Lectures in Econometrics with Applications to Finance[1]

Bruce D. McNevin

July 18, 2022

Contents

[1]This discussion on sufficient statistics and factorization is based on Degroot and Schervish (2011) [7]

[2]This overview of GMM is based on Hall(2015) [11], Cameron and Trivedi(2005) [5], and Greene(2008) [9]

[3]This section is based the IV discussion in Chapter 6 of Cameron and Trivedi [5]

[4]This discussion on testing the CAPM follows the work of Jagannathan, Skoulakis and Wang, [20]. Also, see Jagannathan, Skoulakis and Wang, [19] for a comprehensive survey of GMM applications in finance.

Chapter 1

Basic Models for Evaluating Asset Returns

1.1 Introduction

This lecture is intended as an review of some important basic concepts in financial econometrics. The reader is assumed to be familiar with most of the statistical concepts, though not necessarily the applications in finance.[1]

1.1.1 Efficient Market Hypothesis

The efficient market hypothesis (EMH) states that asset prices fully reflect all available information. Under the EMH, prices change when there is a change in the information set available to investors (i.e. news). By definition, "news" cannot be forecasted. All investors receive new information at the same time, and asset prices change instantaneously. As a result there are no arbitrage opportunities. An investor cannot make an economic profit by trading on the basis of the information set. We can describe an asset price that follows the EMH as follows:

$$E_t[P_{t+1}|I_t] = P_t \tag{1.1.1}$$

where I_t is the information set available to investors at time, t. There are three basic forms of the EMH relating to the degree of information available to investors.

1. Weak efficiency: The information set only includes the history of the prices or returns themselves.

2. Semi-strong Efficiency: The information set includes all information available to the public.

3. Strong Efficiency: The information set includes all information known to any market participant (private information).

[1] Readers requiring additional background on time series should consult, Chatfield [6], or Hamilton [9].

1.1.2 The Random Walk Model

The random walk (RW) statistical model is typically used to describe the movement of asset prices. If asset prices, P_t, follow a random walk then,

$$P_t = \mu + P_{t-1} + \varepsilon_t \tag{1.1.2}$$

where μ is a drift term, and ε_t is $iid(0, \sigma^2)$.

Under the RW model, asset prices cannot be forecast, since the change in price is due to an unobserved random disturbance term. The mean and variance of the RW model at time t conditional on initial price, P_0 are,

$$E[P_t|P_0] = \mu_t t \tag{1.1.3}$$
$$Var[P_t|P_0] = \sigma^2 t \tag{1.1.4}$$

Both the mean and variance are a linear function of time, so asset prices are assumed to be nonstationary.

There are a several relevant variations on the random walk model, depending on the assumptions made about ε_t.[2] A common assumption is that $\varepsilon_t \sim iid(0, \sigma^2)$. The independence assumption means that price increments are uncorrelated, and that nonlinear functions of the increments are also uncorrelated. The identically distributed assumption means that shocks to price are drawn from the same distribution at all points in time.

An alternative, more realistic assumption, is to allow the distribution to change over time. In this case ε_t is assumed to be independent and not identically distributed, or INID. Assuming that ε_t is drawn from different distributions seems more realistic for a financial time series. For instance, it is reasonable to expect that shocks to asset prices are drawn from a different distributions in times of financial crisis.

A third alternative, the somewhat more realistic assumption, is that ε_t are uncorrelated, dependent, and drawn from different distributions over time. One example would be the following,

$$cov(\varepsilon_t, \varepsilon_{t-k}) = 0 \quad \text{for all k} \neq 0. \tag{1.1.5}$$
$$cov(\varepsilon_t^2, \varepsilon_{t-k}^2) \neq 0 \quad \text{for all k} \neq 0. \tag{1.1.6}$$
$$\tag{1.1.7}$$

In this case the innovations are uncorrelated, but dependent since the squared innovations are correlated.

1.1.3 Calculating Returns

Let P_t be the price of an asset, then the simple one period return is defined as,

$$R_t = \frac{P_t}{P_{t-1}} - 1 \tag{1.1.8}$$

[2]This section is a summary of an in-depth discussion of random walk models in Campbell, Lo, and Mackinley [4].

The simple return over k periods is defined as,

$$R_t^k = \frac{P_t}{P_{t-k}} - 1 \tag{1.1.9}$$

If an asset pays dividends in period t, D_t, then the simple one period return is,

$$R_t = \frac{[P_t + D_t]}{P_{t-1}} - 1 \tag{1.1.10}$$

The multi-period return over k years can be written as:

$$R_t^k = \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \ldots \times \frac{P_{t-k+1}}{P_{t-k}} \tag{1.1.11}$$

$$R_t^k = (1 + R_t)(1 + R_{t-1}), \ldots (1 + R_{t-k_1}) \tag{1.1.12}$$

The continuously compounded return of an asset is,

$$r_t \equiv ln(1 + R_1) = ln(\frac{P_t}{P_{t-1}}) \tag{1.1.13}$$

The continuously compounded multi-period return is the sum of the continuously compounded single period returns,

$$r_t^k \equiv ln(1 + R_t^k) = ln((1 + R_t)(1 + R_{t-1}) + ... + (1 + R_{t-k+1})) \tag{1.1.14}$$

$$r_t^k = r_t + r_{t-1} + ... + r_{t-k+1} \tag{1.1.15}$$

1.1.4 A Model of Returns

A common statistical model of returns is,

$$r_{t+1} = \mu_t + \sigma_t \varepsilon_{t+1} \tag{1.1.16}$$

where μ_t is the mean conditional on information as of t, and σ_t is the variance conditional on information as of t. It is common to assume that $\varepsilon_t \sim iidN(0, \sigma^2)$. As we will see in the next section, the assumption of normality is inconsistent with the styled facts of asset returns. Distributions of returns typically have fatter tails than the Normal distribution, (i.e. excess kurtosis). Alternatively, is is often assumed that $\varepsilon_t \sim$iid Student-t with ν degrees of freedom. This assumption is more realistic than the normality assumption as the t-distribution has fatter tails.

Stationarity and Ergodicity

If prices are assumed to follow a random walk, returns (the difference in the log of prices) will be a stationary series. Stationarity is a fundamental requirement for the statistical analysis of returns. It is necessary if we want to use data across different time periods to calculate moments.

Weak Stationarity - a stochastic time series is said to be weakly stationary if following conditions hold,

1. $E[y_t]$ is independent of t.
2. $Var[y_t]$ is a finite, positive constant, independent of t.
3. $Cov[y_t, y_s]$ is a finite function of —t-s— but not of t or s.

A stationary series does not have a trend.

In addition to stationarity, returns must also be ergodic. Ergodicity relates to the consistency of a stationary time series process. Since a time series $\{x_t\}$ is a single realization of some unobserved generating process, it is not entirely clear that a sample mean taken over time will be a consistent estimator of the true mean at time t. That is, at time t,we observe a single observation of x, but $E[x_t]$ is a property of an ensemble of x's at t.

Ergodic Theorem - a stationary process is ergodic if the sample moments converge to the population moments. In the case of the first moment, ergodicity implies

$$\bar{x} = E(x_t) = \mu \quad as \quad N \to \infty \tag{1.1.17}$$

A sufficient condition for this to happen is that

$$\rho_k \to 0 \quad as \quad k \to \infty. \tag{1.1.18}$$

where ρ_k is the correlation of returns at time t and $t-k$. If the sufficient condition holds, the process is called ergodic in the mean.

1.2 Stylized Facts about Asset Returns

Figure 1.1 is a plot of daily equity market excess returns from July 1, 1926 to June 29, 2018. The data which is from the Kenneth French Data Library [3] is a value weighted series of excess returns for all firm listed in the NYSE, AMEX, and NASDAQ over the observation period. The risk free rate, which is also from the Kenneth French Data library, is the yield on one month Treasury bills. Asset returns as a class tend to exhibit a set of characteristics which are often described as stylized facts. The market returns in Figure 1.1 have at least one readily identifiable characteristic - clusters of volatility at different points in time. In this section, we discuss a number of stylized facts about returns, including clustering. To a large extent, all of the lectures in this book arise from research methods that have been developed explain stylized facts about asset returns.[4]

1.2.1 Low or No Autocorrelation

Asset returns are not usually autocorrelated, except for small intraday scales (20 minutes). As a result, returns are difficult to predict. The extent to which returns are predictable depends on:

[3] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[4] The list of stylized facts is based on the work of Cont [7]. Also, see Campbell, Lo, & Mackinlay [4] for a related discussion.

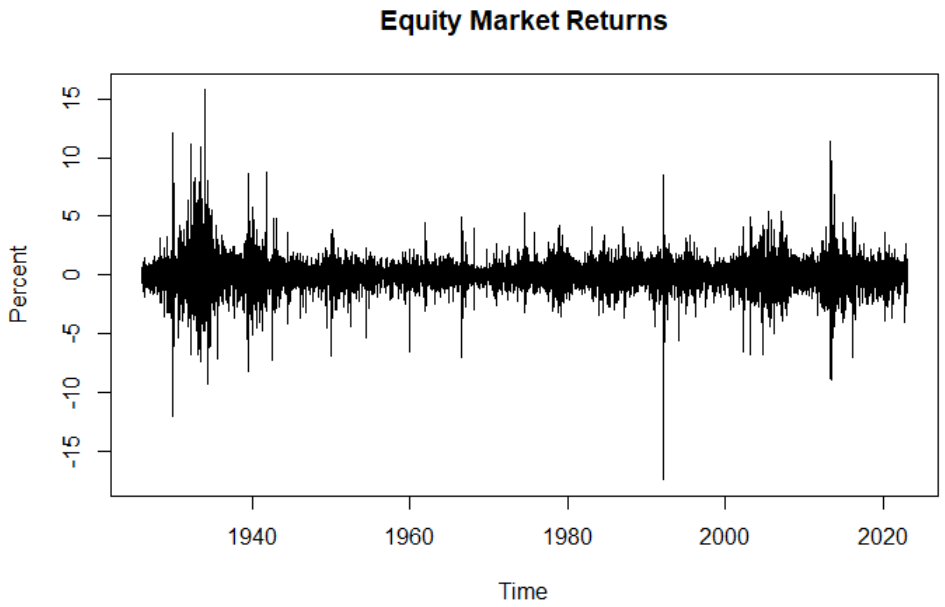

Figure 1.1: Daily Excess Returns to Equities - U.S.

- the forecast horizon,
- the degree of market volatility, and
- the stage of the business cycle.

Predictability tends to rise during crisis periods.

Definition of Autocorrelation Coefficient

The autocorrelation coefficient is the time series version of the correlation coefficient. Given a covariance-stationary series r_t the kth order autocovariance and autocorrelation coefficient respectively, are:

$$\gamma(k) = cov(r_t, r_{t+k}) \tag{1.2.1}$$

$$\rho(k) = \frac{cov(r_t, r_{t+k})}{\sqrt{var(r_t)}\sqrt{var(r_{t+k})}} = \frac{cov(r_t, r_{t+k})}{var(r_t)} = \frac{\gamma(k)}{\gamma(0)} \tag{1.2.2}$$

where, $var(r_t) = var(r_{t+k})$.

The sample analogue for the autocovariance and autocorrelation are,

$$\widehat{\gamma}_k = \frac{1}{T}\sum_{t}^{T-k}(r_t - \bar{r}_T)(r_{t+k} - \bar{r}_t) \tag{1.2.3}$$

$$\widehat{\rho}(k) = \frac{\widehat{\gamma}(k)}{\widehat{\gamma}(0)} \tag{1.2.4}$$

where, $\bar{r}_t = \frac{1}{T}\sum_{i=1}^{T-k} r_t$

Figure 1.2 shows the autocorrelation for the daily market returns shown in Figure 1.1. The data set contains 250 trading day observations per year, so the horizontal axis of the plot shows daily autocorrelations for two years. The dashed horizontal lines represent the 95% confidence interval for zero autocorrelation. There is some evidence of small, but significant short term serial correlation, but only occasional small spikes over the longer term.

Short term positive serial correlation is a common feature of portfolios. Possible sources of serial correlation include time varying risk premia, bid-ask bounce, non-synchronous trading effects such as using stale prices to calculate returns, and partial price adjustment due to trades taking place when traders have incomplete information.[1]

Serial correlation is generally less pronounced for individual firms. As an example, Figure 1.3 shows autocorrelations for AT&T (excess) returns for 500 lags. The degree short term autocorrelation observed in the market returns does not exist for AT&T.

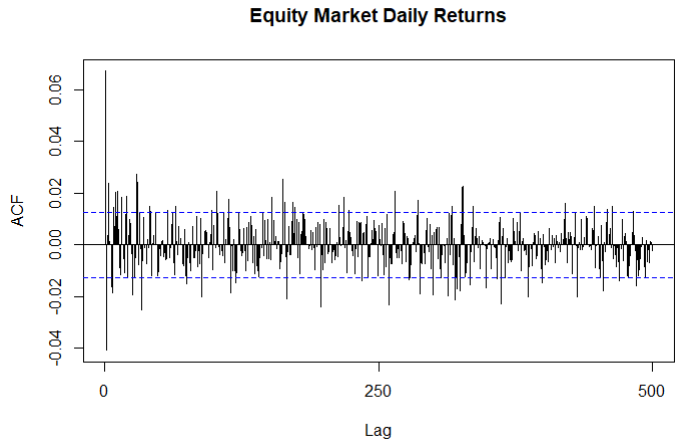

Figure 1.2: Autocorrelations for Daily Excess Returns to Equities - U.S.

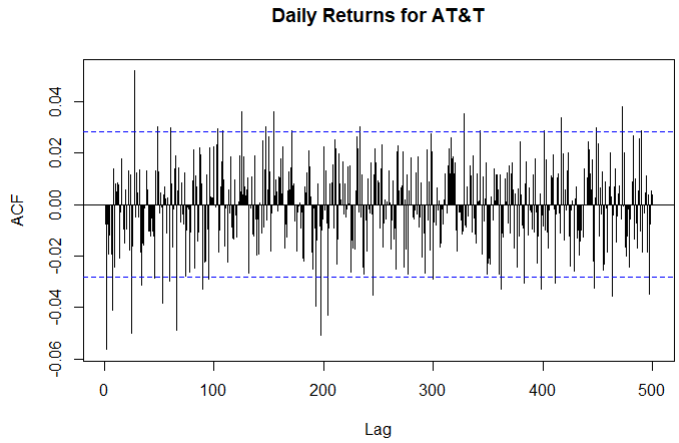

Figure 1.3: Autocorrelations for Daily Excess Returns - AT&T

1.2.2 Measuring Skewness and Kurtosis

For a random variable, x, the skew and kurtosis are

$$S(X) = \frac{E[(x-\mu)^3]}{\sigma^3} \tag{1.2.5}$$

$$K(X) = \frac{E[(x-\mu)^4]}{\sigma^4} \tag{1.2.6}$$

$$\tag{1.2.7}$$

where $S(x)$ is the skew, $K(x)$ is the kurtosis, μ is the mean and σ is the standard deviation of x. The sample analogues are:

$$\widehat{S}(x) = \frac{1}{T\widehat{\sigma}^3}\sum_{t=1}^{T}(x_t-\bar{x})^3 \tag{1.2.8}$$

$$\widehat{K}(x) = \frac{1}{T\widehat{\sigma}^4}\sum_{t=1}^{T}(x_t-\bar{x})^4 \tag{1.2.9}$$

$$\tag{1.2.10}$$

where $\widehat{\mu}$ is the sample mean, and $\widehat{\sigma}^2$ is the sample variance In large samples of normally distributed data the sample skewness and kurtosis are normally distributed with means 0 and 3 and variances of 6/T and 24/T, respectively. Excess kurtosis is defined as $K(x)-3$. The Normal distribution has a kurtosis of 3, so the excess kurtosis of the Normal distribution is zero (mesokurtic). A distribution with positive excess kurtosis (leptokurtic) is said to have fat or heavy tails. Similarly negative excess kurtosis (platykurtic) indicates short tails. A positive skew indicates that the tail on the right side of a distribution is longer than the left side and the bulk of the values lie to the left of the mean. A skew of zero indicates symmetry.

1.2.3 Asymmetry

The distribution of equity returns is often negatively skewed indicating that downturns are typically steeper than recoveries. Sample estimates of skewness for daily US stock returns tends to be negative for stock indexes (a long left hand tail) and close to zero or positive for individual stocks. Table 1.1 contains the skewness of daily returns for the market and a set of media firms for the period January 2013 - June 2017. The market has a small negative skew, which is not apparent in the density plot (Figure 1.4). The firm skew statistics tend to be positive, but some are negative, and some are essentially zero.

1.2.4 Heavy Tailed Distributions

The unconditional distribution of asset returns typically has more mass in the tail areas than would be predicted by a Normal distribution. Figure 1.4 compares the density of returns for the market portfolio with the standard Normal distribution. The market returns have a much narrower peak and wider tails than the Normal distribution.

	Mean	Std. Dev.	Skew	Ex. Kurt
Mkt	0.06	0.80	-0.39	2.17
Alphabet	0.09	1.43	2.15	22.54
Amazon	0.13	1.86	0.47	10.32
AOL	0.11	2.40	0.04	16.51
Apple	0.07	1.54	-0.48	6.49
AT&T	0.01	0.94	-0.36	2.09
CBS	0.05	1.48	0.24	1.40
Disney	0.07	1.15	-0.47	6.90
Facebook	0.17	2.10	2.97	38.08
FOX	0.03	1.40	-0.11	3.95
Microsift	0.09	1.45	0.02	10.71
Netflix	0.26	3.10	3.14	38.51
Sony	0.13	2.08	0.84	6.89
Sprint	0.08	3.13	0.50	10.19
Twitter	-0.04	3.49	-0.55	9.17
Verizon	0.01	1.01	-0.08	1.39

Table 1.1: Summary Statistics of Returns for a Sample of Media Firms

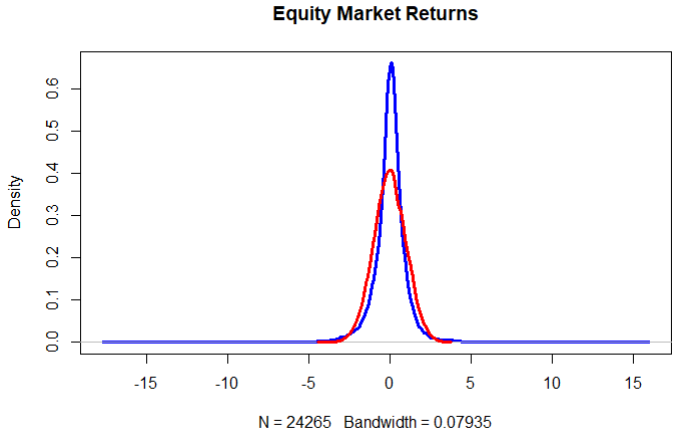

Figure 1.4: Distribution of Daily Excess Market Returns - U.S.
Red = standard Normal distribution, Blue = market portfolio

The excess kurtosis for the market portfolio for the period July 1, 1926 - June 29, 2018 is 16.7. Over the smaller sample period in Tabletab:sumstats, the market portfolio has a kurtosis of 2.17.

Figure 1.6 shows the distribution of the absolution value of the 500 smallest daily returns for the market portfolio. This distribution, with its long tail, has a shape that resembles a Pareto distribution. Fat tails are an important feature of asset returns. As shown in Figure 1.1 and 1.4, the variance of asset returns is high. Understanding tail events is an important part of understanding the behavior of returns as a whole.

The extreme values, or tail, of a distribution of returns are often modeled as either a Pareto or power distribution.[5] A power law is a mathematical relationship between two quantities where a relative change in one quantity results in a proportional change in the other quantity. A well known example is the 80/20 rule or Pareto Principle which states that for many events, roughly 80% of the effects

[5]See Stanley, et.al.(2008) [18] for a discussion of the power law and finance.

come from 20% of the causes. For assets we can write the power law as:

$$Pr[|r_t| > x] \sim x^{-\alpha} \tag{1.2.11}$$

where r_t is the asset return at t, and α is the power-law exponent. The generalized Pareto distribution will be discussed in the lecture on extreme value theory.

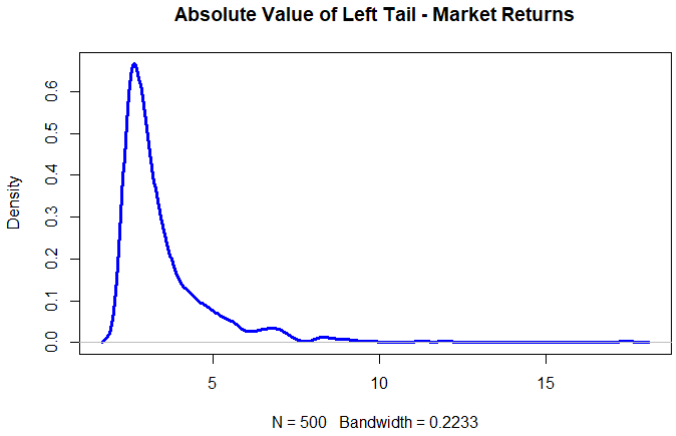

Figure 1.5: Left Tail Distribution of Daily Excess Market Returns - U.S.

1.2.5 Intermittency

The term intermittency refers to the empirical observation that the volatility of returns is high regardless of the time scale. Table 1.2 illustrates intermittency by comparing the volatility of returns for the market portfolio using daily, monthly and annual data. The daily and annual volatility calculations are re-scaled to a monthly level assuming independence of returns across time.[6] As shown in the table, the volatility of the annual returns series exceeds the monthly series which exceeds the daily series. Temporal aggregation does not reduce the volatility of returns.

	Volatility
Daily	4.870663
Monthly	5.332363
Annual	5.891354

Table 1.2: Volatility of Market Returns Across Scale

1.2.6 Gaussianity

The distribution of returns is not the same across scale. That is, as returns are aggregated across scale the distribution becomes more Normally distributed. Figure **??** shows QQ plots for daily, monthly and annual portfolio returns against a theoretical Normal distribution. The QQ plots suggest that the daily and monthly distributions are not Normal but the annual could be Normal. A formal test for departure from normality is needed to provide further insight.

[6]Specifically, Daily vol. times $\sqrt{21}$; Annual vol. divided by $\sqrt{12}$

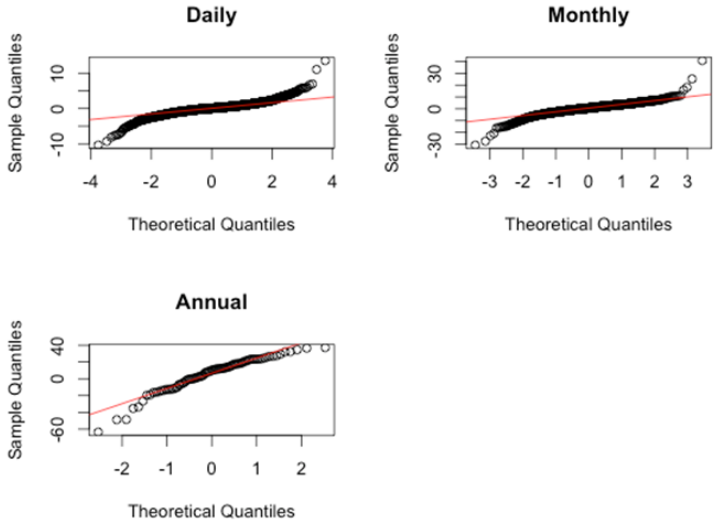

Figure 1.6: QQ Plots of Market Returns by Scale

Testing for Normality

The Shapiro-Wilk (SW) test is a test of the null hypothesis that the data are $iidN(\mu, \sigma^2)$. The test statistic is,

$$SW = \frac{(\sum_{i=1}^{n} a_i x_{(i)}^2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.2.12}$$

where $x_{(i)}$ is the i^{th} order statistic.[7] The null hypothesis is that the data is Normally distributed.

An alternative, perhaps more commonly applied test for Normality, is the Jarque-Bera test (JB). The JB statistic for departure from normality is given by,

$$JB = (T/6)(S^2 + (1/4) * K^2) \tag{1.2.13}$$

where S is skewness and K is kurtosis. Under the null hypothesis that S = 0 and K = 0 the JB statistic is asymptotically distributed as chi-squared with 2 degrees of freedom. A JB statistic greater than 5.99 will reject the null hypothesis at the 95% level of confidence.

Table 1.3 shows the SW and JB statistics for returns of the market portfolio at three scales. Both tests reject the null hypothesis for the daily and monthly scale, but at the annual scale we are unable to reject the null hypothesis that the distribution of returns is Normally distributed. The p-value for the annual JB test statistic is 0.43, and for the annual SW test it is 0.524.

1.2.7 Volatility Clustering

High volatility events tend to cluster in time, so that measures of volatility tend to exhibit positive autocorrelation. To illustrate this, we begin with the following

[7] The ith order statistic is the i^{th} smallest number in the sample of x_i.

	mean	std. dev	skew	kurtosis	SW	JB	Nobs
Daily	0.0293	1.0629	-0.1189	16.7461	0.9210	283586	24265
Monthly	0.6623	5.3324	0.1888	7.9444	0.9160	2910	1104
Annual	8.5044	20.4082	-0.3382	0.0075	**0.9873**	**1.74**	91

Table 1.3: Market Portfolio Statistics by Scale

simple model of returns:

$$r_t = \mu + a_t \tag{1.2.14}$$

where μ is estimated using the sample mean. We consider two measures of volatility, a_t^2, and $|a_t|$. The autocorrelations for these two measures are plotted in Figure 1.7 for daily market returns. Both measures of volatility exhibit strong positive autocorrelation. Notice that the absolute value measure decays much slower than the squared measure. Cont [7] points out that this is often interpreted as long range dependence.

Figure 1.7 in conjunction with 1.2 illustrates an important characteristic of asset returns. There is very little autocorrelation for the returns themselves, but there is significant autocorrelation in the volatility of returns. Returns are serially uncorrelated, but not independent.

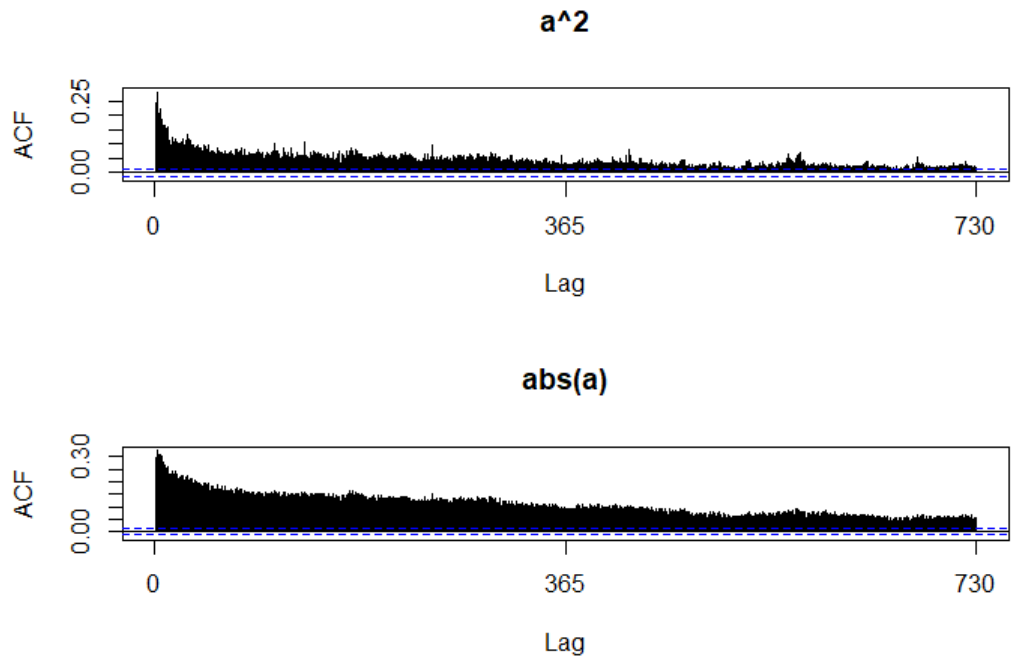

Figure 1.7: ACF for Two measures of Volatility - Daily Market Returns

Unlike returns themselves, the absolute value of returns (or the square), are serially correlated and tend to be predictable. If we re-write returns as

$$r_t = sign(r_t)|r_t| \tag{1.2.15}$$

$$\text{where} \quad sign(r_t) = +1 \quad \text{if} \quad r_t > 0, \tag{1.2.16}$$

$$\text{and} \quad sign(r_t) = -1 \quad \text{if} \quad r_t < 0 \tag{1.2.17}$$

we see that since $|r_t|$ is predictable, the non-predictable part is $sign(r_t)$, or the direction of the market.

1.2.8 Additional Characteristics of Volatility

There are a number of other characteristics of volatility that have been identified. Here we provide a non-exhaustive list:

- Leverage Effects – most measures of volatility of returns are negatively correlated with the returns of the asset. A decline in the stock price of a firm raises the firm's financial leverage, resulting in an increase in the volatility of equity. (Black (1976) [2], Christie (1982) [5]) This strong negative relationship is shown in Figure 1.8 which displays returns for the S&P 500 against the VIX.[8]

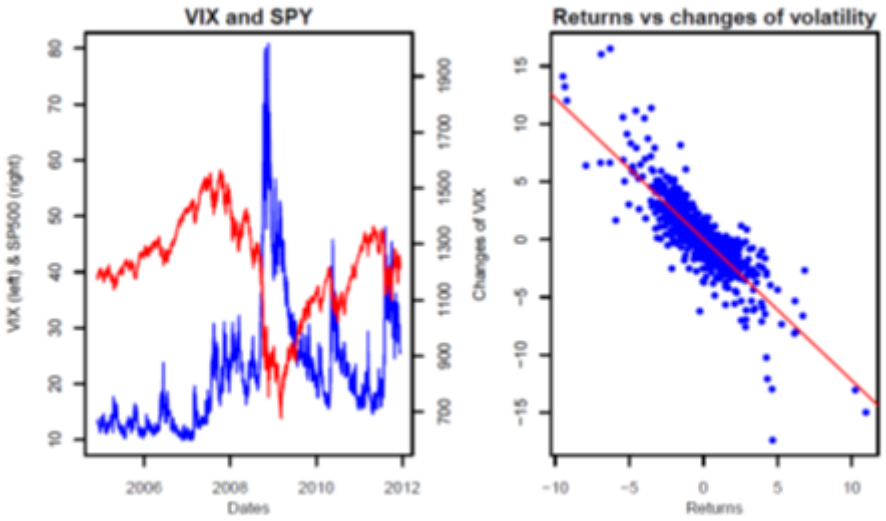

Figure 1.8: Returns and Volatility - S&P 500

- Non-trading Effects, Weekend Effects – when a market is closed information accumulates at a different rate than when it is open. For example, there exists a weekend effect, in which stock price volatility on Monday is not three times the volatility on Friday. (French (1982) [11], French and Roll (1986) [12])

- Expected events – volatility is high at regular times such as news announcements or other expected events, or even at certain times of day. For example, equity returns are less volatile in the early afternoon.(Cornell (1978) [8], Patell and Wolfson (1979) [16])

- Co-movements in volatility – Volatility is positively correlated across markets and assets. (Ramchand and Susmel (1998) [17])

- Evolving – volatility tends to evolve over time in a continuous manner, and jumps are rare.

[8]VIX is an index of implied volatility on options for the S&P 500

- Stationary – volatility tends to vary within some bound, and does not go off to infinity.
- Conditional Heavy Tails - the distribution of returns has heavy tails even after accounting for volatility clustering.

1.3 Modeling the Conditional Mean of Returns

We have shown that the serial correlation in returns is low, but the serial correlation in the volatility of returns can be quite significant. In this section we will review the properties of three basic models, AR(1), MA(1) and the ARMA(1,1), each of which can be used to model the conditional mean of returns.

The AR(1) Process

We begin with the AR(1) process. This simple model is used extensively in financial econometrics. It is defined as follows:

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \tag{1.3.1}$$

Conditional on past returns, asset returns for an AR(1) process have the following moments:

$$E[r_t|r_{t-1}] = \phi_0 + \phi_1 r_t \tag{1.3.2}$$

$$var[r_t|r_{t-1}] = var(a_t) = \sigma_a^2 \tag{1.3.3}$$

Using backward substitution we can write the AR(1) as,

$$r_t = \phi_0(1 + \phi_1 + \phi_1^2 + \ldots) + \phi_1 a_t + \phi_1^2 a_{t-2} + \ldots \tag{1.3.4}$$

If $|\phi_1 < 1$ then

$$1 + \phi_1 + \phi_1^2 + \ldots = \frac{1}{1-\phi_1} \quad \text{is a bounded sequence} \tag{1.3.5}$$

$$\sum_{j=0}^{\infty} |\phi_1^j| < \infty \tag{1.3.6}$$

$|\phi_1| < 1$ is a necessary and sufficient condition for stationarity of the AR(1) process. Assuming weakly stationarity, the unconditional mean and variance of the AR(1) are,

$$r_t = \frac{\phi_0}{1-\phi_1} + \phi_1 a_t + \phi_1^2 a_{t-2} + \ldots \tag{1.3.7}$$

$$\mu = E[r_t] = \frac{\phi_0}{1-\phi_1} \tag{1.3.8}$$

$$\gamma_0 = (1 + \phi_1^2 + \phi_2^4 + \ldots) = \frac{1}{1-\phi_1^2}\sigma_a^2 \tag{1.3.9}$$

Note that $\phi_1^2 < 1$ is required to satisfy the non-negativity requirement of a variance. The autocovariance of a stationary AR(1) process is:

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + a_t \tag{1.3.10}$$

$$\begin{aligned} \gamma_j &= E[(r_t - \mu)(r_{t-j} - \mu)] \\ &= E[(\phi_1(r_{t-1} - \mu) + a_t)(r_{t-j} - \mu)] \\ &= \phi_1 E[(r_{t-1} - \mu)(r_{t-j} - \mu) + a_t(r_{t-j} - \mu)] \end{aligned}$$

$$\gamma_j = \phi_1 \gamma_{j-1} \tag{1.3.11}$$

The autocorrelation coefficient is,

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \phi_1 \frac{\gamma_{j-1}}{\gamma_0} = \phi_1 \rho_{j-1} \tag{1.3.12}$$

$$\tag{1.3.13}$$

where γ_0 is the variance of r_t. Using backward substitution yields,

$$\rho_j = \phi_1^j \tag{1.3.14}$$

- If $0 < \phi_1 < 1$ the ACF decays exponentially.
- If $< -1 < \phi_1 < 0$ the ACF consists of two alternating exponential decays.

Definition: Partial Autocorrelation Function (PACF)

The k^{th} partial autocorrelation of x_t is the autocorrelation of x_t and x_{t+k} after removing the partial correlation of x_t with x_{t+1} to x_{t+k-1}. The definition can be illustrated with the following set of AR regressions,

$$y_t = \alpha_{1,0} + \alpha_{1,1} y_{t-1} + e_{1t}$$

$$y_t = \alpha_{2,0} + \alpha_{2,1} y_{t-1} + \alpha_{2,2} y_{t-2} + e_{2t}$$

$$y_t = \alpha_{3,0} + \alpha_{3,1} y_{t-1} + \alpha_{3,2} y_{t-2} + \alpha_{3,3} y_{t-3} + e_{3t}$$

The partial autocorrelation coefficients are $\alpha_{1,1}, \alpha_{2,2}, \alpha_{3,3}$. For an AR(p) process, the partial autocorrelation function (PACF) will be zero for lags greater than p. As a result, the PACF can be used to identify the number of lags in an AR(p) process.

The MA(1) Process

Let $\{a_t\}$ be a zero-mean white noise process. The MA(1) process is defined as

$$r_t = \mu + a_t - \theta a_{t-1} \tag{1.3.15}$$

The expected value and variance are,

$$E[r_t] = \mu \tag{1.3.16}$$

$$var[r_t] = \sigma_a^2 + \theta^2 \sigma_a^2 = \sigma_a^2(1 + \theta^2) \tag{1.3.17}$$

The 1st order autocovariance is,

$$E[(r_t - \mu)(r_{t-1} - \mu)] = -\theta\sigma_a^2 \tag{1.3.18}$$

The autocovariance for lags greater than one is,

$$\begin{aligned} E[(r_t - \mu)(r_{t-j} - \mu)] &= E[(r_t - \mu)(r_{t-j} - \mu)] = E[(a_t - \theta a_{t-1})((a_{t-j} - \theta a_{t-j-1})] \\ &= E(a_t a_{t-j} - \theta a_{t-1}a_{t-j} - \theta a_t a_{t-j-1} + \theta^2 a_{t-1}a_{t-j-1}) = 0 \quad \text{for} \quad j > 1. \end{aligned} \tag{1.3.19}$$

The autocorrelation is,

$$\rho_0 = 1 \tag{1.3.20}$$

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{-\theta\sigma_a^2}{(1+\theta^2)\sigma_a^2} = \frac{-\theta}{(1+\theta^2)} \tag{1.3.21}$$

$$\rho_j = 0 \quad \text{for} \quad j > 1 \tag{1.3.22}$$

For the MA(1) process, the ACF is zero for lags greater than one.

