses almost completely explain persistence in equity mutual funds' mean and risk-adjusted returns. As a result, mutual funds only beat the market by luck.

Lettau and Ludvigson (2001) explores the ability of conditional versions of the CAPM and the consumption CAPM-jointly the (C)CAPM to explain the cross section of average stock returns. They use the log consumption/wealth ratio as a conditional variable. Their new factor, called "Cay," is the difference between log consumption and a weighted average of log asset wealth and log labor income.

Fama and French (2015) proposed a five factor model by adding profitability and investment factors to the three-factor model. The new factors are RMW_t, the difference between the returns on diversified portfolios of stocks with robust and weak profitability, and CMA_t, the difference between the returns on diversified portfolios of the stocks of low and high investment firms, which they call conservative and aggressive. They find that with the addition of these factors, the HML factor becomes redundant.

Multifactor models provide a framework to describe the cross sectional correlation of average stock returns. The testing of multifactor models faces a joint hypothesis problem. A natural approach is to run time-series regressions and apply the GRS test.

4.2 Factor Models

The first factor model was developed by Spearman (1904) to test the general intelligence factor, or "g" factor. The model is

$$X_i = \mu + \Lambda f_i + e_i,$$

where X_i is individual i's test scores, and $N \times 1$ vector with N tests, μ is the average test scores of all individuals, f_i is a factor score of individual i, *i.e.*, the "g" factor, Λ is the factor loading, an $N \times 1$ vector, common across individuals, and e_i is the error term, *i.i.d.* across individuals and test subjects.

f_i is a latent variable, *i.e.*, there are no operations and criteria for directly measuring it. Because f_i explains the correlations among tests, if we partial it out, the covariances should all equal zero:

$$\text{Var}\,[X_t] - \mathbb{E}[\Lambda\Lambda' f_i^2] = \text{Diagonal Matrix}.$$

Some other examples include the APT, term structure models, *i.e.*, zero-coupon bonds of different maturities, and recommendation systems, *i.e.*, Amazon and Netflix.

4.2.1 Classical Factor Models: Identification and Estimation

The classic factor model is

$$X_t = \mu + \Lambda F_t + e_t,$$

where X_t is an $N \times 1$ vector, μ is the mean of X_t, F_t is the factor score, a $k \times 1$ vector with mean zero, Λ is the factor loading matrix, $\mathbb{E}[e_t] = 0$, $\mathbb{E}[F_t e_t'] = 0$, and e_t are independent in i and t: $\mathbb{E}[e_t e_t'] = \Psi$ is diagonal and $\mathbb{E}[e_t e_s'] = 0\ \forall t \neq s$.

Remark 4.4. Approximate factor models generalize the above model to allow $\mathbb{E}[e_t e_s]$ to be small but not nonzero and Ψ to be approximately diagonal but not exactly diagonal.

Identification of the model is based on the covariance of X_t:

$$\text{Var}\,[X_t] = \Sigma = \mathbb{E}[\Lambda F_t F_t' \Lambda'] + \Psi.$$

However, these equations do not separately identify Λ and F_t, since for any $k \times k$ non-singular matrix P, we always have $(\Lambda P)(P^{-1}F_t) = \Lambda F_t$. Thus, Λ and F_t are observationally equivalent to $\Lambda^* = \Lambda P$ and $F_t^* = P^{-1}F_t$.

The following identification restrictions are often used to identify Λ and F_t:

- (R1) $\mathbb{E}[F_t F_t'] = I$, *i.e.*, the factors are orthogonal.
- (R2) $\Gamma = \Lambda' \Psi^{-1} \Lambda$ is diagonal.

Under (R1), the covariance equation reduces to $\text{Var}\,[X_t] = \Sigma = \Lambda\Lambda' + \Psi$. If the diagonal elements of Γ in (R2) are different, then this allows us to identify the model.

Now suppose we wanted to estimate the model. Let $(X_1, \ldots, X_T)'$ be a sample, where each X_t is an $N \times 1$ vector. Assume that there are k latent factors. The estimation uses Gaussian likelihood:

$$L = -\frac{1}{2} NT \log 2\pi - \frac{T}{2} \log |\Sigma| - \frac{T}{2} \text{tr}(C\Sigma^{-1}),$$

where $\Sigma = \Lambda\Lambda'$, $C = T^{-1} \sum (X_t - \overline{X})(X_t - \overline{X})'$, $\overline{X} = T^{-1} \sum X_t$. The estimates are determined by two sets of nonlinear equations:

$$(C - \hat{\Psi})\hat{\Psi}^{-1}\hat{\Lambda} = \hat{\Lambda}\hat{\Gamma}$$
$$\text{Diag}(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}) = \text{Diag}(C)$$

and the restriction that $\hat{\Gamma} = \hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda}$ is diagonal. There is no analytical solution, so numerical methods are needed.

Remark 4.5. Suppose if we change the units of measurement from X_t to DX_t. Then the MLE becomes $\hat{\Lambda}^* = D\hat{\Lambda}$ and $\hat{\Psi}^* = D\hat{\Psi}D$. This means the estimated factor loadings and error variances are merely changed by the units of measurement.

The classic factor model explains the interdependence of a set of variables in terms of latent factors. The space spanned by the factor scores is identifiable, but the values are not. Common identification restrictions involve assuming the factor scores are orthogonal, but other conditions are also possible. For asymptotic analysis, it is typically assumed that N is small (fixed) and $T \to \infty$.

4.2.2 Principal Component Analysis (PCA)

This method was proposed by Hotelling (1933) to find linear combinations of variables with large variances. The variables are not required to have a factor structure, and in fact they usually do not. Let X_t be an $N \times 1$ vector of variables with 0 mean and known covariance matrix Σ. Let β be an $N \times 1$ vector s.t. $\beta'\beta = 1$. Then the variance of $\beta' X_t = \mathbb{E}[\beta' X_t]^2 = \beta' \mathbb{E}[X_t X_t']\beta = \beta' \Sigma \beta$.

To find the normalized linear combination of X_t with maximum variance, consider the Lagrangian $L = \beta'\Sigma\beta - \lambda(\beta'\beta - 1)$. Then the first order condition (FOC) is $(\Sigma - \lambda I)\beta = 0$. Because β is nonzero, $(\Sigma - \lambda I)$ must be singular, so λ must satisfy $|\Sigma - \lambda I| = 0$.

Note that because Σ is symmetric PSD, this matrix has N positive solutions for λ. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ are the eigenvalues with corresponding eigenvectors $\beta_1, \ldots, \beta_N$. Then λ_1 is the maximum variance, λ_2 is the second highest variance, ... Note that this procedure creates two matrices Λ and B:

$$\Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_N),$$
$$B = (\beta_1, \ldots, \beta_N).$$

Theorem 4.6. *Let the N-component random vector X_t with $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t X_t'] = \Sigma$. Then there exists an orthogonal linear transformation*

$$U_t = B' X_t$$

s.t. the covariance matrix of U_t is $\mathbb{E}[U_t U_t'] = \Lambda$, and the rth component of U_t, $U_{r,t} = \beta_r' X_t$ has maximum variance of all normalized linear combinations uncorrelated with $U_{1,t}, \ldots, U_{[r-1,t]}$.

The principal components are identified from the eigenvectors of a covariance matrix. It is important to know that analysis into principal components is most suitable when all the components of X_t are measured in a common unit.

Theorem 4.7. *Let $X_1, \dots X_T$ be T observations from $\mathcal{N}(\mu, \Sigma)$, where Σ is a matrix with N different eigenvalues. Then, the MLE of $\lambda_1, \dots, \lambda_N$ and $\beta_1, \dots, \beta_N$ consists of the ordered roots of $|\hat{\Sigma} - \lambda I| = 0$ satisfying $(\hat{\Sigma} - \hat{\lambda}_i I)\hat{\beta}_i = 0, \hat{\beta}_i'\hat{\beta}_i = 1$, where $\hat{\Sigma}$ is the MLE of Σ.*

This result gives the MLE of the principal components without reference to any factor structure. Now let us consider the limiting distribution of the estimator.

Theorem 4.8. *Let $X_1, \dots, X_T$ be T observations from $\mathcal{N}(\mu, \Sigma)$, where Σ is a matrix with N different characteristics roots. Then, as $T \to \infty$ and N fixed,*

$$\sqrt{T}(\hat{\lambda}_i - \lambda_i) \xrightarrow{d} \mathcal{N}(0, 2\lambda_i^2)$$
$$\sqrt{T}(\hat{\beta}_i - \beta_i) \xrightarrow{d} \mathcal{N}\left(0, \sum_{k=1, k\neq i}^{N} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \beta_i \beta_k'\right).$$

Furthermore, the $\sqrt{T}(\hat{\lambda}_i - \lambda_i)$ are mutually independent and independent of $\sqrt{T}(\hat{\beta}_i - \beta_i)$, and the asymptotic covariances of the latter satisfy $\text{Acov}(\sqrt{T}(\hat{\beta}_i - \beta_i), \sqrt{T}(\hat{\beta}_k - \beta_k)) = -\frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \beta_i \beta_k'$.

This asymptotic result assumes that $N \ll T$. All principal components are consistently estimated, and they converge at rate $T^{-1/2}$. The PCA transforms the data to a new coordinate system s.t. the greatest variance by any normalized combination of the data comes to lie on the first coordinate:

$$X_t = BB'X_t = BU_t = \beta_1 U_{1,t} + \cdots + \beta_N U_{N,t}.$$

Example 4.9. Consider the bond yields of different maturities. The first three principle components explains a total of 99.96% of the total variance: 97.69, 2.14, 0.12.

Example 4.10. Consider the Fama-French three-factor model. Looking at the average returns, the first three principle components explains 91.31% of the total variance: 83.84, 4.37, 3.10.

Remark 4.11. We can see if the principle component is correlated to the economic factor, *e.g.*, the market factor in the Fama-French three-factor model, by running a regression.

4.2.3 Factor Models of Large Dimensions and the PCA

We consider factor models of large dimensions, *i.e.*, large N and T. We consider estimation, inference, and factor selection. The model is

$$X_t = \Lambda F_t + e_t,$$

where T and N are large, k is small, Var $[e_t]$ is approximately diagonal, and F_t are independent of e_t.

Assume the number of factors is known. We treat Λ and F_t as parameters. The estimation procedure is as follows:

1. Compute the k largest eigenvalues of XX' and multiply them by $\sqrt{T}$ to get $\tilde{F}$.
2. Compute $\tilde{\Lambda}' = \tilde{F}'X/T$.

Notice that the columns of $N^{1/2}\tilde{F}\tilde{V}^{1/2}$ are estimates of the first k principal components of X_t, where

$$\tilde{V} = (TN)^{-1} Diag\{\tilde{\lambda}_1, \dots, \tilde{\lambda}_k\}.$$

This shows that we are simply extracting the principal components of X_t.

Let F_t^0 and λ_i^0 denote the true factor scores and loadings, respectively, with M a generic constant. Consider the following assumptions:

- **Assumption F(0)** (factor score): $\mathbb{E}[\|F_t^0\|]^4 \leq M$ and $T^{-1}\sum F_t^0 F_t^{0\prime} \xrightarrow{p} \Sigma_F > 0$ for an $r \times r$ non-random matrix Σ_F.
- **Assumption L** (factor loading): λ_i^0 is either deterministic s.t. $\|\lambda_i^0\| \leq M$, or it is stochastic s.t. $\mathbb{E}[\|\lambda_i^0\|]^4 \leq 4$, In either case, $N^{-1}\Lambda^{0\prime}\Lambda^0 \xrightarrow{p} \Sigma_\Lambda > 0$ for an $r \times r$ non-random positive definite matrix Σ_Λ, as $N \to \infty$.

Remark 4.12. These are conventional assumptions in econometrics. Assumption L ensures that each factor has a nontrivial contribution of the variance of X_t.

- **Assumption E** (error terms):
  1. $\mathbb{E}[e_{it}] = 0, \mathbb{E}[|e_{it}|^8] \leq M$.
  2. $\mathbb{E}[e_{it}e_{js}] = \sigma_{ij,ts}$, with $|\sigma_{ij,ts}| \leq \overline{\sigma}_{ij}\ \forall(t,s)$ and $|\sigma_{ij,ts}| \leq \tau_{ts}\ \forall(i,j)$, s.t. $N^{-1}\sum_{i,j=1}^{N}\overline{\sigma}_{ij} \leq M$, $T^{-1}\sum_{t,s=1}^{N}\tau_{ts} \leq M$, and $(NT)^{-1}\sum_{i,j,t,s=1}|\sigma_{ij,ts|\leq M}$.
  3. $\forall(t,s)$, $\mathbb{E}[|N^{-1/2}\sum(e_{is}e_{it} - \mathbb{E}[e_{is}e_{it}])|^4] \leq M$.
  4. $\forall t$, $N^{-1/2}\sum\lambda_i e_{it} \xrightarrow{d} \mathcal{N}(0,\Gamma_t)$ as $N \to \infty$, where
  $$\Gamma_t = \lim_{N\to\infty} N^{-1}\sum_{i,j=1}^{N}\mathbb{E}[\lambda_i\lambda_j' e_{it}e_{jt}].$$
  5. $\forall i$, $T^{-1/2}\sum F_i e_{it} \xrightarrow{d} \mathcal{N}(0,\Phi_i)$ as $T \to \infty$, where
  $$\Phi_t = \lim_{T\to\infty} T^{-1}\sum_{s,t=1}^{N}\mathbb{E}[F_t^0 F_s^{0\prime} e_{it}e_{jt}].$$

  Remark 4.13. The second and third conditions require the residuals to be approximately uncorrelated. The last two conditions rule out any correlation between the factor and error components.
  6. **Assumption LFE** (factors and errors): $\{\lambda_i\}, \{F_t\}, \{e_{it}\}$ are three mutually independent groups. Dependence within each group is allowed.
  7. **Assumption IE** (errors): $\forall t \leq T, i \leq N$, $\sum_{s=1}^{T}|\tau_{st}| \leq M$ and $\sum_{i=1}^{N}|\overline{\sigma}_{ij}| \leq M$.