An MA(1) process is always covariance stationary. The mean, variance, and ACF are invariant over time. Note that we did not have to place any restrictions on θ to establish stationarity.

If we do not place any additional restrictions on the MA(1) model, it will not be uniquely identified. The identification problem is a consequence of observing r_t but not a_t.

The following two MA(1) processes have the same ACF (i.e. eq. 1.3.22):

$$r_t = a_t - \theta a_{t-1} \tag{1.3.23}$$

$$r_t = a_t - \frac{1}{\theta}a_{t-1} \tag{1.3.24}$$

Which model should we choose? If we invert the two models we get,

$$a_t = r_t + \theta r_{t-1} + \theta^2 r_{t-2} + \ldots \tag{1.3.25}$$

$$a_t = r_t + \frac{1}{\theta}r_{t-1} + \frac{1}{\theta^2}r_{t-2} + \ldots \tag{1.3.26}$$

If $|\theta| < 1$, then the series of coefficients converges for model A (eq. 1.3.25) and diverges for model B (eq. 1.3.26). Model A is said to be invertible, but model B is not. The invertibility condition ensures that there is a unique MA process for a given ACF.

Invertibility of an MA(q) Process

An MA(q) process defined as

$$r_t = \theta_q(L)a_t \tag{1.3.27}$$

is said to be invertible if there exists a sequence of constants $\{\pi_j\}$ such that

$$\sum_{j=0}^{\infty} |\pi_j| < \infty \tag{1.3.28}$$

and

$$a_t = \sum_{j=0}^{\infty} \pi_j r_{t-j}, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{1.3.29}$$

A process is invertible if the random disturbance at time t can be expressed as a convergent sum of present and past values of r_t. In effect, invertibility means the MA process can be written in the form of an AR process.

Finally, note that the PACF for the MA(1) process never cuts off so it cannot be used to identify the order of an MA(p) process. However, the ACF of an MA(q) process is zero for lags greater than q, so it can be used to identify the order of an MA process.

The ARMA(1,1) Model

The ARMA(1,1) model is defined as,

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t - \theta a_{t-1} \tag{1.3.30}$$

The properties of the ARMA(1,1) model are the same as the AR(1) model with some modifications for the MA component.

$$E[r_t] = \phi_0 + \phi_1 E[r_t] = \frac{\phi_0}{(1-\phi_1)} \tag{1.3.31}$$

$$var[r_t] = \frac{(1 - 2\phi_1\theta + \theta^2)\sigma_a^2}{1 - \phi_1^2} \tag{1.3.32}$$

The stationarity condition is the same as the AR(1): $|\phi_1| < 1$. The ACF of the ARMA(1,1) is:

$$\rho_1 = \frac{\theta\sigma_a^2}{\gamma_0}, \quad \rho_l = \phi_1 \rho_{l-1} \quad \text{for} \quad l > 1. \tag{1.3.33}$$

This ACF is similar to that of an AR(1) model except that the exponential decay starts at lag=2. The PACF is similar to that of an MA(1) model in that it does not cut off at a specific lag. The ACF and PACF are not useful for determining the order of an ARMA model. A common approach for identifying the order is to use an information criteria. That is, estimate multiple models consisting or different orders for the AR and MA processes, and select the model with the minimum AIC or BIC.

1.4 Modeling Volatility - GARCH Models

The traditional time series regression model assumes that the variance is constant over time. Consider the following simple dynamic model:

$$Y_t = \alpha + \beta y_{t-1} + e_t$$

The unconditional variance is constant,

$$E[(y_t - E(y_t))^2] = \sigma^2/(1-\beta)$$

The conditional variance is,

$$Var_{t-1}(Y_t) = E_{t-1}[(y_t - E_{t-1}(y_t))^2] = E_{t-1}[(y_t - \alpha - \beta y_{t-1}]^2 = E_{t-1}[e_t^2]$$

The traditional time series model assumes that the conditional variance is also constant. This assumption is contradicted by the stylized facts, in particular the fact that volatility tends to cluster, and distributions of returns have fat tails.

The ARCH (Engle, 1982), and the Generalized ARCH (Bollerslev, 1982 [3]) allow the conditional variance to change over time. Given our simple model of returns, $r_t = \mu_t + a_t$, the ARCH model specifies the conditional variance as a function of lagged errors,

$$a_t = \sigma_t \varepsilon_t \sigma_t = \alpha_0 + \alpha_1 a_{t-1}^2 + \ldots + \alpha_m a_{t-m}^2 \tag{1.4.1}$$

where $\{\varepsilon_t\}$ is a sequence of iid random variables with a mean of zero and a variance of one. $\alpha_0 > 0$ $\alpha_i \geq 0 \quad$ for $i \neq 0$,

The GARCH(m,s) model is a more generalized structure than the ARCH model that allows for a more parsimonious model. Bollerslev notes that GARCH generalizes ARCH in much the same way that ARMA generalizes the AR model. The GARCH(p,q) model is specified as follows:

$$a_t = \sigma_t \varepsilon_t \tag{1.4.2}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=0}^{p} \alpha_i a_{t-1}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{1.4.3}$$

$$\text{where,} \quad \varepsilon_t \sim \text{iid}(0,1), \alpha_0 >, \alpha_i \geq 0, \beta_j \geq 0, \tag{1.4.4}$$

$$\text{and,} \quad \sum_{i=1}^{max(p,q)} (\alpha_i + \beta_j) < 1. \tag{1.4.5}$$

To understand the properties of the GARCH model, we can focus on the GARCH(1,1) model:

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{1.4.6}$$

$$\text{where,} \quad \alpha_0 > 0, \alpha_1 \geq 0, \beta_1 \geq 0,$$

$$\text{and,} \quad (\alpha_1 + \beta_1) < 1. \tag{1.4.7}$$

Properties of the GARCH(1,1) Model

1. The inequalities in 1.4.7 ensure weak stationarity,
2. If σ_{t-1}^2 is large, σ_t^2 will be large, so that the GARCH specification can exhibit clustering of volatility.
3. If $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ then excess kurtosis exceeds zero, so the process has fat tails.
4. The GARCH specification allows for the evolution of volatility over time.

Forecasting the GARCH(1,1)

The 1 step ahead forecast for a GARCH(1,1) model is:

$$\widehat{\sigma}^2_{t+1} = \alpha_0 + \alpha_1 a_t^2 + \beta_1 \sigma_t^2 \tag{1.4.8}$$

Setting $a_t^2 = e_t^2 \sigma_t^2$

$$\widehat{\sigma}^2_{t+2} = \alpha_0 + (\alpha_1 + \beta_1)\sigma_t^2 + a_1 \sigma_t^2 (e_t^2 - 1) \tag{1.4.9}$$

The two step ahead forecast is,

$$\widehat{\sigma}^2_{t+2} = \alpha_0 + (\alpha_1 + \beta_1)\widehat{\sigma}_t^2 \tag{1.4.10}$$

The k-step ahead forecast is,

$$\widehat{\sigma}^2_{t+k} = \alpha_0 + (\alpha_1 + \beta_1)\widehat{\sigma}^2_{t+k-1}, \quad k > 0 \tag{1.4.11}$$

Generally speaking, since volatility is unobserved it is difficult to evaluate a forecast. Sometimes out of sample forecasts of σ^2_{t+k} are compared to a^2_{t+k}. But while a_t^2 is a consistent estimator of σ_t^2 it is not an efficient estimator, so there is little reason for a single observation of the two to be similar, and they usually are not.

1.4.1 ACF for the GARCH Model

The autocorrelation for a GARCH(p,q) is,

$$\rho_n = \gamma_n \gamma_0^{-1} = \sum_{i=1}^{m} \varphi_i \rho_{n-1} \tag{1.4.12}$$

$$\text{where} \quad \varphi_i = \alpha_i + \beta_i \quad \text{and} \quad m = max(p, q), \tag{1.4.13}$$

$$\text{and} \quad \gamma_n = cov(\varepsilon_t^2, \varepsilon_{t-n}^2) \tag{1.4.14}$$

Example 1.1. In this example, we model the mean and conditional variance of daily returns for the S&P 500. The mean specification is ARMA(1,1). The conditional volatility specification is a GARCH(1,1). The two models are estimated jointly using the 'rugarch' package in R. The output is shown in Figure 1.10. The distribution of standardized residuals from the GARCH model.

The 3 GARCH parameters, ω, α, and β, are all non-negative, and $\alpha + \beta < 1$, so the weak stationarity condition is satisfied. In addition, the stationarity and invertibility conditions are satisfied for the ARMA(1,1) model. The ACF of the normalzed residuals, shown in Figure 1.11, indicates that the model residuals are white noise.

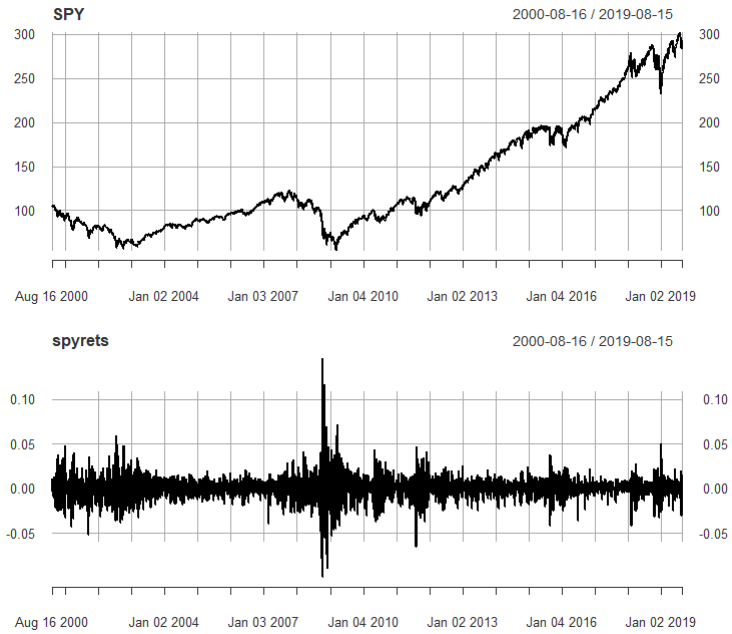

Figure 1.9: S&P 500 - Daily Prices and Returns

```
*----------------------------------*
*          GARCH Model Fit         *
*----------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model     : sGARCH(1,1)
Mean Model      : ARFIMA(1,0,1)
Distribution    : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.000697    0.000093   7.4691 0.000000
ar1     0.730852    0.108317   6.7473 0.000000
ma1    -0.780259    0.099079  -7.8751 0.000000
omega   0.000002    0.000001   2.8810 0.003965
alpha1  0.116277    0.010157  11.4477 0.000000
beta1   0.867079    0.010627  81.5885 0.000000
```

Figure 1.10: GARCH Model output

1.5 The Log Normal Stochastic Volatility Model

Stochastic volatilities (SV) models represent an alternative to the ARCH/GARCH approach of modeling time varying volatility. In contrast to the ARCH/GARCH models which are often described as observation driven, SV models are often described as parameter driven. SV models are essentially state-space models, which is intuitively appealing, since volatility is unobservable. Volatility represents the arrival of new information into the market and consequently it is unobserved.

The basic SV model with continuous volatility is defined as follows:

$$y_t = \varepsilon_t exp(h_t/2) \tag{1.5.1}$$

$$h_t = \alpha + \beta h_{t-1} + \eta_t \tag{1.5.2}$$

where y_t are continuously compounded returns, h_t is log-volatility, ε_t and η_t are

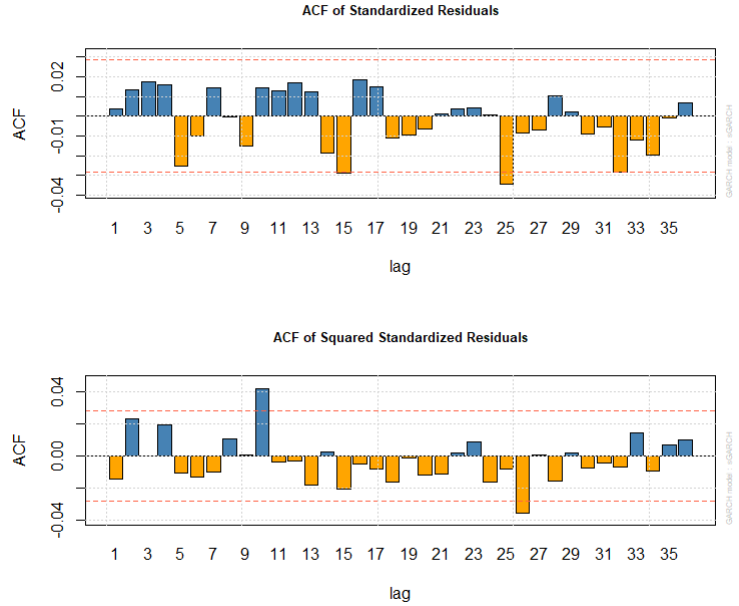

Figure 1.11: ACF of Standardized Residuals and Squared Residuals

two independent Gaussian white noises with variances 1 and σ_η^2, respectively. This model is typically referred to as the log-normal SV model.

Properties of the SV Model

The SV model will be stationary if $|b| < 1$, with:

$$\mu_h = E(h_t) = \frac{\alpha}{(1-\beta)} \tag{1.5.3}$$

$$\sigma_h^2 = var(h_t) = \frac{\sigma_\eta^2}{(1-\beta^2)} \tag{1.5.4}$$

$$\frac{E(y_t^4)}{E[E(y_t^2)]^2} = exp(\sigma_h^2) \geq 3 \tag{1.5.5}$$

$$\rho_{y_t^2}(r) \cong \frac{exp(\sigma_h^2 - 1)}{3exp(\sigma_h^2 - 1)}\beta' \tag{1.5.6}$$

- Since ε_t is always stationary, y_t will be stationary iff h_t is stationary.
- The memory of y_t is defined by the memory of the latent h_t, which is an AR(1) process.
- The kurtosis is not bounded as it is for the GARCH models.
- The SV model is leptokurtotic (fat tails).(Eq.1.5.6)
- If $\beta < 0$, $\rho_{y_t^2}(r)$ can be negative, unlike the ARCH model.

1.6 Applications in Finance

1.6.1 Measuring the impact of news on volatility

The GARCH(1,1) model is widely used to model the volatility of asset returns, and while it captures several of the key empirical properties of returns, it does not capture the asymmetric or leverage effect of volatility. The leverage effect occurs when an unexpected drop in the price of a stock results in a greater increase in volatility than an unexpected increase in the price of the stock.

Several variations of the original GARCH model have been introduced to capture asymmetry. One such model is the exponential GARCH or eGARCH, Nelson (1990) [15]. The eGARCH model is designed to capture the impact of differences in positive and negative news, as well as small and large news events.

The eGARCH(1,1) model is specified as follows,

$$log(\sigma_t^2) = \omega + \beta \cdot log(\sigma_{t-1}^2) + \gamma z_t + \alpha\big[|z_t| + E|z_t|\big] \tag{1.6.1}$$

where $z_t \sim NID(0,1)$[9], and the parameters are not constrained to be non-negative. Under the assumption of Normality, $E[|z_t] = \sqrt{2/\pi}$.

Asymmetry is captured via the sign of z_t. The magnitude effect is captured by the term, $\alpha\big[|z_t| + E|z_t|\big]$. For instance, if $\gamma > 0$ then difference $|z_t| - E|z_t|$ will have a positive impact on $log(\sigma_i^2)$ that increases as the difference increases.

Parameter estimates for the eGARCH(1,1) model using the daily returns for the S&P500 are shown in Figure 1.12. The estimate of α_1 is negative[10], indicating the presence of the leverage effect. When news is negative volatility increases, and when news is positive volatility decreases.

```
Conditional Variance Dynamics
-----------------------------------
GARCH Model     : eGARCH(1,1)
Mean Model      : ARFIMA(1,0,1)
Distribution    : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error    t value Pr(>|t|)
mu      0.000391    0.000085     4.6222    4e-06
ar1     0.238882    0.026311     9.0792    0e+00
ma1    -0.290097    0.026018   -11.1497    0e+00
omega  -0.259333    0.002338  -110.9414    0e+00
alpha1 -0.152734    0.007671   -19.9107    0e+00
beta1   0.972071    0.000085 11449.4300    0e+00
gamma1  0.144939    0.004602    31.4941    0e+00
```

Figure 1.12: eGARCH-S&P 500 Daily Returns

Another GARCH model designed to capture asymmetry is the GJR (Glosten, Jagannathan, Runkle, 1989 [13]) model. This model captures asymmetry by introducing an indicator variable, S_t that is 1 if the news event is negative and zero otherwise. The model specification is,

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_{t-1}^2 + \gamma e_{t-1}^2 S_t^- \tag{1.6.2}$$

[9] $z_t = e_t\sqrt{\sigma_t^2}$, where e_t is the residual from the mean filtration proces. (e.g. ARMA)

[10] The parameter estimate alpha1 in the R package rugarch output, is called γ in equation 1.6.1. Similarly, gamma1 in the R output is α in equation 1.6.1.

where $S_t^- = 1$ if $e_{t-1} < 0$ and 0 otherwise.

Parameter estimates for the eGARCH(1,1) model using the daily returns for the S&P500 are shown in Figure 1.13. The estimate of gamma is positive and significant, suggesting the presence of a leverage effect. Interestingly the estimate of α is not statistically different from zero.

```
Conditional Variance Dynamics
-----------------------------------
GARCH Model     : gjrGARCH(1,1)
Mean Model      : ARFIMA(1,0,1)
Distribution    : norm

Optimal Parameters
-----------------------------------
        Estimate  Std. Error    t value Pr(>|t|)
mu      0.000361    0.000095   3.793244 0.000149
ar1     0.438773    0.264773   1.657166 0.097486
ma1    -0.493150    0.256396  -1.923395 0.054430
omega   0.000002    0.000000  11.227211 0.000000
alpha1  0.000001    0.002653   0.000451 0.999641
beta1   0.883746    0.006162 143.424746 0.000000
gamma1  0.187904    0.012349  15.215699 0.000000
```

Figure 1.13: gjrGARCH-S&P 500 Daily Returns

Engle and Ng, (1993) [9] propose a method for determining the appropriate GARCH specification based on the impact of news on volatility. In the case of the GARCH model, the news-impact curve is defined by holding all information at $t-2$ and earlier constant, and examining the relationship between e_{t-1} and σ_t^2. The curve assumes that all of the conditional variances are evaluated at the unconditional variance of the stock returns. The news-impact curve measures the impact of news at $t-1$ on volatility at t.

For the standard (Gaussian) GARCH model the news-impact curve is a quadratic with its minimum centered at e_{t-1}. For the eGARCH model, the minimum is at e_{t-1} but the curve is not symmetric. Specifically, for the eGARCH model the news-impact curve is,

$$\sigma_t^2 = A \cdot \left[\frac{\gamma + \alpha}{\sigma} \cdot e_{t-1} \right] \quad \text{for} \quad e_{t-1} > 0 \tag{1.6.3}$$

$$\sigma_t^2 = A \cdot \left[\frac{\alpha - \gamma}{\sigma} \cdot e_{t-1} \right] \quad \text{for} \quad e_{t-1} < 0 \tag{1.6.4}$$

where,

$$A = \sigma^{2\beta} exp\left[\omega - \alpha\sqrt{2/\pi}\right] \tag{1.6.5}$$

For the GJR Garch model, the news-impact curve is,

$$\sigma_t^2 = A + \alpha \cdot e_{t-1}^2 \quad \text{for} \quad e_{t-1} > 0 \tag{1.6.6}$$

$$A = \omega + \beta * \sigma^2 \tag{1.6.7}$$

where σ^2 is the unconditional variance.

The news-impact curve for the GARCH(1,1), eGARCH(1,1), and gjrGARCH(1,1) models fit to the daily returns for the S&P 500 are shown in Figure 1.14. The standard GARCH model has a symmetric news impact curve, while the eGARCH and gjrGARCH models display an increase in volatility when there is bad news, and no change in volatility when there is good news. In this example the GJR curve increases faster than the other two models when there is bad news.

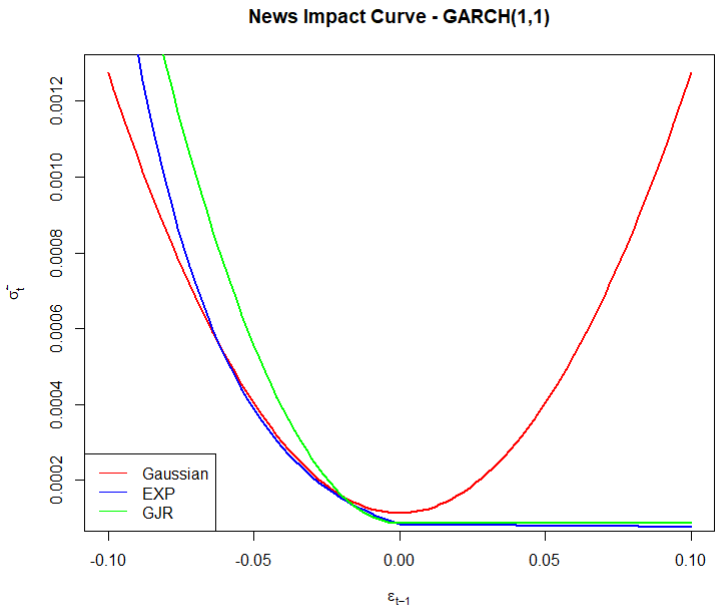

Figure 1.14: News Impact Curves

If a negative shock has a larger impact on volatility than a positive shock the STANDARD GARCH model WILL under forecast the impact of news on volatility for negative shocks and over forecasT the impact of news on volatility for positive shocks. Also, if a large shock has a greater impact than predicted by a quadratic function, the standard GARCH model under predicts volatility after a large shock and over predicts volatility after a small shock.

Engle and Ng propose a set of diagnostic tests based on the news-impact curve. The tests consider whether or not it is possible to predict the squared normalized residual by variables that are not included in the GARCH model. The model specification can be represented as,

$$log(\sigma_t^2) = log(\sigma_0^2) + \delta_a z_0 \tag{1.6.8}$$

where σ_0^2 is the GARCH model being tested, and z_0 is a matrix of additional variables. Under the null hypothesis the parameters $\delta_a = 0$.

The three tests proposed by Engle and Ng are the following,

- The sign bias test,

$$log(\sigma_t^2) = a + \delta_1 \cdot S_{t-1}^- + b \cdot log(\sigma_0^2) \tag{1.6.9}$$

- The positive sign bias test,

$$log(\sigma_t^2) = a + \delta_2 \cdot S_{t-1}^{+} e_{t-1} + b \cdot log(\sigma_0^2) \tag{1.6.10}$$

- The negative sign bias test,

$$log(\sigma_t^2) = a + \delta_3 \cdot S_{t-1}^{-} e_{t-1} + b \cdot log(\sigma_0^2) \tag{1.6.11}$$

where, $S_{t-1}^{-} = 1$ if $e_{t-1} < 0$ and zero otherwise, and $S_{t-1}^{+} = 1$ if $e_{t-1} < 0$, and zero otherwise. For each test the null hypothesis is that $b = 0$. A t-test is applied to the estimate of δ for each individual test. If the correct volatility model is being used, the coefficient estimate for δ will not be significantly different from zero.

Figure 1.15 contains the results of the bias tests for the standard GARCH and eGARCH models using the S&P 500 data. The test rejects the null hypothesis of no sign bias (equation 1.6.9) for the standard GARCH model but not for the eGARCH model. The standard model appears to have positive sign bias (equation 1.6.10 for this particular data set. The eGARCH model also suggests the presence of positive sign bias even though the sign bias test (equation 1.6.9) indicated that there was no sign bias. This illustrates the importance of the specific (positive and negative) sign bias tests. Negative sign bias is not significant at the 95% level of confidence for the eGARCH model, but it is significant for the standard GARCH model.

Standard GARCH

	t-value	prob	sig
Sign Bias	3.0190200	2.549345e-03	***
Negative Sign Bias	0.7452038	4.561852e-01	
Positive Sign Bias	2.5830572	9.822390e-03	***
Joint Effect	33.4091875	2.640100e-07	***

eGARCH

	t-value	prob	sig
Sign Bias	1.622089	0.104850252	
Negative Sign Bias	1.699278	0.089331956	*
Positive Sign Bias	1.982911	0.047434554	**
Joint Effect	11.824960	0.008007471	***

Figure 1.15: Sign Bias Tests - Daily S&P 500 Returns

The joint effect test is a chi-square test for all three coefficients. The null hypothesis for the joint effect test is $\delta_1 = \delta_2 = \delta_3 = 0$. Figure 1.15 indicates that the joint null hypothesis is rejected for both model specifications.

1.7 Exercises

1. Show that,

$$cov(r_t, r_{t-1}) = -\theta\sigma_a^2 \tag{1.7.1}$$

when r_t follows the MA(1) process,

$$r_t = \mu + a_t - \theta a_{t-1} \tag{1.7.2}$$

2. In exercise 1.1 the daily returns of the S&P500 were assumed to follow an AR(1,1) + GARCH(1,1) specification. Is this the minimum AIC specification? Justify your answer.

3. Using daily returns for the S&P 500 for Jan. 02 2000 through December 31 2018 estimate a GJR-GARCH model, and apply the sign bias tests. Interpret the results.

1.8 R Code for Examples

1.8.1 Example 1.1

```
library(PerformanceAnalytics)
library(xts)
library(rugarch)

date<-as.Date(paste(SPY$date), "%m/%d/%Y")
spy<-xts(SPY[,2], order.by=date)
par(mfrow=c(2,1))
plot.xts(spy)
spyrets<-CalculateReturns(spy)
plot(spyrets)

spec = ugarchspec()
fit = ugarchfit(spec = spec, data = na.omit(spyrets))
show(fit)
```

1.8.2 Measuring the impact of news on volatility

```
#Estimate GARCH models
spec = ugarchspec()
fitB = ugarchfit(spec = spec, data = na.omit(spyrets))
show(fitB)
spec1 = ugarchspec(variance.model=list(model="eGARCH"))
fit1 = ugarchfit(data = na.omit(spyrets), spec = spec1)
show(fit1)
# note that newsimpact does not require the residuals (z) as it
```

```
# will discover the relevant range to plot against by using the min/max
# of the fitted residuals.
spec2 = ugarchspec(variance.model=list(model="gjrGARCH"))
fit2 = ugarchfit(data = na.omit(spyrets), spec = spec2)
show(fit2)
#News impact curves
niB=newsimpact(z = NULL, fitB)
ni1=newsimpact(z = NULL, fit1)
ni2=newsimpact(z = NULL, fit2)
plot(niB$zx, niB$zy, ylab=niB$yexpr, xlab=niB$xexpr, type="l", main = "News Impact
lines(ni1$zx, ni1$zy, ylab=ni1$yexpr, xlab=ni1$xexpr, type="l", main = "News Impact
lines(ni2$zx, ni2$zy, ylab=ni2$yexpr, xlab=ni2$xexpr, type="l", main = "News Impact
legend("bottomleft", legend=c("Gaussian","EXP","GJR"),col=c("Red","Blue","Green"),l
#Sign bias tests
signbias(fitB)
signbias(fit1)
signbias(fit2)
```

Bibliography

[1] Anderson, Robert M., Kyong Shik Eom, Sang Buhm Hahn, and Jong-Ho Park. "Autocorrelation and partial price adjustment", The Journal of Empirical Finance, Vol. 24, December 2013, pp. 78-93

[2] Black, Fischer, 1976, "Studies of stock price volatility changes", Proceedings of the 1976 meetings of the American Statistical Association, Business and Economics Statistics Section (American Statistical Association, Washington, DC) 177-181.

[3] Bollerslev, Tim. "Generalized Conditional Autoregressive Heteroskedasticity", Journal of Econometrics 31 (1986) 307-327.

[4] Campbell, John Y., Andrew W. Lo, A. Craig Mackinlay, "The Econometrics of Financial Markets", Princeton University Press, 2nd Edition, 1997.

[5] Christie, Andrew, "The stochastic behavior of common stock variances: Value, leverage, and interest rate effects", Journal of Financial Economics, 1982, 1O, 407-432.

[6] Chatfield, Christopher. "The Analysis of Time Series", Chapman and Hall/CRC, 2003.

[7] Cont, Rama. "Empirical properties of asset returns: stylized facts and statistical issues". Quantitative Finance, Vol.1, 2001, 223-236.

[8] Cornel,B. "Using the option pricing model to measure the uncertainty producing effect of major announcements", Financial Management, 7 (1978) 54-59.

[9] Engle, R.F. and V.K. Ng. "Measuring and testing the impact of news on volatility" Journal of Finance, 48(5) (1993), 1749-78.

[10] Fama, Eugene, "The Behavior of Stock Market Prices", The Journal of Business, Vol. 38, Issue 1, (Jan. 1965), 34-105.

[11] French, Kenneth, "Stock Returns and The Weekend Effect", Journal of Financial Economics, March 1980, 8, 55–69.

[12] French,Kenneth, and Richard Roll,"Stock Return Variances, The arrival of new information and the reaction of traders", Journal of Financial Economics 17 (1985), 5-26.

[13] Glosten, Lawrence R., Ravi Jagannathan and David E. Runkle, "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", The Journal of Finance Vol. 48, No. 5 (Dec., 1993), pp. 1779-1801.

[14] Hamilton, James D., "Time Series" Princeton University Press, 6th Edition, 1994.

[15] Nelson, D., "Conditional heteroskedasticity in asset returns: A new approach", Econometrica, 1990, 59, 347-370.

[16] Patell, James and Mark A. Wolfson. "Anticipated information releases reflected in call option prices", Journal of Accounting and Economics, Volume 1, Issue 2, 1979, pages 117-140.

[17] Ramchand,Latha and Raul Susmel, "Volatility and cross correlation across major stock markets", Journal of Empirical Finance, Volume 5, Issue 4, 1998,Pages 397-416,

[18] Stanley, Eugene, Vasiliki Plerou, and Xavier Gabaix, " A statistical physics view of financial fluctuations: Evidence for scaling and universality", Physica A 387 (2008) 3967-3981.

Chapter 2

Generalized Method of Moments

2.1 Introduction

The Generalized Method of Moments, or GMM, (Hansen, 1982) is an extension of Pearson's method of moments (MoM) which was developed in the 1890's and early 1900's [25], [26]. The method of moments is based on an analogy principle that makes use of the Law of Large Numbers' assurance that sample moments are consistent estimators of population moments.

Assume that we have a set of random variables $(x_1, x_2, \ldots, x_n)$ drawn from the density $f(x|\theta)$ where θ is an unknown parameter. The mean of x is defined as,

$$\mu_1 = \int_{-\infty}^{+\infty} x_i f(x|\theta) dx = h_1(\theta) \tag{2.1.1}$$

By the law of large numbers (LLN),

$$\widehat{\mu}_1 = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \to \mu_1 \quad \text{as} \quad n \to \infty \tag{2.1.2}$$

For large n, the population mean, $h_1(\theta)$ will be well approximated by the sample mean.

$$\widehat{\mu}_1 = h_1(\theta) \tag{2.1.3}$$

The moment estimator of the mean, denoted $\widehat{\theta}$, is the solution to the following equation,

$$\widehat{\mu}_1 = h_1(\widehat{\theta}) \tag{2.1.4}$$

Now we will generalize this concept for a vector of parameters, $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Define the kth population moment as,

$$\mu_k = \int_{-\infty}^{+\infty} x_i^k f(x|\Theta) dx = h_k(\Theta) \tag{2.1.5}$$

The sample analogue for the kth moment is,

$$\widehat{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} x_i^k \tag{2.1.6}$$

In general, the population moments $(\mu_1, \mu_2, \ldots, \mu_k)$ will be functions of the population parameters,

$$\mu_j = h_j(\theta_1, \theta_2, \ldots, \theta_k) \tag{2.1.7}$$

Replacing the left hand side of equation 2.1.7, with the sample moment $\widehat{\mu}_j$ yields the moment estimator, $\widehat{\theta}$.

$$\widehat{\mu}_j = h_j(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k) \tag{2.1.8}$$

We assume that $h_k(.)$ involves some or all of the parameters of the distribution. To estimate the K parameters we define K moment equations,

$$\begin{aligned}
\widehat{\mu}_1 - h_1(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) &= 0 \\
\widehat{\mu}_2 - h_2(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) &= 0 \\
&\vdots \\
\widehat{\mu}_k - h_k(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) &= 0
\end{aligned} \tag{2.1.9}$$

This system has K equations and K unknowns, $\theta_1, \ldots, \theta_k$.

The methods of moments estimators are obtained by solving the system of equations,

$$\begin{aligned}
\widehat{\theta}_1 &= g_1[\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_k] \\
\widehat{\theta}_2 &= g_2[\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_k] \\
&\vdots \\
\widehat{\theta}_k &= g_k[\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_k]
\end{aligned} \tag{2.1.10}$$

For $j = 1, \ldots, k$, if h_j^{-1} exists, then $h_j^{-1} = g_j$, and $\widehat{\theta}_j = g_j(\widehat{\mu}_j)$ is a unique moment estimator. If h_j^{-1} does not exist then any solution of $\widehat{\mu}_j = g_j(\widehat{\theta}_j)$ is a moment estimator.

Although moments based on powers of x provide a natural source of information about the parameters of a distribution, other functions of the data may also be useful. Note also, that there may be more than one set of moments that can be used for estimating the parameters, and there may also be more moment equations available then necessary.

Properties of the MoM Estimator

The MoM estimator has the following properties,

- The law of large numbers (LLN) implies that the moments are consistent estimators of their population counterparts.
- The moments are sample means, therefore the central limit theorem applies so they are asymptotically normal.
- The MoM estimators are not usually efficient estimators, but they will be if the moments are derived from the sufficient statistics.

2.1.1 Some MoM Estimators

As an illustration of the method of moments, suppose that we want to estimate the population mean μ and variance σ^2 of a random variable x_t. These two parameters satisfy the population moment conditions:

$$E[x_i] - \mu = 0 \tag{2.1.11}$$
$$E[x_i^2] - (\sigma^2 + \mu^2) = 0 \tag{2.1.12}$$

Substituting the expected values with their sample analogues gives,

$$\frac{1}{N}\sum_{i=1}^{n} x_i - \mu = 0 \tag{2.1.13}$$
$$\frac{1}{N}\sum_{i=1}^{n} x_i^2 - (\sigma^2 + \mu^2) = 0 \tag{2.1.14}$$

These two equations imply,

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{n} x_i \tag{2.1.15}$$
$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 \tag{2.1.16}$$

Method of Moments - OLS

The OLS estimator can be shown to be a method of moments estimator. Define the population regression as, $y = x\beta + u$, where $y \sim n \times 1$, $\beta \sim k \times 1$, $X \sim n \times k$ and $u \sim n \times 1$. Assume that the first column of the X matrix contains ones.

The following k population moment conditions are assumed to hold:

$$E[Xu] = 0 \tag{2.1.17}$$

This condition states that the disturbance terms and the k variables in the matrix X are orthogonal. This condition can be re-written as:

$$E[X(y - X\beta)] = 0 \tag{2.1.18}$$

The corresponding K sample moments are

$$\frac{1}{n}\sum_{i=1}^{n} x_{i,j}(y_i - x_i'\widehat{\beta}) \text{ for } j = 1, \ldots, K. \tag{2.1.19}$$

Recall that the first column of the X matrix is a column of ones.

The MoM approach requires that we find an estimator for β that sets expression 2.1.19 equal to zero. The OLS estimator $\widehat{\beta} = (X'X)^{-1}Xy$ satisfies this requirement. This can be seen by noting that,

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = X'y - X'X\beta \tag{2.1.20}$$

Method of Moments - MLE

The maximum likelihood estimator is also a method of moments estimator. Suppose we have the following likelihood function:

$$\frac{1}{N} lnL = \frac{1}{N} \sum_{i=1}^{N} lnf(y_i, |x_i, \theta) \tag{2.1.21}$$

Define the population expectations as,

$$E\Big[\frac{\partial f(y_i|x_i,\theta)}{\partial \theta_k}\Big] = 0, k = 1, \ldots, K \tag{2.1.22}$$

The maximum likelihood estimator is obtained by equating the sample analog to zero and solving for the parameters:

$$\frac{1}{N}\frac{\partial lnL}{\partial \theta_k} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial f(y_i|x_i,\theta)}{\partial \theta_k} = 0 \tag{2.1.23}$$

Method of Moments - Instrumental Variable

This application of the method of moments represents a very early statistical solution to the problem of an endogenous explanatory variable. Consider the following supply and demand system of equations (Wright, 1928, [27]):

$$q_t^D = \alpha p_t + u_t^D \tag{2.1.24}$$

$$q_t^S = \beta_1 n_t + \beta_2 p_t + u_t^S \tag{2.1.25}$$

$$q_t^D = q_t^S = q_t \tag{2.1.26}$$

where p_t is price, q_t is quantity, n_t is an exogenous variable, and all variables are expressed as deviations from their means. Suppose that we want to estimate the demand equation. $q_t^D = \alpha p_t + u_t^D$ How do we estimate α given q_t and p_t? Since p_t is endogenous, p_t and u_t^D are correlated.

The OLS estimate of $\widehat{\alpha}$ will be biased and inconsistent. [1] Wright's solution was to use an instrumental variable, Z_t^D, such that,

$$cov(z_t^D, u_t^D) = 0 \tag{2.1.28}$$

$$cov(z_t^D, p_t) \neq 0 \tag{2.1.29}$$

The first condition (eq. 2.1.28) states that the residual and the instrument must be uncorrelated. The second condition (eq. 2.1.29) states that the instrument and the endogenous variable must be correlated. The population moment condition is,

$$cov(q_t^D, z_t^D) - \alpha \cdot cov(z_t^D, p_t) = 0 \tag{2.1.30}$$

[1] The bias can be illustrated as follows:

$$\widehat{\alpha} = \frac{\sigma(q_t^D \cdot p_y)}{\Sigma p_t^2} = \frac{\Sigma(\alpha \cdot p_t + u_t^D \cdot p_y)}{\Sigma p_t^2} = \frac{\Sigma(\alpha \cdot p_t^2 + u_t^D \cdot p_y)}{\Sigma p_t^2} = \alpha + \frac{\Sigma(u_t^D \cdot p_t)}{\Sigma p_t^2} \tag{2.1.27}$$

Since $E(\Sigma(u_t^D \cdot p_t)) \neq 0$, $E(\widehat{\alpha}) \neq \alpha$

If $E[U_t^D] = 0$ then

$$E(q_t^D z_t^D) - \alpha E(z_t^D p_t) = 0 \tag{2.1.31}$$

The method of moments estimator is,

$$\widehat{\alpha} = \left(T^{-1}\sum_{i=1}^{T} z_t^D q_t^D\right) / \left(T^{-1}\sum_{i=1}^{T} z_t^D p_t^D\right) \tag{2.1.32}$$

This is an instrumental variables estimator using instrument z_t^D.

2.1.2 Sufficient Statistics and Factorization[2]

Unlike the maximum likelihood estimator, the method of moment estimator is not usually an efficient estimator. However, when sufficient statistics exist, the method of moments estimator can be functions of them. In this case the methods of moments estimator will also be the maximum likelihood estimator.

Assume that a random sample, $\{x_1, x_2, \ldots, x_n\}$, is drawn from some distribution $f(x|\theta)$, where θ is an unknown parameter. The goal of parameter estimation is to estimate θ from the random sample. A statistic $C_k(x_1, x_2, \ldots, x_n)$ is said to be sufficient for θ if the conditional distribution of $\{x_1, x_2, \ldots, x_n\}$, given $C_k = c$, does not depend on θ for any value of c. The sufficient statistic, C_k, contains all of the information needed to estimate θ. We cannot learn more about θ with any additional knowledge of the probability distribution of {x}.

Fischer-Neyman Factorization Theorem

The factorization theorem facilitates the identification of a sufficient statistic. A statistic $C_k(x_1, x_2, \ldots, x_n)$ is a sufficient statistic for θ if and only if the joint density function $f(x|\theta)$ can be factorized as follows,

$$f(x|\theta) = u(x)v(C_k(x), \theta) \tag{2.1.33}$$

The functions u and v are non-negative. The function u may depend on x, but it only does not depend on θ. The function v depends on θ, but only depends on the observed value x through the value of the statistic $C_k(x)$.

The sufficient statistic(s) of a sample is (are) all that we need to estimate the parameter(s) of the distribution. In addition, the maximum likelihood estimate of a parameter will be a function of the sufficient statistic, and in many cases the sufficient statistic will the maximum likelihood estimate itself.

Exponential Form[3]

If a random variable belongs to the exponential family then the sufficient statistics are particularly easy to identify. The exponential family includes a number of well

[2]This discussion on sufficient statistics and factorization is based on Degroot and Schervish (2011) [7]

[3]See Morris and Lock [23] for further discussion on the properties of the exponential family

known distributions including the beta, normal, binomial, Poisson, and gamma distributions. A random variable x is a member of the exponential (parametric) family of distributions if the probability density function has the form:

$$f(x_i) = exp\Big[A(x_i) + B(\theta) + \sum_{k=1}^{k} C_k(x_i)D_k(\theta)\Big] \tag{2.1.34}$$

where k is the number of parameters in the distribution. The exponential family has the same number of sufficient statistics as it has parameters.

In the case of a single parameter exponential density (k=1) the joint density for n observations of x_i is,

$$f(\mathbf{x}) = \prod_{i-1}^{n} f(x_i) = exp\left[\sum_{i=1}^{n} A(x_i)\right] exp\left[nB(\theta) + D_1(\theta)\sum_{i=1}^{n} C_1(x_i)\right] \tag{2.1.35}$$

where $\mathbf{x}$ is an $n \times 1$ vector consisting of $x_1, x_2, \ldots, x_n$. The term, $\sum_{i=1}^{n} C_1(x_i)$, is the sufficient statistic for θ.

Example 2.1. The Bernoulli distribution is a discrete probability distribution of a random variable, x, which takes a value of 1 with probability, θ, and a value of 0 with a probability of $q = 1 - \theta$.

Its mean and variance are,

$$E[x] = \theta \tag{2.1.36}$$
$$VAR[x] = q \cdot \theta \tag{2.1.37}$$

The probability mass function (pmf) for a single Bernoulli random variable is,

$$P(X = x) = \theta^x(1-\theta)^{1-x} \quad \text{for} \quad x = 0 \quad \text{or, } x = 1 \tag{2.1.38}$$

This pmf, which has a single parameter, θ, can be written in exponential form,

$$P(X = x) = exp\left[xlog\left(\frac{\theta}{1-\theta}\right) + log(1-\theta)\right] \tag{2.1.39}$$

A sample of n independent Bernoulli trials has the following distribution.

$$p(x_1, \ldots, x_n) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{\sum_{i=1}^{n} x_i} \tag{2.1.40}$$

Writing equation 2.1.40 in exponential form shows that,

$$\text{Sufficient Statistic for } \theta = \sum_{i=1}^{n} x_i \tag{2.1.41}$$

Exponential form for a K-parameter density

In general, for a k parameter density in the exponential family, the likelihood function has the form,

$$f(x;\Theta) = A(x) + B(\Theta) + \sum_{k=1}^{K} C_k(x)D_k(\Theta) \tag{2.1.42}$$

where, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$, $A(.), B(.), C_k(.)$ and $D_k(.)$ are functions. If the pdf is of this form, then the k functions $\sum_{i=1}^{n} C_k(x_i)$ are sufficient statistics.

For instance, if k = 2, then the two sufficient statistics are $\sum_{i=1}^{n} C_1(x_i)$ and, $\sum_{i=1}^{n} C_2(x_i)$.

Example 2.2. The normal distribution is a member of the exponential family with two sufficient statistics. The density for a single observation, x_i is,

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} \right] \tag{2.1.43}$$

The joint density for n iid Normal variables is,

$$f(\mathbf{x}) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right]^n exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} \right] \tag{2.1.44}$$

This can be re-written in standard exponential family form (i.e. equation 2.1.42) as follows,

$$f(\mathbf{x}) = exp\left[log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \right] exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - \mu\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\mu^2 \right) \right] \tag{2.1.45}$$

This can be simplified to,

$$f(\mathbf{x}) = exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2} - nlog\sqrt{2\pi\sigma^2} \right] \tag{2.1.46}$$

The sufficient statistics are:

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1.47}$$

$$m_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{2.1.48}$$

The two moment equations are,

$$\widehat{\mu} = m_1 \tag{2.1.49}$$

$$\widehat{\sigma}^2 = m_2 - m_1^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{2.1.50}$$

The estimators for the mean and variance derived from these sample moments (see Pearson's Model) are also the maximum likelihood estimators. (see Exercise 1).

Example 2.3. The beta distribution is also member of the exponential family with two sufficient statistics. It is defined on the range of (0,1). It has two shape parameters α and β which allow it to take a number of very different configurations. For instance, if $\alpha = \beta = 0.5$ the pdf is u-shaped, whereas if $\alpha = \beta = 2$ the pdf is hump-shaped. The beta distribution is the conjugate prior for Bernoulli, binomial, negative binomial and geometric distributions which makes it a very "popular" distribution in Bayesian inference. The beta density is defined as,

$$f(x_i;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1} \tag{2.1.51}$$

where $\alpha > 0$, $\beta > 0$ and $x_i \in (0,1)$. This can be written in exponential form,

$$f(x_i;\alpha,\beta) = exp\left((\alpha-1)log(x_i) + (\beta-1)log(1-x_i) + log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right) \tag{2.1.52}$$

The joint density in exponential form for x_i $(i = 1,\ldots,N)$ is,

$$f(x;\alpha,\beta) = exp\left[(\alpha-1)\sum_{i=1}^{N} log(x_i) + (\beta-1)\sum_{i=1}^{N} log(1-x_i) + nlog\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right] \tag{2.1.53}$$

The sufficient statistics are,

$$\sum_{i=1}^{N} log(x_i) \quad \text{and,} \tag{2.1.54}$$

$$\sum_{i=1}^{N} log(1-x_i) \tag{2.1.55}$$

The reader is asked to estimate α and β using the method of moments in Exercise **??**.

2.2 Introduction to the Generalized Method of Moments[4]

The examples we have seen so far have the same number of moment equations and parameters. There is a single solution to the moment equations, and at that solution, the equations are exactly identified. There are, however, instances where the number of moment equations exceeds the number of parameters, so the system is over-identified.

Assume the existence of $q \geq p$, moment conditions for p parameters,

$$E[f(x_t,\theta_0)] = 0 \tag{2.2.1}$$

[4]This overview of GMM is based on Hall(2015) [11], Cameron and Trivedi(2005) [5], and Greene(2008) [9]

for all t. Where $\theta_0 \sim p \times 1$ is a vector of true parameter values, $f \sim q \times 1$, and x_t is a stationary and ergodic vector of random variables.

The GMM estimator denoted $\widehat{\theta}_{GMM}$, is the value of θ that minimizes

$$Q_T = g_T(\theta)' W_T g_T(\theta) \tag{2.2.2}$$

where g_T is the sample moment ,

$$g_T = T^{-1} \sum_{t=1}^{T} f(x_t, \theta) \tag{2.2.3}$$

and $W_T \sim q \times q$, is a weighting matrix. W_T is positive semi-definite, and is assumed to converge to a positive definite matrix of constants. The GMM estimator is the estimate of θ that sets the function, Q_T as close to zero as possible. There is no exact solution to the moment equations when $q > p$, GMM locates an approximate solution.

If $p = q$ then GMM is the method of moments. The key difference between the two methods is that MoM cannot accommodate $q > p$.

Distribution Properties of the GMM Estimator

The distribution properties of the GMM estimator require the following assumptions:

1. $E[f(x_t, \theta_0)] = 0$
2. Identification: $f(x_t, \theta^1) = f(x_t, \theta^2)$ iff $\theta^1 = \theta^2$
3. The following $q \times p$ matrix exists and is finite with rank p:

$$G(\theta_0) = \frac{1}{T} plim \sum_{t=1}^{T} \frac{\partial f_t}{\partial \theta}\bigg|_{\theta_0}$$

4. W_T converges to a positive definite matrix of constants, W
5. $T^{1/2} \sum_{t=1}^{T} f_t \mid_{\theta_0} \xrightarrow{d} N(0, S_0)$ where, $S_0 = plim T^{-1} \sum_{t=1}^{T} \left[f_t f_t' \right] \Big|_{\theta_0}$

If these assumptions hold, then the GMM estimator, $\widehat{\theta}_{GMM}$, is a consistent estimator of θ_0.

$$T^{1/2}\big(\widehat{\theta}_{GMM} - \theta_0\big) \xrightarrow{d} N(0, V) \tag{2.2.4}$$

where,

$$V = \Big[G(\theta_0)' W_T G(\theta_0)\Big]^{-1} G(\theta_0)' W_T S(\theta_0) W_T G(\theta_0) \Big[G(\theta_0)' W_T G(\theta_0)\Big]^{-1} \tag{2.2.5}$$

The choice of W_T will determine the size of the variance matrix, V. W_T must be a positive definite matrix. If $S(\theta_0)$ is known then the most efficient GMM estimator is found by setting $W_T = S(\theta_0)^{-1}$. This is an intuitive result since it states that

the weighting matrix is inversely related to the asymptotic covariance matrix of the moments.

Setting $W_T = S(\theta_0)^{-1}$ simplifies the asymptotic variance of θ_{GMM},

$$V = \left[G(\theta_0)'S(\theta_0)^{-1}G(\theta_0)\right]^{-1} \tag{2.2.6}$$

When $S(\theta_0)$ is unknown, which is usually the case, the most efficient estimator of θ is found by setting $W_T = \widehat{S}^{-1}$, where $\widehat{S}$ is a consistent estimator of $S(\theta_0)$.

Estimating the GMM model

Hansen proposed a two step procedure for obtaining the optimal GMM estimator.

1. Set $W_T = 1$ and estimate the parameters using the GMM estimator (eq. 2.2.2). The resulting parameters will be consistent.

2. Using the parameter estimates from step 1, calculate $\widehat{S}_T$ and re-estimate θ using the GMM estimator where $W_T = \widehat{S}_T^{-1}$.

An early estimator of S_T was,

$$\widehat{S}_T = \widehat{\Omega}_0 + \sum_{j=1}^{h}(\widehat{\Omega}_j + \widehat{\Omega}_j') \tag{2.2.7}$$

where,

$$\widehat{\Omega}_j = T^{-1}\sum\nolimits_{t=1}^{T-j} f_t(v_{t+j}, \widehat{\theta_T})f_t(v_t, \widehat{\theta_T})' \tag{2.2.8}$$

As originally proposed by Hansen, estimation consisted of a single pass of these two steps. In practice, however, it is common to iterate through these two steps until the change in the vector of parameter estimates, $(\widehat{\theta}^{K+1} - \widehat{\theta}^{K})$ is negligible. The iterative procedure is believed to improve the finite sample properties of the estimator.

First Order Conditions

The first order conditions (FOC) for the minimization of Q_T are

$$G_T(\widehat{\theta}_T)W_T g_T(\widehat{\theta}_T) = 0 \tag{2.2.9}$$

where,

$$G_T(\theta) = T^{-1}\sum\nolimits_{t=1}^{T} \frac{\partial f_t(v_t, \theta)}{\partial \theta'} \tag{2.2.10}$$

Generally the FOC are solved using a numerical optimization technique such as BFGS.[6]

[6] BFGS, or the Boyden–Fletcher–Goldfarb–Shannon algorithm is an iterative method for solving unconstrained nonlinear optimization problems. See Chapter 10 of Cameron and Trivedi [5] for a discussion on numerical optimization techniques.

The HAC Estimator

The estimator for S_T described in equation 2.2.7 is a consistent estimator, but there is no guarantee that it will be positive semi-definite. Consequently, $\widehat{V}$ may not be positive semi-definite, and for a finite sample size the estimated variance of $\widehat{\theta}$ may be negative. [7]

Newey and West [24], proposed a heteroskedastic-autocorrelation consistent (HAC) estimator for S_T that will always be positive semi-definite.

$$\widehat{S}_{HAC} = \widehat{\Omega}_0 + \sum_{j=1}^{h} \omega_{jm}(\widehat{\Omega}_j + \widehat{\Omega}'_j) \tag{2.2.11}$$

where

$$\omega_{jm} = 1 - [j/(m+1)] \tag{2.2.12}$$

ω_{jm} is a kernel weight function, and m is the bandwidth (also called the lag truncation parameter). The kernel weight function ensures positive semi-definiteness. In finance applications, where the data consists of time series, it is common to specify S as a HAC estimator.

The kernel weight can have a number of alternative specifications. Andrews (1991) [1] describes three common choices for the kernel in the context of HAC estimators: Truncated, Bartlett, and Parzen. He also considers two additional kernels, the Tukey-Hanning kernel, and the Quadratic Spectra.[8]

$$\text{Truncated:} \quad \omega_{TR}(x) = \begin{cases} 1 & \text{for} \quad |x| \leq 1, \\ 0 & \text{otherwise6}|x|^3 \end{cases} \tag{2.2.13}$$

$$\text{Bartlett:} \quad \omega_{BT}(x) = \begin{cases} 1 - |x| & \text{for} \quad |x| \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{2.2.14}$$

$$\text{Parzen:} \quad \omega_{PR}(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{for} \quad 0 \leq |x| \leq 1/2, \\ 2(1 - |x|)^3 & \text{for} \quad 1/2 \leq |x| \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{2.2.15}$$

$$\text{Tukey - Hanning:} \quad \omega_{TH}(x) = \begin{cases} (1 + cos(\pi x))/2| & \text{for} \quad |x| \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{2.2.16}$$

$$\text{Quadratic Spectra :} \quad \omega_{QS}(x) = \frac{25}{12\pi^2 x^2}\left[\frac{sin(6\pi x/5)}{6\pi x/5} - cos(6\pi x/5)\right] \tag{2.2.17}$$

Andrews showed that the Quadratic-Spectra kernel is asymptotically optimal in terms of the mean square error and confidence level performance.[9] A Monte Carlo

[7] Also, iterative techniques used to estimate the optimal GMM estimate of $\widehat{\theta}$ when S_T is estimated using 2.2.7 may behave poorly.

[8] Equation 2.2.12 is a Bartlett kernel.

[9] Andrews considers confidence levels of 99%, 95% and 90%.

study comparing the five kernels shows that there are only small differences among the kernels. The Monte Carlo simulation also indicates that the Bartlett kernel used by Newey and West is somewhat inferior to the other four kernels.

All five kernel types are available choices in the R 'gmm' package.

Moment Selection

Increasing the number of moments in a GMM model will not reduce asymptotic efficiency, and generally improves asymptotic efficiency. It does so by reducing the variance,

$$V = \left[G(\theta_0)' S(\theta_0)^{-1} G(\theta_0) \right]^{-1} \tag{2.2.18}$$

The improvement in asymptotic efficiency may, however, come at a cost since increasing the number of moments will likely increase small sample bias and may also increase the small sample variance of the estimator.

The benefits of adding more moments depends on the application. For maximum likelihood there is no benefit since it is already fully efficient. Also, if the increase in moments is accompanied by a corresponding increase in parameters there will be no increase in efficiency. However, in the case of instrumental variables there can be considerable benefit.

The J statistic (Hansen, 1982 [14]), or the over-identifying restrictions test (OIR) is a specification test used for models where there are more moment conditions than parameters. It is a test of the closeness of the sample moment conditions to zero. When the model parameters are exactly identified ($p = q$), the objective function, Q_T, is exactly zero. However, when the model parameters are over-identified by the moment equations ($q > p$), the moment equations imply substantive restrictions. If the hypothesis that led to moment equations is incorrect then some of the sample moment restrictions will be violated. We test the closeness of the sample moment conditions to zero. The OIR test statistic is,

$$J = g_T(\widehat{\theta}_T)' \widehat{S}_T^{-1} g_T(\widehat{\theta}_T) \tag{2.2.19}$$

The J statistic is distributed χ^2_{q-p}. If the model is miss-specified or some of the moment conditions do not hold, the J statistic will be large relative to a χ^2_{q-p}. While the J statistic acts as a test for model miss-specification, it does not provide any information about how the model is miss-specified.

Finite Sample Properties of the GMM Estimator

While the GMM estimator is a consistent estimator, small samples may be biased. For instance, the inverse of the sample covariance matrix of the moments may be adversely impact by a single large value. This , in turn, will impact the estimate of W.

Hansen, Heaton and Yaron (1996) [15] examine the finite sample properties of three GMM estimators. Given the moment conditions,

$$E[f(x_t, \theta_0)] = 0 \tag{2.2.20}$$

We have seen that an efficient GMM estimator of the parameter vector θ is vector $\widehat{\theta}$ that minimizes the following function,

$$\left[T^{-1}\sum_{t=1}^{T} f(x_t,\widehat{\theta})\right]' [\widehat{S}(\theta)]^{-1}\left[T^{-1}\sum_{t=1}^{T} f(x_t,\widehat{\theta})\right] \tag{2.2.21}$$

where, $\widehat{\theta}$ is a consistent estimator of θ, and $\widehat{S}(\theta)$ is the estimated covariance matrix of the moments.

The first estimator Hansen et. al. consider is the *two-step* estimator. In this case, a consistent estimator of θ is found by setting S to the identity matrix (step 1), and estimating equation 2.2.22.

$$\left[T^{-1}\sum_{t=1}^{T} f(X_t,\widehat{\theta})\right]'\left[T^{-1}\sum_{t=1}^{T} f(X_t,\widehat{\theta})\right] \tag{2.2.22}$$

Given the parameters estimates from step one, $\widehat{S}$ is calculated and equation 2.2.21 is estimated to get the final estimates of θ.

The second estimator is the *iterative GMM* procedure. This procedure loops through second step of the two step procedure j times, updating $S(\theta_T^j)$ to $S(\theta_T^j)$ each time it estimates θ. The procedure continues until either j gets very large, or the change in θ_T^j is very small. Denote this estimator as θ_T^∞. Note that this procedure has no asymptotic advantage over the standard two step procedure. The issue for Hansen, et.al. is whether or not it improves the finite sample properties. In fact, a Monte Carlo test showed a very slight improvement over the two-step procedure.

The third estimator is called the *continuous updating estimator* or CUE. This procedure simultaneously estimates $\widehat{S}(\theta)$ and $\widehat{\theta}$. That is, the weighting matrix is continuously changed during the minimization process. In this case θ_c^T minimizes,

$$\left[T^{-1}\sum_{t=1}^{T} f(X_t,\widehat{\theta})\right]' [\widehat{S}(\theta)]^{-1}\left[T^{-1}\sum_{t=1}^{T} f(X_t,\widehat{\theta})\right] \tag{2.2.23}$$

One advantage that the CUE has over the other two estimators is that it is invariant to scaling of the moment conditions. The Monte Carlo analysis, however, did not find that any one estimator was dominant. The CUE was found to have less median bias then the other two estimators, but the sample distributions of the CUE estimates had fatter tails. The J test for over identifying retrictions was found to generally be more reliable for the CUE. The default estimator in the R 'gmm' package is the two-step method, but the package offers all three estimators as options.

2.3 GMM and Instrumental Variables[11].

Earlier we discussed Wright's [27] seminal work on instrumental variables (IV) and the method of moments. Now we will discuss the IV model in the context of the

[11]This section is based the IV discussion in Chapter 6 of Cameron and Trivedi [5]

GMM estimator. We begin with the basic linear model:

$$y_i = X_i'\theta + u_i \tag{2.3.1}$$

where $x_i' \sim 1 \times k$ and $E[u_i|X_i] = 0$. The standard model assumes that u_i and X_i are uncorrelated (exogeneity). If the exogeneity assumption is violated the OLS parameter estimates will be inconsistent. In the discussion of Wright's model, we saw that consistent parameter estimates can be obtained by introducing an instrumental variable. In IV estimation, an instrument z_i, must satisfy two conditions:

1. Exogeneity: $E[u_i|z_i] = 0$.
2. Relevance: z_i and x_i are correlated.

The exogeneity condition implies the population moment condition,

$$E[z_i(y_i - x_i'\theta)] = 0 \tag{2.3.2}$$

The objective function for the GMM estimator is,

$$Q_N = \left[\frac{1}{N}(y - X'\theta)Z\right] W_N \left[\frac{1}{N}Z'(y - X\theta)\right] \tag{2.3.3}$$

where $Z \sim N \times r$, $X \sim N \times k$, $W_N \sim r \times r$, and $\theta \sim k \times 1$. The first order conditions are,

$$\frac{\partial Q_n}{\partial \theta} = -2\left[\frac{1}{N}X'Z\right] W_N \left[Z'(y - X'\theta)\right] = 0 \tag{2.3.4}$$

Solving for θ gives the GMM IV estimator,

$$\widehat{\theta}_{GMMIV} = (X'ZW_nZ'X)^{-1}X'ZW_nZ'y \tag{2.3.5}$$

Replacing y with $y = X\theta + u$, and multiplying by N^{-1}

$$\widehat{\theta}_{GMMIV} = \theta + ((N^{-1}X'Z)W_N(N^{-1}Z'X))^{-1}(N^{-1}X'Z)W_N(N^{-1}Z'u) \tag{2.3.6}$$

Consistency requires that $plim(N^{-1}Z'u) \to 0$. The GMM estimator is asymptotically Normal with mean θ and variance:

$$V = N[X'ZW_NZ'X]^{-1}[X'ZW_N\widehat{S}W_NZ'X][X'ZW_NZ'X]^{-1} \tag{2.3.7}$$

where

$$\widehat{S} = lim\frac{1}{N}\sum_{i=1}^{N} E[u_i^2 z_i z_i'] \tag{2.3.8}$$

When the errors are homoskedastic, $E[u_i^2|z_i] = \sigma^2$, and S is consistently estimated using,

$$\widehat{S} = s^2 Z'Z/N \quad \text{where} \quad s^2 = (N-k)\sum_{i=1}^{N} \widehat{u}_i^2 \tag{2.3.9}$$

and, $\widehat{u} = y - X'\theta_{GMMIV}$.

When heteroskedasticity is present, S is consistently estimated by

$$\widehat{S} = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i^2 z_i z_i' = Z'DZ/N \tag{2.3.10}$$

where D is a diagonal matrix with entries $\widehat{u}_i^2$.

If the model is exactly identified ($p = q$), the choice of W_N is immaterial since all choices of W_N result in the same estimator. In this instance X'Z is a square invertible matrix so $[X'ZW_NZ'X]^{-1} = (X'Z)^{-1}(W_N)^{-1}(Z'X)^{-1}$. The GMM estimator simplifies to the standard IV estimator,

$$\widehat{\theta}_{GMMIV} = (X'Z)^{-1}Z'y \tag{2.3.11}$$

$$\widehat{V} = (Z'X)^{-1}\widehat{S}(Z'X)^{-1} \tag{2.3.12}$$

If the model is over identified ($q > p$), the optimal weighting matrix is $W_N = \widehat{S}$. If hetereoskedasticity is present, the GMM estimator is a two step estimator with,

$$\widehat{\theta}_{GMMIV} = (X'Z\widehat{S}^{-1}Z'X)^{-1}X'Z\widehat{S}^{-1}Z'y \tag{2.3.13}$$

$$\widehat{V} = (X'Z)\widehat{S}^{-1}(Z'X) \tag{2.3.14}$$

If $q > p$ and the residuals are homoskedastic, the optimal weighting matrix is $W_N = (Z'Z)^{-1}$. This estimator is the same as the 2SLS estimator and the estimation process has one step.

$$\widehat{\theta}_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'y) \tag{2.3.15}$$

$$\widehat{V} = (X'Z(Z'Z)^{-1}Z'X)^{-1} \tag{2.3.16}$$

2.3.1 GMM with Many Moment Conditions

Han and Phillips (2006) [12] examine the asymptotic properties of the GMM estimator when the number of moments is allowed to increase with the sample size, and the moments are allowed to be weak. One situation where this might arise is in the application of GMM to instrumental variables. That is, there may be weak instruments.

When the moment conditions are weak the GMM estimator will not be consistent. As Han and Phillips explain, the basic reason for the inconsistency is that moment condition has a low signal to noise ratio. Given the moment conditions,

$$E[f(w_i, \theta_0)] = 0 \tag{2.3.17}$$

where w_i are iid and θ_0 are the true parameters.

The GMM estimator minimizes $g(\theta)'g(\theta)$ where

$$g(\theta) = N^{-1}\sum\nolimits_{i=1}^{n} f(w_i, \theta) \tag{2.3.18}$$

$g(\theta)$ can be decomposed into its mean, and the variation about the mean,

$$g(\theta) = E[g(\theta)] + [g(\theta) - E[g(\theta)]] \tag{2.3.19}$$

Any information about θ_0 found in the sample of w_i comes from $E[g(\theta)]$. The other part of equation 2.3.19 is noise. When $E[g(\theta)]$ is flat and close to zero in a neighborhood of θ_0 the moment condition cannot identify θ_0 and is therefore a weak moment.

In the standard GMM framework $E[g(\theta)] = 0$ for θ_0, and as $n \to \infty$ this does not diminish. The asymptotics for weak moments are different in that the signal diminishes to zero at a rate $\sqrt{n}$ and the noise term also diminishes at the rate $\sqrt{n}$. The result is that the signal never dominates the noise, and the GMM estimate is inconsistent.

2.4 Testing the CAPM with GMM[14]

MacKinlay and Richardson (1991) [22] show that GMM can be used to test a portfolio of N assets for mean-variance efficiency. Their test is robust in the sense that GMM can account for heteroskedasticity and autocorrelation in the returns. In addition, the GMM approach is not constrained by having to assume Normality, which is the standard assumption one makes when using maximum likelihood,

Let R_t be a vector of **N** asset returns in excess of the risk free rate at time,t, and let f_t be a vector of **K** economy-wide risk factors. For instance, in the case of the standard market model, f_t represents returns for the market portfolio, r_{mt}. For the Fama-French model $f_t = (r_m t, SML_t, HML_t)$, where SML denotes returns for portfolio's defined by firm size (market capitalization), and HML denotes returns for portfolio's defined by book value. The mean and variance of the factors, are μ and $\mathbf{\Omega}$. The standard beta representation of the asset pricing model is

$$E[R_t] = \beta\delta \tag{2.4.1}$$

where δ is a vector of factor risk premia, and β is a matrix of factor loadings. The matrix β is defined as,

$$\beta \equiv E[R_t(f_t - \mu)']\Omega^{-1} \tag{2.4.2}$$

The factor loadings can also be defined as a time series parameter,

$$R_t = \phi + \beta f_t + \epsilon_t \tag{2.4.3}$$

where, $E[\epsilon_t] = 0$ and $E[f_t\epsilon_t] = 0$

Equation 2.4.1 imposes the following restriction on equation 2.4.3,

$$\phi = \beta(\delta - \mu) \tag{2.4.4}$$

[14]This discussion on testing the CAPM follows the work of Jagannathan, Skoulakis and Wang, [20]. Also, see Jagannathan, Skoulakis and Wang, [19] for a comprehensive survey of GMM applications in finance.

Substituting this into the asset pricing model gives the following three equations.

$$R_t = \beta(\delta - \mu + f_t) \tag{2.4.5}$$
$$E[\epsilon_t] = 0 \tag{2.4.6}$$
$$E[f_t\epsilon_t] = 0 \tag{2.4.7}$$

These three definitions can be used to define the moment conditions:

$$E[R_t - \beta(\delta - \mu + f_t)] = 0 \tag{2.4.8}$$
$$E[(R_t - \beta(\delta - \mu + f_t)f_t)] = 0 \tag{2.4.9}$$
$$E[f_t - \mu] = 0 \tag{2.4.10}$$

When the economy-wide factor, f_t, is the excess return on a traded asset the risk premium is equal to the mean of the factor ($\delta = \mu$) and the sample mean of the factor can be used to estimate the risk premium. In this case we do not have to include equation 2.4.10 in the set of estimated moment equations.[15] The traded factor typically used in CAPM studies is the portfolio of all traded stocks. For instance, the Kenneth French data set defines the market portfolio as all stocks traded on the NYSE,AMEX and NASDAQ. Alternatively, the Fama-French three factor model also includes returns from size and book-to-market portfolios as additional traded assets.

For traded factors the moment conditions can be write as,

$$E[R_t - \beta f_t)] = 0 \tag{2.4.11}$$
$$E[(R_t - \beta f_t)f_t)] = 0 \tag{2.4.12}$$

In this case, if $k = 1$ so that the only factor is the market portfolio, there are N parameters and 2N moment conditions. The J statistic can be applied to test for over-identification.

Alternatively, when all of the assets are traded we can test the hypothesis that $\alpha = 0$ by defining the moment conditions as,

$$E[f(\alpha, \beta)] = 0_{2N\times 1} \tag{2.4.13}$$

where

$$f(\alpha, \beta) = \begin{bmatrix} \varepsilon_t \\ z_{mt}\varepsilon_t \end{bmatrix} = \begin{bmatrix} z_t - \alpha - \beta z_{mt} \\ z_{mt}(z_t - \alpha - \beta z_{mt}) \end{bmatrix}$$

Here we included α in the moment conditions, so we can test the hypothesis that $\alpha = 0$. This model has the same number of parameters as moment conditions so that the sample analogues are exactly satisfied. The result is the least squares estimator. Since the GMM equation is exactly specified the J test irrelevant. This is because $J = 0$ when the system of equations is exactly identified.

Alternatively, MacKinlay and Richardson point out that mean-variance efficiency can be tested by setting $\alpha = 0$. In this case, there are N parameters and $2N$ restrictions. The model is over identified, and the J-statistic can be used to test the validity of the restrictions.

[15] If the factor is not a traded asset this restriction does not hold, and the three moments need to be estimated. An example of a non-traded factor is inflation.

Example 2.4. In this example we test the hypothesis that $\alpha = \mathbf{0}$ for a set of 8 "new media" firms. These are non-traditional media firms that have entered the media industry in the past 10 years, and have effectively changed the nature of the industry.[16] Summary statistics of the daily returns for each firm, along with the market portfolio are shown in Table 2.1. The sample of daily returns covers the period from November 7 2013 to June 16, 2017. Netflix has the highest average daily return in the sample, and Twitter the lowest. Twitter has the highest standard deviation, and Alphabet has the lowest. The market portfolio, Apple, and Twitter have a negative skew. The excess kurtosis is positive for all firms, and also the market portfolio.

Company	**Ticker**	**Mean**	**Std. Dev.**	**Skew**	**Ex. Kurt**
Apple	AAPL	0.07	1.54	-0.48	6.49
Amazon	AMZN	0.13	1.86	0.47	10.32
Facebook	FB	0.17	2.10	2.97	38.08
Alphabet	GOOG	0.09	1.43	2.15	22.54
Microsoft	MSFT	0.09	1.45	0.02	10.71
Netflix	NFLX	0.26	3.10	3.14	38.51
Twitter	TWTR	-0.04	3.49	-0.55	9.17
Yahoo	YHOO	0.10	1.85	0.07	3.03
	Mkt	0.06	0.80	-0.39	2.17

Table 2.1: Summary Statistic for New Media Company, Daily Returns (%)

Output from the gmm command in the 'gmm' package in R is shown in Table 2.2. This example used the default options, which include the optimal 2 step estimation procedure, and the quadratic spectral kernel. By default the command also displays the J-test, which in this case is zero since the model is exactly identified. The results for

Results from the hypothesis test $\mathbf{H_0} : \alpha = \mathbf{0}$ are shown in Table 2.3. We are unable to reject the null hypothesis that the intercepts are jointly equal to zero.