  Remark 4.14. Assumption LFE is a restrictive assumption since it requires the factors to be strictly exogenous. Assumption IE strengthens E.

Now we look at the results. Let $C_{NT} = \min\{T^{1/2}, N^{1/2}\}$.

1. **Factor Space: Convergence Rate**. For any fixed $k > 1$, under Assumptions F(0), L, and E:
$$C_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\|\tilde{F}_t - H'F_t^0\|^2\right) = O_p(1),$$
where
$$H = (\Lambda^{0\prime}\Lambda^0/N)(F^{0\prime}\tilde{F}/T)\tilde{V}^{-1}$$
$$\tilde{V} = \frac{1}{TN}Diag\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k\},$$
with the diagonal elements being the eigenvalues of XX' arranged in decreasing order.

2. **Asymptotic Distribution: Factor Scores and Loadings**. Let H be as above, $Q = V^{1/2}\Psi\Sigma_\Lambda^{-1/2}$, where V is a diagonal matrix containing the eigenvalues of $\Sigma_\Lambda^{-1/2}\Sigma_F\Sigma_\Lambda^{-1/2}$ in decreasing order and Ψ are the corresponding orthonormal eigenvectors. Under Assumptions F(0), L, E, and LFE, as $N, T \to \infty$:

   If $\sqrt{N}/T \to 0$, then $\forall t$,
   $$\sqrt{N(\tilde{F}_t - H'F_t^0)} \xrightarrow{d} \mathcal{N}(0, V^{-1}Q\Gamma_t Q'V^{-1}).$$

   If $\sqrt{T}/N \to 0$, then $\forall i$,
   $$\sqrt{T(\tilde{\lambda}_i - H^{-1}\lambda_i^0)} \xrightarrow{d} \mathcal{N}(0, (Q')^{-1}\Psi_i Q^{-1}).$$

Remark 4.15. The scores and loadings are both estimated up to an invertible matrix transformation. Knowing $H'F_t^0$ is often as good as knowing F_t^0. The convergence rates of $\tilde{F}_t$ and $\tilde{\lambda}_i$ are $\sqrt{N}$ and $\sqrt{T}$, respectively.

3. **Asymptotic Distribution: The Common Component**. Let $A_{it} = \lambda_i^{0\prime}\Sigma_\Lambda^{-1}\Gamma_t\Sigma_\Lambda^{-1}\lambda_i^0$ and $B_{it} = F_t^{0\prime}\Sigma_F^{-}1\Phi_i\Sigma_F^{-1}\Sigma_F^{-1}F_t^0$, where Φ_i is the variance of $T^{-1/2}\sum_{t=1}^T F_t^0 e_{it}$.
   (a) Under Assumptions F(0), L, E, LFE, and IE
   $$(N^{-1}A_{it} + T^{-1}B_{it})^{-1/2}(\tilde{C}_{it} - C_{it}^0) \xrightarrow{d} \mathcal{N}(0,1).$$
   (b) If $N/T \to 0$, then $\sqrt{N}(\tilde{C}_{it} - C_{it}) \xrightarrow{d} \mathcal{N}(0, A_{it})$.
   (c) If $T/N \to 0$, then $\sqrt{T}(\tilde{C}_{it} - C_{it}) \xrightarrow{d} \mathcal{N}(0, B_{it})$.
4. **The Factor Space: Maximum Deviation**. Suppose Assumptions F(0), L, E, and LFE hold, then
   $$\max_{1\le t\le T} \|\tilde{F}_t - H'F_t^0\| = O_p(T^{-1/2}) + O_p((T/N)^{1/2}).$$

   Remark 4.16. This result proves an upper bound on the maximum deviation of the estimated factors from the space spanned by the true ones.
5. (Determine the Number of Factors). Define the information criteria
   $$PCP(k) = S(k) + k\overline{\sigma}^2 g(N,T)$$
   $$IC(k) = \ln S(k) + kg(N,T),$$
   where $S(k)$ is the estimated SSR allowing for k factors, $\overline{\sigma}^2 2$ is equal to $S(k_{\max})$ for a pre-specified value $k_{\max}$. The second criterion uses the log scale, and as a result, $\overline{\sigma}^2$ drops out. The optimal k is taken to the minimizer of the information criterion. We have $\hat{k}_{PCP}$ and $\hat{k}_{IC}$.

   Theorem 4.17. *Suppose Assumptions F(0), L, E, and LFE hold. If (i)* $g(N,T) \to 0$ *and (ii)* $C_{NT}^2 g(N,T) \to \infty$ *as* $N,T \to \infty$*, then* $\Pr[\hat{k}_{PCP} = r] \to 1$ *and* $\Pr[\hat{k}_{IC} = r] = 1$.

5 Volatility in Time and Space

We have established that aggregate stock returns are weakly predictable and portfolio returns are cross sectionally correlated, exhibiting an approximate factor structure. Now, we will examine asset return volatility. We focus on basic models and concepts, preparing for further analysis.

The motivation is the volatility is strongly persistent (useful for forecasting) and cross sectionally correlated (important for portfolio management). This is true for companies in different sectors and different countries.

The readings for this chapter are

- Clark, P. (1973). "A Subordinated Stochastic Process Model With Finite Variance for Speculative Prices," Econometrica 41, 135-155.
- Engle, R.F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation," Econometrica 50:987-1008.
- Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh. (2016). "The common factor in idiosyncratic volatility: Quantitative asset pricing implications," Journal of Financial Economics, 2016, 119, 249-283.
- Engle, R.F. and S. Campos-Martins. (2020). "Measuring and Hedging geopolitical Risk," Working paper.

5.1 Volatility Concepts

Let $\{r_t\}_{t=1}^T$ be a sample of returns, *e.g.*, on the S&P500 index, over evenly spaced time intervals. The *realized variance* is the sum of squared returns. For a period of n time intervals with returns $r_{t-n}, \ldots, r_{t-1}$, the realized variance is $RV = \sum_{i=1}^n (r_{t-i} - \overline{r})^2$.

Remark 5.1. Daily realized variance is a summation over squared intraday returns, *e.g.*, over 5-minute intervals. Realized variance is a finite sample concept, not a population concept.

Realized volatility is the square root of the realized variance: $S = \sqrt{RV}$.

Remark 5.2. In practice, S is often multiplied by a constant to bring the measure to an annualized scale. For instance, if the RV is daily realized variance, then an annualized realized volatility is given by $\sqrt{252 \cdot RV}$.

Conditional variance is the variance of a future return that is conditional on an information set, such as the history of returns. *Conditional volatility* is the square root of the conditional variance.

A *stochastic volatility (SV) model* specifies the volatility as a latent process. A typical stochastic volatility model is

$$\begin{aligned} r_t &= \mu + \exp(h_t/2)\varepsilon_t, \\ h_t &= \omega + \rho h_{t-1} + \eta_t, \\ \varepsilon_t &\sim \mathcal{N}(0,1) \ i.i.d., \\ \eta_t &\sim \mathcal{N}(0,1) \ i.i.d., \end{aligned}$$

where η_t is independent of $z_s \ \forall t, s$.

Implied volatility is the volatility level implied by options prices.

Remark 5.3. The implied volatility is different from realized volatility if the volatility risk is priced. Empirical estimates show that implied volatility tends to be higher than realized volatility. In other words, there is a positive volatility risk premium.

5.2 Volatility in Time

5.2.1 ARCH(1) Model

This is the simplest model for conditional variance:

$$\begin{aligned} r_t &= \mu + h_t^{1/2}\varepsilon_t \\ \varepsilon_t &\sim \mathcal{N}(0,1) \ i.i.d., \end{aligned}$$

where $h_t = \omega + \alpha(r_{t-1} - \mu)^2$.

The volatility parameters are $\omega > 0$ and $\alpha \geq 0$. Note that the volatility level $h_t^{1/2}$ is known at time $t-1$. If we define $e_t = r_t - \mu$, the error of forecasting e_t^2 is

$$v_t = e_t^2 - \mathbb{E}[e_t^2 \mid r_{t-1}, r_{t-2}, \dots] = e_t^2 - h_t.$$

Combining this with the definition of h_t, we have $e_t^2 = \alpha e_{t-1}^2 + v_t$, where v_t is uncorrelated with e_{t-1}^2 because it is a forecasting error. Therefore, the demeaned square return follows an AR(1) process.

α governs the moments and persistence of e_t^2. For e_t^2 to be stationary, we need $|\alpha| < 1$. For Var $\left[e_t^2\right] < \infty$, we need $3\alpha^2 < 1$. Also, $\rho_\tau = \text{Cor}[e_t^2, e_{t-\tau}] = \alpha^\tau$. This structure, *i.e.*, a single parameter governing both moments and persistence, is too rigid to describe the return process successfully.

A natural extension is to consider an ARMA(1, 1) model instead of an AR(1) model, for squared returns. This leads to the GARCH(1, 1) model.

5.2.2 GARCH(1, 1) Model

The model is

$$\begin{aligned} r_t &= \mu + h_t^{1/2}\varepsilon_t, \\ \varepsilon_t &\sim \mathcal{N}(0,1) \ i.i.d., \end{aligned}$$

where $h_t = \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1}$. The parameters satisfy $\omega \geq 0, \alpha \geq 0, \beta \geq 0$. The model is stationary if $\alpha + \beta < 1$. By recursive substitution, the conditional variance is an exponentially weighted average of past square returns $(r_{t-\tau} - \mu)^2$. Note that the sign of $r_t - \mu$ does not affect the volatility level, *i.e.*, GARCH(1, 1) is a symmetric model.

The moments of returns satisfy:

- $\mathbb{E}[r_t] = \mu$.
- Var $[r_t] = \mathbb{E}[(r_t - \mu)^2] = \mathbb{E}[h_t]\mathbb{E}[\varepsilon_t^2] = \mathbb{E}[h_t] = \omega/(1 - \alpha - \beta) := \sigma^2$.
- $\mathbb{E}[(r_t - \mu)^3] = 0$.

- $\mathbb{E}[(r_t-\mu)^4] = \mathbb{E}[h_t^2]\mathbb{E}[\varepsilon_t^4] = 3\mathbb{E}[h_t^2]$. The expression for $\mathbb{E}[h_t^2]$ can be found by squaring the equation $\omega + (\alpha\varepsilon_{t-1}^2 + \beta)h_{t-1}$ and taking the expectation, which shows $3\mathbb{E}[h_t^2]/\sigma^4 > 3$. Therefore, $\mathbb{E}[(r_t-\sigma)^4]$ is finite iff $2\alpha^2 + (\alpha+\beta)^2 < 1$. This property follows because the unconditional distribution is a mixture of normal distributions with random weights.

Define

$$\begin{aligned} e_t &= r_t - \mu \\ s_t &= e_t^2 = h_t\varepsilon_t^2 \\ v_t &= e_t^2 - h_t. \end{aligned}$$

Then,

$$s_t = \omega + (\alpha+\beta)s_{t-1} + v_t - \beta v_{t-1}.$$

The demeaned squared return follows an ARMA(1, 1) process. The correlation satisfies $\text{Cor}[s_t, s_{t+\tau}]$ with $C(\alpha+\beta) = \frac{\alpha(1-\alpha\beta-\beta^2)}{(\alpha+\beta)(1-2\alpha\beta-\beta^2)}$. The decay rate is $(\alpha+\beta)^\tau$. In practice, $\beta \approx 1, \alpha \approx 0$.

In general, the conditional variance $\forall n \geq 1$ is

$$\text{Var}\left[r_{t+n} \mid r_t, r_{t-1}, \dots\right] = \sigma^2 + (\alpha+\beta)^{n-1}(h_{t+1} - \sigma^2).$$

Because r_t is serially uncorrelated, we also have

$$\text{Var}\left[r_{t+1} + \dots r_{t+n} \mid r_t, r_{t-1}, \dots\right] = n\sigma^2 + \frac{1-(\alpha+\beta)^n}{1-\alpha-\beta}(h_{t+1} - \sigma^2).$$

Define $\theta = (\mu, \omega, \alpha, \beta)$. Let $f(h_1, r_1, \dots, r_T; \theta_0)$ be the joint density of $h_1, r_1, \dots, r_T$, which satisfies

$$f(h_1, r_1, \dots, r_T; \theta_0) = f(h_1, r_1; \theta_0)\prod_{t=2}^{T} f(r_t \mid r_{t-1}, \dots, r_1, h_1; \theta).$$

Then the log-likelihood function, conditional on the initial observation r_1 and the initial volatility level $h_1^{1/2}$, is

$$L(\theta) = \sum_{t=2}^{T} \log f(r_t \mid r_{t-1}, \dots, r_1, h_1; \theta).$$

We can show that

$$L(\theta) = K - \frac{1}{2}\sum_{t=2}^{T}\log h_t - \sum_{t=2}^{T}\frac{(r_t-\mu)^2}{2h_t},$$

where $h_t = \omega + \alpha(r_{t-1}-\mu)^2 + \beta h_{t-1}$ for $t = 2, \dots, T$. The likelihood function can be maximized numerically w.r.t. θ.

5.3 Volatility in Space

The starting point is a standard model for returns:

$$r_{i,t} = \alpha_i + \lambda_i' f_t + \sqrt{h_{i,t}}e_{i,t},$$

where f_t and λ_i are factors and loadings, $h_{i,t}$ is the conditional variance of $r_{i,t}$, *e.g.*, $h_{i,t} = \omega + \alpha(r_{t,i}-\mu)^2 + \beta h_{i,t-1}$, $e_{i,t}$ is uncorrelated over i and t with $\mathbb{E}_{t-1}[e_{i,t}] = 0$ and $E_{t-1}[e_{i,t}e_{j,t}] = 0$. Then consider the case that $e_{i,t}$ is uncorrelated but not necessarily independent. In particular, define $\psi_{i,t}$ as a volatility shock:

$$\psi_{i,t} = e_{i,t}^2 - 1 == \frac{(r_{i,t} - \alpha_i - \lambda_i' f_t)^2 - h_{i,t}}{h_{i,t}}.$$

It is possible that $E_{t-1}[\psi_{i,t}\psi_{j,t}] > 0$. Engle and Campos-Martins proposed a model to capture time-varying positive comovement between these volatility shocks:

$$\begin{aligned} e_{i,t} &= \sqrt{g(s_i, x_t)}\varepsilon_{i,t}, \\ g(s_i, x_t) &= s_i(x_t - 1) + 1, \\ \varepsilon_{i,t} &\sim \mathcal{N}(0,1) \text{ } \textit{i.i.d.} \text{ in } i, t, \end{aligned}$$

where x_t is a latent volatility factor with $E_{t-1}[x_t] = 1$ and $E_{t-1}[(x_t - 1)^2] = v_t$, $s_i \in [0, 1]$ is a factor loading, $\varepsilon_{i,t}$ and x_t are two mutually independent sequences, and the functional form of $g(s_i, x_t)$ ensures $\mathbb{E}_{t-1}[e_{i,t}^2] = 1$, which is compatible with a GARCH specification. This model implies

$$\mathbb{E}_{t-1}[e_{i,t}^2 e_{j,t}^2] = \mathbb{E}_{t-1}[g(s_i, x_t)g(s_j, x_t)] = s_i s_j v_t,$$
$$\mathbb{E}_{t-1}[e_{i,t}^4] = 3s_i^2 + 2.$$

We can write $e_{i,t}^2 - 1 = s_i(x_t - 1) + \eta_{i,t}$ with $\eta_{i,t} = g(s_i, x_t)(\varepsilon_{i,t}^2 - 1)$. This is a linear factor model for $e_{i,t}^2 - 1$, with factor score equal to $(x_t - 1)$ and loading equal to s_i. Therefore, the PCA provides consistent estimates if data on $e_{i,t}^2$ are available.

6 The Stochastic Discount Factor

The readings for this chapter are

- Hansen, L.P. and R. Jagannathan. 1997. Assessing Specification Errors in Stochastic Discount Factor Models. Journal of Finance 52, 587-613.
- Hansen, L.P. and S.F. Richard. 1987. The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. Econometrica 55, 587-613.

The CAPM and static multi-factor models are designed to explain average asset returns, not their dynamic properties. Also, these theories take the expected return on the market portfolio (or factor portfolios) as given. Consequently, they are completely silent about the determination of the aggregate equity premium. Stochastic discount factor (SDF) models address these issues.

Consider a portfolio that costs one unit and pays R_{t+1} units in the next period. Then

$$1 + \mathbb{E}_t[m_{t+1}R_{t+1}],$$

where m_{t+1} is called the *stochastic discount factor* (SDF). The SDF exists if there are no arbitrage opportunities, a valid SDF must be non-negative, and the SDF is unique iff the market is complete.

6.1 Example Models

We will first look at a simple model then a more general model.

Example 6.1 (Time-separable utility)**.** Suppose a representative investor solves

$$\max \mathbb{E}_t \left[\sum_{j=1}^{\infty} \delta^j U(c_{t+j}) \right].$$

Then, the SDF is

$$m_{t+1} = \delta \frac{U'(C_{t+1})}{U'(C_t)}.$$

For example, if $U(c_t) = (c_t^{1-\gamma} - 1)/(1 - \gamma)$, then $m_{t+1} = \delta(c_{t+1}/c_t)^{-\gamma}$. A log-linearization leads to $m_{t+1} = \theta_1 - \theta_2 \ln(c_{t+1}/c_t)$. This is the CCAPM model, where the SDF is affine in consumption growth.

Example 6.2 (State Non-Separable Preferences)**.** The Epstein-Zin-Weil objective function is defined recursively by

$$U_t = \{(1 - \delta)c_t^\rho + \delta(\mathbb{E}_t[U_{t+1}^\alpha])^{\rho/\alpha}\}^{1/\rho},$$

where $1/(1-\rho)$ equals the elasticity of intertemporal substitution and $(1-\alpha)$ measures risk aversion. Assume the budget constraint for a representative agent is $W_{t+1} = R_{m,t+1}(W_t - c_t)$ with $R_{m,t+1}$ the return on the market portfolio and W_t the wealth level. We can show that this leads to the SDF

$$m_{t+1} = \left\{ \delta \left(\frac{c_{t+1}}{c_t} \right)^{-(1-\rho)} \right\}^{\alpha/\rho} \left(\frac{1}{R_{m,t+1}} \right)^{1-\alpha/\rho},$$

and a log-linearization leads to the Consumption and Market-based CAPM:

$$m_{t+1} = \theta_1 + \theta_2 \ln(c_{t+1}/c_t) + \theta \ln R_{m,t+1}.$$

The SDF is affine in consumption growth and market return.

Example 6.3 (Time non-separable preferences)**.** Suppose a representative investor solves

$$\max \mathbb{E}_t \left[\sum_{j=1}^{\infty} \delta^j \frac{(c_{t+j} - X_{t+j})^{1-\gamma} - 1}{1-\gamma} \right],$$

where X_t is the external consumption habit. Let $S_t = (c_t - X_t)/c_t$ be the surplus consumption ratio. Then, the SDF is

$$m_{t+1} = \delta \left(\frac{S_{t+1} c_{t+1}}{S_t c_t} \right)^{-\gamma}$$

and under certain assumptions, the log-linearized SDF is

$$m_{t+1} = \theta_1 + \theta_2 \ln S_t + \theta f(\ln S_t) + (\theta_4 + \theta_5 f(\ln S_t)) \ln(c_{t+1}/c_t).$$

This SDF is affine in consumption growth, where the time–varying coefficient $f(\ln S_t)$ captures state-dependent risk aversion.

Example 6.4 (Conditional CCAPM)**.** This is a generalization of the CCAPM to capture time-varying risk premium. The SDF is

$$m_{t+1} = (\theta_1 + \theta_2 z_t) + (\theta_3 + \theta_4 z_t) \ln(c_{t+1}/c_t),$$

where z_t is a state variable for the risk premium, *e.g.*, corporate bond spread, consumption to wealth ratio, or labor income to consumption ratio.

Example 6.5 (Factor models)**.** These models are motivated by cross sectional multifactor pricing models and admit the following common expression:

$$m_{t+1} = \theta_1 + \theta_2 R_{m,t+1} + \theta_3' F_{t+1},$$

where $R_{m,t+1}$ is the return to the market portfolio and F_{t+1} is a vector of factor portfolios. Later, we will show there is an equivalence between a linear factor model of the SDF and a linear factor model written in a return-beta form with the same factors. Therefore, in this case, an SDF can be tested by testing the corresponding multifactor pricing model.

Example 6.6 (Non-parametric approach)**.** This approach begins with a flexible nonparametric approximation. For example, Rosenberg and Engle (2002) apply a generalized Chebyshev polynomial approximation for the SDF:

$$m_{t+1}(R_{t+1}) = \theta_0 T_0(R_{t+1}) \exp\left(\sum_{j=1}^{N} \theta_j T_j(R_{t+1}) \right).$$

6.2 The Hansen-Jagannathan (HJ) Bound

This bound provides an admissible region for the unconditional mean and standard deviation of the SDF using security market data. It is based on the following two assumptions: (1) the law of one price (portfolios with the same payoff must have the same price), and (2) no arbitrage (nonnegative payoffs that are positive with positive probability must have positive prices).

Let $x \in \mathbb{R}^N$ be a vector of random payoffs, and q be a price vector (omitting the t-index. Then $q = \mathbb{E}_t[mx]$.

- **Assumption 1**: $\mathbb{E}[|m|^2] < \infty, \mathbb{E}[|x|^2] < \infty, \mathbb{E}[xx^\top]$ is nonsingular, and $\mathbb{E}[|q|] < \infty$.
- **Restriction 1**: $\mathbb{E}[q] = \mathbb{E}[mx]$, the law of one price.
- **Restriction 2**: $m > 0$, sufficient to role out arbitrage.

6.2.1 The HJ Bound with a Riskfree Asset

Assume x contains a riskless asset. We use m to denote the true SDF, unobservable to the econometrician. Define α_0 as the solution to $\mathbb{E}[xx^\top \alpha_0] = \mathbb{E}[q]$, *i.e.*, $\alpha_0 = \mathbb{E}[xx^\top]^{-1}\mathbb{E}[q]$. Let $m^* = x^\top \alpha_0$. Then $m^* = x^\top \mathbb{E}[xx^\top]^{-1}\mathbb{E}[xm]$. Therefore, m^* is the (population) least-squares projection of m on x.

Some properties of m^*:

1. m^* is a valid SDF for pricing x in the sense that $\mathbb{E}[xm^*] = \mathbb{E}[q]$.
2. Because the payoff includes a riskfree payoff, by Restriction 1, $\mathbb{E}[m^*] = \mathbb{E}[m] = 1/R_f$, where $1/R_f$ is the discount rate.

3. Because m^* is the least-squares projection of m onto x, we have $\mathbb{E}[x(m-m^*)]=0$, *i.e.*, the projection residual is orthogonal to the explanatory variable. Consequently

$$\text{Var}[m] = \text{Var}[m^*] + \text{Var}[m^* - m] \geq \text{Var}[m^*]. \tag{3}$$

Therefore, $\text{Var}[m^*]$ is a lower bound for all admissible stochastic discount factors that satisfy Restriction 1.

Using Equation (3), we obtain

$$\frac{\sigma(m)}{\mathbb{E}[m]} \geq \frac{\sigma(m^*)}{\mathbb{E}[m^*]}, \tag{4}$$

where the right hand side is the HJ bound. We state this result as a lemma.

Lemma 6.7. *Suppose x includes a riskfree payoff. Then, all admissible SDFs satisfying Assumption 1 and Restriction 1 are inside the bound specified by Equation* (4)*.*

The structural economic model determines the left hand side. The right hand side can be computed using data on x and q. We reject the SDF if it violates the bound.

We will relate the HJ bound to the mean-variance frontier. Recall the mean-variance frontier comprises of linear combinations of a finite number of assets with price 1. Suppose these assets from a set $\mathcal{N}$. We have the following corollary:

Corollary 6.8. *Suppose x includes a riskfree payoff. Then, the HJ bound on $\sigma(m)/\mathbb{E}[m]$ under Assumption 1 and Restriction 1 is given by the Sharpe ratio of the mean-standard deviation frontier for $\mathcal{N}$:*

$$\frac{\sigma(m)}{\mathbb{E}[m]} \geq \textit{Sharpe ratio of the efficient frontier for } \mathcal{N}.$$

This result demonstrates that a steep slope of the mean-standard deviation frontier for asset payoffs implies a potentially dramatic bound on the volatility of the SDF.

6.2.2 The HJ Bound Without a Riskfree Asset

Let x^a denote the $(N+1)$-dimensional random vector formed by augmenting x with a unit payoff. Since $\mathbb{E}[xx^\top]$ is nonsingular and no linear combination of x is equal to one with probability one, $\mathbb{E}[x^a x^{a\top}]$ is also nonsingular. Because the market is incomplete, the price of the riskless asset is not uniquely determined.

Let v be a candidate for the discount rate and π_v be the corresponding extension of π from $\mathcal{N}$ to $\mathcal{N}^a$. Then $\exists m_a^*$ with $\mathbb{E}[m_v^* x] = \mathbb{E}[q]$ and $\mathbb{E}[m_v^*] = v$. The volatility bound for this v is then $\sigma(m) \geq \sigma(m_v^*)$. For each v, we compute the Sharpe ratio. After repeating the computation for all allowed v, we obtain an admissible region for $\sigma(m)/\mathbb{E}[m]$.