[16] Some prominent "old media" firms include CBS, FOX, Disney, and Sony.

gmm(g = z zm, x = h)
Method:twoStep
Kernel:Quadratic Spectral

Coefficients:

Estimate	Coefficient	Std. Error	t value	$Pr(>\|t\|)$
MSFT_(Intercept)	0.0267	0.0363	0.7359	0.4618
FB_(Intercept)	0.0873	0.0521	1.6768	0.0936
TWTR_(Intercept)	-0.0978	0.1083	-0.9036	0.3662
AAPL_(Intercept)	0.0380	0.0421	0.9018	0.3672
YHOO_(Intercept)	0.0150	0.0467	0.3213	0.7480
AMZN_(Intercept)	0.0796	0.0569	1.3987	0.1619
NFLX_(Intercept)	0.1066	0.0863	1.2356	0.2166
GOOG_(Intercept)	0.0294	0.0393	0.7487	0.4541
MSFT_zm	1.1091	0.0489	22.6630	0.0000
FB_zm	1.2080	0.0677	17.8450	0.0000
TWTR_zm	1.0897	0.1294	8.4228	0.0000
AAPL_zm	0.9786	0.0519	18.8650	0.0000
YHOO_zm	1.2136	0.0669	18.1300	0.0000
AMZN_zm	1.1690	0.0841	13.9030	0.0000
NFLX_zm	1.2863	0.1210	10.6300	0.0000
GOOG_zm	1.0495	0.0513	20.4580	0.0000

J-Test: df = 0

J-test	P-value	
Test E(g)=0:	2.600222	7e-26 *******

Table 2.2: GMM Estimates of CAPM - New Media Firms

Linear hypothesis test

Hypothesis:
MSFT_((Intercept) = 0
FB_((Intercept) = 0
TWTR_((Intercept) = 0
AAPL_((Intercept) = 0
YHOO_((Intercept) = 0
AMZN_((Intercept) = 0
NFLX_((Intercept) = 0
GOOG_((Intercept) = 0

Model 1: restricted model
Model 2: z zm

Res.Df Df Chisq Pr(¿Chisq)
1 913
2 905 8 7.9166 0.4417

Table 2.3: Hypothesis Test - New Media Firms

2.5 GMM Estimation of a Stochastic Volatility Model

GMM has been used by a number of researchers to estimate the stochastic volatility model.[19] The log Normal stochastic volatility model is specified as,

$$r_t = \sigma_t \epsilon_t \tag{2.5.1}$$

$$ln(\sigma_t^2) = \omega + \beta ln(\sigma_{t-1}^2) + \sigma_u u_t \tag{2.5.2}$$

where r_t is the return on an asset, ϵ_t and u_t are independent of each other, and each is standard Normal. When $0 < \beta < 1$ and $\sigma_u \geq 0$, r_t is stationary and ergodic, and the unconditional moments of any order exist. The parameters in this model are $\theta = (\omega, \beta, \sigma_u)$ where ω is the level of log-variance, β is the persistence of log-variance, and σ_u is the volatility of log-variance. Note that the only observable in this specification is the time series of returns, r_t. In addition, note the time subscript on the volatility indicating that this is a model of time varying volatility.

This model captures an important characteristic of asset returns, in that r_t is uncorrelated but dependent over time. The dependence is captured by the autoregressive specification of $ln(\sigma_t^2)$.

Andersen and Sorensen[3] use Monte Carlo analysis to evaluate the finite sample properties of the GMM parameter estimates for the stochastic volatility model. They select a set of lower order moment conditions in order to capture the fat tails that characterize the distribution of returns. The moments were originally proposed by Jacquirer, Polson and Rossi (JPR) [21]. [20] Andersen and Sorensen ran simulations for various subsets of the following moments:[21]

$$E|r_t| = (2/\pi)^{1/2} E(\sigma_t) \tag{2.5.3}$$

$$E(r_t^2) = E(\sigma_t^2) \tag{2.5.4}$$

$$E(r_t^3) = 2\sqrt{(2/\pi)} E(\sigma_t^3) \tag{2.5.5}$$

$$E(r_t^4) = 3E(\sigma_t^4) \tag{2.5.6}$$

$$E|r_t r_{t-j}| = (2/\pi) E(\sigma_t \sigma_{t-j}) \quad \text{for} \quad j = 1 \ldots 10 \tag{2.5.7}$$

$$E[r_t^2 r_{t-j}^2] = E(\sigma_t^2 \sigma_{t-j}^2) \quad \text{for} \quad j = 1 \ldots 10 \tag{2.5.8}$$

where for any positive integer j, and any positive constants r and s:

$$E(\sigma_t^r) = exp(r\mu/2 + r\sigma^2/8) \tag{2.5.9}$$

$$E(\sigma_t^r \sigma_t^s) = E(\sigma_t^r) E(\sigma_t^s) exp(rs\beta^j \sigma^2/4) \tag{2.5.10}$$

where $\mu = \omega/(1-\beta)$ and $\sigma^2 = \sigma_u^2/(1-\beta^2)$ are the unconditional mean and variance of equation 2.5.2.

[19] See Andersen and Sorensen, 1996 [3] for a discussion of the different approaches that have been used to estimate the stochastic volatility model.

[20] Andersen and Sorensen use the specification of JPR as a benchmark. JPR showed that the GMM estimators of the log Normal stochastic volatility model have poor finite sample properties compared with the Bayesian estimator.

[21] See JPR for the first derivatives of the moments.

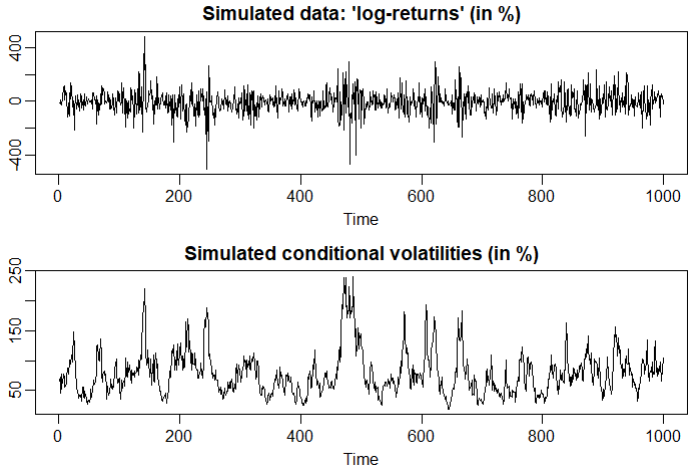

Figure 2.1: Simulated data for Monte Carlo

Example 2.5. In this example we compare the results of a Monte Carlo analysis using the same 3 and 5 moments conditions from Andersen and Sorensen. We simulate a series of T=1000 observations using the following parameter vector, $\theta = (\omega, \beta, \sigma_u) = (-0.736, 0.9, 0.363)$, and estimate the model using the GMM two-step procedure with a quadratic spectral kernel. This procedure is repeated 1000 times. The simulated series for log returns and log volatility are shown in Figure 2.1. Following AS [3] the exactly identified model uses moments, M1, (eq:2.5.3), M2,(eq:2.5.4), and M5,(eq:2.5.7). The discrepancy functions for the 3 moment conditions are:

$$g_1 = (1/T)\sum |y_t| - (2/\pi)^{1/2} exp(\mu/2 + \sigma^2/8) \tag{2.5.11}$$

$$g_2 = (1/T)\sum y_t^2 - exp(\mu + \sigma^2/4) \tag{2.5.12}$$

$$g_5 = (1/T)\sum |y_t y_{t-1}| - (2/\pi) exp(\mu + \sigma^2/4)(1 + \beta^1) \tag{2.5.13}$$

In addition to the g_1 and g_2, the 5 moment specification also includes the following discrepancy functions:

$$g_4 = (1/T)\sum y_t^4 - 3 * exp(2\mu + 2\sigma^2) \tag{2.5.14}$$

$$g_6 = 1/T)\sum |y_t y_{t-2}| - (2/\pi) * exp(2(\mu/2 + \sigma^2/8) + (\sigma^2/4) * \beta^2) \tag{2.5.15}$$

$$g_{15} = (1/T)\sum (y_t^2 y_{t-1}^2) - exp(2(\mu/2 + 4 * \sigma^2/8) + \beta\sigma^2)) \tag{2.5.16}$$

The procedure used to derive the estimate was as follows:

1. Simulate a sample of 1,000 observations using the parameters $\theta = (\omega, \beta, \sigma_u) = (-0.736, 0.9, 0.363)$

2. Set the initial value of ω to zero, and initialize β and σ_u with random draws from a uniform distribution with a range of $0 < x < 1$.

3. Estimate the model for a set of moments, which for this example is either 3 or 5.

4. Repeat steps 1 - 3, N times. In this example, N=1,000, 5,000, and 10,000 times.

The model specification places two constraints on the parameter space, $0 < \beta < 1$, and $\sigma_u > 0$. In this example, if the parameter estimates were outside these bounds they were discarded. Also, if the covariance matrix was singular the results were discarded. The results of the Monte Carlo simulation, showing average parameter estimates, and the root mean squared error (RMSE) of the parameter estimate, are reported in Table 2.4.[22] The RMSE is generally smaller for the 5 moment specification. The two specifications differ the most in the estimation of $\widehat{\sigma}_u$, with the RMSE for the 3 moment model almost 4 times larger than that of the 5 moment model. Surprisingly the RMSE does not change much as T increases for either specification. AS found a substantial decrease in the RMSE at T increased. They also encountered a large number of iterations where the model did not converge. They determined that this happened most often when $\widehat{\beta}$ approached one, so they set a maximum value of 0.9999 to reduce the number of non-convergences.

# of Moments	**3**	**5**
T=1000		
$\widehat{\omega}$	-0.4661 (0.5597)	-0.4130 (0.3700)
$\widehat{\beta}$	0.5628 (0.4162)	0.5545 (0.4055)
$\widehat{\sigma}_u$	0.7438 (0.7372)	0.3135 (0.2006)
T=5000		
$\widehat{\omega}$	-0.4812 (0.5618)	-0.3952 (0.3836)
$\widehat{\beta}$	0.5605 (0.4163)	0.5713 (0.3901)
$\widehat{\sigma}_u$	0.7809 (0.7585)	0.3046 (0.1998)
T=10000		
$\widehat{\omega}$	-0.4943 (0.5845)	-0.3978 (0.3816)
$\widehat{\beta}$	0.5636 (0.4132)	0.5697 (0.3912)
$\widehat{\sigma}_u$	0.7922 (0.7779)	0.3066 (0.2000)

Table 2.4: Monte Carlo Simulation, Average Parameter Estimate, Root Mean square Error in Parentheses

[22]The code used to estimate the parameters is provided at the end of the chapter.

2.6 Exercises

1. Show that the maximum likelihood estimates for mean and variance of the Normal distribution are the same as the method of moment estimates.

2. Show the the OLS estimate of α in equation 2.1.24 is an inconsistent estimator.

3. Let $\{x_1, x_2, ..., x_n\}$ denote a random sample from a beta distribution The population mean and variance for the beta distribution are,

$$E[x_i] = \frac{\alpha}{\alpha + \beta} \tag{2.6.1}$$

and,

$$Var[x_i] = \frac{(\alpha)(\beta)}{(\alpha_+\beta)^2(\alpha + \beta + 1)} \tag{2.6.2}$$

Estimate α and β using the method of moments.

2.7 R Code for Examples

2.7.1 Example 2.4

```
library(gmm)
z <- as.matrix(newmediafirms[,3:10])
zm <- as.matrix(newmediafirms[,2])
t <- nrow(zm)
h <- matrix(zm,t ,1)
res <- gmm(z ~ zm, x = h)
summary(res)

library(car)
linearHypothesis(res,cbind(diag(8),matrix(0,8,8)),rep(0,8))
```

2.7.2 Example 2.5

Three moments:

```
library(stochvol)
sim <- svsim(1001, mu = -0.736, phi = 0.9, sigma = 0.363)
x1<-Lag(sim$y, 1)
x<-cbind(sim$y,x1)
x<-na.omit(x)
par(mfrow = c(2, 1))
plot(sim)

library(gmm)
#Define Moment Function
g1 <- function(theta, x) {
mu<-theta[1]/(1-theta[2])
sigmasq<-theta[3]^2/(1-theta[2]^2)
moments<-c(
m1 <- sqrt(2/pi)*exp(mu/2 + sigmasq/8),
m2 <- exp(mu + sigmasq/2),
m5 <- (2/pi)*exp(2*(mu/2 + sigmasq/8) + theta[2]*sigmasq/4)
)
obsdata <- data.frame(cbind(abs(x[,1]),x[,1]^2,abs(x[,1]*x[,2]) ))
f<-obsdata - t(moments)
return(f)
}
theta<-matrix(data=NA,nrow=3,ncol=1)
outparm<-matrix(NA,nrow=10000,ncol=5)
#Simulate T times
for(i in 1:10000){
sim <- svsim(1001, mu = -0.736, phi = 0.9, sigma = 0.363)
theta[1]<- 0
```

```
theta[2]<- runif(1)
theta[3]<- runif(1)
x2<-Lag(sim$y, 1)
x1<-cbind(sim$y,x2)
x<-na.omit(x1)
res1 <- gmm(g1,x=as.matrix(x),t0=theta)
dum<-ifelse(0.0001>res1$coefficients[2] || res1$coefficients[2]>0.9999,0,1)
outparm[i,1:3]<-res1$coefficients
outparm[i,4]<-dum
outparm[i,5]<-res1$vcov[1,1]
print(i)
}

#Apply parameter constraints and remove NA's
outparm<-outparm[outparm[,3]>0,,drop=FALSE]
outparm<-outparm[outparm[,2]>0.0001,,drop=FALSE]
outparm<-outparm[outparm[,2]<0.9999,,drop=FALSE]
outparm<-outparm[!is.na(outparm[,1]),]
outparm<-outparm[!is.na(outparm[,1]),]
outparm<-outparm[!is.infinite(outparm[,5]),]


mean(outparm[,1])
mean(outparm[,2])
mean(outparm[,3])

#RMSE
parm1<-sqrt(mean((outparm[,1]-(-0.736))^2))
parm2<-sqrt(mean((outparm[,2]-(0.9))^2))
parm3<-sqrt(mean((outparm[,3]-(0.363))^2))

#Plot
plot(density(outparm[,1]),main="Omega")
plot(density(outparm[,2]),main="Beta")
plot(density(outparm[,2]),main="Sigma_eta")
```

5 moments:

```
g1 <- function(theta, x) {
mu<-theta[1]/(1-theta[2])
sigmasq<-theta[3]^2/(1-theta[2]^2)
moments<-c(
m1 <- sqrt(2/pi)*exp(mu/2 + sigmasq/8),
m2 <- exp(mu + sigmasq/2),
m4 <- 3*exp(2*mu + 2*sigmasq),
```

```
m6 <- (2/pi)*exp(2*(mu/2 + sigmasq/8) + (sigmasq/4)*theta[2]^2),
m15 <- exp(2*(mu/2+ 4*sigmasq/8) +  theta[2] * sigmasq) )
obsdata <- data.frame(cbind(abs(x[,1]),x[,1]^2,x[,1]^4,abs(x[,1]*x[,3]),x[,1]^2*x[,
f<-obsdata - t(moments)
return(f)
}
```

Bibliography

[1] Andrews, Donald W.K., "Heteroskedasticity and Autocorrelation Consistent Covariance MAtrix Estimation", Econometrics, Vol. 59, No. 3, (May 1991) , pp. 817-858.

[2] Andrews, Donald W.K., "Consistent moment selection procedures for generalized method of moments estimation", Econometrics, Vol. 67, No. 3, (May 1999) , pp. 543-564.

[3] Andersen, Torben G. and Bent E. Sorensen. "GMM Estimation of a Stochastic Volatility Model" A Monte Carlo Study", Journal of Business and Economic Statistics, Vol. 14, No. 3. (Jul., 1006), pp. 328-352.

[4] Bontemps, Christian and Nour Meddahi. "Testing Normality: A GMM Approach", Journal of Econometrics 124 (2005), pp.149-186

[5] Cameron, A. Colin and Pravin K. Trivedi. "Microeconometrics:methods and applications", Cambridge University Press, 2006.

[6] Declerq, David and Patrick Duvaut, "Hermite Normality Tests", Signal Processing, 69(2):101-116, Oct, 1998.

[7] Degroot, Morris H. and Mark J. Schervish. "Probability and Statistics" 4th Edition(2011), Pearson.

[8] Erickson, Timothy and Toni Whited. "Two-step GMM estimation of the Errors-in-Variables model using high-order moments", Econometric Theory, 18, 2002, 776-799.

[9] Greene, William H., "Econometric Analysis", 6th edition (2008), Pearson Hall, New Jersey.

[10] Hall, Alistair R, and Peixe, F.P.M. (2003), 'A Consistent Method for the Selection of Relevant Instruments', Econometric Reviews, Vol. 22, No.3 (Aug. 2003), 269–88.

[11] Hall, Alistair R., "Econometricians Have Their Moments: GMM at 32", Economic Record, Vol. 91, Special Issue, June 2015 1-15.

[12] Han, Chirok, and Peter C.B. Phillips, "GMM with Many Moment Conditions", Econometrica, Vol. 74. No. 1 (January 2006), 147-192.

[13] Hansen, Bruce E., and Kenneth D. West. "Generalized Method of Moments and Macroeconomics", Journal of Business and Economic Statistics, October 2002, Vol. 20, No. 4, pp 460-469.

[14] Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators" Econometrica, Vol. 50, No. 4 (July 1982), pp. 1029-1054.

[15] Hansen, Lars Peter, John HEaton, and Amir Yaron, "Finite Sample Properties of Some GMM Estimators", Journal of Business and Economic Statistics, Vol. 14, No. 3 (July 1996), pp. 262-280.

[16] Hansen, LArs Peter, "Generalized Method of Moments Estimation: A Time Series Perspective", International Encyclopedia of Social and Behavioral Sciences, 2001, Elsevier.

[17] Hansen, Lars Peter, "Generalized Mthod of Moments Estimation", The NEw Palgrave Dictionary of Economics, Second Edition, 2008. Edited by Steven N. Durlauf and Lawrence E. Blume.

[18] Hogg, Robert V., and Allan T. Craig. "Sufficient statistics in elementary distribution theory", Sankhyā: The Indian Journal of Statistics, Volume 7, No. 3, (Dec. 1956), pp 209-216.

[19] Jagannathan, Ravi, Georgios Skoulakis, and Zhenyu Wang, "Generalized Method of Moments: Applications in Finance", Journal of Business and Economic Statistics, Vol. 20, No. 4 (Oct. 2002), pp 470-481.

[20] Jagannathan, Ravi, Georgios Skoulakis, and Zhenyu Wang, "The Analysis of the Cross-Section of Security Returns", The Handbook of Financial Econometrics, Volume 2, 2009, Editors: Yacine Ait-Shahalia and Lars Peter Hansen. Elsevier.

[21] Jacquirer, Eric, Nicholas G. Polson, and Peter E. Rossi, "Bayesian Analysis of Stochastic Volatility Models", Journal of Business & Economic Statistics, Vol. 12, No. 4 (Oct. 1994), pp 69 - 87.

[22] MacKinlay, A. Craig and Matthew P. Richardson, "Using GMM to Test Mean-Variance Efficiency", The Journal of Finance, Vol. 66, No. 2 (June 1991), pp. 511-527.

[23] Morris, Carl N. and Kari F. Lock, "Unifying the Named Natural Exponential Families and Their Relatives", The American Statistician, Vol. 63, No. 3, (August 2009), pp. 247-253.

[24] Newey, Whitney K. and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", Econometrica, Vol. 55, No. 3 (May 1987), pp. 703-708.

[25] Pearson, Karl. "Contribution to the mathematical theory of evolution", Philosophical Transactions of the Royal Society of London 185, Series A, 71-110.

[26] Pearson, Karl. "On the systematic fitting of curves to observations and measurements", Biometrika, Volume 1, N0. 3, April 1902.

[27] Wright, Philip. "The Tariff on Animal and Vegetable Oils", Appendix B, 1928, New York: Macmillan.

1

Chapter 3

Fractional Differencing and Long Memory Models

3.1 Introduction

In this lecture we generalize the ARIMA model to allow for long memory. The autocorrelation of a time series with long memory is more persistent than that of a standard stationary ARMA process. In Chapter 1 we saw that the ACF for a stationary series decreases exponentially as lag length, k, increases. There are, however, stationary time series that have an ACF that declines at a slower rate, and clearly cannot be considered $I(0)$. These series are characterized as long memory. Heuristically we can define a time series as having a long memory if the autocorrelation function of the series decreases hyperbolically, rather than exponentially. As we will see, stationary long memory models are neither $I(1)$ nor $I(0)$. Instead, they are $I(d)$ where $|d| < 0.5$

Consider the following ARIMA model,

$$\mathbf{\Delta^d x_t} = \mathbf{(1-B)^d x_t} = \mathbf{e_t} \tag{3.1.1}$$

where B is a lag operator,d is the degree of integration, and e_t is a white noise process with $E(e_t) = 0$, $E(e_t^2) = \sigma^2$, and $E(\sigma_s, \sigma_t) \neq 0$ for $s \neq t$. The series x_t said to be integrated of order d, or x_t $I(d)$.
If d = 1 then,

$$\mathbf{x_t} = \mathbf{x_{t-1}} + \mathbf{e_t} \tag{3.1.2}$$

and the series x_t is a random walk.
If d = 0 then,

$$\mathbf{x_t} = \mathbf{e_t} \tag{3.1.3}$$

and the series x_t is a white noise stationary process.

The autocorrelation function (ACF) for the random walk series is shown in Figure 3.1. It decreases very slowly with k in this example. Note, however, that

the covariance is not fixed for $|t-s|$ as it is for a stationary series.[1]

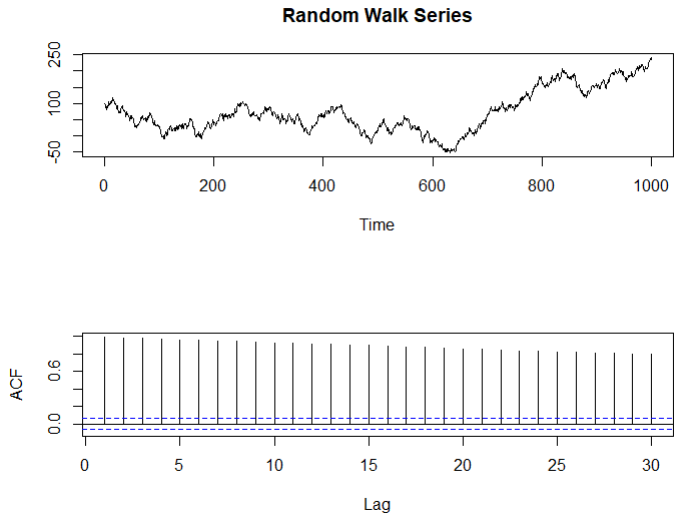

Figure 3.1: Random Walk Series and Autocorrelation Function

Figure 3.2 shows the first difference of the random walk series, along with the ACF. This series is a white noise process, and the ACF is never significantly different from zero.

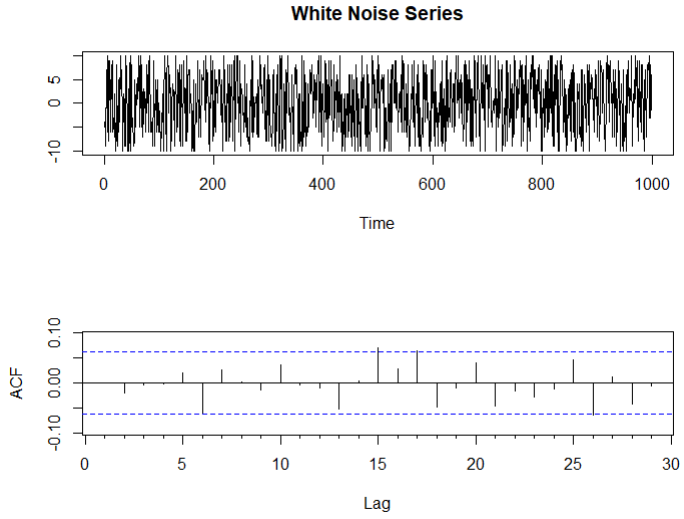

Figure 3.2: White Noise Series and Autocorrelation Function

First differencing the random walk series results in a series that is stationary. This, of course, is a standard approach for transforming a nonstationary time series

[1]In Chapter 1, the derivation of the ACF assumed that the time series was stationary. The autocorrelation function for a random walk is $\frac{1}{\sqrt{(1+k/t)}}$, where k = lag and t = time. For small values of k relative to t, the autocorrelation will be close to one.

into a stationary series. We transform the data from an $I(1)$ process, to an $I(0)$ process. If a process is fractional d, however, first differencing may not be an inappropriate transformation of the series.

Standard unit root tests have been shown to have less power against a fractional d process (see, for instance, Diebold and Rudebusch [6]). As a result, if a process is fractional d, d=1 will be chosen more often then the correct $d < 1$ using a standard such as the Dickey-Fuller test.

3.2 Long Memory Models

3.2.1 Defining Long Memory

There are several definitions of a long memory model. Definitions 1, 2, and 3 below are discussed in Granger, 2004.[12][2]

Definition 1: When d>0 the autocovariance follows a power law:

$$\gamma_k \equiv c_k k^{2d-1} \text{ as k} \to \infty \tag{3.2.1}$$

where γ is the autocovariance, c_k is a slowly varying function at infinity, and d is the order of integration.

In this case, the autocorrelation function (ACF) is proportional to a power law,

$$\rho_k \propto k^{2d-1} \text{ as k} \to \infty \tag{3.2.2}$$

The autocorrelation function decays hyperbolically when 0 >d <1, as opposed to the exponential decay of the ACF of a stationary process.[3] Figure 3.3 illustrates this by comparing the ACF for a series with d=0.4[4] to a stationary AR(1) series.[5]

Definition 2: A time series x_t has long memory if the following measure is non-finite:

$$\lim_{T\to\infty} \sum_{k=-T}^{T} |\rho_k| \tag{3.2.3}$$

where ρ_k is the autocorrelation with lag k. In this case, the autocorrelations decay so slowly that they are not summable. An equivalent way of stating this is that for a long memory process the spectral density $f(\omega)$ is unbounded at low frequencies. [6] This leads to the third definition.

[2]Also, See Samorodnitsky, 2006, [20] for an indepth discussion on long range dependence and the various definitions.

[3]Note when $d = 0$ the series is said to have short memory.

[4]$\Delta^{0.4} x_t = e_t, e_t \sim iid(0, \sigma^2)$

[5]$x_t = 0.9 * x_{t-1} + e_t, e_t \sim iid(0, \sigma^2)$. The ACF for the series is 0.9^k, where k = lag.

[6]This is in contrast to a stationary and invertible ARMA process which has autocorrelations that are bounded at low frequencies, and has short memory. That is, $|\rho_k| \leq cm^{-k}$ for large k, where $0 < m < 1$. see [3]

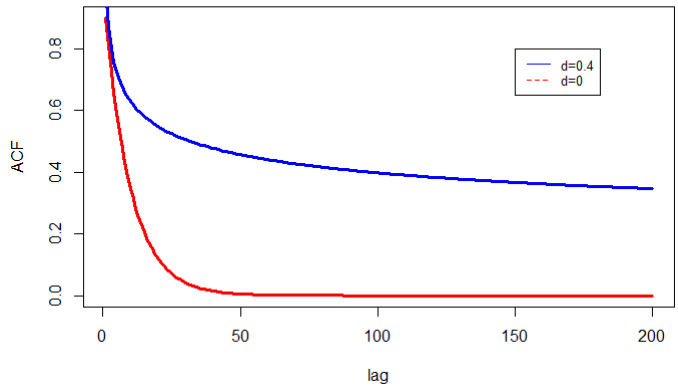

Figure 3.3: ACF for process with d=0.4 vs. d=0

Definition 3: As the frequency of a time series approaches zero, the spectral density approaches infinity.

$$f_x(\omega) = \infty \text{ as } \omega \to 0^+ \tag{3.2.4}$$

where ω is the frequency. Figure 3.4 shows the spectral densities of a standard white noise series ($d = 0$)[7] along with that of a fractionally integrated white noise series with $d = 0.4$.[8] The spectral density for the standard white noise process has no discernible pattern across frequencies. The fractionally integrated series ($d = 0.4$), shows a clear upward trend in the spectrum as the frequency decreases. It indicates that there is a long cycle (memory) in the series.

Definition 4: Alternatively, the memory of a process can be defined by the rate of growth in the variance of partial sums of the process.(see Inoue and Diebold, 1991 [6]). Define the partial sum S_T as,

$$S_T = \sum_{t=1}^{T} x_t \tag{3.2.5}$$

The memory of x_t is defined as,

$$var(S_T) = \mathcal{O}(T^{2d+1}) \tag{3.2.6}$$

where d is the order of integration. The variance of partial sums for a short memory process increases proportional to the number of terms in the sum. In the case of the long memory process the variance of partial sums grows much faster. This definition of long memory is closely related to the rescaled range developed by Hurst and discussed in Application **??**.

[7] $\Delta^1 x_t = e_t$,e_t $iid(0, \sigma^2)$

[8] The ACF of the fractionally integrated series is shown in Figure 3.3.

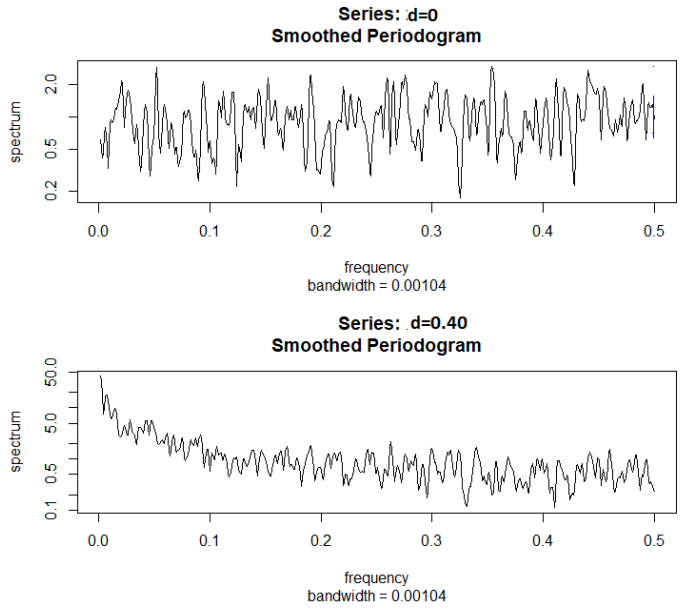

Figure 3.4: Spectral density, d=0 vs. d=0.4

3.2.2 Defining the ARIMA(0,d,0) Model

The fractional difference operator, d, is defined by a binomial expansion of Δ^d:

$$\Delta^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{n}{k} - B^k$$
$$= 1 - dB - \frac{1}{2}d(1-d)B^2 - \frac{1}{6}d(1-d)(2-d)B^3 - \ldots \quad (3.2.7)$$

Note that this is an infinite sum. In cases where d is an integer the expansion is finite.[9]

Define the ARIMA(0,d,0) as $\Delta^d = e_t$. Let x_t be an ARIMA(0,d,0) stochastic process. Hosking (1981) shows that the following hold:

1. When d <1/2, x_t is a stationary process and has an infinite order moving average representation.

$$x_t = \psi(B)\alpha_t = \sum_{k=0}^{\infty} \psi_k \alpha_{t-k} \quad (3.2.8)$$

where $\psi_k = \frac{(k+d-1)!}{k!(d-1)!}$, as $k \to \infty, \psi_k \sim \frac{k^{d-1}}{(d-1)!}$

[9]The Binomial Theorem states that:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Applying the theorem to polynomials of the lag operator $(1-L)^d$ where d is a positive integer we see that the coefficients will be zero for k >d, so the series is truncated at order d.

2. When d >-1/2, x_t is invertible and has an infinite order autoregressive representation.

$$\pi(B)x_t = \sum_{k=0}^{\infty} \pi_k x_{t-k} = \alpha_t \tag{3.2.9}$$

where $\pi_k = \frac{(k-d-1)!}{k!(-d-1)!}$, as $k \to \infty, \pi_k \sim \frac{k^{-d-1}}{(-d-1)!}$

3. The spectral density of x_t is

$$s(\omega) = 2sin(-\frac{1}{2}\omega)^{-2d} for 0 < \omega \leq \pi \tag{3.2.10}$$

as $\omega \to 0, s(\omega) \sim \omega^{-2d}$

4. The autocorrelation of x_t is

$$\rho_k \sim \frac{(-d)!}{(d-1)!} k^{-1-2d} \tag{3.2.11}$$

5. The partial correlation of x_t is

$$\phi_{kk} = \frac{d}{k-d} \tag{3.2.12}$$

6. The partial linear regression coefficients are:

$$\phi_{kj} = -\binom{n}{k} \frac{(j-d-1)!(k-d-j)!}{(-d-1)!(k-d)!} \tag{3.2.13}$$

7. For $1 \ll j \ll k$ as $j, k, \to \infty$ and $j/k \to 0$

$$\phi_{jk} \sim -j^{-d-1}/(-d-1)! \tag{3.2.14}$$

Based on the work of Hosking [14] and Granger and Joyeaux [11] , we know the following about d for the ARIMA(0,d,0) process:

- When $-1/2 > d < 1/2$, the ARIMA(0,d,0) process is invertible and stationary, and the autocorrelation function decays hyperbolically. The parameters π_k and ψ_k both decay hyperbolically, as opposed to exponentially as is the case for an AR(p,0,q) process. An ARIMA(0,d,0) process that is nonstationary can be differenced until $|d| < 1/2$. Then it will be stationary and invertible.

- When $0 < d < 1/2$ the ARIMA(0,d,0) process has long memory. The autocorrelations are all positive and decay monotonically and hyperbolically to zero as the lags increase. In this case the spectral density is concentrated at low frequencies, and $s(\omega)$ is a decreasing function of ω. $s(\omega) \to 0$ as $\omega \to 0$ but $s(\omega)$ is integrable.

- When d=0 the ARIMA(0,d,0) is white noise and the spectral density is constant.

- When $-1/2 < d < 0$ the ARIMA(0,d,0) process exhibits anti-persistence. The sum of the absolute values of ρ_k tend to be finite, so the series has short memory based on Definition 2. The correlations are negative but they decay monotonically and hyperbolically to zero as the lags increase. $s(\omega)$ is an increasing function of ω so that the spectral density is dominated by high frequency components.

- When d = 1/2, the process will be nonstationary and invertible. The spectral density is $s(\omega) = 1/(1/2sin(2\omega)) \sim \omega^{-1}$ as $\omega \to 0$ In this case the process is said to be '1/f noise'.

- When d = -1/2 the process will be stationary but not invertible.[10]

- When $1/2 < d < 1$ the process is mean reverting and nonstationary. In a sense this range provides a continuum from a purely nonstationary process where d=1, to a purely stationary process where $|d| < 1/2$.

3.2.3 ARFIMA

The ARFIMA(p,d,q) model (see [11], and [14]) is a generalized ARIMA(p,d,q) process that allows for fractional d. This model combines the ARIMA(0,d,0) model with the ARIMA(p,d,q) of Box and Jenkins while allowing d to be a real number. Let y_t be an ARFIMA(p,d,q) process,

$$\Phi(B)(1-B)^d(x_t - \mu) = \Theta(B)e_t \tag{3.2.15}$$

where d is the fractional differencing parameter. Then,

1. x_t is stationary if $d < 1/2$ and all of the roots of the equation $\Phi(B) = 0$ lie outside the unit circle.

2. x_t is invertible if $d > -1/2$ and all of the roots of the equation $\Theta(B) = 0$ lie outside the unit circle.

   If x_t is stationary and invertible with spectral density $s(\omega)$ and correlation function ρ_k. Then,

3. $\lim \omega 2^d s(\omega)$ exists as $\omega \to 0$ and is finite.

4. $\lim k^{(1-2d)}\rho_k$ exists as $k \to \infty$ and is finite.

[10] Since the series is not invertible, a forecast of the process cannot be expressed as a convergent sum of past values of the process.

3.3 Estimation

There are a number of methods available for estimating d. In this section we discuss two of the more common approaches. First, we examine the Geweke, Porter-Hudak method of estimating d alone. Then, we consider the quasi-maximum likelihood approach which estimates the complete ARFIMA model. [11]

Estimating d with GPH

The GPH method exploits the fact that a random variable of the form

$$(1-B)^d(x_t-\mu)=e_t \tag{3.3.1}$$

can be written in frequency domain as

$$f_x(\omega)=|1-e^{-i\omega}|^{-2d}f_e(\omega) \tag{3.3.2}$$

where e_t is a stationary and invertible process, and $f_x(\omega)$ and $f_e(\omega)$ are the spectral densities of x_t and e_t, respectively. This expression can be rewritten as

$$ln(f_x(\omega))=(4sin^2\frac{\omega}{2})^{-d}+ln(f_e(\omega)) \tag{3.3.3}$$

$$ln(f_x(\omega_i))=ln(f_e(0))-d*ln((4sin^2\frac{\omega_i}{2}))+ln(f_e(\omega_i)/f_e(0)) \tag{3.3.4}$$

The GPH method is a regression based on equation 3.3.4 using the spectral ordinates ω_i from the periodogram of x_t, $I(f(\omega_i))$ to find d. For $i=1,2,...M$, where $M \ll T$ the regression is defined as:

$$lnI(\omega_i)=\beta_0+\beta_1 ln(4sin^2(\frac{\omega_i}{2}))+\eta \tag{3.3.5}$$

where $d=-\beta_1$. An estimate of d allows one to filter the time series to make it stationary. Once we have a stationary series we can estimate an ARMA(p,q) model.

Statistical properties of the GPH d estimator

GPH show that

$$(\hat{d}-d)/\sqrt{var(\hat{d})} \xrightarrow{d} N(0,1) \tag{3.3.6}$$

where $var(\hat{d})$ is from the regression.

The choice of M is important since $\hat{d}$ will be biased if M is too large. M is typically chosen as a function of the sample size, T. A common rule is to set $M=T^\alpha$ for $\alpha \in [0.5, 0.8]$.[12]

Example 3.1.

[11] In the application at the end of this chapter (section 3.5.1) we describe a third approach for estimating, d, called the rescale-range method.

[12] See Agiakloglou,1992 [1] for a detailed discussion on bias in the estimator of d.

Figure 3.5 shows the average monthly yield on 3 month T-Bills for January 1934 through May 2018. The yield in January 1934 was 0.7%. It peaked in May 1981 at 16.3%, and as of May 2018 the rate was 1.9%. The Augmented Dickey -Fuller test with a maximum of 14 lags fails to reject the null hypothesis that the series is nonstationary, and suggests that first differencing is required to make it stationary. Figure 3.6 show the ACF for the yield. The autocorrelations drop off slowly suggesting that d is less than 1. The periodogram[13] for the series is shown in Figure 3.7. The spectrum is highest at low frequencies indicating the presence of long memory.

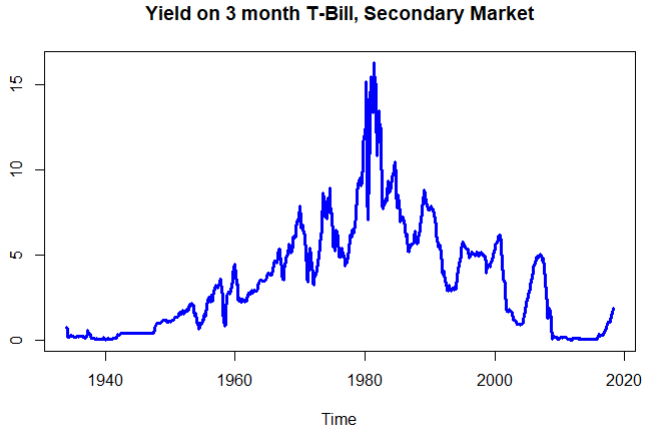

Figure 3.5: Yield on 3 month T-Bills, January 1934-May 2018

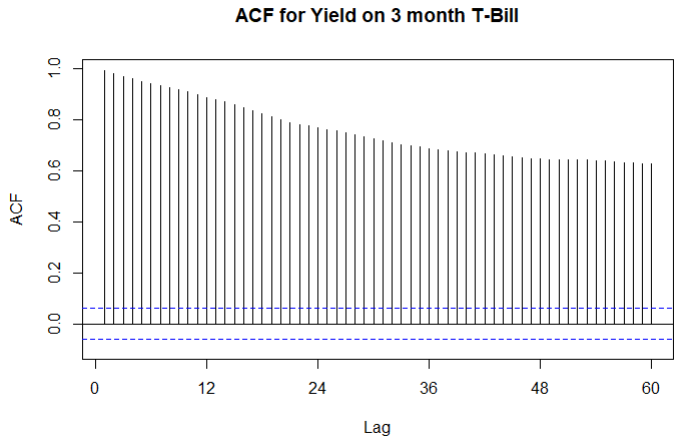

Figure 3.6: ACF of 3 month T-Bill Yield

[13]The periodogram was smoothed using a Daniell kernel with m=14. The Daniell kernel is a centered moving average whose width is 2m+1.

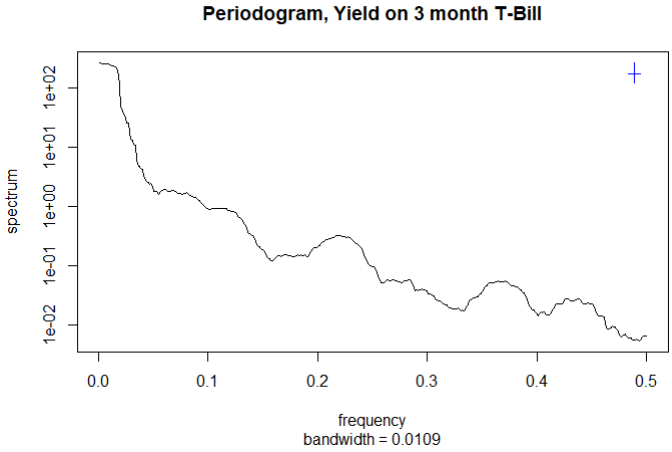

Figure 3.7: Periodogram of 3 month T-Bill Yield

The "fracdiff" library in R has a command, fdGPH, which estimates d using the GPH method.[14] The GPH regression is shown in 3.3.7.

$$lnI(\omega_i) = \beta_0 - d * ln(4sin^2(\frac{\omega_i}{2})) + \eta \tag{3.3.7}$$

The left hand side variable is an (Mx1) vector containing the elements of the periodogram. [15] The expression for the periodogram is,

$$lnI(\omega_i) = \hat{\gamma}(0) + 2\sum_{k=1}^{n-1} \hat{\gamma}(k)cos(\omega_i(k)) \tag{3.3.8}$$

where $\hat{\gamma}(0)$ is the sample variance of x_t, $\hat{\gamma}_i(k)$ is the sample autocorrelation for lag k, $\omega_i = (2\pi i)/T$, and T = number of observations in the univariate time series x_t, and $i = 1, ..., M, M < T$.

Applying the fdGPH command to the 3 month yield on T-Bills results in an estimate of $d = 0.85$ with a standard deviation of 0.12. The series is a mean reverting nonstationary series. Applying the difference filter $(1-B)^d$ to the original series where $d = 0.85$ results in the series shown in Figure 3.8. The ACF for this series is shown in Figure 3.9. The autocorrelation drops off to zero after 6 lags indicating that the series is stationary.

Applying the fdGPH command to the fractionally differenced series finds $d = -0.045$ with a standard deviation of 0.12. The differenced series exhibits a small level of anti-persistence but it is stationary. Note that this time series covers a long period of time with a number of economic and political events which could have caused structural breaks. We will revisit the series later in the chapter when we discuss the impact of structural breaks on long memory. The periodogram for the fractionally differenced series is shown in Figure 3.10.

[14]https://CRAN.R-project.org/package=fracdiff

[15]Recall that the periodogram is the sample analog of the spectral density.

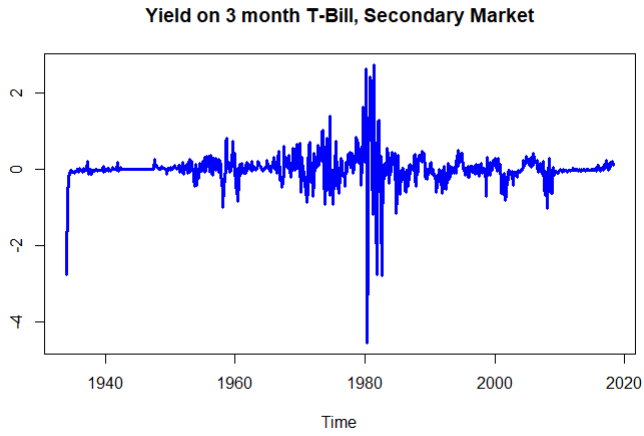

Figure 3.8: Fractionally differenced 3 month T-Bill Yield, d=0.85

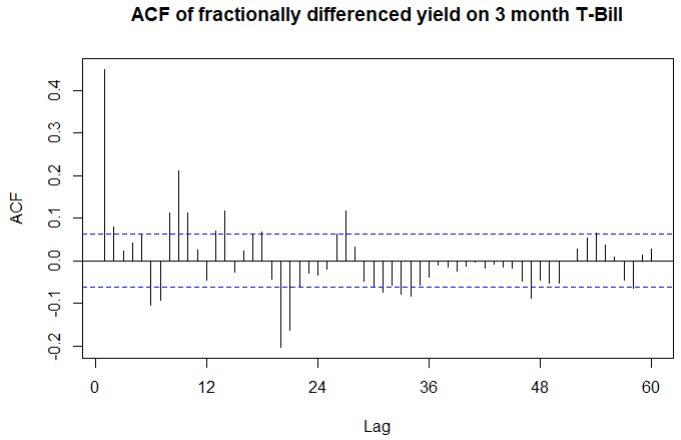

Figure 3.9: ACF for 3 month T-Bill Yield

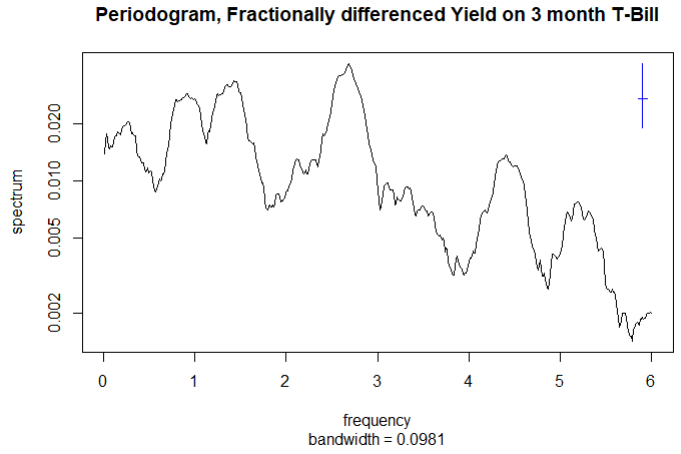

Figure 3.10: Periodogram of fractionally differenced 3 month T-Bill Yield

Exact Likelihood Estimation of an ARFIMA(p,d,q) Model

If d is estimated with the GPH spectral density regression, then estimating the ARFIMA(p,d,q) is a multi-step process. This is often referred to as a semi-parametric approach. First d is estimated, then the series is filtered, and lastly the ARMA(p,q) is estimated. Filtering involves using the estimate of d to find u_t:

$$u_t = (1 - B)^d x_t \tag{3.3.9}$$

As Sowell points out, $(1 - B)^d$ is an infinite polynomial, so u_t can only be defined when there is an infinite realization of x_t. One solution is to use the Binomial Theorem and the estimate of d to create a series, u_t that is truncated at the sample size of y_t. [16]

An alternative approach for estimating the ARFIMA, first detailed by Sowell [21], is to estimate all of the parameters simultaneously using maximum likelihood. This is typically referred to as the parametric approach. Assume that y_t is a stationary fractionally integrated time series with the properties of the ARFIMA(p,d,q) model outlined above. Therefore $0 \leq d < 0.5$.

$$\Phi(B)(1 - B)^d(x_t - \mu) = \Theta(B)e_t \tag{3.3.10}$$

Let YX be a sample of T observations, where $X \sim N(0, \Sigma)$, X is a Tx1 vector and Σ is a TxT matrix. The density of X is:

$$f(X) = (2\pi)^{-T/2}|\Sigma|^{1/2} exp^{(0.5X'\Sigma^{-1/2}X)} \tag{3.3.11}$$

Calculating the solution to this equation requires the inverse of the variance-covariance matrix, Σ. For moderate size matrices this can be done using the Cholesky decomposition. When the data set is large, Cholesky is not an efficient method. An alternative method is the Levinson-Durbin algorithm which takes advantage of the fact that the variance-covariance matrix is in Toeplitz form:[17]

$$\Sigma = [\gamma(i - j)], i, j = 1, 2, ..., T \tag{3.3.12}$$

An important step in implementing either the Cholesky or Levinson-Durbin procedure is estimating the autocovariances, which must be specified in terms of the parameters of the model. Sowell presents a closed form solution which involves transforming x_t to a spectral density, and deriving the autocovariances using:

$$\gamma(s) = \frac{1}{2\pi} \int_0^{2\pi} f_y(\omega) e^{-i\omega s} ds \tag{3.3.13}$$

The spectral density of x_t is derived in two steps. First the spectral density of $u_t = (1 - B)^d x_t$ is calculated. This is the spectral density for an ARMA(p,q).

[16] An equivalent approach proposed by GPH uses the Fourier transform. See Sowell [21] for details.

[17] The Toeplitz matrix has constants on each of the diagonals.

Then, the spectral density of x_t is calculated. This is the spectral density of an ARIMA(p,d,q) model. The autocovariance functions is written as:

$$\gamma(s) = \sigma^2 \sum_{l=-q}^{q} \sum_{j-1}^{p} \psi(l)\zeta_j C(d, p+l-s, \rho_j) \tag{3.3.14}$$

where
$C(d,h,\rho) = \frac{1}{2\pi}\int_0^{2\pi}[\frac{\rho^{2p}}{(1-\rho e^{-i\omega})} + \frac{1}{(1-\rho^{-1}e^{-i\omega})}](1-e^{-i\omega})^{-d}(1-e^{i\omega})^{-d}e^{-i\omega h}d\omega$

$\zeta_j = \rho^i \prod_{i=1}^{p}[(1-\rho_i\rho_j)\prod_{m\neq j}(\rho_j - \rho_m)]^{-1}$

$\Psi(l) = \sum_{s=max(0,l)}^{min(q,q-l)} \theta_s\theta_{s-l}$

The steps involved in the derivation help provide the intuition behind these equations. First, u_t is written in terms of the Wold decomposition:

$$u_t = \frac{\Theta(B)}{\Phi(B)}(1-B)^{-d}e_t \tag{3.3.15}$$

The roots of $\Phi(B)$ are assumed to lie outside the unit circle so we can write:

$$\Phi(x) = \prod_{j=1}^{p}(1-\rho_j x) \tag{3.3.16}$$

where, $|\rho_j| < 1$
The spectral density for the stationary ARMA(p,q) process is defined as follows:

$$f_u(\omega) = \sigma^2 \sum_{l=-q}^{q} \psi(l)\eta^l \sum_{j-1}^{p} \eta^p \zeta_j \left[\frac{\rho_j^{2p}}{(1-\rho_j\eta)} - \frac{1}{(1-\rho_j^{-1}\eta)}\right] \tag{3.3.17}$$

where $\eta = e^{-i\omega}$.
The spectral density of x_t is:

$$f_x(\omega) = (1-\eta)^d(1-\eta^{-1})^{-d}f_u(\omega) \tag{3.3.18}$$

If p=0, so that the model is ARIMA(0,d,q), the spectral density simplifies quite a bit. First, $C(d,h,\rho)$ simplifies because the term inside the braces is one. Also, the j indexed sum and the ζ_j do not have to be calculated.

Estimation of the log-likelihood involves the following steps:

1. Calculate the autoregressive polynomial.
2. Calculate the $\zeta's$
3. Calculate the different C(d,h, ρ) values. Note that d is restricted to real values less than 1/2, h can be any integer, and ρ can be any complex number in the unit circle.
4. Evaluate the covariance matrix.
5. Calculate the determinant of the inverse of the covariance matrix.
6. Evaluate the log likelihood function.

Sowell also presents an alternative approach for estimating C which is more efficient.

An Autoregressive Approximation of the ARIMA(p,d,q) Model

Since the exact maximum likelihood method is computationally intensive, several authors have suggested an autoregressive approximation. An approximation by Haslett and Rafferty [13] is available in the 'fracdiff' package.

Given the ARFIMA process for a series x_t

$$\Phi(B)(1-B)^d x_t = \Theta(B)e_t \tag{3.3.19}$$

The one step ahead forecast can be approximated as,

$$\widehat{x_t} = \Theta(B)\Phi(B)^{-1}\sum_{j=1}^{t-1}\phi_{tj}x_{t-j} \tag{3.3.20}$$

The variance of the forecast error of y_t is:

$$v_t = var(x_t - \widehat{x_t}) = \sigma_y^2\kappa\prod_{j=1}^{t-1}(1-\phi_{jj}^2) \tag{3.3.21}$$

where σ_y^2 is the $var(x_t)$, κ is the ratio of innovations variance to the variance of the ARMA(p,q) process defined by $\Theta(B)$ and $\Phi(B)$, and ϕ_{tj} are the partial regression coefficients defined by Hoskings (see Item 6). To eliminate the need to calculate a large number of ϕ_{ij} the following approximation is used:

$$\sum_{j=1}^{t-1}\phi_{tj}x_{t-j} \approx \sum_{j=1}^{m}\phi_{tj}x_{t-j} - \sum_{j=M+1}^{t-1}\pi_{tj}x_{t-j} \tag{3.3.22}$$

where π is the $\pi(B)$ used by Hoskings (see item 2). Note that $-\pi$ is used as an approximation of ϕ for large j. This is based on the work of Hoskings, and can be seen by comparing the results in item 2 with those of item 7 in the discussion of the ARIMA(0,d,0) model. Haslett and Rafferty make one additional simplification by assuming that the π_j terms are constant for $j \geq M+1$, and replace the individual terms with an average to get the following expression:

$$\sum_{j=M+1}^{t-1}\pi_{tj}x_{t-j} \approx M\pi_M d^{-1}[1-(\frac{M}{t})^d]\bar{x}_{M+1,t-1-M}) \tag{3.3.23}$$

where, $\bar{x}_{M+1,t-1-M} = \frac{1}{t-1-2M}\sum_{j=M+1}^{t-1-M}x_j$

The quasi-maximum likelihood function is:

$$lnL = constant - \frac{1}{2}nlog[\widehat{\sigma}_e^2(\theta)] \tag{3.3.24}$$

where, $\widehat{\sigma}_e^2(\theta) = \frac{1}{n}\sum_{t=1}^{n}\frac{(x_t-\widehat{x}_t)^2}{v_t}$

Example 3.2.

In this example we simulate a stationary series with d = 0.42, $\theta = 0.5$, and, $\phi = -0.5$, and apply both the exact likelihood method of Sowell, and the quasi-maximum likelihood method of Haskins and Rafferty.

The ACF is shown in Figure 3.11. It drops off much slower than a series with d=0. The smoothed periodogram is shown in Figure 3.12. Note the presence of the low frequency spectrum even though the series is stationary. Table 3.1 shows the results based on the exact maximum likelihood method. The estimate of $\widehat{d} = 0.37$, compared with the actual value of 0.42. Table 3.2 shows the results of the quasi-maximum likelihood estimator. The estimate of d=0.38. The two estimation methods give relatively close parameter estimates.

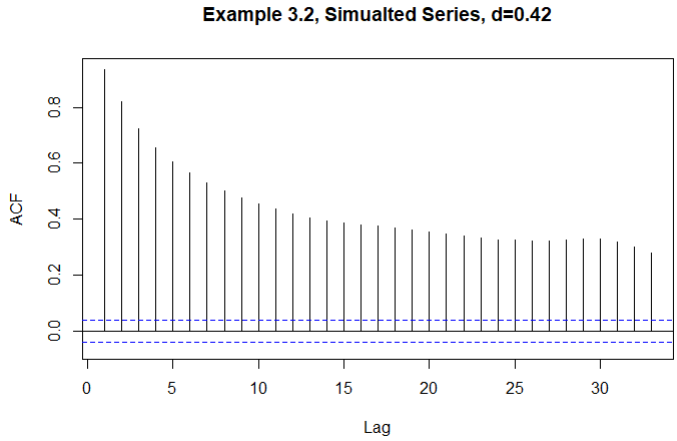

Figure 3.11: ACF for Simulated Series

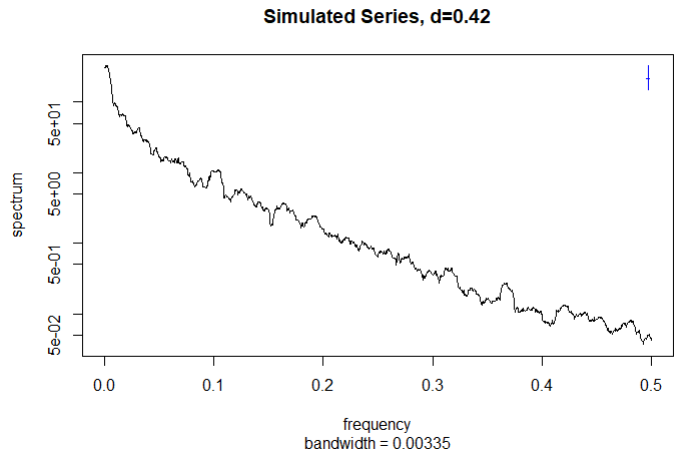

Figure 3.12: ACF for Simulated Series

	Coef.	SE.
phi(1)	0.547384	0.046664
theta(1)	-0.51774	0.019745
d	0.370546	0.043086
Fitted mean	5.23071	1.52687
logl	-14.999	
sigma$\hat{2}$	1.01178	

Table 3.1: Exact Maximum Likelihood Estimates

	Estimate	Std. Error	z value	Pr($>\|z\|$)
d	0.38462	0.01468	26.19	2.00E-16
ar	0.53558	0.0255	21	2.00E-16
ma	-0.51697	0.02024	-25.54	2.00E-16

Table 3.2: Quasi-Maximum Likelihood Estimates

3.3.1 Testing the order of integration

The FD-F test, developed by Dolado, Gonzalo & Mayoral (DGM) [9] is a simple test based on the Dickey-Fuller test. It evaluates the null hypothesis that a series is $I(d_0)$ against the alternative that it is $I(d_1)$, where d_0 and d_1 are real numbers.

The testing regression is:

$$\Delta^{d_0} x_t = \phi \Delta^{d_1} x_{t-1} = e_t \tag{3.3.25}$$

where, $a_t \sim I(0)$

The test is based upon the t-statistic of the coefficient estimate of ϕ. The null and alternative hypothesis are:

$$H_0 : \phi = 0 \; x_t \text{ is } FI(d_0)$$
$$H_1 : \phi < 0 \; x_t \text{ is } FI(d_1)$$

$\Delta x_t^{d_0}$ and $\Delta x_t^{d_1}$ are differenced according the null and alternative hypotheses, respectively.

When $\phi = 0$, the series is fractional white noise, defined as $\Delta x_t^{d_0} = e_t$, implying that x_t is $FI(d_0)$ under the null hypothesis.

When $\phi < 0$, x_t is an $FI(d_1)$ process which can be expressed as

$$\Delta^{d_1} = C(B)e_t \text{ where, } C(B) = ((1-B)^{d_0-d_1} - \phi)^{-1} \tag{3.3.26}$$

All of the roots of the polynomial are outside the unit circle if $-2^{(1-d_1)} < \phi < 0$.

When $d_0 = 1$ and $d_1 = 0$ this test is the standard DF test of $I(1)$ vs. $I(0)$. The test of interest is $d_0 = 1$ vs. $0 \leq d_1 \leq 1$. A test that x_t is a random walk against an alternative that x_t is a mean reverting stationary series. [18]

[18] As DGM point out the test can be applied to any $H_0 : FI(d_0)$ vs. $H_1 : FI(d_1)$ for $d_0 > d_1$.

DGM show that the asymptotic properties of the test are determined by how far the alternative hypothesis is from the null hypothesis. Given the nonstationarity assumed under the null ($d_0 = 1$), if the alternative is nonstationary ($0.5 \leq d_1 < 1$) the test is standard Normal. If the alternative is stationary ($0 \leq d_1 < 0.5$) the limit distribution of the test statistic is non-standard.
The FD-F test is standard Normal if either of the following is true:

1. The process is nonstationary under the null ($d_0 > 0.5$) and ($d_1 - d_0 < 0.5$)

2. The processes are stationary under both hypotheses.

Otherwise the distribution is non-standard. DGM provide significance tables for non-standard cases.

The DGM test is a Wald-type test derived under the assumption that d_1 is known, but as they show, any $T^{1/2}$-consistent estimator of d_1 can be used.

Example 3.3.

In this example we estimate d for a daily index of cumulative excess market returns for the period July 1, 1926 - December 30, 2017. The data, which is provided by Kenneth French, consists of the returns for all companies listed on the New York, American and NASDAQ stock exchanges. French reports daily returns in excess of the yield on 1 month T-Bills. We used this data to calculate an index of cumulative returns. The index is shown in Figure 3.13. Applying GPH, the estimate of d is found to be 0.9833 with a standard deviation of 0.022. This result suggests that the index is nonstationary and mean reverting. We can test the random walk hypothesis against an alternative of d_1 using the FD-F test. The results of the FD-F test are shown in Table 3.3. The test suggests that we can reject the null hypothesis that the index is a random walk.

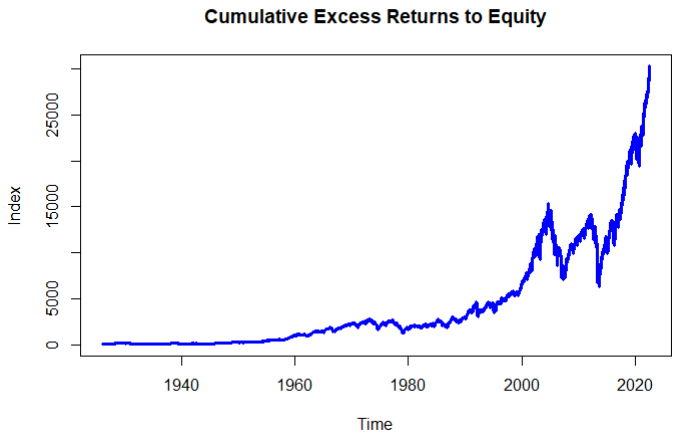

Figure 3.13: Cumulative Excess Returns on Equities

Coefficients:

	Estimate	Std. Error	t-Stat	p-value
Intercept	2.15E-04	6.83E-05	3.142	0.00168
$est.\phi$	8.62E-02	6.40E-03	13.455	2.00E-16

Table 3.3: Regression Results for FD-F Test

3.4 Determinants of Long Memory

The origins of long memory in financial and economic data has always been difficult to rationalize. Granger [10] explained the presence of long memory as a property of aggregation. Consider the aggregation of $i = 1, ..., N$ cross-section $AR(1)$ processes,

$$x_{it} = \alpha_i x_{i,t-1} + e_{it} \tag{3.4.1}$$

where, e_{it} is noise, e_{it} is orthogonal to e_{jt}, and a_i is orthogonal to e_{jt} for all i,j, and t. Granger showed that the τ^{th} autocovariance of the sum, $x_t = \sum_{i=1}^{N} x_{it}$ is,

$$\gamma(\tau)_x = A\tau^{1-q} \tag{3.4.2}$$

and, $x_t \sim I(1 - q/2)$
Diebold and Inoue [7] show that regime switching and long memory can easily be confused.

3.5 Applications

3.5.1 Long-Term Memory in Stock Market Prices

Andrew Lo [18] tests for long range dependence in equity prices using a variation of the rescaled-range test (R/S) developed by Hurst [15]. Hurst a hydrologist, studied the long term storage capacity of reservoirs. He is best known for his study of the ebbs and flows of the Nile River. Hurst developed the R/S statistic as a measure of the long range dependence that he observed in his studies of the Nile and other rivers.

Given a time series of asset returns, $x_1, x_2, \ldots, x_n$, the R/S statistic is defined as,

$$R_n = \frac{1}{s_n}\left[\max_{1\le k\le n}\sum_{j=1}^{k}(x_j - \bar{x_n}) - \min_{1\le k\le n}\sum_{j=1}^{k}(x_j - \bar{x_n})\right] \tag{3.5.1}$$

where,

$$s_n = \left[\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x_n})^2\right]^{1/2} \tag{3.5.2}$$

Calculating the RS Statistic

To calculate R_0:

1. Calculate the sample mean, m, for the series.
2. Create a mean adjusted series, $y_t = x_t - m$ for $t = 1, 2, \ldots, n$.
3. Calculate the cumulative deviation series,

$$z_t = \sum_{t=1}^{t} y_1 \text{ for } t = 1, 2, \ldots, n \tag{3.5.3}$$

4. Create a range series:

$$R_t = max(z_1, z_2, \ldots, z_t) - min(z_1, z_2, \ldots, z_t) t = 1, 2, \ldots, n \tag{3.5.4}$$

5. Create a standard deviation series:

$$S_t = \frac{1}{T}\left[\sum_{t=1}^{t}(x_t - m)^2\right]^{0.5} \tag{3.5.5}$$

6. Calculate the rescaled range series:

$$(R/S)_t = R_t/S_t \text{ for } t = 1, 2, \ldots, n \tag{3.5.6}$$

The R/S statistic is sensitive to short-range influences, so it is possible to conclude that there is long range dependence in a series when in fact it is actually a symptom of short-run autocorrelation. As a result, early studies of equity returns applied the R/S test and concluded that long-memory existed! Lo(1991) modified the test to include short run dependence in the statistic, and showed that the earlier equity studies were incorrect. He was unable to find long memory in equity returns. The modified test by Lo corrects for short term dependence:

$$R_q = \frac{1}{s_q}\left[\max_{1\le k\le n}\sum_{j=1}^{k}(x_j - \bar{x_n}) - \min_{1\le k\le n}\sum_{j=1}^{k}(x_j - \bar{x_n})\right] \tag{3.5.7}$$

where,

$$s_q^2 = s_n^2\left(1 + \frac{2}{T}\sum_{j=1}^{2} w_{qj}\rho_j)^2\right) \quad w_{qj} = 1 - \frac{j}{q+1} \quad q < T \tag{3.5.8}$$

ρ_j for $j = 1, \ldots, q$ are the sample autocorrelation for x_t.

There are problems with the asymptotics of Lo's version of the RS statistic when x_t is fat tailed. This led Lo to conclude that the test should only be used as a preliminary test, similar to a Portmanteau. That is, it should be used to complement a more comprehensive analysis of long term memory.

Estimating the Hurst Exponent

The Hurst exponent (H) is used to measure long memory dependence in time series. It is related to the R/S statistic as follows:

$$E(R/S)_t = c * t^H \tag{3.5.9}$$

The H coefficient is the slope in the log-log transformation:

$$ln(R/S)_t = c + H * ln(t) \tag{3.5.10}$$

How does H relate to d?

$$d = H - 0.5 \tag{3.5.11}$$

- $H = 0.5$ indicates a random (white noise) series.
- A value in the range $0.5 < H < 1$ indicates a time series with long term positive autocorrelation. The series is trending.
- A value in the range 0¡H¡0.5 indicates a series with long term switching between high and low values in adjacent pairs (negative autocorrelation).

3.5.2 A Nonlinear Long Memory Model for US Unemployment

In their 2002 paper, van Dijk, et.al.[8] model the unemployment rate in the US as a fractionally integrated smooth transition autoregression (FI-STAR). They chose this approach to capture two key features of the unemployment rate: persistence, and asymmetry. The asymmetry is apparent in Figure 3.14 which contains monthly unemployment rates from January 1948 through December 2018. The slope of the unemployment rate series during recessions is much steeper than the slope during economic expansions. The unemployment rate tends to increase quickly and decrease slowly. One possible explanation for this is that labor is a quasi-fixed factor of production. Employers are hesitant to layoff workers until they are absolutely certain that they must. The reason for the hesitance may be that employers have invested time and money in workers and they do not want to lose their investment. More generally, the asymmetry may be the result of asymmetric costs associated with firing and hiring workers.

Figure 3.15 illustrates the high degree of persistence in the unemployment rate data. The ACF is statistically different from zero for lags up to 50 months. The decline in the ACF is much slower than one would expect from an $I(0)$ stationary series.[19] The shape is much more similar to the hyperbolic decline that characterizes a fractionally integrated series.

[19]The ACF for an I(0) stationary series typically declines exponentially.

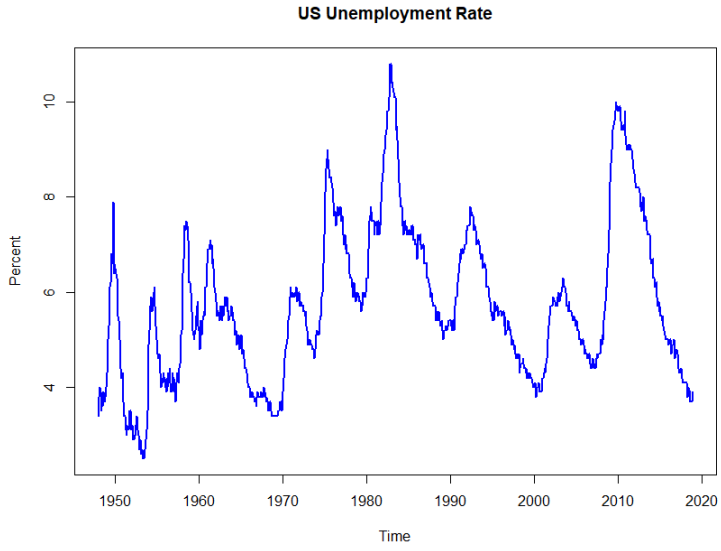

Figure 3.14: US Unemployment Rate

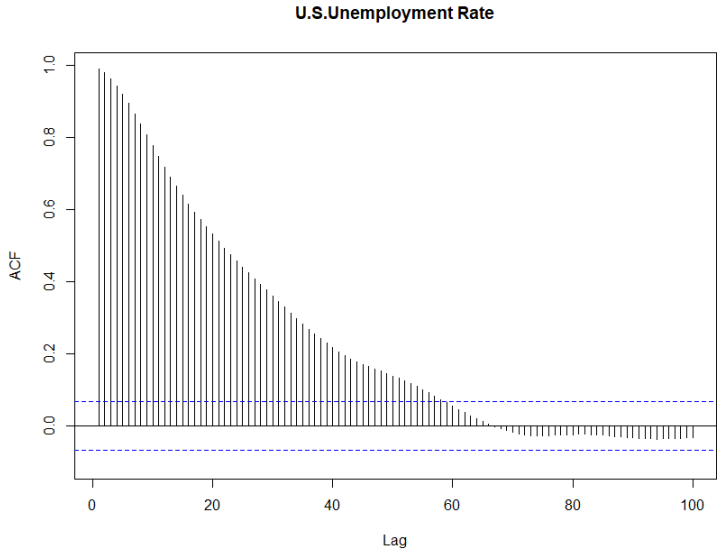

Figure 3.15: ACF - US Unemployment Rate

The FI-STAR functional form of van Dijk, et al. is,

$$\begin{aligned} x_t &= (\phi_{1,0} + \phi_{1,1}x_{t-1} + \cdots + \phi_{1,p}x_{t-p})(1 - G(s_t, \gamma, c)) \\ &+ (\phi_{2,0} + \phi_{2,1}x_{t-1} + \cdots + \phi_{2,p}x_{t-p})G(s_t, \gamma, c) + e_t \end{aligned} \tag{3.5.12}$$

where, $\gamma > 0$, e_t is white noise, and G is a logistic function,

$$G(s_t, \gamma, c) = (1 + exp(-\gamma(s_t - c)/\sigma_{s_t}))^{-1} \tag{3.5.13}$$

y_t is the observed series, and x_t is a fractionally differenced series.

$$(1-L)^d y_t = x_t \tag{3.5.14}$$

We know from our earlier discussion, that y_t is covariance stationary if $0 < d < 0.5$, and it is invertible if $d > -0.5$. As the transition variable, s_t, increases, $G(s_t, \gamma, c)$ changes monotonically from zero to one. σ_{s_t} is the standard deviation of s_t.

The smoothness of the change in G is determined by the slope parameter of the logistic function, γ. When $\gamma = 0$, $G(s_t, \gamma, c) = 0.5$ for all s_t. In this case the model parameters are fixed across all states. When $\gamma \to \infty$, $G(s_t, \gamma, c)$ is an indicator function, and the change in parameters from one state to the other is immediate.

The FI-STAR model has a long term, and a short term component. Equation 3.5.12 is the short term component. The effective AR parameters in this equation change with the state. The weights of the two sets of parameters are determined by the logistic transition equation.

Equation 3.5.14 is the long term component. It is constant in the sense that the transform from y_t to x_t is determined by the fixed parameter, d.

We estimate the model using two approaches. In the first approach we estimate the model in three steps: 1) estimate the difference parameter, d; 2) estimate the parameters for the transition function, $G(\gamma, c)$, and 3) estimate the AR parameters for the differenced series. In the second approach we estimate all of the model parameters simultaneously.