Theorem 6.9. *Suppose x contains risky assets only. Then, all admissible SDFs satisfying Assumption I, $\mathbb{E}[m_v^* x] = \mathbb{E}[q]$, and $\mathbb{E}[m_v^*] = v$ are inside the bound specified by*

$$\frac{\sigma(m_v)}{\mathbb{E}[m_v]} \geq \frac{[(\mathbb{E}[q] - v\mathbb{E}[x])^\top \Sigma^{-1} (\mathbb{E}[q] - v\mathbb{E}[x])]^{1/2}}{v}.$$

Remark 6.10. The bound is nonparametric, depending only on means and covariances. Different testing assets produce different bounds.

6.3 The HJ Distance

The goal is to assess specification errors in SDF models, since most models are likely misspecified. The problem is that the true SDF is not observed and may be non-unique. How can we compare (or rank) different asset pricing models, *i.e.*, different SDF proxies?

Hansen and Jagannathan (1997) provide a distance measure for this purpose.

Suppose the vector of security market payoffs used in an econometric analysis is denoted by x. which is used to generate a collection of payoffs using portfolio weights in $\mathbb{R}^n$:

$$P = \{p \mid p = a^\top x \text{ for some } a \in \mathbb{R}^n\}.$$

Let q denote the vector of securities prices and $\mathcal{F}$ the information set used for pricing. Then, $q = \mathbb{E}[mx \mid \mathcal{F}]$, which implies

$$\mathbb{E}[q] - \mathbb{E}[mx] = 0. \tag{5}$$

Let M denote the set of all random variables with finite second moments satisfying Equation (5). Clearly, M depends on x. Let y denote some "proxy" variable for a SDF that, strictly speaking, does not satisfy Equation (5). Then define the following least-squares measure of misspecification:

$$\delta^2 = \min_{m \in M} \mathbb{E}[(y-m)^2].$$

The bound δ^2 is unconditional and, as a result, is time-invariant.

- **Assumption 1**: $\mathbb{E}[|m|^2] < \infty$, $\mathbb{E}[|x|^2] < \infty$, $\mathbb{E}[xx^\top]$ is nonsingular, and $\mathbb{E}[|q|] < \infty$.
- **Assumption 2**: $\forall \alpha \in \mathbb{R}^n$, $\alpha^\top \mathbb{E}[q] > 0$ if $\alpha^\top x \geq 0$, and $\Pr[\alpha^\top x] > 0$. Further, for $\alpha \in \mathbb{R}^n$, $\alpha^\top \mathbb{E}[q] \geq 0$ if $\alpha^\top x = 0$.
- **Assumption 3**: If $\alpha^\top x = \alpha^{*\top} x$ and $\alpha^* \mathbb{E}[q] = \alpha^{*\top} \mathbb{E}[q]$ for some $\alpha, \alpha^* \in \mathbb{R}^n$, then $\alpha = \alpha^*$.

Remark 6.11. Assumption 2 is a statement of the principle of arbitrage applied to expected prices. It guarantees a nonnegative SDF s.t. Equation (5) holds.

Theorem 6.12. *Under Assumptions 1-3, the squared HJ distance between the SDF proxy y and the set of true SDFs satisfying Equation* (5) *is given by*

$$\delta^2 = \mathbb{E}[xy - q]^\top \mathbb{E}[xx^\top]^{-1} \mathbb{E}[xy - q].$$

Remark 6.13. The specification error δ equals the *maximum pricing error* among all portfolios with the second moment equal to 1. For example, if $\delta = 0.3$, then the expected pricing error is 30 cents on a portfolio with return standard deviation equal to 1 dollar. Note that this value is quite large.

Now consider estimating the specification error. Suppose there are T observations: $\{x_t, q_t, y_t\}_{t=1}^T$.

- **Assumption 4**: The process $\{x_t, q_t.y_t\}$ is stationary and ergodic.

The estimator is

$$\hat{\delta}^2 = \frac{1}{T} \left[\sum_{t=1}^{T-1} x_{t+1} y_{t+1} - q_t \right]^\top \left(\sum_{t=1}^{T-1} x_{t+1} x_{t+1}^\top \right)^{-1} \left[\sum_{t=1}^{T-1} x_{t+1} y_{t+1} - q_t \right].$$

Theorem 6.14. *Under Assumptions 1-4, $\hat{\delta}^2$ converges a.s. to δ^2.*

Theorem 6.15. *Let $\eta = \mathbb{E}[x_{t+1} y_{t+1} - q_t], h_{1t+1} = (x_{t+1} y_{t+1} - q_t) - \eta, \Omega = \mathbb{E}[x_{t+1} x_{t+1}^\top], h_{2t+1} = x_{t+1} x_{t+1}^\top - \Omega$. Assume $\eta \neq 0$. Under Assumptions 1-4,*

$$\sqrt{T}(\hat{\delta}^2 - \delta^2) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} h_{t+1} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V),$$

where

$$h_t = 2\eta^\top \Omega^{-1} h_{1t} - \eta^\top \Omega^{-1} h_{2t} \Omega^{-1} \eta,$$
$$V = \sum_{j=-\infty}^{\infty} \mathbb{E}[h_t h_{t-j}].$$

Example 6.16. We can evaluate a variety of SDF models: CAPM, CRRA, EZ, FF3, FF5, DEF, and CAY. Empirical evidence shows that all models are strongly rejected by the J test, a test for model misspecification. All confidence intervals overlap, so the differences are not significantly different. As a result, we cannot rank different models.

6.4 Estimation and Inference

We apply the generalized method of moments (GMM) to estimate and test models of the SDF. Assume m_t is a valid SDF. Then, for any return vector R_t, we have $1 = \mathbb{E}[m_{t+1} R_{t+1} \mid \mathcal{F}_t]$, where $\mathcal{F}_t$ is the information set at t. Explicitly,

$$1 = \int m_{t+1} R_{t+1} f(m_{t+1}, R_{t+1} \mid \mathcal{F}_t) dR_{t+1} dm_{t+1}.$$

If an economic specifies the joint conditional distribution of m_{t+1} and R_{t+1}, we can compute the integral explicitly and estimate the SDF by minimizing the pricing errors. However, an economic model usually only provides some conditional moment restrictions, not the joint distribution.

Example 6.17. Let $m_t(\theta)$ be a model of the SDF with θ being a finite-dimensional parameter vector. Let $R_t = (R_{1,t}, \ldots, R_{q,t})$ be a vector of returns. Then $1 = \mathbb{E}[m_{t+1}(theta)R_{t+1} \mid \mathcal{F}]$. Thus $\forall z_t \in \mathcal{F}_t$, we have $\mathbb{E}[m_{t+1}(\theta)R_{i,t+1}z_t] = 0 \ \forall i \in [q]$. Moreover, for any well-behaved function h, $\mathbb{E}[(m_{t+1}(\theta)R_{i,t+1} - 1)g(z_t)] = 0 \ \forall i \in [q]$.

Example 6.18 (Nagel and Singleton (2011))**.** Their goal is to estimate and test several conditionally affine SDFs that take the form $m_{t+1}(\theta) = (\beta_1 + \gamma_1 s_t) + (\beta_2 + \gamma_2 s_t)\Delta c_{t+1}$, where $\Delta c_{t+1} = \log(c_{t+1}/c_t)$ is the consumption growth, and s_t is either the consumption-wealth ratio, the corporate bond spread, or the labor income-consumption ratio. R_{t+1} is the real returns on 4 Fama-French portfolios and the 3-month t-bill rate, and $z_t = (1, R_t, s_t, \Delta c_t)$.

6.4.1 The GMM Estimator

The GMM estimator is based on unconditional moment restrictions. Suppose these moment conditions are $\mathbb{E}[u(x_t, \theta_0)] = 0$, where $u : \mathbb{R}^N \to \mathbb{R}$, x_t is a random vector, and $\theta_0 \in \mathbb{R}^K$ is a parameter vector. x_t can include lagged values of variables, so $u(x_t, b_0)$ can be serially correlated. In Example 6.17, $u(x_{t+1}, \theta_0) = (m_{t+1}(\theta)R_{t+1} - 1) \otimes z_t \in \mathbb{R}^{pq}$, where $R_{t+1} \in \mathbb{R}^q, z_t \in \mathbb{R}^p$. Define

$$g_T(b) = \frac{1}{T}\sum_{t=1}^{T} u(x_t, \theta).$$

The GMM estimator is computed by setting $g_T(\theta)$ as close to zero as possible with respect to a weighting matrix W_T:

$$\hat{\theta}(W_T) = \arg\min_{\theta} g_T(\theta)^\top W_T g_T(\theta).$$

A central object in the GMM theory is the covariance matrix of $T^{1/2}g_T(\theta)$:

$$S_0 = \lim_{n\to\infty} \mathrm{Var}[T^{1/2}g_T(\theta)] = \sum_{s=-\infty}^{\infty} \mathbb{E}[u(x_t, \theta_0)\theta(v_{t-s}, \theta_0)^\top].$$

This is the long run covaraince matrix of $u(x_t, \theta_0)$. In Example 6.17, $u(x_t, \theta_0)$ is serially uncorrelated, so $S_0 = \mathbb{E}[u(x_t, \theta_0)u(x_t, \theta_0)^\top]$. If the model is correctly specified, then under some regularity conditions,

$$\sqrt{T}(\hat{\theta}(W_T) - \theta_0) \xrightarrow{d} \mathcal{N}(0, V(W_0)),$$

where $V(W_0) = [G_0^\top W_0 G_0]^{-1}(G_0^\top W_0 S_0 W_0 G_0)[G_0^\top W_0 G_0]^{-1}$, $W_0 = \lim W_T$ (non-random and positive definite), and $G_0 = \partial\mathbb{E}[u(x_t, \theta_0)]/\partial\theta'$.

6.4.2 The J Test

By the delta method,

$$\sqrt{T}g_T(\hat{\theta}(W_T)) \xrightarrow{d} \mathcal{N}(0, Q(W_0)),$$

where $Q(W_0) = (I - G_0[G_0^\top W_0 G_0]^{-1}G_0^\top W_0)S_0(I - G_0[G_0^\top W_0 G_0]^{-1}G_0^\top W_0)^\top$. Let $\hat{Q}$ be a consistent estimator of $Q(W_0)$, $\hat{Q}^+$ be its pseudoinverse. Then under the null hypothesis,

$$J = T g_T(\hat{\theta})^\top \hat{Q}^+ g_T(\hat{\theta}) \xrightarrow{d} \chi^2_{N-K}.$$

This is the J-statistic for correct model specification.

Remark 6.19. The covariance matrix $V(W_0)$ is minimized when $W_0 = S_0^{-1}$. The estimator with $W_T = \hat{S}^{-1}$ is called the optimal GMM estimator. However, it is important to be aware that

1. The resulting estimator is optimal within a small family, *i.e.*, among estimators using the same moment estimators.
2. In practice, the weighting matrix is very hard to estimate, often leading to disappointing finite sample properties.
3. When the model is misspecified, different weighting schemes imply different pseudo-true values.

Asset returns are often strongly cross-sectionally correlated, so $\hat{S}$ is often singular. We can use $W_T = I$ or $W_T = (\mathrm{Diag}\,\hat{S})^{-1}$.

6.5 Applications

6.5.1 CAPM

Consider the usual system for the CAPM, where $R_t = \alpha + \beta R_{mt} + e_t$. If the CAPM holds, then $\mathbb{E}[R_t] = \gamma\beta$, where $\gamma = \mathbb{E}[R_{mt}]$. The parameters γ, β, α can be estimated using the GMM:

$$\mathbb{E}\begin{bmatrix} R_t - \alpha - \beta R_{mt} \\ (R_t - \alpha - \beta R_{mt})R_{mt} \\ R_t - \gamma\beta \end{bmatrix} = 0.$$

Let

$$g_T(\gamma, \beta, \alpha) = \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} R_t - \alpha - \beta R_{mt} \\ (R_t - \alpha - \beta R_{mt})R_{mt} \\ R_t - \gamma\beta \end{pmatrix}.$$

The GMM estimator with an identity weighting matrix is given by

$$(\hat{\gamma}, \hat{\beta}, \hat{\alpha}) = \arg\min_{\gamma,\beta,\alpha} g_T(\gamma, \beta, \alpha)^\top g_T(\gamma, \beta, \alpha).$$

We have the following:

$$\begin{aligned} \hat{\beta} &= \frac{\sum(R_{mt} - \overline{R}_m)(R_t - \overline{R})}{\sum(R_{mt} - \overline{R}_m)^2}, \\ \hat{\alpha} &= \overline{R} - \hat{\beta}\overline{R}_m, \\ \hat{\gamma} &= (\hat{\beta}^\top\hat{\beta})^{-1}\hat{\beta}^\top\overline{R}, \end{aligned}$$

where $R = T^{-1}\sum R_t$. The null hypothesis of the CAPM is $\mathbb{E}[R_t] - \gamma\beta = 0$, which can be tested using the J-test: $Tg_T(\hat{\gamma}, \hat{\beta}, \hat{\alpha})^\top \hat{Q}^* g_T(\hat{\gamma}, \hat{\beta}, \hat{\alpha})$, where $\hat{Q}$ is a consistent estimate of the variance of $\sqrt{T}g_T(\hat{\gamma}, \hat{\beta}, \hat{\alpha})$. Under the null hypothesis, this test converges to χ^2_{N-1}.

6.5.2 Example 6.18

We compare different estimators:

- Unconditional: based on $\mathbb{E}[m_{t+1}(\theta)R_{t+1} - 1] = 0$, where the elements of p are 1 for gross returns and 0 for excess returns.
- Fix IV: based on $\mathbb{E}[(m_{t+1}(\theta)R_{t+1}) \otimes z_t]$, where $z_t = (1, R_t, s_t, \Delta c_t)$.
- Optimal IV Sieve: based on $\mathbb{E}[z_t^*(m_{t+1}(\theta)R_{t+1} - 1)] = 0$, where z_t^* contains the optimal instruments, and the conditional expectations are estimated using the sieve method.

Empirically, we find

- The Δc_{t+1} coefficients differ drastically between the conditional and unconditional cases.
- The weighting matrix matters.
- The J test overwhelmingly rejects the model in the conditional case, but not in the unconditional case.
- The results from fixed IV (using optimal weighting matrix) is similar to that of the optimal IV sieve.

7 Continuous-Time Models

7.1 Discrete-Time Martingales

Define $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$.A sequence of random variables $\{Y_t\}$ is a *martingale* w.r.t. the sequence of random variables $\{X_t\}$ if:

1. $\forall t \geq 1, \exists f_t : \mathbb{R}^t \to \mathbb{R}$ s.t. $Y_t = f_t(X_1, \ldots, X_t)$.
2. $\{Y_t\}$ satisfies the fundamental martingale identity $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = Y_{t-1}$ $\forall t \geq 1$.

If a sequence of random variable $\{A_t\}$ is s.t. $\forall t$, we have $A_t \in \mathcal{F}_{t-1}$, then we say $\{A_t\}$ is *non-anticipating w.r.t.* $\{\mathcal{F}_t\}$. The process $\{\tilde{Y}_t\}$ defined by setting $\tilde{Y}_0 = Y_0$ and $\tilde{Y}_t = Y_0 + A_1(Y_1 - Y_0) + A_2(Y_2 - Y_1) + \cdots + A_t(Y_t - Y_{t-1})$ $\forall t \geq 1$ is called the *martingale transform* of $\{Y_t\}$ by $\{A_t\}$.

Theorem 7.1 (Martingale transform theorem)**.** *If $\{Y_t\}$ is a martingale w.r.t. $\{\mathcal{F}_t\}$, and if $\{A_t\}$ is a sequence of bounded random variables that are non-anticipating w.r.t. $\{\mathcal{F}_t\}$, then the sequence of martingale transforms $\{\tilde{Y}_t\}$ is itself a martingale w.r.t. $\{\mathcal{F}_t\}$.*

A random variable τ that takes values in $\mathbb{N} \cup \{0, \infty\}$ is called a *stopping time* for the sequence $\{\mathcal{F}_t\}$ if $\{\tau \leq t\} \in \mathcal{F}_t \ \forall t$.

Remark 7.2. In practice, to avoid stopping at ∞, we often use $t \wedge \tau := \min\{t, \tau\}$, which is a bounded stopping time.

Theorem 7.3 (Stopping time theorem)**.** *If $\{Y_t\}$ is a martingale w.r.t. the sequence $\{\mathcal{F}_t\}$, then the stopped process $\{Y_{t\wedge\tau}\}$ is also a martingale w.r.t. $\{\mathcal{F}_t\}$.*

If the integrable random variables $Y_t \in \mathcal{F}_t$ satisfy $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \geq Y_{t-1} \ \forall t \geq 1$, then we say $\{Y_t\}$ is a *sub-martingale* adapted to $\{\mathcal{F}_t\}$. The martingale transform and stopping time theorems apply to submartingales.

7.2 Continuous-Time Martingales and the Ito Integral

A continuous time stochastic process $\{B_t\}$ is called a *standard Brownian motion* on $[0, T)$ if it has the following four properties:

1. $B_0 = 0$.
2. The increments of B_t are independent, that is, for any finite set of times $0 < t_1 < t_2 < \cdots < t_n < T$, the random variables $B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent.
3. For any $0 \leq s < t < T$, the increments $B_t - B_s \sim \mathcal{N}(0, t - s)$.
4. For all ω in a set of probability one, $B_t(\omega)$ is a continuous function w.r.t. t.

If a collection of σ-algebras $\{\mathcal{F}_t\}$ satisfy $\mathcal{F}_s \subset \mathcal{F}_t \ \forall s \leq t$, then we call it a *filtration*. If the random variables $\{X_t\}$ are s.t. X_t is $\mathcal{F}_t$-measurable, then we say $\{X_t\}$ is *adapted* to $\{\mathcal{F}_t\}$.

Suppose $\{X_t\}$ is adapted to $\{\mathcal{F}_t\}$. We say $\{X_t\}$ is a *martingale* if:

1. $\mathbb{E}[|X_t|] < \infty \ \forall t$ and
2. $\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s \ \forall s, t$ s.t. $0 \leq s \leq t < \infty$.

The Ito integral is denoted by

$$\int_0^T f(\omega, t)\, dB_t,$$

where B_t is the standard Brownian motion, and ω indicates that f can depend on the history of B_t.

Theorem 7.4. *For any continuous function $f : \mathbb{R} \to \mathbb{R}$ that satisfies $\mathbb{E}[(f(B_{t+h} - f(B_t))^2] \leq Ch$ for some $C < \infty, h > 0$, if we take the partition of $[0, T]$ given by $t_i = i\Delta t$ with $\Delta t = T/2^{-n}$ for $0 \leq i \leq n$, then*

$$\lim_{n\to\infty} \sum_{i=1}^{n} f(B_{t_{i01}})(B_{t_i} - B_{t_{i-1}}) = \int_0^T f(B_s)\, dB_s,$$

where the limit is understood in the sense of convergence in probability.

More generally, for any adapted, measurable function $f : \Omega \times [0, T] \to \mathbb{R}$ with

$$\Pr\left[\int_0^T f^2(\omega, t)\, dt < \infty\right] = 1,$$

the integral

$$\int_0^T f(\omega, t)\, dB_t$$

exists and is unique.

7.3 Ito's Formula

Theorem 7.5. *If the function $f : \mathbb{R} \to \mathbb{R}$ has a continuous second derivative, then*

$$f(B_t) = f(B_0) + \int_0^t f'(B_s)\, dB_s + \frac{1}{2}\int_0^t f''(B_s)\, ds.$$

In practice, we often express this result as

$$df(B_t) = f'(B_t)\, dB_t + \frac{1}{2} f''(B_t)\, dt,$$

where the initial value is $f(B_0)$.

Theorem 7.6. *If $f(t, B_t)$ has a continuous first derivative in the first argument and a continuous second derivative in the second argument, then*

$$f(t, B_t) = f(0,0) + \int_0^t \frac{\partial}{\partial x} f(s, B_s)\, dB_s + \int_0^t \frac{\partial}{\partial s} f(s, B_s)\, ds + \frac{1}{2}\int_0^t \frac{\partial^2}{\partial x^2} f(x, B_s)\, ds.$$

In practice, we often express this result as

$$df(t, B_t) = f_x\, dB_s + \left[f_t + \frac{1}{2} f_{xx} \right] dt,$$

where the initial value is $f(0, B_0)$.

We say that a proces $\{X_t\}_{0 \le t \le T}$ is *standard* if $\{X_t\}$ has the representation

$$X_t = X_0 + \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s \ \forall 0 \le t \le T,$$

where a, b are adapted, measurable processes with

$$\Pr\left[\int_0^T |a(\omega, s)|\, ds < \infty \right] = 1, \Pr\left[\int_0^T |b(\omega, s)|\, ds < \infty \right] = 1.$$

Theorem 7.7. *If $f(t, B_t)$ has a continuous first derivative in its first argument and a continuous second derivative in its second argument, and $\{X_t\}_{0 \le t \le T}$ is a standard process with the integral representation*

$$X_t = \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s \ \forall 0 \le t \le T,$$

then

$$f(t, X_t) = f(0,0) + \int_0^t f_x dX_s + \int_0^t f_s\, ds + \frac{1}{2}\int_0^t f_{xx} b^2(\omega, s)\, ds.$$

In practice, we often write these two equations as

$$dX_t = a(\omega, t)\, dt + b(\omega, t)\, dB_t,$$
$$df(t, X_t) = f_x\, dX_t + f_t\, ft + \frac{1}{2} f_{xx} b^2(\omega, t)\, dt.$$

Theorem 7.8. *If $f(X_t, Y_t)$ has continuous second derivatives in both arguments, and X_t and Y_t are standard processes with integral representations*

$$X_t = \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s$$
$$Y_t = \int_0^t \alpha(\omega, s)\, ds + \int_0^t \beta(\omega, s)\, dB_s,$$

then

$$df(X_t, Y_t) = f_x\, dX_t + f_y\, dY_t + \frac{1}{2} f_{xx} b^2(\omega, s)\, dt + \frac{1}{2} f_{yy} \beta^2(\omega, s)\, dt + f_{xy} b(\omega, s) \beta(\omega, s)\, dt.$$

Theorem 7.9. *If $f(X_t, Y_t)$ has continuous second derivatives in both arguments, and X_t and Y_t are standard processes with integral representations*

$$X_t = \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s^1$$
$$Y_t = \int_0^t \alpha(\omega, s)\, ds + \int_0^t \beta(\omega, s)\, dB_s^2,$$

where B_s^1, B_s^2 are two independent Brownian motions, then

$$df(X_t, Y_t) = f_x\, dX_t + f_y\, dY_t + \left[\frac{1}{2} f_{xx} b^2(\omega, s) + \frac{1}{2} f_{yy} \beta^2(\omega, s)\right] dt.$$

Theorem 7.10. *If $f(X_t, Y_t)$ has continuous second derivatives in both arguments, and X_t and Y_t are standard processes with integral representations*

$$X_t = \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s^1$$
$$Y_t = \int_0^t \alpha(\omega, s)\, ds + \int_0^t \beta(\omega, s)\, dB_s^2,$$

where B_s^1, B_s^2 are two standard Brownian motions with

$$\mathbb{E}[dB_s^1 dB_s^2] = \rho\, dt,$$

then

$$df(X_t, Y_t) = f_x\, dX_t + f_y\, dY_t + \left[\frac{1}{2} f_{xx} b^2(\omega, s) + \frac{1}{2} f_{yy} \beta^2(\omega, s) + \rho f_{xy} b(\omega, s)\beta(\omega, s)\right] dt.$$

7.4 Quadratic Variation

A finite ordered set of grid points $\pi_n = \{t_0 \le t_1 \le \cdots \le t_n\}$ with $t_0 = 0$ and $t_n = t$ is called a *partition* of $[0, t]$. The *mesh* $\mu(\pi)$ of a partition π is the maximum grid size of this partition. For any partition π_n of $[0, t] \subset [0, T]$ and for any process $\{X_t\}$ on $[o0, T]$, the π_n-*quadratic variation* of the process $\{X_t\}$ is defined to be

$$Q_{\pi_n}(X_t) = \sum (X_{t_i} - X_{t_{i-1}})^2.$$

If Q_{π_n} converges in probability to a process $\{V_t\}$ for any sequence of partitions $\{\pi_n\}$ of $[0, t]$ s.t. $\mu(\pi_n\} \to 0$ as $n \to \infty$, then we say that $\{V_t\}$ is the *quadratic variation* of $\{X_t\}$, often denoted as $\langle X \rangle_t$.

Theorem 7.11. *If X_t is a standard process with the process with the representation*

$$X_t = \int_0^t a(\omega, s)\, ds + \int_0^t b(\omega, s)\, dB_s, \tag{6}$$

then the quadratic variation of X_t exists and is given by

$$\langle X \rangle_t = \int_0^t b(\omega, s)^2\, ds$$

for $t \in [0, T]$.

Remark 7.12. In Equation (6), the first term represents the drift and the second term represents the volatility. We will see this in later stochastic differential equations.

7.5 Continuous-Time Models

We will focus on the Black-Scholes (1973) and Merton (1973) models. We have the following assumptions:

1. There is no market imperfection. That is, there are no taxes, transactions costs, or short sales constraints, and the trading is continuous and frictionless.
2. There is unlimited opportunity for riskless borrowing and lending at the continuously compounded rate of return r. A \$1 investment in such an asset over the time interval τ grows to $\exp(r\tau)$. Alternatively, if $D(t)$ is the date t price of discount bound maturing at date T with face value 1, then for $t \in [0, T]$, the bond price dynamics are given by $dD(t) = rD(t)\, dt$.

3. The stock price follows a geometric Brownian motion, which is the solution to the following Ito SDE on $t \in [0, T]$:
$$dP(t) = \mu P(t)\, dt + \sigma P(t)\, dB(t), \tag{7}$$
where $P(0) = P_0 > 0$, $B(t)$ is a standard Brownian motion, and at least one investor observes σ without error.