Two Step Estimation

We begin our analysis by testing for non-stationarity. An augmented Dickey-Fuller test with 9 lags rejected the null hypothesis that the series is non-stationary. Next, we estimate the difference parameter using the "hurstexp" command in the pracma library. The simple rescaled range method in the "hurstexp" command estimates that $H = 0.803$, so the difference operator $d = 0.303$. Since $d < 0.5$, the series is stationary.

The differenced unemployment rate is displayed in Figure 3.16.[20]. The asymmetry present in the original series is also present in the differenced series.

[20]The series was differenced using the "diffseries" command in the "fracdiff" library

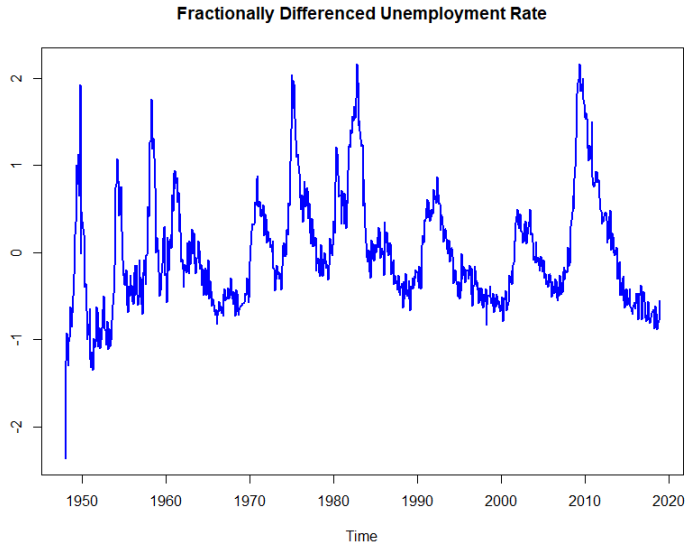

Figure 3.16: Fractionally Differenced US Unemployment Rate, d = 0.3

As Dijk, et.al. note, estimation of the AR parameters is linear given the parameters d, γ, c. This suggests a grid search where we select d, γ, c and minimize Q in 3.5.15, below. The AR parameters are estimated by minimizing the following function:

$$Q = \sum_{t}^{T} \left(x_t - \hat{\phi}_1(d,\gamma,c) G_t(d,\gamma,c) \; - \hat{\phi}_2(d,\gamma,c)(1 - G_t(d,\gamma,c) \right)^2 \qquad (3.5.15)$$

where $\hat{\phi}_i, i = 1, 2$ denote the two AR functions in equation 3.5.12.

Figure 3.17 shows the grid search results when $d = 0.43$. There are quite a few values of γ and c that have an SSR between 30 and 35. The minimum SSR for this grid search was 30.6. with $\gamma = 16$ and $c = 2$. Since the value for C is at the edge of the search grid, we the grid was extend along the c axis. The results are shown in Figure 3.18.

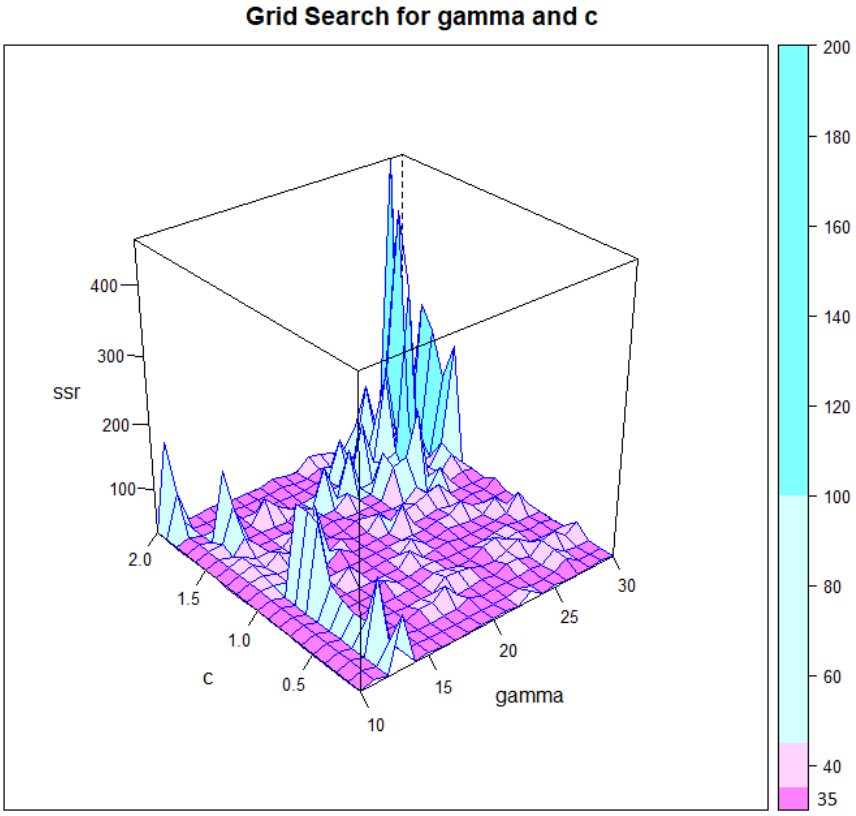

Figure 3.17: Sum of Square Residuals - Grid Search, d=0.3

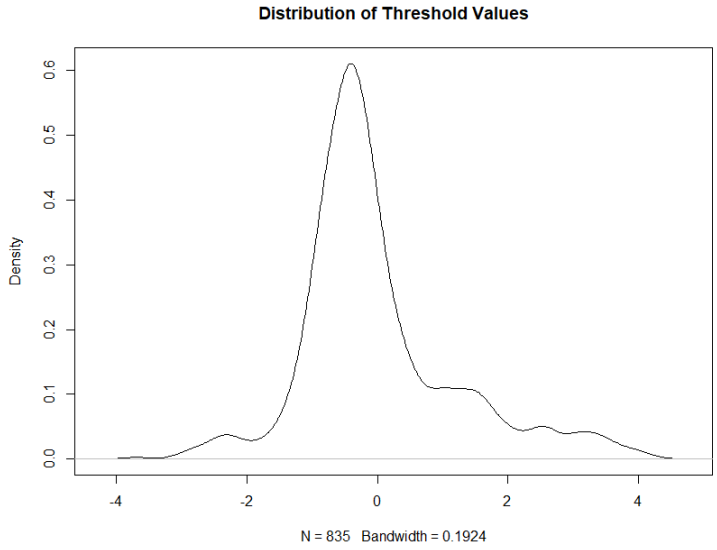

Figure 3.18: Distribution of Threshold Values

The AR parameters for the two regimes can be estimate using the 'lstar' command in the 'tsDyn' package. The user provides the differenced series (x_t), the

number of lags for the AR processes, the state variable (s_t), and lag for s_t. [21]

[21] R help describes additional options available to the end user.

3.6 Exercises

1. Derive the ACF for a random walk.

3.7 R Code for Examples

Example 3.1

```
library(fracdiff)\\
fdGPH(x,bandw.exp = 0.5)\\
k1 $\leftarrow$ kernel("daniell",m=14)\\
spectrum(x,k1)
```

where x is the series being differenced, and *bandw.exp* is the bandwidth parameter. The default value for the bandwidth parameter is 0.5. The bandwidth is used to set the width of the frequency interval. It is defined as $M = trunc(T^{band.w)})$. If the bandwidth is set too wide the periodogram may smooth out important peaks in the spectral density. If the bandwidth is too narrow the periodogram will be very noisy.

R code for example 3.2

```
library(fracdiff)
library(forecast)
library(arfima)
sim1 = fracdiff.sim(n=2500, ar=0.5,ma= -0.5, d= 0.42)
Acf(as.ts(sim1\$series))
k1 = kernel("daniell",m=14)
spectrum(sim1\$series,k1)
arfima(sim1\$series,order=c(1,0,1),back=T) \textbf{(exact ML)}
fracdiff(sim1\$series,nar=1,nma=1) \textbf{(quasi ML)}
```

R code for example 3.3

```
lequity = log(equity)
dlequity = diff(lequity)
out = fdGPH(lequity)
flequity = diffseries(lequity,out\$d)
lagfl = lag(flequity)
alldat = ts.intersect(dlequity,lagfl,dframe=TRUE)
summary(lm(alldat[,1]~alldat[,2], na.action=NULL))
```

FI STAR Example

```
fstar<-function(p){
x<-rinput
lag<-4
N<-nrow(x)-lag
H<-p[1]
gamma<-p[2]
c<-p[3]
a0<-p[4]
```

```
a1<-p[5]
a2<-p[6]
a3<-p[7]
a4<-p[8]
b0<-p[4]
b1<-p[5]
b2<-p[6]
b3<-p[7]
b4<-p[8]
dunrate<-diffseries(x[,2], d=H-0.5)
res<-matrix(data=NA,nrow=N,ncol=1)
ssr<-matrix(data=NA,nrow=N,ncol=1)

j<-1
for( i in 5:835) {
G<-exp(-gamma*(x[i,3]-c)/rinput[i,4])
up<-a0+a1*dunrate[i-1]+a2*dunrate[i-2]+a3*dunrate[i-3]+a4*dunrate[i-4]
dn<-b0+b1*dunrate[i-1]+b2*dunrate[i-2]+b3*dunrate[i-3]+b4*dunrate[i-4]
res[j]<-x[i,2]-(up*G+dn*(1-G))
ssr[j]<-res[j]*res[j]
j<-j+1
}
return(sum(ssr))
}
gridvec$out<-rep(NA)
#simple grid search
for(j in 1:420){
p<-c(0.93,gridvec[j,2],gridvec[j,3],0,0.5,0.3,0.2,0.1,0,0.5,0.3,0.2,0.1)
outstar<-optim(par=p,fstar,method=c("Nelder-Mead"))
print(outstar$value)
gridvec[j,4]<-outstar$value
print(j)
}
library(lattice)
gamma<-gridvec[,2]
c<-gridvec[,3]
ssr<-gridvec[,4]
wireframe(ssr~gamma*c,drape=TRUE,col="Blue",scales=list(arrows=FALSE),at=c(25,40,10
```

Bibliography

[1] Agiakloglou,C., P. Newbold, and M. Wohar. "Bias in an estimator of the fractional difference parameter", Journal of Time Series Analysis, 14:235-246, 1992.

[2] Agiakloglou,C., P. Newbold. "Lagrange multiplier tests for fractional difference", Journal of Time Series Analysis, 15:253-262, 1994.

[3] Baillie, Richard T. "Long memory processes and fractional integration in econometrics", Journal of Econometrics, 5 1996.

[4] Bollerslev, Tim, Hans Ole Mikkelsen, "Modeling and pricing long memory in stock market volatility", Journal of Econometrics, 73 (1996) 151-184.

[5] Chan, Ngai Hang and Wilfredo Palma. "Estimation of Long Memory Time Series Models:A Survey of Different Likelihood-Based Methods", Econometric Analysis of Financial and Economic Time Series/Part Advances in Econometrics, Volume 20, 89–121.

[6] Diebold, Francis X., and Glen Rudebusch."On the power of Dickey-Fuller tests against fractional alternatives", Economic Letters, 35 (1991) 155-160.

[7] Diebold, Francis X., and Atsushi Inoue."Long memory and regime switching", Journal of Econometrics, 105 (201) 131-159.

[8] van Dijk, Dick, Phillip Hans Frances, and Richard Paap, "A Nonlinear Long Memory Model for US Unemployment", Journal of Econometrics, 110 2002, 135-165.

[9] Dolado, JJ. J. Gonzalo, and L. Mayoral. "A Fractional Dickey-Fuller Test for Unit Roots", Econometrica 70:1963-2006, 2002.

[10] Granger, C.W.J., 1980. "Long-memory relationships and the aggregation of dynamic models". Journal of Econometrics 14, 227–238.

[11] Granger, C.W.J. and Roselyne Joyeux. "An Introduction to Long Memory Time Series Models and Fractional Differencing", Journal of Time Series Analysis, Vol.1, No. 1, 1980.

[12] Granger, C.W.J., and Namwon Hyung, "Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns", Journal of Empirical Finance, 11 (2004) 399-421.

[13] J. Haslett and A. E. Raftery (1989) Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource (with Discussion); Applied Statistics 38, 1–50.

[14] Hosking, J.R.M., "Fractional Differencing", Biometrika Vol. 68 No.1 pp. (April 1981) 165-176.

[15] Hurst, H.E., "Long term storage capacity of reservoirs", Trans. Am. Soc. Eng. 116: 770-799.

[16] Ito, Ryoko."Long Memory and Fractional Differencing: Revisiting Clive W. J. Granger's Contributions and Further Developments", European Journal of Pure and Applied Mathematics, Vol. 10, No. 1, 2017, 82-103.

[17] Jensen, Mark. "Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter", Journal of Forecasting 18, 17-32 (1999).

[18] Lo, Andrew W., "Long-Term Memory in Stock Market Prices", Econometrica, Vol. 59, No. 5 (Sep., 1991), pp. 1279-1313.

[19] Rosenblatt, M., "A central limit theorem and a strong mixing condition", Proceedings of the National Academy of Sciences (1956) 42, 43-47.

[20] Samorodnitsky, Gennady, "Long Range Dependence", Foundations and Trends in Stochastic Systems, Vol.1 No. 3 (2006), 163-257.

[21] Sowell, Fallaw. "Maximum likelihood estimation of stationary univariate fractionally integrated time series models", Journal of Econometrics 53 (1992) 165-188.,

[22] Whitcher, Brandon and Mark Jensen. "Wavelet of a local long memory parameter', Exploration Geophysics (2000) 31, 094-103.

Chapter 4

Dynamic Linear Models

In this lecture we discuss the estimation of dynamic linear models, also known as linear state space models.[1] We begin with a discussion of the Kalman Filter. As you will see it is quite natural to describe the Kalman Filter in a Bayesian framework. Following the discussion of the Kalman filter we will discuss a more general, Bayesian approach to estimating dynamic linear models. We end the lecture with a review of several applications in finance.

4.1 Introduction

The dynamic linear model consists of two equations, the state equation, and the measurement equation. The state equation (4.1.1) determines the state of nature, β_t, at each point in time. The state of nature is a Markov process in the sense that $(\beta_t|\beta_{t-1}, ..., \beta_0 = \beta_t|\beta_{t-1})$ The state vector is unobserved. We will assume that G_t is known and does not vary over time, and so throughout the lecture we will refer to it simple as G. In most applications of interest, G is assumed to be the identity matrix.[2] The disturbance term, w_t, is Normal iid noise.

$$\beta_t = G_t\beta_{t-1} + w_t, w_t \sim N(0, W_t) \tag{4.1.1}$$

Note that if G is the identity matrix, then the state of nature follows a random walk.

The measurement equation (4.1.2) contains two sets of observed values, y_t, $t = 1, 2, ..., T$ which is the observed value of interest, and F_t which is a txk matrix of explanatory variables. y_t, may be a univariate or multivariate.

$$y_t = F_t\beta_t + v_t, v_t \sim N(0, V_t) \tag{4.1.2}$$

In this discussion, unless stated otherwise, we will assume that y_t is univariate. F_t may vary over time. The relationship between F_t and β_t is linear. The disturbance term e_t is Normal iid noise. The two disturbance terms in the system, e_t and

[1] Readers interested in additional reading on state space models should see Durbin and Koopman [7] which is an excellent book on the topic.

[2] The 'dlm' package in R allows G to vary over time.

w_t, are independent. We are primarily interested in a method for estimating the unobserved state of nature, β_t, which changes over time.

In the early 1960's, Kalman [10], and Kalman and Bucy [11] introduced an optimal method (know as the Kalman Filter) for estimating the state space model (equations 4.1.2 and 4.1.1). The Kalman filter was originally introduced in the engineering literature as a method for filtering noisy data to identify and predict an underlying signal. It was subsequently adopted in the statistics and econometrics [3] literature as an alternative to standard regression analysis which assumes that the state of nature, or regression coefficients, are fixed over time.

The Kalman filter is a sequential forecasting and updating procedure which lends itself naturally to a Bayesian interpretation. In each time period, the filter is used to update the estimate of the state of nature, and to forecast the value of next periods observed variable, y_{t+1}. We will also discuss smoothing in this lecture. Smoothing is essentially a retrospective adjustment to the Kalman Filter estimates of the state parameters.

It can be useful to think of the Kalman filter, and the smoothing process, as providing estimates of the state parameter conditional on the available observed data $y_1, ..., y_T$. Let the estimate as $\widehat{\beta}_{k|T} = \widehat{\beta}_{k|T}(y_1, ..., y_T)$. Then, we can define the following

1. Prediction $k \geq T$
2. Filtering $k = T$
3. Smoothing: $k \leq T$

4.1.1 Some useful model specifications

Dynamic linear models (DLM) constitute a large class of models with time varying parameters. Two useful models are the 1st order DLM, and the 2nd order (DLM). [4] The first order DLM, also called the local level model is,

Local Level Model

$$y_t = \beta_t + v_t, v_t \sim N(0, V_t) \tag{4.1.3}$$
$$\beta_t = \beta_{t-1} + w_t, w_t \sim N(0, W_t) \tag{4.1.4}$$

where v_t and w_t are independent. This is a state space model where F=G=1.The observed variables y_t, is the sum of a trend component (β) and a noise term. The trend component follows a random walk. The level parameter β_t is a locally weighted mean. The local level model is first difference stationary,

$$\Delta y_t = \Delta\beta_t + \Delta v_t \tag{4.1.5}$$
$$\Delta y_t = w_{t-1} + v_t - v_{t-1} \tag{4.1.6}$$

[3] Duncan and Horn [6] and, Meinhold and Singpurwalla [14] are two early examples of statisticians describing the work of Kalman and Bucy.

[4] See West and Harrison, [21] for a detailed discussion of these two models, and dynamic linear models in general

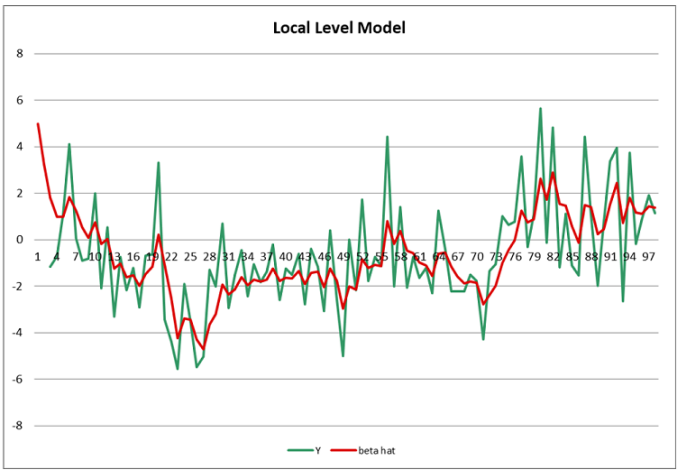

Figure 4.1: Local Level model

An illustration of the local level model is shown in Figure 4.1. The green line is simulated data representing y_t. The red line is an estimate of the local level parameter β_t. The estimate of β_t is much less volatile than y_t, and tends to go lie in the center of the y_t values. The model was estimated using the Kalman filter. Example 4.1 discusses estimation of the local level model using the 'dlm' package in R.

Linear Growth Model

The second order DLM, or the linear growth model, has the following specification,

$$y_t = \beta_{1,t} + v_t, v_t \sim N(0, V_t) \tag{4.1.7}$$
$$\beta_{1,t} = \beta_{1,t-1} + \beta_{2,t} + + w_{1,t} \tag{4.1.8}$$
$$\beta_{2,t} = \beta_{2,t-1} + w_{2,t}, \tag{4.1.9}$$
$$(w_{1t,}, w_{2,t}) \sim N(0, W_t) \tag{4.1.10}$$

where $\beta_{2,t}$ represents the growth of the level of the series. In state space form $F_t = \begin{bmatrix} 1 & 0 \end{bmatrix}'$ and $G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$

The ARMA Model in State Space Form

In addition to the 1^{st} and 2^{nd} order polynomial models, a number of common models can be expressed in state space form, including the AR(p), MA(q), ARMA(p,q), and multivariate regression models. In this section we discuss the equivalence of the ARMA(p.q) and the state space model. While showing equivalence between the ARMA and the state space model is somewhat involved, it is worth going through the exercise as it provides a nice foundation for understanding the remainder of the lecture. Later on in the lecture we will discuss estimation of the multivariate regression model in state space form allowing for time varying regression parameters.

Several approaches exist for describing the ARMA model in state space form. We follow the approach of Hamilton [9]. Consider the following AR(1) model,

$$y_{t+1} = \phi y_t + e_{t+1} \tag{4.1.11}$$

where $e_t$is iid, $N(0, \sigma^2)$. Using recursive substitution we can show that

$$y_{t+m} = \phi^m y_t + \phi^{m-1} e_{t+1} + \phi^{m-2} e_{t+2} + ... + \phi^1 e_{t+m-1} + e_{t+m} \tag{4.1.12}$$

for m= 1,2,..., The optimal m step ahead forecast of y_t is,

$$E[y^m | y^{m-1}, y^{m-2}, ...] = \phi^m y_t \tag{4.1.13}$$

The process is stable if $|\phi| < 1$.

The measurement equation is

$$Y_t = F'\beta + v_t \tag{4.1.14}$$

where $y_t \sim nx1$, F' is an nxr matrix of coefficients, and $w_t \sim N(0, W)$ is is measurement error.[5] Future values of y_t are determined by the state variable,

$$y_{t+m} = F'G^m \beta_t \tag{4.1.15}$$

The state equation is,

$$\beta_{t+1} = G\beta_t + w_{t+1} \tag{4.1.16}$$

where $G \sim rxr$ matrix, $w_t \sim rx1$, and $w_t \sim$N(0,W_t) are iid.

We can also re-write the state equation using recursive substitution,

$$\beta_{t+m} = G^m \beta_{t+1} + G^{m-1} w_{t+1} + G^{m-2} w_{t+2} + ... + G^1 w_{t+m-1} + w_{t+m} \tag{4.1.17}$$

for m = 1,2,..., and G^m is G multiplied by itself m times. Then,

$$E[\beta_{t+m} | \beta_t, \beta_{t-1}, ...] = G^m \beta_t \tag{4.1.18}$$

Future values of the state vector only depend on past values through the current value β_t. The system is stable if all of the eigenvalues of G lie inside the unit circle. The state space representation of the ARMA model captures the dynamics of the measurement equation through the state equation.

We will now illustrate the equivalence between the ARMA(p,q) and the state space model by writing a p^{th} order autoregression in state space form.

$$(y_{t+1} - \mu) = \phi_1 (y_t - \mu) + \phi_2 (y_{t-1} - \mu) + ... + \phi_p (y_{t-p+1} - \mu) + e_{t+1} \tag{4.1.19}$$

This can also be written as,

$$\begin{bmatrix} (y_{t+1} - \mu) \\ (y_t - \mu) \\ \vdots \\ (y_{t-p+2} - \mu) \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} (y_t - \mu) \\ (y_{t-1} - \mu) \\ \vdots \\ (y_{t-p+1} - \mu) \end{bmatrix} + \begin{bmatrix} e_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.1.20}$$

[5] Hamilton's version of the model also includes a matrix of predetermined or exogenous variables [9].

The first row of equation 4.1.20 is the same as equation 4.1.19. The other rows are the identities, $(y_{t-j} - \mu) \equiv (y_{t-j} - \mu)$ for $j = 1, \ldots, p-2$. Equation 4.1.20 has the same form as the state equation 4.1.16, where r=p, and

$$\beta_t = [(y_t - \mu), (y_{t-1} - \mu), \ldots, (y_{t-p+1} - \mu)] \tag{4.1.21}$$

$$w_{t+1} = (e_{t+1}, 0, \ldots, 0)' \tag{4.1.22}$$

$$G = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 1 & \ldots & 1 & 0 \end{bmatrix} \tag{4.1.23}$$

The measurement equation is,

$$y_t = \mu + F'\beta_t \tag{4.1.24}$$

where the first row of F' is the first row of a (pxp) identity matrix.

Now, suppose that F in the measurement equation 4.1.24 is replaced with a vector of parameters so that it is written as,

$$y_t = \mu + [1, \theta_1, \theta_2, \ldots, \theta_{p-1}]\beta_t \tag{4.1.25}$$

What kind of dynamic system does this specification describe? Assume that β_t evolves as an AR(p) vector so that it can be written as,

$$\begin{bmatrix} \beta_{1,t+1} \\ \beta_{2,t+1} \\ \vdots \\ \beta_{p,t+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 1 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \vdots \\ \beta_{p,t} \end{bmatrix} + \begin{bmatrix} w_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.1.26}$$

The first row of 4.1.26 shows that $\beta_{1,t+1}$ can written as an AR(p) process,

$$(1 - \phi_1 L - \phi_2 L^2 -, ..., -\phi_p L^p)\beta_{1,t+1} = w_{t+1} \tag{4.1.27}$$

The j^{th} row of equation 4.1.26 indicates that

$$\beta_{j,t} = L^j \beta_{1,t+1} \tag{4.1.28}$$

Equations 4.1.25 and 4.1.28 imply that for $j = 1, 2, 3, ..., p$,

$$y_t = \mu + [1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_{p-1} L^{p-1}]\beta_{1,t} \tag{4.1.29}$$

Subtracting μ from both sides of 4.1.29 and multiplying both sides by $(1 - \phi_1 L - \phi_2 L^2 -, ..., -\phi_p L^p)$ gives,

$$\begin{aligned} &(1 - \phi_1 L - \phi_2 L^2 -, ..., -\phi_p L^p)(y_y - \mu) = \\ &\mu + [1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_{p-1} L^{p-1}] \times \\ &(1 - \phi_1 L - \phi_2 L^2 -, ..., -\phi_p L^p)\beta_{1,t} = \\ &[1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_{p-1} L^{p-1}] w_{1,t} \end{aligned} \tag{4.1.30}$$

Thus, equations 4.1.25 and 4.1.26 are equivalent to an ARMA(p,p-1).

The 'dlm' package in R provides the end user with the ability to specify ARMA dlm models with a single command.

4.1.2 Bayes Theorem

The Kalman filer is a recursive procedure for the inference of β_t. It is easiest to understand in a Bayesian context. Bayes theorem states that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)}{P(B)} \tag{4.1.31}$$

$$P(B|A) = P(A|B)P(B) \tag{4.1.32}$$

We can rewrite this basic formulation of Bayes Theorem in terms of a prior distribution, likelihood, and posterior distribution:

$$p(\theta|y_t) = \frac{p(y_t|\theta)p(\theta)}{p(y)} \tag{4.1.33}$$

$$p(y) = \int_\theta p(y_t|\theta)p(\theta)d\theta \tag{4.1.34}$$

$$p(\theta|y_t) \propto p(y_t|\theta)p(\theta) \tag{4.1.35}$$

Equation 4.1.33 is the basis for Bayesian inference. It contains known and unknown quantities. The known quantity is the data, denoted y_t. The unknown quantities are the parameters, denoted θ. Bayesian inference refers to the updating of prior beliefs into posterior beliefs conditional on observed data. $p(\theta)$ is the prior distribution for the state of nature. It represents the modelers beliefs prior to observing the data. $p(y_t|\theta, x_t)$ is the likelihood and describes the data generating process of y_t. $p(\theta|y_t)$ is the posterior distribution, and describes the probability distribution of the state of nature after the data has been observed. $p(y)$ is the marginal likelihood. It is a normalizing constant in the sense that it guarantees that the area under the posterior curve integrates to one.

The posterior kernel (4.1.35) states that the posterior distribution is proportional to the product of the likelihood and the prior. Bayesian inference is often done using the kernel, but calculating moments of the posterior requires the use of the full model specification, including the marginal likelihood.

Definition - Density Kernel
The probability of a random variable ,X, often has the form $kg(X)$ where k is a numerical constant whose role is to ensure that $kg(X)$ integrates to one. The remaining portion of the density, $g(X)$ is called the kernel of the density function. To illustrate, the univariate Normal distribution (4.1.36) and its kernel (4.1.37) are shown below.

$$p(x) = \sqrt{\left(\frac{\tau}{2\pi}\right)} e^{-\left(\frac{-\tau}{2}\right)(x-\mu)^2} \tag{4.1.36}$$

$$p(x) \propto e^{-\left(\frac{-\tau}{2}\right)(x-\mu)^2} \tag{4.1.37}$$

where τ is the precision, or the inverse of the variance.

4.2 The Kalman Filter

The Kalman filter is a sequential process which can be used for estimating a dynamic linear model. The filter updates estimates of the state variables β_t each period as new information arrives. The sequential nature of the Kalman filter lends itself naturally to a Bayesian interpretation. For instance, we can describe a step for the Kalman filter as follows

$$\textbf{p(state of nature|data)} \propto \textbf{p(data|state of nature)} \times \textbf{p(state of nature)}$$

Using the notation from the state space model we can write the Bayes formula as

$$p(\beta_t \mid y_t, \mathbf{Y_{t-1}}) \propto p(y_t \mid \beta_{t-1}, \mathbf{Y_{t-1}}) \times p(\beta_{t-1} \mid \mathbf{Y_{t-1}}) \tag{4.2.1}$$

where $\mathbf{Y_{t-1}}$ is a vector representing the entire history of y_t observations as of time t-1.
At time, t-1 our knowledge of the state of nature can be described as

$$\beta_{t-1}|y_{t-1} \sim N(\widehat{\beta}_{t-1}, \Sigma_{t-1})) \tag{4.2.2}$$

This is the posterior value of β_t at time, t-1. Note that the distribution of the posterior is Normal, the mean is an expectation, and Σ is the variance.

The Kalman updating procedure may be thought of as consisting of two stages:

1. The estimate of β_t prior to observing y_t.
2. The estimate of β_t after observing y_t.

At time t-1, prior to observing y_t, the state equation 4.1.1 is used to estimate β_t. The prior $(\beta_t|y_{t-1})$ is defined as[6]

$$\beta_t \sim N(G\widehat{\beta}_{t-1}, R_t = G\Sigma_{t-1}G' + W_t) \tag{4.2.3}$$

After observing y_t we can calculate the likelihood, and the posterior. To calculate the likelihood we define the forecast error for y_t based on the forecast at t-1,

$$e_t = y_t - \widehat{y}_t = y_t - F_t G\widehat{\beta}_{t-1} \tag{4.2.4}$$

Note that F_t, G, and β_{t-1} are known, so once y_t is observed, the forecast error is known.

Bayes formula (4.2.1) can be rewritten as,

$$p(\beta_t \mid e_t, \mathbf{Y_{t-1}}) \propto p(e_t \mid \beta_t, \mathbf{Y_t}) \times p(\beta_t \mid \mathbf{Y_{t-1}}) \tag{4.2.5}$$

The measurement equation 4.1.2 can be used in conjunction with the forecast error to write

$$e_t = F_t(\beta_t - G\widehat{\beta}_{t-1}) + v_t \tag{4.2.6}$$

[6]The mean and variance in 4.1.1 come from applying the following result to the measurement equation: if $X \sim N(\mu, \Sigma)$ then $CX \sim N(C\mu, C\Sigma C')$ where C' is the transpose of C.

Since $v_t \sim N(0, V)$,

$$e_t \sim N(F_t(\beta_t - G\widehat{\beta}_{t-1}), V_t) \tag{4.2.7}$$

Bayes theorem can then be used to obtain the posterior at time t. Note that we need to calculate the marginal likelihood, $p(y_t)$ which requires the evaluation of an integral.

$$p(\beta_t \mid e_t, \mathbf{Y_{t-1}}) = \frac{p(e_t \mid \beta_t, \mathbf{Y_{t-1}}) \times p(\beta_t \mid \mathbf{Y_{t-1}})}{\int_{all\beta_t} p(e_t, \beta_t \mid \mathbf{Y_{t-1}}) d\beta_t} \tag{4.2.8}$$

Once we have the posterior, $p(\beta_t \mid e_t, \mathbf{Y_{t-1}})$, the filter proceeds to the next time period using the posterior from time t as the new prior in equation 4.2.3.

The calculation of the posterior can be simplified using the approach of Meinhold and Nozer [14]. Their method makes use of the definition and properties of a bivariate Normal density. Before discussing their method, we briefly review the bivariate Normal density function.

4.2.1 Bivariate Normal Density

Let X_1 and X_2 be bivariate normal random variables, The bivariate density can be written as,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]$$

where μ_1 and μ_2 are the mean of X_1 and X_2, respectively. $\Sigma_{i,j}$, $i, j = 1, 2$ are the variance and covariance of X_1 and X_2. One of the properties of the bivariate Normal distribution is that given the joint density of X_1 and X_2 there is a closed form solution for the conditional densities $X_1 \mid X_2$ and $X_2 \mid X_1$. The converse also holds, meaning that given the conditional densities we can write the joint density. The conditional of $X_1 \mid X_2$ is defined as,

$$X_1 \mid X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{4.2.9}$$

As pointed out by Meinhold and Nozer, the conditional density $(X_1 \mid X_2 = x_2)$ can be interpreted in terms of a regression of X_1 on x_2. The term $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ is a regression function and $\Sigma_{12}\Sigma_{22}^{-1}$ is a regression coefficient.

4.2.2 Deriving the Posterior Distribution

We will now substitute the mean and variance for β_t and e_t which were discussed in Section 4.2 into the equations of the bivariate Normal distribution. Let $X_1 \Leftrightarrow e_t$ and $X_2 \Leftrightarrow \beta_t$, then based on the result in equation 4.2.3, $\mu_2 \Leftrightarrow G\widehat{\beta}_{t-1}$ and $\Sigma_{22} \Leftrightarrow R_t$. Based on 4.2.7, and the definition of $X_1 \mid X_2$ we can write

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \Leftrightarrow F_t(\beta_t - G\widehat{\beta}_{t-1}) \tag{4.2.10}$$

$$\mu_1 + \Sigma_{12}R_t^{-1}(\beta_t - G\widehat{\beta}_{t-1}) \Leftrightarrow F_t(\beta_t - G\widehat{\beta}_{t-1}) \tag{4.2.11}$$

Therefore, $\mu_1 \Leftrightarrow 0$ and $\Sigma_{12} \Leftrightarrow F_t R_t$.

Next, we find the variance of e_t as follows:

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \Leftrightarrow \Sigma_{11} - F_t R_t F_t' \Leftrightarrow V_t \tag{4.2.12}$$

and so, $\Sigma_{11} \Leftrightarrow V_t + F_t R_t F_t'$.

Now that we have identified the distributions, we can write the joint distribution for e_t and β_t using the converse property of the bivariate Normal distribution.

$$\begin{pmatrix} \beta_t \mid Y_{t-1} \\ e_t \mid Y_{t-1} \end{pmatrix} \sim N\left[\begin{pmatrix} G\widehat{\beta}_{t-1} \\ 0 \end{pmatrix}, \begin{pmatrix} R_t & R_t F_t' \\ F_t R_t & V_t + F_t R_t F_t' \end{pmatrix} \right]$$

To obtain the posterior distribution, $p(\beta_t|Y_t)$, we use fact that the conditional distribution can be derived from the joint distribution.

$$(\beta_t \mid e_t, Y_t) \sim N(G_t\widehat{\beta}_t + R_t F_t'(V_t + F_t R_t F_t')^{-1} e_t, \\ R_t - R_t F_t'(V_t + F_t R_t F_t')^{-1} F_t R_t) \tag{4.2.13}$$

In summary at a point in time, t, we have the following posterior estimates for the mean and variance of β_t conditional on e_t.

$$\widehat{\beta}_t = G\widehat{\beta}_{t-1} + R_t F_t'(V_t + F_t R_t F_t')^{-1} e_t \tag{4.2.14}$$

$$\Sigma_t = R_t - R_t F_t'(V_t + F_t R_t F_t')^{-1} F_t R_t \tag{4.2.15}$$

Given initial values at t=0 for $\widehat{\beta}_0$ and Σ_0 we can use the recursive procedure outlined above to estimate $\widehat{\beta}_t$ at each point in time. Note that the Kalman Filter does not include estimates of the variance terms, V_t and W_t. Estimation of these parameters will be discussed later.

4.2.3 Interpreting the Posterior

The posterior mean of $\widehat{\beta}_t$ is the sum of two parts, the mean of the prior for $\widehat{\beta}_t$ and the one step ahead forecast error times a correction factor. The correction factor which is called the 'Kalman gain', has the same form as the coefficient from a regression of β_t on e_t. We can see this by noting that

$$\Sigma_{12}\Sigma_{22}^{-1} = R_t F_t'(V_t + F_t R_t F')^{-1} \tag{4.2.16}$$

The Kalman filter can be viewed as a series of regression functions of β_t on e_t with a new intercept ($G_t\widehat{\beta}_t$) and slope ($R_t F_t'(V_t + F_t R_t F')^{-1}$) in each time period. In effect, the Kalman filter is a learning process, where each new data point provides potentially new information about the model parameters.[7]

[7] As pointed out by Meinhold and Nozer [14] the "regression" only involves one observation of e_t and β_t at each point in time.

Properties of the Kalman Filter

- Linear - the Kalman filter discussed in this lecture estimates the state β_t for a linear dynamic system.
- Recursive - estimating the current state β_t does not require the entire history of observed data, y_T.
- Optimal - if all of the noise terms are Gaussian the Kalman filter minimizes the mean square error of the parameter estimates. If the noise terms are non-Gaussian the Kalman filter will be the best linear estimator, but a nonlinear estimator may be better.

Example 4.1.

We estimate a first order dynamic linear model using the Kalman filter. The measurement and state equations are as follows:

$$y_t = \beta_t + v_t, v_t \sim N(0, V_t)$$
$$\beta_t = \beta_{t-1} + w_t, w_t \sim N(0, W_t)$$

The observed data y_t is a univariate series consisting of a trend component, μ_t, and a noise component , e_t. The trend component follows a random walk. The Kalman filter is applied to simulated data where, $W = 0.25, V = 1, y_0 = 0, N = 100, F = 1, G = 1$. Note that these parameters are fixed in time. The simulated data, along with the Kalman filter estimate of β_t are plotted in Figure 4.4. The posterior estimate of β_t is highly correlated ($\rho = 0.95$) with y_t, and generally follows the trend of the data. The R code for this example is provided at the end of the chapter. The estimation was done with the "dlm" package.[8] Note that the variances for the measurement and state equations was provided. Section 4.4.1 discusses the use of maximum likelihood to estimate these parameters.

Example 4.2.

In this example, instead of using a packaged routing, we write our own Kalman filter and estimate time varying "betas" for the CAPM, where the dependent is a value weighted portfolio of excess returns for the firms in the durable goods sector. The market rate is the market return in excess of the risk-free rate, where the market is represented as a value-weighted return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ. The risk free rate is the one-month Treasury bill rate from Ibbotson Associates and provided by K. French. All of the data is available in the Kenneth French Data Library.[9]

The data set consists of of monthly returns beginning in July 1926 and ending in May 2018. The model specification is,.

$$r_t - r_{ft} = \alpha_t + \beta_t * (r_{mt} - r_{ft}) + v_t, \quad v_t \sim N(0, V_t) \tag{4.2.17}$$

$$\alpha_t = \alpha_{t-1} + w_{\alpha,t}, \quad w_{\alpha,t} \sim N(0, W_{\alpha,t}) \tag{4.2.18}$$

[8] There are several R packages that provide Kalman filtering, including KFAS, DSE and FKF.
[9] `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

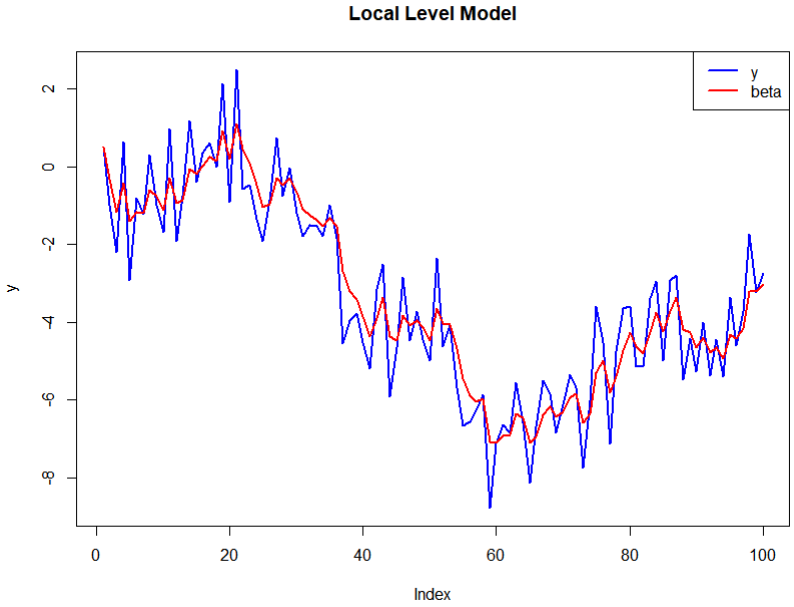

Figure 4.2: Kalman Filter - local level model

$$\beta_t = \beta_{t-1} + w_{\beta,t}, \quad w_{\beta,t} \sim N(0, W_{\beta,t}) \tag{4.2.19}$$

where r_t is the return for the durable goods index at time t, $r_f t$ is the risk free rate, and r_{mt} is the market return at time t. We assume that, $G_t = F_t = 1$, and W and V are constant over time.

The program for the filter is at the end of the chapter. Note that we have not estimated any of the parameters in the set $\Phi = (V_t, W_t, G, F_t)$. At this point we either assume they are known, or, as in the case of variance terms, we make random draws from a uniform distribution. A plot of the 'beta' estimates is shown in Figure 4.3. The red line is the least squares estimate of the "beta" for the entire data set.

4.3 Smoothing

In dynamic models, it is more common to use the smoothed distribution of the state vector than the basic forward filter. The Kalman filter uses data up to and including time, t, to estimate the state vector at time t. The smoother looks backwards in time at all of the data, and revises the estimates of the state vector at each time period. The smoother is retrospective in the sense that it reverses time and estimates $p(\beta_t \mid \beta_{t+1}, Y_T)\ \forall T$.

There are a number of different smoothers. In this lecture we discuss two smoothers which are commonly used. First we discuss the Rauch, Tung and Striebel (RTS) [16] smoother. The RTS smoother is sometimes called a fixed interval smoother since it is based on a fixed time interval. In this discussion the interval is the entire span of the observed data, y_T. [10] The other smoother that we

[10] Other types of smoothers include the fixed point smoother, and the fixed lag smoother. These two types of smoothers have applications in engineering.

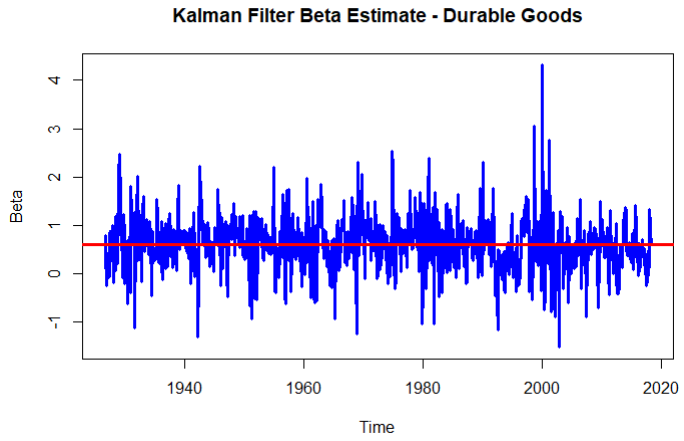

Figure 4.3: Kalman Filter Estimates - Durable Goods Betas

discuss is the DeJong and Shephard simulation smoother[3]. Simulation smoothers utilize Gibbs sampling to draw from the posterior distribution.

4.3.1 The RTS Smoother

The RTS smoother can be derived using the bivariate Normal distribution framework. [11] First we define the joint distribution [12]

$$\begin{pmatrix} \beta_t \mid Y_t \\ \beta_{t+1} \mid Y_t \end{pmatrix} \sim N\left[\begin{pmatrix} \widehat{\beta}_t \\ G\widehat{\beta}_t \end{pmatrix}, \begin{pmatrix} \Sigma_t & \Sigma_t G' \\ G\Sigma_t & R_{t+1} \end{pmatrix} \right]$$

Based on the properties of the bivariate Normal distribution,

$$p(\beta_t \mid \beta_{t+1} = b_{t+1}, Y_t) = \widehat{\beta}_t + \Sigma_t G' R_t^{-1}(\beta_{t+1} - G\widehat{\beta}_t) \tag{4.3.1}$$

In order to derive the smoother we need to add an additional step and take expectations across all values of β_{t+1}. The reason for this can be seen by noting that equation 4.3.1 is conditioned on a specific (unobserved) value of β_{t+1}. Applying the law of iterated expectations gives:

$$\begin{aligned} &E[\beta_t|Y_T] = E[E[\beta_t \mid \beta_{t+1}, Y_T] \mid Y_T] \\ &= E[E[\beta_t \mid \beta_{t+1}, Y_t]|Y_T] \\ &= E[\widehat{\beta}_t + \Sigma_t G' R_t^{-1}(\beta_{t+1} - G\beta_t)|Y_T] \\ &= \beta_{t|t} + \Sigma_{t|t} G' R_{t+1}^{-1}(\beta_{t+1|T} - \beta_{t+1|t}) \end{aligned} \tag{4.3.2}$$

[11] See Särkka [17] for a derivation of the smoother and its variance, as well as a comprehensive discussion of Bayesian smoothing in general.

[12] The $cov(\beta_t, \beta_{t+1}|Y_T) = cov(\beta_t, G\beta_t + w_t|Y_T) = cov(\beta_t, G\beta_t|Y_T) + cov(\beta_t, w_t|Y_T) = Gcov(\beta_t, \beta_t|Y_T) = G\Sigma_{t|t}$

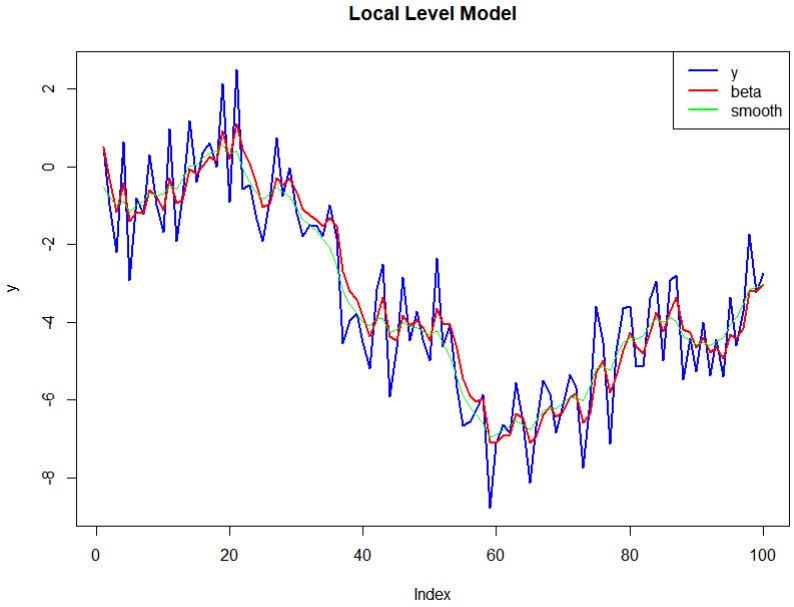

Figure 4.4: Smoothed posterior - local level model

where $t \leq T$. Note that $\widehat{\beta}_t, G\widehat{\beta}_t, \Sigma_t,$ and R_t are all calculated during the forward pass (Kalman filter), so they are known when smoothing. As a result, conditioning on T is the same as conditioning on t.

Looking at equation 4.3.2 we see that the smoother at t is equal to the posterior from the Kalman filter at t, plus the spread between the smoothed state parameter at $t+1$, $\beta_{t+1|T}$, and the prior estimate of of the state parameter at t, $\beta_{t+1|t}$, times an adjustment factor. The adjustment factor consists of the prior and posterior variances.

The variance of the smoother estimate shown in equation 4.3.3 is also derived using the bivariate Normal distribution and taking expected values.

$$\Sigma_{t|t+1} = \Sigma_{t|t} + L_t(\Sigma_{T|t} - \Sigma_{t+1|t})L_t' \tag{4.3.3}$$

where $R_{t+1|t} = G\Sigma_t G' + W_{t+1}$ and $L_t = \Sigma_{t|t}G'R_{t+1|t}^{-1}$

Derivation of the variance is given as an exercise at the end of the chapter.

Example 4.3. In this example we apply the smoother to the simulated data from example 4.1. The smoothed parameter estimates (Figure 4.4) are a bit less correlated with y_t ($\rho = 0.83$) but they provide a better sense of the underlying trend or "signal" in the data. The additional R commands used for this example are provided at the end of the Chapter in Section 4.8 .

Example 4.4. In this example we extend example 4.2 and write our own smoother to go along with the Kalman filter. We apply the smoother to the beta estimates for the durables goods sector. The results are shown in Figure 4.5. While a number of software packages offer Kalman smoothers we see that it doesn't take too many lines of code to write our own. At this point it is important to understand the role of the variance.

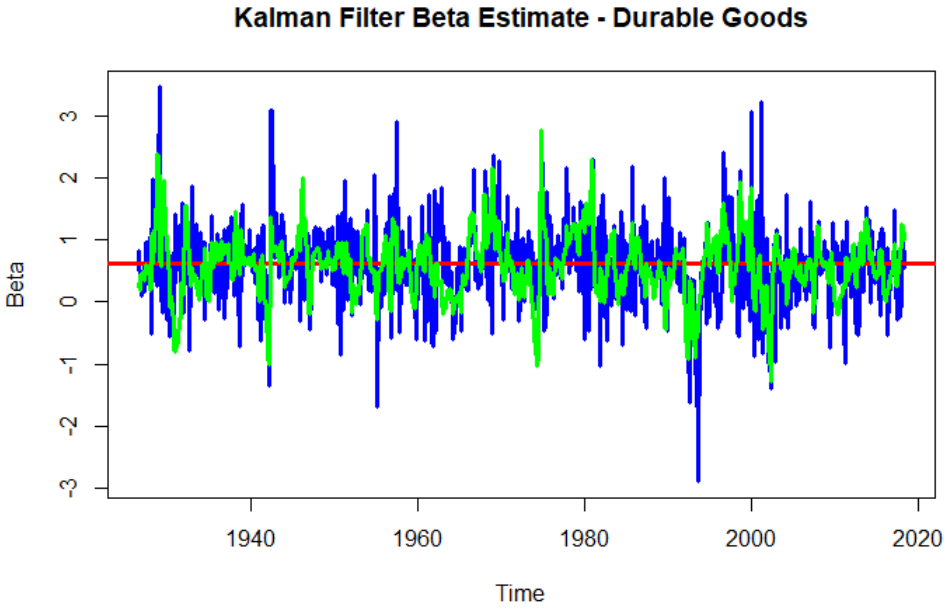

Figure 4.5: Smoothed Betas - Durable Goods

4.3.2 A Simulation Smoother

An alternative class of smoothers which has generated a lot of interest is the simulation smoother. Early development of the simulation smoother was done by Fruwirth-Schnatter (1994) [8], Carter & Kohn (1994) [12], and DeJong & Shephard (1995) [3]. The simulation smoother uses Gibbs sampling to sample from the joint posterior distribution, $p(\beta, \theta|Y_T)$, where β is a time series of the latent states, and θ contains the time series of unknown parameters which we will are assume are the variances of the measurement and state equations, V and W. The Gibbs sampler is a Markov Chain Monte Carlo simulation method, and will be discussed in detail in the lecture on MCMC. For the time being, it will suffice to understand that the Gibbs sampler is a sequential procedure that draws parameter estimates from the conditional densities $p(\beta|\theta, Y)$ and $p(\theta|\beta, y)$. The sequence of draws will eventually converge to draws from the method marginals densities, $p(\theta)$ and $p(\beta)$. Convergence requires that the processes that characterizes the draw have the proper Markovian properties. [13]

There are several approaches for drawing samples from the posterior distribution. First, the latent states can be drawn for each time period. That is, draw β_t from $p(\beta_t, \theta_t|Y_t)$ for $t = 1, ..., T$. This approach can be inefficient, and both Fruwirth-Schnatter (1994) [8], Carter & Kohn (1994) [12] suggest using a multi-state approach based on the following identity,

$$p(\beta|y, \theta) = p(\beta_t|y, \theta)p(\beta_{t-1}|y, \beta_t, \theta) \dots p(\beta_1|y, \beta_2, \dots, \beta_t, \theta) \tag{4.3.4}$$

The smoother developed by Fruwirth-Schnatter, which applies when $p(\beta|y, \theta)$ is

[13] Basically the stochastic processes must be recurrent and nt have absorbing states for the conditional densities to converge to marginal densities. The Gibbs sampler is discussed in Chapter 5

Gaussian, consists of the following steps,[14]

- Apply the Kalman filter to obtain $\widehat{\beta}_t$ and Σ_t for all t.
- Sample the most recent value of the state vector $\beta_T^{(s)}$ from

$$p(\beta_T|Y_T,\theta) \sim N(\widehat{\beta}_{T|T}, \Sigma_{T|T}) \tag{4.3.5}$$

- Sample β_t^s for $t = T-1,\ldots,1$ from the density $p(\beta_t|\beta_{t+1}, y_t, \theta)$.

DeJong and Shephard (1995) [3] proposed an alternative simulation smoother that involves sampling from the posterior distribution of the model disturbances rather then sampling the posterior of the state variables. Conditional on the latent parameter set, $\omega = (\omega_0^{'}, \omega_1^{'}, \ldots, \omega_n^{'}.)$, y_t is assumed to be generated by the following state space model,

$$Y_t = F_t\beta_t + T_t u_t \tag{4.3.6}$$
$$\beta_{t+1} = G_t\beta_t + H_t u_t \tag{4.3.7}$$

where, $u_t \sim N(0,\sigma^2)$, and the coefficient matrices may depend implicitly on ω. The simulation smoother draws $\eta \sim p(\eta|y,\omega)$, where $\eta = (\eta_0^{'}, \eta_1^{'}, \ldots, \eta_n^{'})$, and $\eta_t = Z_t u_t$. The definition of Z_t, which determines the distribution used to draw η_t, is arbitrary. For instance,

- If $Z_t = I$ then η_t is drawn from $p(\eta|y,\omega)$.
- If $Z_t = T_t$ then samples of η_t are drawn from the measurement equation disturbance terms.
- If $Z_t = H_t$ then samples of η_t are drawn from the state equation disturbance terms.

The first step in applying the simulation smoother is to run the Kalman filter for $t = 1,2,\ldots,n$. The Kalman equations are as follows:

$$\begin{aligned} e_t &= y_t - F_t\beta_t \\ D_t &= F_t P_t F_t^{'} + T_t T_t^{'} \\ K_t &= (G_t P_t F_t^{'} + H_t T_t^{'}) D_t^{-1} \\ \beta_{t+1} &= G_t\beta_t + K_t e_t \\ P_{t+1} &= G_t P_t L_t^{'} + H_t J_t^{'} \end{aligned}$$

where $L_t = G_t - K_t F_t$ and $J_t = H_t - K_t T_t$. e_t is the innovation vector, D_t is the scaled innovation covariance matrix, and K_t is the Kalman gain. The values of e_t, D_t, and K_t are saved for calculating the smoother. In the smoother phase, which reverses time and runs for $t = n, n-1, ..., 1$, η_t is calculated and stored,

$$\eta_t = Z_t(T_t^{'} D_t^{-1} e_t + J_t^{'} r_t) + \epsilon_t \tag{4.3.8}$$

[14]The a more in depth discussion of the simulation smoother in the lecture on Bayesian inference

where,

$$\begin{aligned} C_t &= Z_t(I - T_t' D_t^{-1} T_t - J_t' U_t J_t) Z_t' \\ V_t &= Z_t(T_t' D_t^{-1} F_t + J_t' U_t L_t) \\ r_{t-1} &= F_t' D_t^{-1} e_t + L_t' r_t - V_t' C_t^{-1} \epsilon_t \\ \epsilon_t &\sim N(0, \sigma^2 C_t) \\ U_{t-1} &= F_t' D_t^{-1} F_t + L_t' U_t L_t + V_t' C_t^{-1} V_t \\ r_n &= 0, U_n = 0 \end{aligned} \tag{4.3.9}$$

η_t is a draw from the Gaussian density $p(Z_t u_t | y, \eta_{t+1}, ..., \eta_n, \omega)$ with condition mean of, $F_t(T_t' D_t^{-1} e_t + J_t' r_t)$, and a conditional covariance matrix, $\sigma^2 C_t$. DeJong and Shephard's smoother has several advantages over the state sampler. First, it has fewer storage requirements than the state sampler which requires all one step ahead estimates of β_t and the corresponding covariance, P_t. It does not require the inversion of P_t. It has enhanced numerical stability since it can operate in square root form. It is easier to draw $p(\omega | y, \eta)$ then it is to draw $p(\omega | y, \beta)$

To better understand how the smoother works, consider the case where $Z_t = H_t$, and let $\Omega_t = H_t H_t'$. As noted earlier, the definition of Z_t determines the distribution used to draw η_t. In this case, the equations that define the smoother (4.3.9) simplify to,

$$C_t = \Omega_t - \Omega_t U_t \Omega_t, \quad \epsilon_t \sim N(0, \sigma^2 C_t), \quad V_t = \Omega_t U_t L_t \tag{4.3.10}$$

$$r_{t-1} = Z_t' D_t^{-1} e_t + L_t' r_t - V_t' C_t^{-1} \epsilon_t \tag{4.3.11}$$

$$U_{t-1} = F_t' D_t^{-1} F_t + L_t' U_t L_t + V_t' C_t^{t-1} V_t$$

A draw of η_a is from the distribution $p(H_t u_t | y, \omega)$, with conditional mean $\eta_t = \Omega_t r_t + \epsilon_t$, and conditional covariance matrix of $\sigma^2 C_t$.

4.4 Parameter Estimation

As previously noted, complete specification of the dynamic linear model requires an estimate of the variance for both the state and measurement equation. The Kalman filter provides a mean and variance at each point in time for the unobserved state variables. The filter also provides one step ahead forecasts at each time period along with a mean and variance. However, the filter does not estimate any of the model parameters, which includes $\Phi = (F_t, G_t, V_t, W_t, \widehat{\beta}_0, \Sigma_0)$. A number of approaches have been proposed to estimate the parameters, including maximum likelihood (MLE), expectation maximization (EM), and Gibbs sampling. MLE and EM are discussed in this lecture and the Gibbs approach is covered in the lecture on Bayesian inference.

4.4.1 Maximum Likelihood

Estimation of the parameters using ML is an iterative procedure. Assume that the state variable at $t = 0$ is Normal, $\beta_0 \sim N(0, \Sigma_0)$. In addition, assume that the

disturbance terms for the measurement and state equations, v_t and w_t, are jointly Normal.[15] Given the forecast error for the measurement equation and its variance,

$$e_t = y_t - \widehat{y}_t = y_t - F_t G \widehat{\beta}_{t-1} \tag{4.4.1}$$

and

$$\Omega_t = V_t + F_t R_t F_t' \tag{4.4.2}$$

The likelihood function is

$$logL(\Phi) = -\frac{1}{2}\sum_{t=1}^{n} log|\Omega_t(\Phi)| - 0.5\sum_{t=1}^{n} e_t'(\Phi)\Omega_t(\Phi)^{-1}e_t(\Phi) \tag{4.4.3}$$

Estimation proceeds as follows (see Gupta and Mehra):

1. Initialize the parameters, Φ^0.
2. Run the Kalman filter to obtain forecast errors and an estimate of Ω_t^0.
3. Update the parameter estimates by maximizing the likelihood function (4.4.3) using a numerical optimization routine such as Newton-Raphson using the values from step 2.
4. Iterate through steps 2 and 3 until the change in $LogL(\Phi)$ is small.

4.4.2 The EM Algorithm

Shumway and Stoffer (1982) [19] proposed using the Expectation Maximization (EM) algorithm in conjunction with the Kalman filter and smoother to estimate state space models. EM is typically used for estimation when a model contains latent variables. In the case of the linear state space model the state variable, β_t is latent. Suppose, however, that we actually observed the states in addition to the measurement data. That is, at each point in time observe pairs (y_t, β_t). The joint density over all T is,

$$f(\beta_t, Y_t) = f(\beta_0)\prod_{t-1}^{T} f(\beta_t|\beta_{t-1})\prod_{t=1}^{T} f(y_t|\beta_t) \tag{4.4.4}$$

The joint density is the product of three densities,

1. The initial value for β
2. The density of the state variable at t conditioned on its value at t-1
3. The density of the measurement variable, y_t, conditioned on the state variable at t.

[15] Estimating the parameters using maximum likelihood was first proposed by Schewppe [18].

The log likelihood is,

$$
\begin{aligned}
-2lnL(\Phi) &= ln|\Omega_0|_(\beta_0 - G\mu_0)'\Omega_0^{-1}(\beta_0 - \mu_0) \\
&+Tln|W| + \sum_{t=1}^{T}(\beta_t - G\beta_{t-1})'W^{-1}(\beta_t - G\beta_{t-1}) \\
&+Tln|V| + \sum_{t=1}^{T}(\beta_t - F_t\beta_t)'V^{-1}(\beta_t - F_t\beta_t)
\end{aligned} \tag{4.4.5}
$$

where μ_0 is the mean of β_0.

If the states were observed, then the model parameters could be estimated by maximizing the likelihood function in the standard fashion. However, as the states are unknown Shumway and Stoffer [16] propose using the EM approach. That is, maximize the expected likelihood conditional on a set of parameters $\Phi^{(j)}$,

$$
Q(\Phi|\Phi^{(j)}) = E\big\{ -2lnL(\Phi|Y_t, \Phi^j)\big\} \tag{4.4.6}
$$

The conditional expectation is defined in terms of smoothed state estimates,

$$
\begin{aligned}
Q(\Phi|\Phi^{(j)}) &= ln|\Omega_0| + tr\big\{\Omega_0^{-1}[\Sigma_0^t + (\beta_{0|T} - \mu_0)(\beta_{0|T} - \mu_0)']\big\} \\
&+Tln|W| + tr\big\{W^{-1}[S_{11} - S_{10}\Phi' - \Phi S_{10}' + \Phi S_{00}\Phi]\big\} \\
&+Tln|V| + tr\big\{V^{-1}\sum_{t=1}^{T}(y_t - F_t\beta_{t|T})(y_t - F_t\beta_{t|T})' + F_t\Sigma_{t|T}F_t']\big\}
\end{aligned} \tag{4.4.7}
$$

where the smoother are defined as,

$$
S_{11} = \sum_{t=1}^{T}(\beta_{t|T}\beta_{t|T}' + \Sigma_{t|T}) \tag{4.4.8}
$$

$$
S_{10} = \sum_{t=1}^{T}(\beta_{t|T}\beta_{t-1|T}' + \Sigma_{t-1|T}) \tag{4.4.9}
$$

$$
S_{00} = \sum_{t=1}^{T}(\beta_{t-1|T}\beta_{t-1|T}' + \Sigma_{t-1|T}) \tag{4.4.10}
$$

Φ^j are the current values of the parameters, and the conditional subscripts indicate smoothed values, e.g. $\beta_{0|T}$ is β_0 after smoothing.

Minimizing equation 4.4.7 with respect to the parameters at iteration j results in the following update equations,

$$
G^{(j)} = S_{10}S_{00}^{-1} \tag{4.4.11}
$$

$$
W^{(j)} = T^{-1}(S_{11} - S_{10}S_{00}^{-1}S_{10}') \tag{4.4.12}
$$

$$
V^{(j)} = T^{-1}\sum_{t=1}^{T}[(y_t - F_t\beta_{t|T})(y_t - F_t\beta_{t|T})' + F_t\Sigma_{t|T}F_t'] \tag{4.4.13}
$$

The updates for the initial mean and variance are

[16] This discussion follows Shumway and Stoffer (2006) [20] which summarizes the results in Shumway and Stoffer (1982) [19].

$$\beta_0^j = \beta_{0|T} \text{ and } \Sigma_0^{(j)} = \Sigma_{0|T}$$

Given these equations, the EM algorithm iterates between estimating expected values of the parameters and maximizing the (incomplete) likelihood function (4.4.3). The steps are as follows:

1. Set $j = 0$ and initialize Φ^j

2. Calculate the likelihood function (4.4.3) using Φ^j .

3. Expectation step: Filter forward and smooth backward using parameter estimates Φ^j and then calculate equations 4.4.8 - 4.4.10.

4. Maximization Step: Update the parameter estimates using equations 4.4.11 - 4.4.13 to obtain Φ^{j+1}

5. Repeat steps 2 to 5 until convergence.

Example 4.5.

In this example we use the 'dlm' package to estimate the betas for the durable goods sector that was estimated in Example 4.2. The variances of the state and measurement equation are estimated using maximum likelihood. Figure 4.6 shows the smooth time varying "beta". The horizontal line is the OLS estimate for the entire data set. The red lines are the 95% confidence interval. Interestingly, the time varying estimate of beta was below the OLS estimate for most of the estimation period. The two exceptions are the Great Depression and the Great Recession.

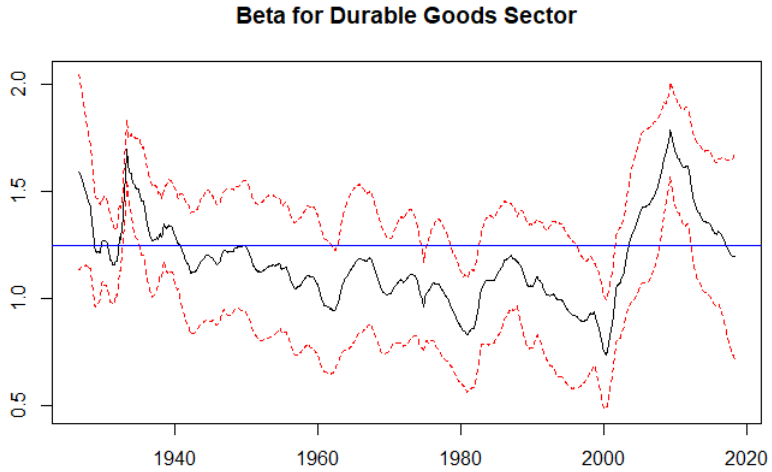

Figure 4.6: Smoothed CAPM Beta - Durable Goods Sector

4.5 Forecasting

As we saw earlier, the recursive nature of the Kalman filter results in one step ahead forecasts at each time period, t. Now we will use the filter to calculate l-step ahead forecasts.[17] Assume we have a data set consisting of $Y_T = \{y_1, ..., y_T\}$ observations. Our objective is to forecast l periods into the future to $T + l$, where $l = 1, 2, ...L$. Assume that we want the forecast value that minimizes the mean square forecast error. This is, we want to find $\widehat{y}_{T+l}$ so that,

$$MSE_{T+l} = minE[(\widehat{y}_{T+l} - y_{T+l})(\widehat{y}_{T+l} - y_{T+l})'|Y_T] \tag{4.5.1}$$

Since y_t is a random variable, the value of Y_{t+l} that minimizes equation 4.5.1 is the expected value conditional on Y_T, [18]

$$\bar{y}_{T+l} = E[y_{T+l}|Y_T] \tag{4.5.2}$$

The measurement equation l periods into the future is,

$$y_{T+l} = F_{T+l}\beta_{T+l} + v_t \tag{4.5.3}$$

Given the measurement equation, and equation 4.5.2,

$$\bar{\beta}_{T+l} = E[\beta_{T+l}|Y_T] \tag{4.5.4}$$

$$\bar{\Sigma}_{T+l} = E[(\bar{\beta}_{T+l} - \beta_{T+l}))(\bar{\beta}_{T+l} - \beta_{T+l})'] \tag{4.5.5}$$

The MSE forecast is,

$$\bar{y}_{T+l} = F_{T+l}\bar{\beta}_{T+l} \tag{4.5.6}$$

The MSE variance is[19]

$$E[(\bar{y}_{T+1} - y_{T+1})(\bar{y}_{T+1} - y_{T+1})'] = F_{T+1}\bar{\Sigma}_{T+l}F'_{T+1} \tag{4.5.7}$$

Durbin and Koopman [7] show that the l-step ahead forecast of the mean and variance for the state parameters can be estimated using the Kalman recursion formulas by setting the disturbance of the state equation and the Kalman gain to zero. Given these assumptions the l-step ahead forecast of the state parameter is

$$\bar{\beta}_{T+l} = G\bar{\beta}_{T+l-1} \tag{4.5.8}$$

and the variance is,

$$\bar{\Sigma}_{T+l} = G\bar{\Sigma}_{T+l-1}G' \tag{4.5.9}$$

[17] This discussion in this section is based the work of Durbin and Koopman [7], Chapter 4, Section 4.11.

[18] As noted in Durbin and Koopman, [7], this follows from the fact that in general, $E((X - \lambda)(X - \lambda)')$ is minimized for $\lambda = \mu$, where μ is the mean of X.

[19] This can be shown by expanding $E[(\bar{y}_{T+1} - y_{T+1})(\bar{y}_{T+1} - y_{T+1})']$, taking expectations, and noting that $\bar{\beta}\beta' = \beta\bar{\beta}'$

Example 4.6. In this example, we model home price appreciation (HPA) as a local state space model, and forecast it for 36 months. HPA is defined as the monthly growth rate of the Case-Shiller National Home Price Index. The model is estimated using data from January 1987 to April 2018. The model is then used to forecast 'out-of-sample' growth for May 2018-May 2019.

The model is,

$$hpa_t = \beta_t + v_t, v_t \sim N(0, V_t) \tag{4.5.10}$$

$$\beta_t = \beta_{t-1} + w_t, w_t \sim N(0, W_t) \tag{4.5.11}$$

where v_t and w_t are independent.

Historical home price appreciation along with the 36 month projections is shown in Figure 4.7. Historical growth ranges from -12.5% to +14.5 percent. The projected growth rates range from 3% to 10%. The forecast command in the dlm package allows the user to draw a sample from the forecast distribution. In this example the sample size is 10.

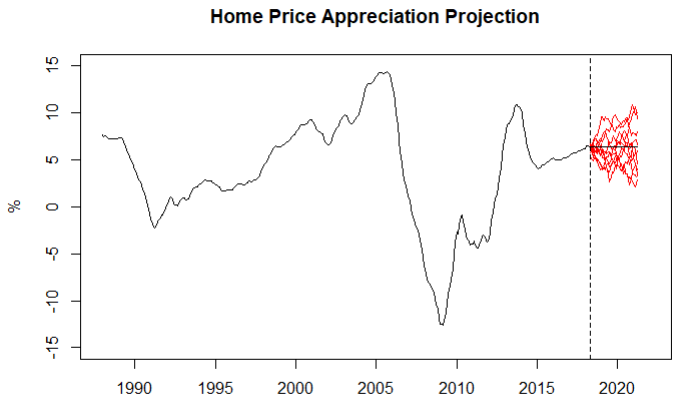

Figure 4.7: Case-Shiller National Home Prices - Year-over-year growth

4.6 Applications in the Literature

4.6.1 The Macroeconomy and the Yield Curve

In this section we discuss two papers by Diebold, et. al. that examine the dynamic nature of the Nelson-Siegel model of the yield curve, and it's relationship to the economy. In "The Macroeconomy and the Yield Curve", Diebold, Rudebusch & Aruoba (2006) (ADR) [5] use the Nelson-Siegel [15] 3 factor model described in Diebold & Li (2006) [4] to analyze the relationship between the yield curve and inflation, real economic activity, and monetary policy.

Nelson and Siegel fit the forward rate curve at a given date with the following three factor model,

$$f_t(\tau) = \beta_{1,t} + \beta_{2,t} exp(-\lambda_t \tau) + \beta_{3,t}(\lambda_t) exp(-\lambda_t \tau) \tag{4.6.1}$$

By averaging over the forward rates, equation 4.6.1 can be expressed in terms of spot rates.[20] The cross-section of yields at a point in time is:

$$y_\tau = \beta_{1,t} + \beta_{2,t}\left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau}\right) + \beta_{3,t}\left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{4.6.3}$$

where τ is the maturity, and $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}$ and λ are parameters.

The parameter λ_t governs the exponential decay rate:

- Small values of λ_t produce slow decay and can better fit the long maturity bonds.
- Large values of λ_t produce fast decay and better fit the curve at short maturities.
- λ_t also governs where the loading on $\beta_{b3,t}$ achieves its maximum.

Diebold and Li (2006) [4] interpret $\beta_{it}, i = 1, ..., 3$ as latent dynamic factors. The loadings for each factor are plotted across maturity in Figure 4.8.