4. There are no arbitrage opportunities.

Remark 7.13. The solution to Equation (7) is given by
$$P(t) = P_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma B(t)\right).$$

Let G denote the price of a call option. We assume that it depends only on the interest rate and the stock price, that is $G = G(P, t)$. Then by Ito's formula, $dG = \mu_g G\, dt + \sigma_g\, dB(t)$, where
$$\mu_g = \frac{1}{G}\left(\mu P \frac{\partial G}{\partial P} + \frac{\partial G}{\partial t} + \frac{\sigma^2 P^2}{2}\frac{\partial^2 G}{\partial P^2}\right),$$
$$\sigma_g = \frac{1}{G}\left(\sigma P \frac{\partial G}{\partial P}\right).$$

We have two risky assets, a stock and an option, driven by a single Brownian motion. Thus, a certain linear combination of the two assets can eliminate the influence of the Brownian motion, producing a riskfree asset. Let $I_g(t)$ and $I_p(t)$ denote the dollar amount invested in the option and the stock, respectively. The instantaneously dollar return of the portfolio is
$$\begin{aligned} dI &= \underbrace{\frac{I_p}{P}}_{\text{shares of stock}} dP + \underbrace{\frac{I_g}{G}}_{\text{shares of option}} dG \\ &= \frac{I_p}{P}(\mu P(t)\, dt + \sigma P(t)\, dB(t)) + \frac{I_g}{G}(\mu_g G\, dt + \sigma_g G\, dB(t)) \\ &= (I_p \mu + I_g \mu_g)\, dt + (I_p \sigma + I_g \sigma_g)\, dB(t). \end{aligned}$$
We choose $I_p\sigma + I_g\sigma_g = 0$, so
$$dI = I_g\left(\mu_g - \mu\frac{\sigma_g}{\sigma}\right) dt.$$
Notice that this requires being able to buy and sell assets in continuous-time. Then the same dollar amount invested in the bond produces the following return:
$$\frac{I}{D}\, dD(t) = \frac{I}{D} rD\, dt = r(I_g + I_p)]\, dt = rI_g\left(1 - \frac{\sigma_g}{\sigma}\right) dt.$$
The above two returns must equal to each other to rule out arbitrage:
$$\mu_g - \mu\frac{\sigma_g}{\sigma} = r\left(1 - \frac{\sigma_g}{\sigma}\right).$$
This implies
$$\mu_g - r = \frac{\mu - r}{\sigma}\sigma_g,$$
so
$$\frac{\sigma^2 P^2}{2}\frac{\partial^2 G}{\partial P^2} + rP\frac{\partial G}{\partial P} + \frac{\partial G}{\partial t} - rG = 0. \tag{8}$$
Equation (8) can be solved subject to the boundary conditions
$$G(P(T), T) = \max\{P(T) - X, 0)\},$$
$$G(0, t) = 0.$$
The solution is:
$$\begin{aligned} G(P(t), t) &= P(t)\Phi(d_1) - Xe^{-r(T-t)}\Phi(d_2) \\ d_1 &= \frac{\log P(t)/X + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}} \\ d_2 &= \frac{\log P(t)/X + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}, \end{aligned}$$
where $\Phi(\cdot)$ is the standard normal CDF, and $(T - t)$ is the time to maturity.

Remark 7.14. Because of the assumption that the market is complete, the option is redundant. Hence, its price does not depend on the agents' preferences.

7.5.1 The Martingale Approach

We have showed: There exists a portfolio that eliminates the risk from the Brownian motion: the risk is zero for every state and all t. Consider a probability measure that assigns zero probability to the same events as the real-world probability measure. Denote it by Q^*. Then, a portfolio eliminates risk under Q iff it eliminates risk under Q^*. To apply the arbitrage argument, it does not matter which probability measure we are working with. It turns out to be computationally useful to consider the risk-neutral measure. Define

$$B^*(t) = B(t) - \frac{\mu - r}{\sigma} t.$$

Then, Girsanov's theorem says that $\exists Q^*$ under which $B^*(t)$ is a Brownian motion. Thus,

$$dP(t) = rP(t)\,dt + \sigma P(t)\,dB^*(t).$$

The stock and bond now both have r as their instantaneous expected rate of return under the new measure. Thus, to rule out arbitrage, the option must also have r as its instantaneous expected rate of return. Hence

$$G(t) = e^{-r(T-t)}\mathbb{E}_{Q^*}[\max\{P(T) - X, 0\}],$$

where conditional expectation is taken w.r.t. the risk-neutral measure. In practice, if we can simulate the risk-neutral measure, then we can obtain the option price.

7.5.2 Violations of the Model

The Black-Scholes model will fail under any of the following four violations of its assumptions. These levels become increasingly serious and difficult to remedy as we move down the list:

1. The local volatility of the underlying asset, the riskless interest rate, or the asset payout rate is a function of the concurrent underlying asset price or time.
2. The local volatility of the underlying asset, the riskless interest rate, or the asset payout rate is a function of the prior path of the underlying asset price.
3. The local volatility of the underlying asset, the riskless interest rate, or the asset payout rate is a function of a state-variable which is not the concurrent underlying asset price or the prior path of the underlying asset price; or the underlying asset price, interest rate or payout rate can experience jumps in level between successive opportunities to trade.
4. The market has imperfections such as significant transactions costs, restrictions on short selling, taxes, noncompetitive pricing, etc.

Remark 7.15. Although, violations of types 1 and 2 still leave the arbitrage reasoning-the essence of the Black-Scholes argument-intact, type 2 violations lead perhaps to insurmountable computational problems. Violations of type 3 are far more serious, since they destroy the arbitrage foundations of the Black-Scholes model and have left researchers so far with two unpalatable alternatives: either an equilibrium model in which investor preferences explicitly enter, or other securities in addition to the underlying and riskless assets must be included in the arbitrage strategy. Violations of type 4 are the worst, because their effects are notoriously difficult to model and they typically lead only to bands within which the option price should lie.

7.6 Estimating the Parameters of the Black-Scholes Model

From an econometrician's perspective, if we know the price of an option, we can estimate the parameters of an option. The main challenge is that we have a continuous-time model, but only discrete-time observations.

Consider estimating μ and σ^2 in Equation (7). A natural approach is to obtain a discrete approximation to the continuous-time model and then apply MLE to this approximation. Suppose h is a time interval, *e.g.*, 5 minutes or a day. Applying the Euler discretization to the model, we have

$$P(t+h) - P(t) \approx \mu P(t)h + \sigma P(t)(B(t_h) - B(t)),$$

or

$$\frac{P(t_h) - P(t)}{P(t)} \approx \mu h + \sigma(B(t+h) - B(t)),$$

and the approximation error decreases to zero as $h \to 0$. Suppose the prices are available for this interval frequence, *i.e.*, we have $P_k = P(kh),\ k = 0, 1, \ldots, n$ with $n = T/h$l Denote the simple returns by $R_k(h) := P_k/P_{k-1} - 1$. Then since $B(t_h) - B(t) \sim \mathcal{N}(0, h)$ *i.i.d.*, the approximating log-likelihood for $\{R_k(h)\}$ is

$$L(\mu, \sigma) = -\frac{n}{2} \log 2\pi\sigma^2 h - \frac{1}{2\sigma^2 h} \sum_{k=1}^{n} (R_k(h) - \mu h)^2.$$

The MLE is

$$\hat{\mu} = \frac{1}{nh} \sum_{k=1}^{n} R_k(h),$$
$$\hat{\sigma}^2 = \frac{1}{nh} \sum_{k=1}^{n} (R_k(h) - \hat{\mu})^2.$$

We notice that this estimator is inconsistent. For a fixed $h > 0$, by the LLN,

$$\hat{\mu} \xrightarrow{p} \frac{1}{h} \mathbb{E}[R_k(h)],$$
$$\hat{\sigma}^2 \xrightarrow{p} \frac{1}{h} \operatorname{Var}[R_k(h)].$$

Now compare $\hat{\mu}, \hat{\sigma}^2$ with the model's parameters. By the solution to the SDE, we have

$$P_k = P_0 \exp\left(\sigma B(kh) + \left(\mu - \frac{1}{2}\sigma^2\right)(kh)\right).$$

Therefore, $\forall h > 0$, $R_k(h)$ is log-normally distributed with

$$\mathbb{E}[R_k(h)] = e^{\mu h} - 1,$$
$$\operatorname{Var}[R_k(h)] = e^{2\mu h}(e^{\sigma^2 h} - 1).$$

Thus, $\hat{\mu} \not\xrightarrow{p} \mu$ and $\hat{\sigma}^2 \not\xrightarrow{p} \sigma^2$. Furthermore, the estimates are biased for fixed h; however, the bias converges to 0 as $h \to 0$. Now we study the asymptotic distribution of $(\hat{\mu} - \mu)$ as $T \to \infty$ and $h \to 0$:

$$\begin{aligned}\sqrt{nh}(\hat{\mu} - \mu) &= \sqrt{nh}\left(\frac{1}{n}\sum_{k=1}^{n} R_k(h) - \frac{\mathbb{E}[R_k(h)]}{h} + \frac{\mathbb{E}[R_k(h)]}{h} - \mu\right) \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{R_k(h) - \mathbb{E}[R_k(h)]}{\sqrt{h}} + \sqrt{nh}\left(\frac{\mathbb{E}[R_k(h)]}{h} - \mu\right).\end{aligned}$$

The first term converges to $\mathcal{N}(0, \sigma^2)$, and the second term is $O(\sqrt{nh}h) = O(T^{1/2}h)$. Therefore, if $T \to \infty$ and $T^{1/2}h \to 0$, *i.e.*, we sample more frequently as the sample size increases, then $\sqrt{nh}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

Remark 7.16. In this example, the bias is $O(h)$.

Now, instead of simple returns, we use continuously compounded returns $r_k(h) = \log P_k / \log P_{k-1}$ for estimation. Applying Ito's formula, we have

$$d \log P_t = \alpha\, dt + \sigma\, dB(t),$$

where $\alpha = \mu - \sigma^2/2$. Thus, $\log P_k/P_{k-1}$ is normally distributed with mean αh and standard error $\sigma\sqrt{h}$. Then, the sample standard deviation of $r_k(h)/\sqrt{h}$ is an unbiased estimate of σ. In fact, it is the MLE of σ. The log-likelihood function is given by

$$L(\mu, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2 h) - \frac{1}{2\sigma^2 h} \sum_{k=1}^{n} (r_k(h) - \alpha h)^2,$$

and the MLEs are

$$\hat{\alpha} = \frac{1}{nh} \sum_{k=1}^{n} r_k(h)$$
$$\hat{\sigma}^2 = \frac{1}{nh} \sum_{k=1}^{n} (r_k - \hat{\alpha} h)^2.$$

The estimates are unbiased and efficient.

7.7 Estimation and Inference

Suppose our model is

$$dX_t = \mu(X_t, t; \theta)\, dt + \Sigma(X_t, t; \theta)\, dW_t,$$

where W_t is a vector of independent Brownian motions, the drift vector $\mu \in \mathbb{R}^K$ and the diffusion matrix $\Sigma \in \mathbb{R}^{K \times K}$ are functions of the vector X_t, time t, and an L-dimensional parameter vector θ whose true value is θ_0. This model is a multivariate diffusion: μ, Σ depend on the underlying process only through the most recent value, and there are no latent processes such as volatility processes. The goal is to estimate θ.

There are four estimation approaches:

1. Simulated MLE: First, obtain a discrete-time approximation to the continuous-time model. Next, use simulations to compute the likelihood function for this discrete-time model.
2. MCMC: First, obtain a discrete-time approximation to the continuous-time model. Next, use MCMC to sample from the posterior distribution of θ, given this discrete-time approximation and some priors on parameters. Finally, compute the mean (or mode) of this posterior distribution as an estimate for θ.
3. Analytic approach: Compute the likelihood analytically without any discretization by exploiting the structure of the model.
4. Simulated methods of moments (SMM).

Remark 7.17. We will focus on the first and third approaches. Later, we will discuss the second approach in the context of stochastic volatility models.

7.7.1 Simulated MLE

Let the available sample be $X_{t_0}, \ldots, X_{t_N}$. Suppose we have daily or weekly observations. The joint density of $X_{t_0}, \ldots, X_{t_N}$ is

$$\begin{aligned} f(X_{t_0}, \ldots, X_{t_N}; \theta_0) &= f(X_{t_0}; \theta_0) \prod_{n=0}^{N-1} f(X_{t_{n+1}} \mid X_{t_n}, \ldots, X_{t_0}; \theta_0) \\ &= f(X_{t_0}; \theta_0) \prod_{n=0}^{N-1} f(X_{t_{n+1}} \mid X_{t_n}; \theta_0), \end{aligned}$$

where the last equality uses the DGP's Markovian property. Notice that we cannot analytically compute $f(X_{t_{n+1}} \mid X_{t_n}; \theta_0)$ because the DGP is in continuous time. One approach is to obtain a discrete time approximation to the model, and then compute the transition density of this approximation. Consider just the interval $[t_0, t_1]$. Divide this interval into M sub-intervals of length $h = 1/M$. Denote these intervals by $[t_0, t_0 + h], \ldots, [t_0 + (M-1)h, t_1]$. Note that we only observe data at times t_0 and t_1.

The estimation procedure is as follows:

1. Pick an initial value for θ.
2. For each t_n with $n = 1, \ldots N$, start at t_n and apply the Euler approximation $M - 1$ times to produce a chain of values:

   $$X_{t_n} \to X_{t_n + h} \to \cdots \to X_{t_n + (M-1)h}.$$

   Denote the realization of $X_{t_n+(M-1)h}$ by $z^{(1)}$. Repeat this for S times to obtain a sample of realizations:

   $$\{z^{(1)}, \ldots, z^{(S)}\}.$$

3. Compute a sample average to approximate $f(X_{t_{n+1}} \mid X_{t_n}; \theta_0)$:

   $$f(X_{t_{n+1}} \mid X_{t_n}; \theta_0) = \int f(X_{t_{n+1}} \mid X_{t_n+(M-1)h}; \theta_0) f(X_{t_n+(M-1)h} \mid X_{t_n})\, dX_{t_n+(M-1)h}$$

   $$f_{M,S}(X_{t_{n+1}} \mid X_{t_n}; \theta) = \frac{1}{S} \sum_{s=1}^{S} \phi(X_{t_{n+1}}; z^{(s)} + \mu(z^{(s)}; \theta)h, \Sigma(z^{(s)}; \theta_0)\Sigma(z^{(s)}; \theta_0)^\top h),$$

   where ϕ is the standard normal density.

4. Compute the approximate log-likelihood

$$L_{M,S}(\theta) = \sum_{n=1}^{N-1} \log f_{M,S}(X_{t_{n+1}} \mid X_{t_n}; \theta).$$

5. Repeat steps 1-4 for different θ and search for the value that maximizes the above approximate likelihood:

$$\hat{\theta}_{M,S} = \arg\max_{\theta} L_{M,S}(\theta).$$

Remark 7.18. As we vary the parameters, we should use the same error terms (ε_t) in the Euler discretization step to generate $z^{(s)}$. This produces densities that are smooth w.r.t. the parameters. Two tuning parameters are involved: M controls the discretization bias, while S determines the quality of the simulation approximation. In theory, we need $M \to \infty$ and $S \to \infty$ for their effects to be negligible. In practice, a small M tends to work well for daily observations and a large S, *e.g.*, $S > 5000$ tends work for empirically relevant applicatons.

We have the following assumptions:

1. The drift μ and diffusion Σ functions are infinitely differentiable with continuous and bounded derivatives of all orders.
2. The covariance matrix $\Sigma\Sigma^\top$ is positive definite.
3. $\theta \in \Theta$, where Θ is a compact set that contains the true θ_0 in its interior.
4. The likelihood function is twice continuously differentiable in θ in a neighborhood of the true parameter vector θ_0. Furthermore, the average Hessian has full rank and is bounded for all parameters $\theta \in \Theta$.
5. $\forall \lambda \in \mathbb{R}^K$, $N^{-1}\lambda^\top I_N(\theta)\lambda > C > 0$, where

$$I_N(\theta) = \mathbb{E}\left[\sum_{n=0}^{N-1} \frac{\partial \ln f(X_{t_{n+1}|X_{t_n};\theta})}{\partial \theta} \frac{\partial \ln f(X_{t_{n+1}} \mid X_{t_n}; \theta)}{\partial \theta^\top}\right].$$

6. $I_N(\Theta)^{-1/2}\partial \sum_{n=0}^{N-1} \ln f(X_{t_{n+1}} \mid X_{t_n}; \theta_0)/\partial\theta \xrightarrow{d} \mathcal{N}(0, I)$. (This assumption rules out unit-root type processes.)

Theorem 7.19. *Given Assumptions 1-5, as $M \to \infty$ and $S \to \infty$, $\hat{\theta}_{M,S}$ converges to the maximum likelihood estimator $\hat{\theta}$, which in turn converges to the true value θ_0 as $N \to \infty$.*

Theorem 7.20. *Given Assumptions 1-6, as $M, N, S \to \infty$ with $N/S^{1/2} \to 0$ and $N/M \to 0$, we have*

$$I_N(\theta)^{-1/2}(\hat{\theta}_{M,S} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I),$$

where I is an identity matrix.

Remark 7.21. The conditions $N/M \to 0$ an $N/S^{1/2} \to 0$ ensure that the Eueler approximation and the simulation have negligible effects. They can be relaxed to $N^{1/2}/M \to 0$ and $N/S \to 0$, respectively.

7.7.2 Estimation Based on Analytical Approximations

The SMLE method uses simulations to approximate the transition density. Alternatively, for some models, we can approximate the transition analytically using a series expansion. This is similar to approximating a function using a Taylor series, or a Fourier series. The main difference is that the function to be approximated is now a transitional density. Here, we use Hermite polynomials for our approximation.

Consider a scalar process. The extension to the multivariate case is NOT straightforward. The main step is to compute the transition density $f(X_{t_{n+1}} \mid X_{t_n}; \theta)$ using a Hermite polynomial approximation. For a density $f(x)$, we obtain an approximation using a truncated series:

$$f(x) \approx \phi(x) \sum_{j=0}^{J} \beta_j H_j(x),$$

where H_j is the jth order Hermite polynomial. The approximation is most effective if the target density is close to the standard normal. However, $f(X_{t_{n+1}} \mid X_{t_n}; \theta)$ is far from $\mathcal{N}(0, 1)$ for two reasons: the volatility

$\sigma(X_t;\theta)$ is not constant and $f(X_{t_{n+1}} \mid X_{t_n};\theta)$ is peaked around X_{t_n} if h is small. The proposed solution is to first transform the density using a change of variables technique, apply the approximation, and then reverse the transformation.

Define the Lamperti transform as

$$Y_t = \gamma(X_t,\theta) = \int^{X_t} \frac{1}{\sigma(X_u;\theta)}\,du.$$

Applying Ito's lemma, we have

$$\begin{aligned} dY_t &= \frac{1}{\sigma(X_t;\theta)}\,dX_t - \frac{1}{2}\frac{\partial\sigma(X_t;\theta)}{\partial X_t}\,dt \\ &= \left[\frac{\mu(X_t;\theta)}{\sigma(X_t,t;\theta)} - \frac{1}{2}\frac{\partial\sigma(X_t;\theta)}{\partial X_t}\right]dt + dW_t \\ &= \left[\frac{\mu(\gamma^{-1}(Y_t;\theta);\theta)}{\sigma(\gamma^{-1}(Y_t;\theta);\theta)} - \frac{1}{2}\sigma_X(\gamma^{-1}(Y_t;\theta);\theta)\right]dt + dW_t. \end{aligned} \tag{9}$$

The process Y_t has unit diffusion. Consequently, the conditional distribution of Y_t is approximately normal. Y_t is closer to a normal random variable than X_t is. However, it is still not practical to expand the density of Y_t, because it gets peaked around the conditional value y_0 as the sampling interval gets small. To fix this, introduce an additional transformation:

$$Z_t = h^{-1/2}(Y_t - y_0).$$

Once an approximation to the transition density of Z_t are constructed, we reverse the change of variables to obtain the approximation for X_t.

The details on the implementation are as follows:

1. Start with the first time interval $[t_0, t_1]$.
2. Let $f_Y(y \mid y_0;\theta)$ denote the conditional density of $Y_{t_1}|_{Y_{t_0}=y_0}$ evaluated at $Y_{t_1} = y$. Define f_X and f_Z in a similar way.
3. $(X \to Y \to Z)$: Construct a Hermite polynomial approximation to $f_Z(z \mid y_0;\theta)$ as follows:
$$f_Z^{(J)}(z \mid y_0;\theta) = \phi(z)\sum_{j=0}^{J}\beta_j(y_0;\theta)H_j(z),$$
$$\beta_j(y_0;\theta) = \frac{1}{j!}\int_{-\infty}^{\infty} H_j(z) f_Z(z \mid y_0;\theta)\,dz,$$
where $\phi(z)$ is the standard normal density, and $\beta_Z^{(j)}(y_0;\theta)$ can be computed approximately using simulation or a Taylor expansion based on Equation 9.
4. $(Z \to Y)$: Apply the change of variable technique to obtain an approximation to $f_Y(h, y \mid y_0;\theta)$ from $f_Z^{(J)}(z \mid y_0;\theta)$:
$$f_Y^{(J)}(y \mid y_0;\theta) = h^{-1/2} f_Z^{(J)}(h^{-1/2} + (y - y_0) \mid y_0;\theta).$$
5. $(Y \to X)$: Compute $f_X(x \mid x_0;\theta)$ using the Jacobian formula:
$$f_X^{(J)}(x \mid x_0;\theta) = \sigma(x;\theta)^{-1} f_Y^{(J)}(\gamma(x,\theta) \mid \gamma(x_0,\theta);\theta).$$
6. Repeat steps 3-5 for all sampling intervals. Compute the approximation log-likelihood using $f_X^{(J)}(x \mid x_0;\theta)$ using
$$L_J(\theta) = \sum_{n=0}^{N-1} \log f_X^{(J)}(X_{t_{n+1}} \mid X_{t_n};\theta).$$
7. Maximize $L_J(\theta)$ numerically to obtain the estimator.

Remark 7.22. Under some regularity conditions, $f_X^{(J)}(h, x \mid x_0;\theta)$ converges to $f_X(h, x \mid x_0;\theta)$ as J increases. The maximizer of the approximate likelihood converges to the MLE. This approach, though elegant, is more restrictive than the SMLE approach. At the same time, using a series to approximate an intractable function is a powerful idea, and it is useful in much broader contexts.

8 Bayesian Inference

8.1 General Statistical Theory

In a Bayesian framework, a parameter is a random variable with an unknown distribution. A researcher approaches the inference problem with a model and a prior belief about the parameter, before looking at the data. This prior belief is then updated by the data using the likelihood function implied by the model. The updated distribution is called the posterior distribution. The updating follows Bayes rule:

$$\Pr[\theta \mid y] = \frac{\Pr[y \mid \theta]\pi(\theta)}{\Pr[y]},$$

where θ is the parameter, y is the data sample, $\Pr[y|\theta]$ is the likelihood function, $\pi(\theta)$ is the prior, $\Pr[y] = \int \Pr[y|\theta]\pi(\theta)\, d\theta$ is the marginal distribution, and $\Pr[\theta|y]$ is the posterior distribution. Generally, we use

$$\Pr[\theta \mid y] \propto \Pr[y \mid \theta]\pi(\theta)$$

because the marginal distribution is not of interest (and difficult to compute).

A *Bayes estimator* for θ is an estimator or decision rule that minimizes the Bayes risk among all estimators. A Bayes estimator depends on three factors: the prior belief, the model (equivalently the likelihood), and the loss function. The resulting Bayes estimator is optimal in the sense that no other estimator can yield a smaller loss under these three conditions.

Suppose the prior distribution for θ is $\pi(\theta)$. Let $\hat{\theta}(y)$ be an estimator of θ based on a sample y. Let $L(\theta, \hat{\theta}(y))$ be a loss function. Then the risk under this loss function is

$$R(\theta, \hat{\theta}) = \mathop{\mathbb{E}}_{\Pr[y|\theta]}\left[L(\theta, \hat{\theta}(y)\right].$$

The *Bayes risk* of $\hat{\theta}$ is defined as

$$r_\theta(\hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta)\, d\theta.$$

The *Bayes estimator* minimizes the above risk among all estimators.

Example 8.1. Suppose we use squared error loss. Then

$$\int L(\theta, \hat{\theta}) \Pr[\theta \mid y]\, d\theta = \int (\theta - \hat{\theta}(h)^2 \Pr[\theta \mid y]\, d\theta.$$

Clearly, the Bayes estimator is the posterior mean $\hat{\theta}(y) = \mathbb{E}[\theta \mid y]$.

Suppose the distribution of θ is absolutely continuous. Let $B_\varepsilon(a)$ denote an open ε-neighborhood of a. Then, the *zero-one loss* function is given by

$$L = 1 - 1_\theta(B_\varepsilon(a)),$$

where $1_\theta(B_\varepsilon(a)) = 1$ iff $\theta \in B_\varepsilon(a)$.

Example 8.2. Suppose $\Pr[\theta \mid y]$ is continuous with a mode at θ_M. Then, the Bayes estimator under a zero-one loss function, $\hat{\theta}(y, \varepsilon)$, satisfies

$$\lim_{\varepsilon \to 0} \hat{\theta}(y, \varepsilon) = \theta_M.$$

Thus, the posterior mode is the limit of Bayes estimators under a zero-one loss function.

The Bernstein-von Mises theorem provides the large sample properties of the posterior distribution. It plays the role of the frequentist-CLT for the Bayesian setting. In particular:

- If there are many i.i.d. observations governed by a smooth, finite-dimensional statistical model, the Bayesian estimate and the maximum likelihood estimate will be close to each other.
- Furthermore, the posterior distribution of the parameter vector around the posterior mean or mode will be close to the distribution of the maximum likelihood around the true value.

A statistical model $\{P_\theta \mid \theta \in \Theta\}$ is called *differentiable in quadratic mean* if there exists a measurable vector-valued function g_{θ_0} s.t. as $\theta \to \theta_0$,

$$\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^\top g_{\theta_0} \sqrt{p_{\theta_0}}\right) d\mu = o(\|\theta - \theta_0\|^2).$$

Theorem 8.3 (Bernstein-von Mises theorem)**.** *Let the experiment $\{P_\theta \mid \theta \in \Theta\}$ be differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that there exists a sequence of uniformly consistent tests ϕ_n for testing $H_0 : \theta = \theta_0$ against $H_1 : \|\theta - \theta_0\| > \varepsilon \ \forall \varepsilon > 0$. Furthermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then, the corresponding posterior distributions satisfy*

$$\|P_{\sqrt{n}(\theta_n - \theta_0)|y_1,\ldots,y_n} - N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})\| \to 0,$$

where $\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta_0}^{-1} g_{\theta_0}(y_i)$ with g_{θ_0} being the score of the model.

8.2 General Sampling Methods

In practice, the posterior distribution $\Pr[\theta \mid y]$ is often a complicated object, so we need effective sampling methods. Write out the parameter vector as

$$\theta = (\theta_1, \ldots, \theta_p).$$

Suppose the posterior density (omitting the dependence on y to simplify notation) is

$$\Pr[\theta_1, \ldots, \theta_p]. \tag{10}$$

Let

$$\Pr[\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p] \tag{11}$$

denote the conditional density of θ_j given the remaining parameters. MCMC addresses the sampling problem by breaking the joint distribution (10) into its complete set of conditionals (11), which are typically of lower dimensions and easier to sample from. The Gibbs sampler and the Metropolis-Hastings methods are two of the most commonly used samplers.

8.2.1 Gibbs Sampler

Without loss of generality, suppose $p = 2$. The steps are as follows:

1. Choose a set of initial values, say θ_1^0 and θ_2^0.
2. Draw $\theta_1^1 \sim \Pr[\theta_1 \mid \theta_2^0]$.
3. Draw $\theta_2^1 \sim \Pr[\theta_2 \mid \theta_1^1]$.
4. Repeat steps 2 and 3 N times with the values of the conditioning variables updated sequentially.

8.2.2 Metropolis-Hastings

Suppose we want to obtain a sample from the distribution of $\theta = (\theta_1, \ldots, \theta_p)$. The Metropolis method uses a proposal distribution to generate draws. Let $q(y \mid x)$ denote the density of the proposal distribution, which needs to be symmetric in x, y. The steps are as follows:

1. Draw an initial value, θ^0, and set $t = 0$.
2. Draw $\theta^* \sim q(\cdot \mid \theta^0)$.
3. Calculate the ratio $r = \Pr[\theta^*] / \Pr[\theta^0]$.
4. If $r \geq 1$, set $\theta^1 = \theta^*$. Otherwise, set $\theta^1 = \theta^*$ w.p. r and $\theta^1 = \theta^0$ w.p. $1 - r$.
5. Increase t by one and proceed to step 2 to draw from $q(\cdot \mid \theta^{t-1})$. Continue.

To implement this in practice, we need to be able to sample from $q(\cdot \mid \theta^t)$ and to evaluate the ratio $r = \Pr[\theta^*] / \Pr[\theta^t]$. We do note sample directly from $\Pr[\theta]$.

The Metropolis-Hastings method is a generalization of the Metropolis algorithm. $q(x \mid y)$ is no longer required to be symmetric. To implement this method, only the third step needs to be modified, where we set

$$r = \frac{\Pr[\theta^*] q(\theta^t \mid \theta^*)}{\Pr[\theta^t] q(\theta^* \mid \theta^t)}.$$

A large number of Metropolis-Hastings algorithms have been proposed, *e.g.*, the independence M-H and the random walk M-H samplers.

8.2.3 Markov Chains

We now examine why the Gibbs and M-H samplers are valid using a Markov chain framework. A homogeneous Markov chain can be uniquely characterized by a transition kernel, defined as

$$K(x, y) = \Pr[X_{t+1} = y \mid X_t = x].$$

Suppose the distribution at time t is given by p_t and at $t + 1$ by p_{t+1}. Then, these two distributions are obviously related by

$$p_{t+1}(y) = \int K(x, y)p_t(x)\, dx.$$

If $p_{t+1} = p_t = p$, then the distribution p is called a stationary distribution of the chain. Formally, p is called a *stationary distribution* of a Markov chain if

$$p(y) = \int K(x, y)p(x)\, dx.$$

A Markov chain converges to its stationary distribution if the following conditions hold:

- The chain is irreducible, *i.e.*, it is possible to get to any state from any state.
- It is aperiodic, *i.e.*, $\Pr[X_{t+m} = x \mid X_t = x] > 0\ \forall m$ large enough.

Theorem 8.4 (Law of large numbers for Markov chains, Johannes and Polson)**.** *Suppose X_t is an ergodic chain with stationary distribution π and suppose f is a real-valued function with $\mathbb{E}_\pi[|f|] < \infty$. Then for all X_t with any initial starting value X_0,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} f(X_t) = \int f(x)\pi(x)\, dx.$$

Theorem 8.5 (Central limit theorem for Markov chains, Johannes and Polson)**.** *Suppose X_t is an ergodic chain with stationary distribution π and suppose f is a real-valued function with $\mathbb{E}_\pi[|f|] < \infty$. Then $\exists \sigma_f \in \mathbb{R}$ s.t. for all X_t with any initial starting value X_0,*

$$\sqrt{N}\left(\frac{1}{N}\sum_{t=1}^{N} f(X_t) - \int f(x)\pi(x)\, dx\right) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2).$$

8.3 Applications

We estimate continuous-time asset pricing models from a Bayesian perspective. We start with simple models, and then make them progressively more complex. We assume the observations are at a daily frequency and that the Euler discretization provides an adequate approximation to the continuous time model.

The essential steps in the MCMC estimation are as follows:

1. Write out the price dynamics and state evolution in a state space form.
2. Characterize the joint distribution by its complete set of conditionals.
3. Use the Metropolis of Gibbs sampler to generate draws from the posterior.

8.3.1 GBM

The price S_t follows the SDE:

$$d \log S_t = \mu\, dt + \sigma\, dW_t^{\mathbb{P}}.$$

In discrete time, we have

$$Y_t = \log \frac{S_t}{S_{t-1}} = \mu + \sigma\varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0, 1)$. For a sample $Y = (Y_1, \ldots, Y_T)$, the likelihood is given by

$$f(Y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^T \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(Y_t - \mu)^2\right).$$

Suppose we use independent priors, s.t. $\Pr[\mu, \sigma^2] = \Pr[\mu]\Pr[\sigma^2]$, Consequently,

$$\Pr[\mu, \sigma^2 \mid Y] \propto f(Y \mid \mu, \sigma^2)\Pr[\mu]\Pr[\sigma^2].$$

Remark 8.6. Because $\Pr[\mu, \sigma^2 \mid Y]$ is not a standard distribution, it is hard to sample from it directly. Instead, we apply the Gibbs sampler to sample from the conditionals $\Pr[\mu \mid \sigma^2, Y]$ and $\Pr[\sigma^2 \mid \mu, Y]$ iteratively.

Suppose we use a normal prior for μ: $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and an inverse Gamma prior for σ^2: $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$. Then,

$$\mu \mid \sigma^2, Y \sim \mathcal{N}\left(\frac{\sigma_0^2 \sum_{t=1}^T y_t + \sigma^2 \mu_0}{T\sigma_0^2 + \sigma^2}, \frac{1}{T/\sigma^2 + 1/\sigma_0^2}\right),$$
$$\sigma^2 \mid \mu, Y \sim \mathcal{IG}\left(\frac{T}{2} + \alpha, \frac{1}{2}\sum_{t=1}^T (Y_t - \mu)^2 + \beta\right).$$

8.3.2 GBM and Black-Scholes

Now we incorporate option data into data estimation. The price of a call option with strike K is given by

$$C_t = BS(\sigma, S_t) = S_t \Phi(d_1) - e^{r(T-t)} K \Phi(d_1 - \sigma\sqrt{T-t}),$$

where we assume the continuously compounded interest rate r is observed. The discrete-time model is:

$$\log \frac{S_t}{S_{t-1}} = \mu + \sigma\varepsilon_t$$
$$C_t = BS(\sigma, S_t) + \varepsilon_t^c,$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ and $\varepsilon_t^c \sim \mathcal{N}(0, \sigma_c^2)$. In this discretization, we assume that option prices are observed with normally distributed errors. We have

$$\Pr[C_t \mid S_t, \mu, \sigma^2, \sigma_c^2] \propto \exp\left(-\frac{1}{2\sigma_c^2}(C_t - BS(\sigma, S_t))^2\right).$$

The conditionals for the posterior distribution are $\Pr[\mu \mid \sigma^2, S], \Pr[\sigma_c^2 \mid \sigma^2, S, C], \Pr[\sigma^2 \mid \mu, \sigma_c^2, S, C]$. We use a Normal prior for μ and inverse Gamma priors for σ_c^2, σ^2. Note that the first two priors can be sampled directly. The third distribution depends both on stock and option price equations. We use M-H to sample σ^2. Note that

$$\Pr[\sigma^2 \mid \mu, S] \propto f((S \mid \mu, \sigma^2) \Pr[\sigma^2] \sim \mathcal{IG}.$$

The M-H algorithm works as follows:

1. Draw $(\sigma^2)^{j+1} \sim \sigma^2 \mid \mu^{j+1}, S$.
2. Accept $(\sigma^2)^{j+1}$ with probability

$$r = \min\left\{\frac{\Pr[C \mid (\sigma_c^2)^{j+1}, (\sigma^2)^{j+1}, S]}{\Pr[C \mid (\sigma_c^2)^{j+1}, (\sigma^2)^j, S]}, 1\right\}.$$

8.3.3 Merton's Jump Diffusion Model (Multivariate)

A k-vector of asset prices solves:

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t^{\mathbb{P}} + d\left(\sum_{j=1}^{N_t^{\mathbb{P}}} S_{\tau_{j-}}(\exp Z_j^{\mathbb{P}} - 1)\right),$$

where $W_t^{\mathbb{P}}$ is a vector of standard Brownian motion, $\Sigma = \sigma\sigma^\top$ is the diffusion matrix, $N_t^{\mathbb{P}} \sim \text{Pois}(\lambda)$ is a Poisson process, and $Z_j^{\mathbb{P}} \sim \mathcal{N}(\mu_J, \Sigma_J)$ is the jump size.

Solving the SDE, we have

$$\log \frac{S_t}{S_{t-1}} = \mu + \sigma\varepsilon_t + \sum_{j=N_{t-1}^{\mathbb{P}}+1}^{N_t^{\mathbb{P}}} Z_j^{\mathbb{P}},$$

where the drift vector has been redefined to account for the variance correction. Assuming there is at most one jump per time interval, we obtain the following model:

$$Y_t := \log \frac{S_t}{S_{t-1}} = \mu + \sigma\varepsilon_t + J_t Z_t,$$

where $J_t \sim \text{Bern}(\lambda)$. Define $\theta = (\mu, \Sigma, \lambda, \mu_J, \Sigma_J)$. The jump size and location, J_t, Z_t, are latent state variables. We treat them as additional parameters. This is called data-augmentation. Let $X_t = (J_t, Z_t)$. The MCMC algorithm will sample from $\Pr[\theta, X \mid Y]$. We use the conjugate priors:

$$\begin{aligned}
\Sigma &\sim \text{Inverse Wishart}; \mu \mid \Sigma \sim \mathcal{N}(a, b\Sigma) \\
\lambda &\sim \text{Beta} \\
\Sigma_j &= \text{Inverse Wishart}; \mu_J \mid \Sigma_j \sim \mathcal{N}(a_J, b_J\Sigma).
\end{aligned}$$

The complete set of conditionals are $\Pr[\mu, \Sigma \mid X, Y], \Pr[\mu_J, \Sigma_J \mid J, Z], \Pr[\lambda \mid J], \Pr[J \mid \theta, Z, Y], \Pr[Z \mid J, \theta, Y]$. The first two distributions can be sampled directly, as in the Black-Scholes case. Then,

$$\begin{aligned}
\Pr[\lambda \mid J] &\propto \beta(\alpha^*, \beta^*) \\
\Pr[Z_t \mid J_t, \theta, Y_t] &\propto \exp\left(-\frac{1}{2}[r_t^\top \Sigma^{-1} r_t + (Z_t - \mu_J)^\top \Sigma_J^{-1}(Z_t - \mu_J)]\right) \\
\Pr[J_t = 1 \mid \theta, Z_t, Y_t] &\propto \lambda \exp\left(-\frac{1}{2}(Y_t - \mu - Z_t)^\top \Sigma^{-1}(Y_t - \mu - Z_t)\right),
\end{aligned}$$

where $r_t = Y_t - \mu - Z_t J_t$.

8.3.4 Time-Varying Equity Premium

The Black-Scholes model assumes that the drift and diffusion terms are constant. Now we allow the drift to vary over time:

$$\begin{aligned}
\frac{dS_t}{S_t} &= \left(r_t + \mu_t + \frac{1}{2}\sigma^2\right) dt + \sigma \, dW_t^s \\
d\mu_t &= k_\mu(\theta_\mu, -\mu_t) \, dt + \sigma_\mu \, dW_t^\mu,
\end{aligned}$$

where r_t is the observed risk-free rate, and the two Brownian motions can be correlated.