- The loading on factor $\beta_{1,t}$ is 1. Since this loading is fixed across maturity $\beta_{1,t}$ can be interpreted as a long term factor.
- The loading on factor $\beta_{2,t}$ is a function that starts at 1 and decays monotonically and and rather quickly to zero. $\beta_{2,t}$ can be viewed as a short term factor, since it loads more heavily on short rates.
- The loading on $\beta_{3,t}$ starts at zero, increases, and then decreases to zero. It can be viewed as a medium term factor.

Alternatively, Diebold & Li describe the three time varying factors as the level (L_t), slope (S_t), and curvature (C_t) of the yield curve at a point in time.

- **Level,** L_t - $\beta_{1,t}$ corresponds to the level of the yield curve. An increase in $\beta_{1,t}$ raises the yield level across the curve.
- **Slope,** S_t - $\beta_{2,t}$ corresponds to the slope of the yield curve. An increase in $\beta_{2,t}$ increases short rates more than long rates, changing the slope.
- **Curvature,** C_t - $\beta_{3,t}$ corresponds to the curvature of yield curve. An increase in $\beta_{3,t}$ causes the medium term yields to increase more than short and long term yields.

[20] The spot rate is defined as,

$$y(n) = \frac{1}{n}\int_0^n f(s)ds \tag{4.6.2}$$

where y = spot rate at t, f is the forward rate, and n is the number of periods ahead when the forward contract expires

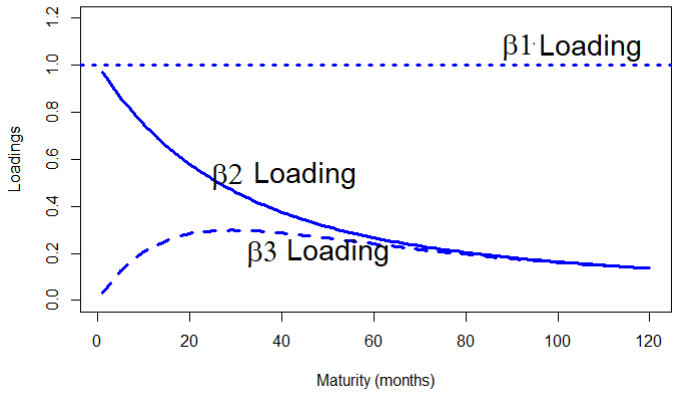

Figure 4.8: Nelson-Seigel Factor Loadings

By re-interpreting the Nelson-Seigel model as a dynamic model, Diebold and Li were able to replicate many of the empirical facts that characterize movements in the term structure of interest rates over time. In addition, they found that although one month ahead forecasts of the yield curve using the model were no better than a random walk, 1 year ahead forecasts outperformed a number of different models in out-of-sample tests.

Their model specification is,

$$y_\tau = L_t + S_t\left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau}\right) + C_t\left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{4.6.4}$$

where $L_t, C_t,$ and S_t are time varying versions of $\beta_1, \beta_2,$ and β_3, respectively. The model can be written in state space form. The state equation is,

$$\begin{bmatrix} L_t - \mu_L \\ S_t - \mu_S \\ C_t - \mu_C \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} L_{t-1} - \mu_L \\ S_{t-1} - \mu_S \\ C_{t-1} - \mu_C \end{bmatrix} + \begin{bmatrix} \eta_t(C) \\ \eta_t(S) \\ \eta_t(C) \end{bmatrix} \tag{4.6.5}$$

The measurement equations, which relates the N yields to the unobserved states at a point in time is shown in equation 4.6.6.

$$\begin{bmatrix} y(\tau_1) \\ y(\tau_2) \\ \vdots \\ y(\tau_N) \end{bmatrix} = \begin{bmatrix} 1 & \left(\frac{1-e^{1-\lambda\tau_1}}{\lambda\tau_1}\right) & \left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau_1} - e^{-\lambda\tau_1}\right) \\ 1 & \left(\frac{1-e^{1-\lambda\tau_2}}{\lambda\tau_2}\right) & \left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau_2} - e^{-\lambda\tau_2}\right) \\ \vdots & \vdots & \vdots \\ 1 & \left(\frac{1-e^{1-\lambda\tau_N}}{\lambda\tau_N}\right) & \left(\frac{1-e^{1-\lambda\tau}}{\lambda\tau_N} - e^{-\lambda\tau_N}\right) \end{bmatrix} \begin{bmatrix} L_t \\ S_t \\ C_t \end{bmatrix} + \begin{bmatrix} \epsilon_t(\tau_1) \\ \epsilon_t(\tau_2) \\ \vdots \\ \epsilon_t(\tau_N) \end{bmatrix} \tag{4.6.6}$$

where t= 1,2,...,T. In matrix form the state and measurement of the model are,

$$(f_t - \mu) = A(f_{t-1} - \mu) + \eta_t \tag{4.6.7}$$

and

$$y_t = \Lambda f_t + \epsilon_t \tag{4.6.8}$$

where η_t and ϵ_t are orthogonal Normal white noise terms.

$$\begin{bmatrix} \eta_t \\ \epsilon_t \end{bmatrix} \sim WN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & H \end{bmatrix} \right) \tag{4.6.9}$$

ADR (2006) use the Nelson-Siegel 3 factor model described in Diebold & Li to analyze the relationship between the yield curve and inflation, real economic activity, and monetary policy. They define a 'macro-yield' model [21] using equations 4.6.7 and 4.6.8 with $f' = (L_t, S_t, C_t, CU_t, FFR_t, INFL_t)$ where

- CU_t is manufacturing capacity utilization
- FFR_t is the Federal Funds Rate
- $INFL_t$ is the rate of inflation

For this specification, $\Lambda \sim T \times 6$ matrix with zeroes in the last three columns so that the yield still only loads on the three curve factors. The measurement equation is a 6 equation first order VAR system consisting of the 3 yield factors and the 3 economic factors. This specifications allows the macroeconomic factors to contribute to the forecast of the yield factors.

The VAR functional form provides a useful framework to evaluate the relationship between the macro variables and the yield curve. Using impulse-response functions ADR evaluate 4 groups of shocks: macro reponse to macro shocks; macro response to yield curve shocks; yield curve response to macro shocks; yield curve response to yield curve shocks. Some key results are summarized below:

Macro response to yield curve shock:

- Macro variables have a negligible response to curvature shocks.
- An increase in the slope is matched by a one-to-one increase in FFR_t.
- An increase in the level of the yield positively shocks all three macro variables.
- An increase in L_t ,which is correlated with inflation, lowers the real rate of interest measured as $FFR_t - L_t$, causing the economy to expand in the near term.

Yield curve response to macro shocks:

- Curvature shows very little response to macro shocks;
- The slope factor responds positively to shocks in all three macro variables.
- An increase in FFR_t immediately flattens the curve.
- Positive shocks to capacity utilization and inflation result in delayed positive responses.
- Shocks to macro variables effect the level of the yield curve.

[21] They also estimate a 'yield only' model that is similar to the Diebold & Li model, and a 'macro only' model.

4.7 Exercises

1. Simulate a data set of 500 observations assuming a 2nd order DLM model with μ_t= 0.85 , and W_t=0.2. Then, write a program to estimate the model using a Kalman filter.

2. Derive the Kalman filter mean and variance for a random walk without drift.

3. Derive the variance of the Kalman smoother discussed in section 4.3. Hint: $Var(x) = E_y(Var(X|Y = y)) = +Var_y(E(X|Y = y))$

4. Using R write a program for an RTS smoother to accompany the Kalman filter code in Example 2.2.

5. Write an R program to estimate the 95% confidence intervals for the (RTS) smoothed parameters.

4.8 R Code for Examples

4.8.1 Example 4.1

```
library(dlm)
#Simulate data for a local level model
W = 0.25
V = 1.0
n = 100
x0 = 0.00
w = rnorm(n,0,sqrt(W))
v = rnorm(n,0,sqrt(V))
x = y = rep(0,n)
x[1] = x0 + w[1]  #Initial State equation
y[1] = x[1] + v[1]  # Initial Observation equation
for (t in 2:n) {
x[t] = x[t-1] + w[t] #Update state equation with random draw
y[t] = x[t] + v[t] } #Update observation equation with random draw plus new state v
#Estimate the dlm for the local level model
dlm1 <- dlmFilter(y,dlmModPoly(order=1,dV=1.0,dW=0.25))
#Plot the data, y, and the fitted "beta".
plot(y,type='l', lwd=2, main="Local Level Model",col="blue")
lines(dlm1$m[-1],col="red",lwd=2)
legend("topright", legend=c("y", "beta"), col=c("blue","red"),lty=1:1 ,lwd=2)
```

4.8.2 Example 4.2

```
#The Kalman Filter
y<-durables
X<-cbind(rep(1), market)
T<-length(y)
k<-2 #unknown states}
#Define containers
beta<-matrix(data=NA,nrow=T,ncol=k) #store series of state parameters
W<-matrix(data=NA,nrow=k,ncol=k)
VCV<-matrix(data=NA,nrow=T,ncol=2*k) #VCV matrix for state disturbance terms
bnew<-matrix(data=NA,nrow=k,ncol=1) #posterior
bold<-matrix(data=NA,nrow=k,ncol=1) #prior
sigold<-matrix(data=NA,nrow=k,ncol=k) #prior variance
signew<-matrix(data=NA,nrow=k,ncol=k) #Posterior variance
#Set initial values
bold[1]<-0
bold[2]<-1
sigold[1,1]<-0.1
sigold[1,2]<-0
sigold[2,1]<-0
```

```
sigold[2,2]<-0.1
e<-0
beta[1,1]<-bold[1]
beta[2,1]<-bold[2]
#Apply Filter
for(t in 2:T){
v<-runif(1)
W[1,1]<-runif(1)*0.5
W[1,2]<-0
W[2,1]<-0
W[2,2]<-runif(1)*0.5
F<-matrix(X[t,],nrow=1,ncol=k)
e <- y[t] - F%*%bold #Forecast error
R<-sigold  + W
gain<-R%*%t(F)%*%solve(v+F%*%R%*%t(F))  #Kalman gain
bnew<-bold + gain%*%e #posterior state estimate
signew<-R - gain%*%F%*%R #posterior variance of state estimate
beta[t,1]<-bnew[1]
beta[t,2]<-bnew[2]
bold<-bnew
VCV[t,1]<-signew[1,1]
VCV[t,2]<-signew[1,2]
VCV[t,3]<-signew[2,1]
VCV[t,4]<-signew[2,2]
sigold<-signew
}
bout<-ts(beta[,2],freq=12,start=c(1926,8))
plot(bout,type="l",lwd=3,col="blue",main="Kalman Filter Beta Estimate - Durable Goo
abline(h=0.61,lwd=3,col="red")
```

4.8.3 Example 4.3

```
skal <- dlmSmooth(dlm1)
lines(dropFirst(skal$s),col="green")
legend("topright", legend=c("y", "beta","smooth"), col=c("blue","red","green"),lty=
```

4.8.4 Example 4.4

```
#Smoother
SVCV<-matrix(data=NA,nrow=T,ncol=2*k)
SVCV[T,]<-VCV[T,]
bsnew<-matrix(data=NA,nrow=T,ncol=2)#smoothe state
bsnew[T,]<-beta[T]
L<-matrix(data=NA,nrow=k,ncol=k)

for(t in {T-1}:1){
```

```
W<-matrix(c(runif(1)*0.025,0,0,runif(1)*0.025),nrow=k,ncol=k)
sigold<-matrix(VCV[T,],nrow=k,ncol=k)#prior variance
R<- G%*%sigold%*%t(G)  + W
R<-(R+t(R))/2
L = sigold%*%t(G)%*%solve(R)
bold<-matrix(beta[t,],nrow=k,ncol=1)
bnew<-matrix(bsnew[t+1,],nrow=k,ncol=1)
bsnew[t,]<-c(bold + L%*%(bnew - G%*%bold))
signew<-matrix(SVCV[t+1,],nrow=k,ncol=k)
SVCV[t,]<-c(sigold + L%*%(signew - R)%*%t(L))
}
bout<-ts(beta[-1,2],freq=12,start=c(1926,8))
plot(bout,type="l",lwd=3,col="blue",main="Kalman Filter Beta Estimate - Durable Goo
abline(h=0.61,lwd=3,col="red")
```

4.8.5 Example 4.5

```
buildTVP <-function(parm) {
parm$ <- exp(parm)
return( dlmModReg(X=x, dV=parm[1], dW=c(parm[2],parm[3])) )
}
start.vals= c(0,0,0)
TVP.mle = dlmMLE(y=y, parm=start.vals,  build=buildTVP, hessian=T)
TVP.dlm <- buildTVP(TVP.mle\$par)
TVP.f <- dlmFilter(y, TVP.dlm)
TVP.s <- dlmSmooth(TVP.f)
```

4.8.6 Example 4.6

```
y<-ts(CH3CaseShiller[,3],freq=12,start=c(1988,1))
#Define a 1st Order DLM with unknown W and V
#Define model for maximum likelihood
build <- function(parm) {
dlmModPoly(order = 1, dV = exp(parm[1]), dW = exp(parm[2]))}
#Estimate W and V using MLE  Using the 'dlm' package
fit <- dlmMLE(y, rep(0,2), build )
names(fit)
unlist(build(fit$par)[c("V","W")])
# Define the model
simPoly<-dlmModPoly(order=1,dV=3.284197e-10,dW=1.863845e-01)
unlist(simPoly)
#Run the Kalman Filter
simFilt <- dlmFilter(y, simPoly)
str(simFilt, 1)
n <- length(y)
```

```
attach(simFilt)
#Plot the observed series
plot(y,type="l",xlab="time",ylab="")
# Plot the filter state series (means)
lines(dropFirst(simFilt$m),col=2)
title(paste("V and W Estimated Using ML"))

#36 month ahead forecasts
set.seed(1)
unFore <- dlmForecast(simFilt, nAhead = 36, sampleNew = 10)
plot(window(y, start = c(1988, 1)), type = 'l',
xlim = c(1988, 2022), ylim = c(-15, 15.0),
xlab = "", ylab = "%", main="Home Price Appreciation Projection")
names(unFore)
attach(unFore)
invisible(lapply(newObs, function(x) lines(x, col = "red")))
lines(f, type = 'l')
abline(v = mean(c(time(f)[1], time(y)[length(y)])), lty = "dashed")
```

Bibliography

[1] Arnold Tom, Mark Bertus, and Jonathan Godbey. "A Simplified approach to understanding the Kalman filter technique", The Engineering Economist, 53: 140-155, 2008.

[2] Briers, Mark, Arnaud Doucet, and Simon Maskell. "Smoothing algorithms for State Space Models", AN Inst Stat Math (2010) 62:61-89.

[3] DeJong, Piet and Neil Shephard. "The Simulation smoother for time series models", Biometrika (1995) 82, 2, pp. 339-50.

[4] Diebold, Francis X. and Canlin Li. "Forecasting the term structure of government bond yields", Journal of Econometrics 130 (2006), 337-364.

[5] Diebold, Francis X., Glenn D. Rudebusch, and S. Boragan Aruoba. 2006. "The macroeconomy and the yield curve: a dynamic latent factor approach." Journal Of Econometrics 131, 309-338.

[6] Duncan, D.B. and S.D. Horn. "Linear Dynamic Recursive Estimation from the Viewpoint of Regresion Analysis", The American Statistician, Dec. 1972, Vol. 67 No. 340.

[7] Durbin, J. and S.J. Koopman. "Time Series Anaysis by State Space Methods", 2nd Ed. 2012, Oxford University Press.

[8] Fr*ü*hwirth-Schnatter, Sylvia. "Data Augmentation and Dynamic Linear Models", Journal of Time Series Analysis, Vol.15, No. 2, 1994, pp. 183-202.

[9] Hamilton, James, "State Space Models", Handbook of Econometrics, Volume IV, Editors R.F. Engel and D.L. McFadden. 1994 Elsevier Science

[10] Kalman, R.E. "A New Approach to Linear Filtering and Prediction Problems", Journal of Basic Engineering, March 1960 35-45.

[11] Kalman, R.E. and R.S.Bucy. "New Results in Linear Filtering and Prediction Theory", Journal of Basic Engineering, March 1961 95-108.

[12] Carter, C.K. and R Kohn, "On Gibbs Sampling for State Space Models", Biometrika, Vol.81, No. 3 (Aug. 1994), pp. 541-553.

[13] Koopman, Siem Jan, and Charles S. Bos, "State Space Models with a Common Stochastic Variance", Journal of Business & Economic Statistics, Vol. 22, No.3 (July 2004), pp. 346-357.

[14] Meinhold, Richard and Nozer D. Singpurwalla. "Understanding the Kalman Filter", The American Statistician, May 1983, Vol.37 No.2.

[15] Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curve. Journal of Business 60, 473–489.

[16] Rauch, H.E., C.T. Striebel, and F. Tung. "Maximum likelihood estimates of linear dynamic systems", AIAA Journal 3:8, pp 1445-1480, 1965.

[17] Särkka, Simo, "Bayesian Filtering and Smoothing", Cambridge University Press, 2013.

[18] Schweppe, C. F. (1965). Evaluation of likelihood functions for Gaussian signals. IEEE Trans. Info. Theory 11, 61-70.

[19] Shumway, R.H. and D.S. Stoffer, "An Approach to Time Series Smoothing and Forecasting using the EM Algorithm", Journal of Time Series Analysis, Volume 3, Issue 4, July 1982.

[20] Shumway, R.H. and D.S. Stoffer,"Time Series Analysis and its Applications" 2nd Edition, Springer, 2006.

[21] West, M. and P. J. Harrison, "Bayesian Forecasting & Dynamic Models", Springer, 1997, 2nd Ed.

Chapter 5

Cointegration

5.1 Introduction

Up until this point our discussions have primarily focused on applications using stationary series. If a time series is found to be integrated of order d, than we can take the d^{th} difference to create a stationary series, and then construct a model. In this chapter we expand the breadth of models to include stationary processes that are linear combinations of non-stationary processes. This is known as cointegration, and as we will see there are several useful models that make use of the concept of cointegration. The first model is the error correction model. In my experience, the error correction model is one of the most useful tools available for forecasting time series over long horizons. By combining long run equilibrium relationships with short run deviations it takes advantage of information that exists in the levels of non-stationary series along with the stationary change in level. A second useful application of cointegration is pairs trading which involves identifying asset prices that move together over time and create a stationary relationship (cointegration). Statistically meaningful deviations from the long run relationship present trading opportunities

5.2 Cointegration

We begin this section by reviewing some the differences between stationary and non-stationary series as outlined by [1].

- If $x_t \sim I(0)$ then 1) the variance of x_t is finite; 2) an innovation to x_t has a finite life; 3) the autocorrelation of x_t decreases steadily and have a finite sum; 4) the expected crossing of $x_t = 0$ occurs within a finite period of time; 5) the spectrum of x_t, $f(\omega)$ has the property, $0 < f(f) < \infty$.

- If $x_t \sim i(1)$ with $x_0 = 0$ then 1) the variance of x_t goes to infinity as t goes to infinity; 2) an innovation to x_t has a permanent impact on x_t; 3) the autocorrelation is one for $k \to \infty$; 4) the time between expected crossings of $x_t = 0$ is infinite; 5) the spectrum of x_t has the approximate shape $f(\omega) \sim A\omega^{-2d}$ so that for small ω, so $f(0) = \infty$.

The variance of an I(1) series comes from the low frequencies of the series which tend to be smooth and consist of long waves compared to a I(0) series. As a result of the difference in the relative size the variances of an I(1) and I(0) series adding the two together will almost always result in an I(1) series.

Suppose that we have two non-stationary time series $X = (x_1, x_2, ;x_n)$ and $Y = (y_1, y_2, ;y_n)$ both of which are I(d). In addition assume that a and b are constants and $b \neq 0$. In general it will be the case that,

$$z_t = y_t - a \times x_t, \text{is also I(d)} \tag{5.2.1}$$

However, as Engel and Granger [1] point out, it is also possible that $z_t \sim I(d-b)$, where $b > 0$.

Definition

If X and Y are each integrated I(d), and a linear combination of X and Y is integrated I(d-b) where $b > 0$ then the series are said to cointegrate.

$$z_t = y_t - \alpha x_t, \text{where,} z_t \sim I(d-b) \tag{5.2.2}$$

For the time being we will focus on the case where we have two I(1) series. Later we will generalize to the case of N series with multiple cointegrating relationships. For the case where X and Y are both I(d) with d = b = 1. If X and Y cointegrate then $z = I(d-b) = I(0)$. A combination of two non-stationary processes are combined to create a stationary process. Or, two series with long run trends and infinite variances combine to form a stationary series that crosses zero often. The cointegrating vector, z, is the equilibrium error of X and Y.

5.2.1 Testing for Cointegration

Testing for cointegration between two series is a two step process:

1. Test each series for non-stationarity. A popular test of non-stationarity is the augmented Dickey-Fuller (ADF) test, which is available in most statistical software packages. If the null hypothesis cannot be rejected then each series is non-stationary.

2. Regress series X and Y against one another, and test the residuals for non-stationarity. If the null hypothesis for the ADF test is rejected the two series are assumed to cointegrate.

The ADF has 3 cases: 1) no intercept or trend; 2) an intercept; 3) an intercept and a deterministic trend; Each case has a different set of critical values. The augmented DF adjusts the original DF for autocorrelation. The specification for each of the 3 cases is as follows:

$$Case1 : \Delta y_t = \gamma \times \Delta y_{t-1} + \Sigma_{j=1}^{p} \phi_j \Delta_{t-j} + \eta_t \tag{5.2.3}$$

Case 1: The unit root test is carried out by testing the null hypothesis $\gamma = 0$ against $\gamma < 0$.

$$Case2 : \Delta y_t = \mu + \gamma \times \Delta y_{t-1} + \Sigma_{j=1}^{p} \phi_j \Delta_{t-j} + \eta_t \tag{5.2.4}$$

Case 2: The unit root test is carried out by testing the null hypothesis $\gamma = 0$ against $\gamma < 0$.

$$Case3 : \Delta y_t = \mu + \beta \times trend + \gamma \times \Delta y_{t-1} + \Sigma_{j=1}^{p} \phi_j \Delta_{t-j} + \eta_t \tag{5.2.5}$$

Case 3: The unit root test is carried out by testing the null hypothesis $\gamma = 0$ against $\gamma < 0$, or $\gamma = \beta = 0$ a

Example 5.1. As an illustration we will test the asset prices for Budweiser (BUD) and Molson-Coors (TAP) to see if they cointegrate. It seems reasonable to assume that they move together as both manufacture and market beer under well known brand names. But no attempt has been made to investigate these firms, and while they both sell beer one or both may be quite diversified. Figure 5.1 shows daily adjusted prices for both stocks from Jan 2010 through May 2021.[1]. Both stock prices follow a similar path, but BUD appears to be more volatile in recent years

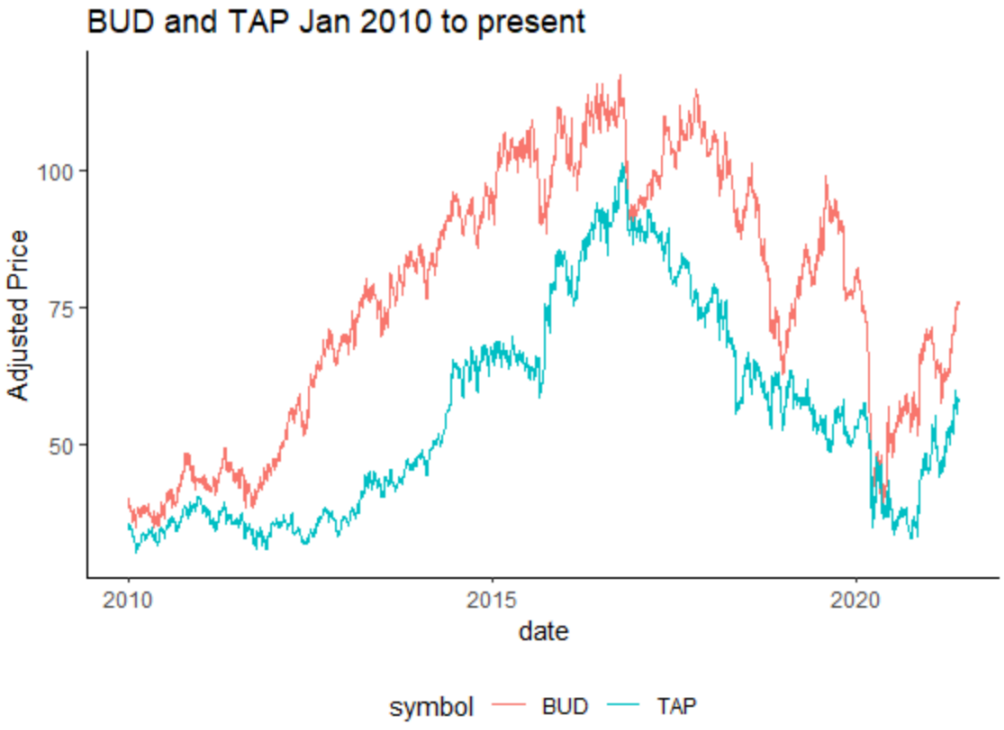

Figure 5.1: Adjusted Daily Stock Prices

ADF test results for the two stocks are shown in Table 5.1. The test assumed a "drift" term , and the number of lags were selected using the AIC statistic. In both cases we are unable to reject the null hypothesis that the series is a random

Value of test-statistic is:

	tau2	phi1
BUD	-1.8135	1.8122
TAP	-1.5683	1.3263

Critical values for test statistics:

	1pct	5pct	10pct
tau2	-3.43	-2.86	-2.57
phi1	6.43	4.59	3.78

Table 5.1: ADF Test Results for BUD and TAP

walk. Figure 5.2 shows the residuals from the regression of TAP prices against BUD prices. The residuals do not have a trend but there are long periods of time when they are far away from zero. This is particularly true for the first 7 years. Results for the ADF test on the residuals are shown in Table 5.2. The null hypothesis is rejected, and we conclude that BUD and TAP cointegrate.

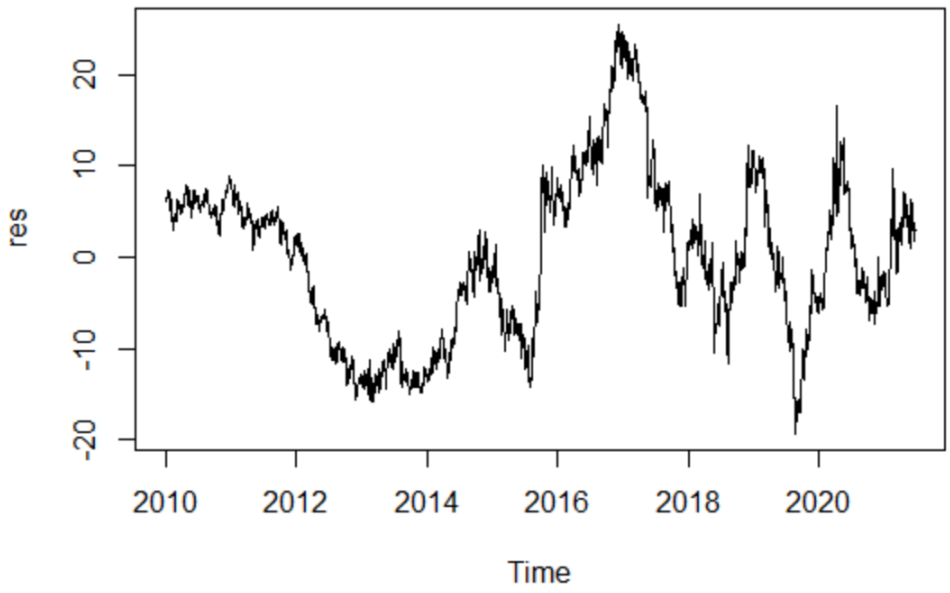

Figure 5.2: Residual series used to test for cointegration

5.3 Error Correction Models

In my experience, the error correction model (ECM) is one of the most useful tools available for forecasting time series over long horizons. By combining long run

[1]The adjusted price is adjusted for corporate actions such as dividend payments and stock splits

Value of test-statistic is:

	tau1
Residuals	-2.6617

Critical values for test statistics:

	1pct	5pct	10pct
tau1	-2.58	-1.95	-1.62

Table 5.2: ADF Test for Residuals

equilibrium relationships with short run deviations it takes advantage of information that exists in the levels of non-stationary series along with the stationary change in level. An error correction model is a dynamic system of two or more time series which reflects a long run equilibrium as well as short term deviations from the equilibrium. The ECM is specified as follows:

$$\Delta Y_t = \Delta X_t + \alpha \epsilon_{t-1} + u_t \tag{5.3.1}$$

where X_t ans Y_t are both integrated of order 1, I(1), and

$$\epsilon_{t-1} = Y_{t-1} = \beta X_{t-1} \tag{5.3.2}$$

The change in Y_t is due to the change in X_t and ϵ_{t-1} which is viewed as the disequilibrium term in the sense that it measures the extent to which Y_t and X_t deviate from their long run equilibrium values. If ϵ_t is non-zero the system is out of equilibrium. Note also that α determines the rate at which ϵ_t returns to zero. Looking at the ECM specification, we see that the model is a combination of levels and first differences of X and Y. For the ECM to be internally consistent X and Y must cointegrate. That is, if X and Y are I(1) then ΔX and ΔY are I(0). Therefore the equilibrium term, e_{t-1}, must also be I(0). It should be pointed out that the ECM pre-dates the concept of cointegration by about 25 years, so this insight by Granger and Engel was significant in reconciling the concept of cointegration with the ECM.

Granger Representation Theorem

ECM implies cointegration and cointegration implies ECM.

If X and Y are I(1) then ΔX and ΔY are I(0). Therefore the equilibrium term, e_{t-1}, must also be I(0). The converse also hold: if X is generated by an ECM then X is cointegrated.

5.3.1 Pairs Trading

A common application of cointegration in finance is pairs trading. Recall that equity prices are typically I(1). The idea behind a pairs trading strategy is to identify pairs of equities whose prices have a long run relationship (cointegrate). When the relationship deviates from equilibrium, buy the undervalued stock and short sell the overvalued stock. In this strategy, you track the spread between the stocks, and trade when the spread widens beyond the equilibrium level.

Highly-correlated pairs often (but not always) come from the same sector because they face similar systematic risks. [2]

5.4 VECM

5.4.1 Pesaran Model

[2]Caution: Two completely unrelated I(1) series may have a high correlation.

5.5 R Code for Examples

5.5.1 Example 5.1

```
library(tidyquant)
library(tibble)
library(urca)
options("getSymbols.warning4.0"=FALSE)
options("getSymbols.yahoo.warning"=FALSE)

tickers = c("BUD", "TAP")

prices <- tq_get(tickers,
from = "2010-01-01",
to = "2021-05-28",
get = "stock.prices")

ggplot(data = prices, aes(x = date, y = adjusted, color = symbol)) +
geom_line(aes(group = symbol)) +
theme_classic() +
theme(legend.position = "bottom") +
ylab("Adjusted Price") +
ggtitle("BUD and TAP Jan 2010 to present")

tappr <- subset(prices,symbol=="TAP")
bud<- subset(prices, symbol=="BUD")
summary(ur.df(bud$adjusted,type=c("drift"),selectlags=c("AIC")))
summary(ur.df(tappr$adjusted,type=c("drift"),selectlags=c("AIC")))
reg_pr<-lm(tappr$adjusted~bud$adjusted)
summary(reg_pr)
names(reg_pr)
res<-ts(reg_pr$residuals, frequency = 250, start=c(2010-01-01))
plot(res)
summary(ur.df(res,type=c("none"),selectlags=c("AIC")))
```

Bibliography

[1] Granger, C. W., and R. Engle. 1987. "Cointegration and Error-Correction: Representation, Estimation, and Testing." Econometrica 55:251–76.

Chapter 6

Regime Change

6.1 Introduction

Economics and financial times series are often subject to temporary (and sometimes permanent) changes in their statistical properties that can best be described as regime changes. These changes are often abrupt, and quite apparent in a plot of the data. For instance, and examination of daily returns for the S&P 500 shows very clear clusters of high volatility. (see Figure 1.1).

Regime changes, whether the are temporary jumps or longer term shifts in the mean or variance of an economic series constitute nonlinearities. A standard linear model cannot capture a regime change, so an alternative approach is needed.

The following simple model is often used to convey the basic ideas behind a switching regime model. Suppose that y_t is an observable stationary time series, and s_t is a random variable that takes a value of 1 if the economy is in an expansion, and 0 when the economy is in a recession. Assume that the following model describes y_t:

$$y_t = \mu_{st} + \phi_{st} y_{t-1} + \sigma_{st} + e_t \text{ where } e_t \sim iidN(0,1) \tag{6.1.1}$$

In this model the intercept, AR coefficient and volatility all change over time depending on the state. (i.e. $s = 0$ or $s = 1$) The regime indicator s_t is unobserved, and assumed to be a first Markov chain. That is, the probability of being in either of the two states at time t is conditional on the state of the previous period, and not the entire history of states.

$$Prob(s_t = j | s_{t-1} = i) = p_{ij}, \quad i, j = 0, 1 \tag{6.1.2}$$

Since we haveassumed that there are 2 states, there a total of 4 transition probabilities.

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

6.2 Hamilton's Switching Regime Model

6.2.1 The EM Algorithm

6.3 Structural Change and Unit Root Tests

6.3.1 Andrews-Zivot Test

6.4 Regime Changes and Financial Markets

Ang and Timmerman

Bibliography

[1] Ang, Andrew and Alan Timmerman, "Regime Changes and Financial Markets", Annual Review of Financial Economics, 2012, 4:313-337.

[2] Hamilton, James, "Analysis of Time Series Subject to Changes in Regime", Journal of Econometrics, July/August 1990.

[3] Hamilton, James D. "Specification Testing in Markov-Switching Time-Series Models." Journal of Econometrics, vol. 70, no. 1, Jan. 1996, pp. 127–157.

[4] Hamilton, James, "Macroeconomic Regimes and Regime Shifts", Handbook of Macroeconomics, Vol.2.

1

Chapter 7

Introduction to Simulation

This lecture is designed to provide the reader with the basic level of understanding needed to simulate from a posterior distribution in Bayesian inference. The lecture begins with a discussion on generating uniform random variables. In many instances this is the first step in generating non-uniform variates. This is discussed in section 2. The remainder of the lecture is focused on simulating from distributions with an unspecified functional form.

7.1 Simulating Uniform Random Variables

A common version of the pseudo-random number generator (PRNG) is defined in equation 7.1.1. Given an initial value or seed, this algorithm will produce a set of N uniform random variables in the range $[0, 1]$.

$$\frac{x_i}{M} = (kx_{i-1} + c) mod\, M \tag{7.1.1}$$

where k, c, and M are constants. Note that mod signifies the modulo operation, so that $X mod\, M$ returns the remainder after X is divided by M.

As an illustration, Figure 7.1 contains the output of 500 random uniform numbers using equation 7.1.1 and setting the $seed = 1234$, $k = 200$, $M = 6$ and $c = 0.1$. A Portmanteau test on the series is unable to reject the null hypothesis that of independence, and the coverage of the unit square seems reasonable. However, the coverage is not uniform. In general, the PRNG can have areas where it under-samples, and over-samples. Two areas where there is under-sampling are outlined in red in Figure 7.1.

The quasi-random number generator (QRNG) is an alternative to the pseudo-random number generator. The Halton sequence, defined in equation 7.1.2 is an example of a QNRG.

$$S_{t+1} = \left[S_t, S_t + \frac{1}{k^t}, S_t + \frac{2}{k^t}, \ldots, S_t + \frac{(k-1)}{k^t} \right] \tag{7.1.2}$$

where k is a prime number.

As an illustration of how the Halton sequence generates a set of uniform random numbers, begin with a prime number, say 3. Divide the unit line into thirds, then

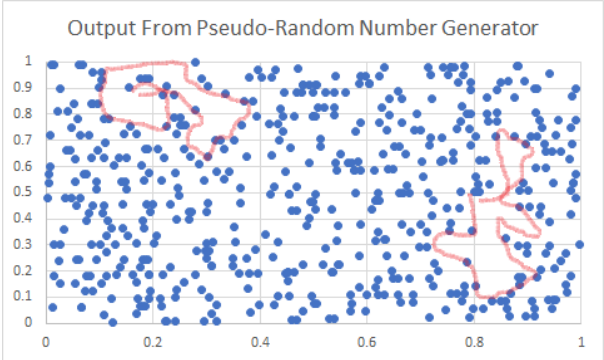

Figure 7.1: Uniform random numbers generated using equation

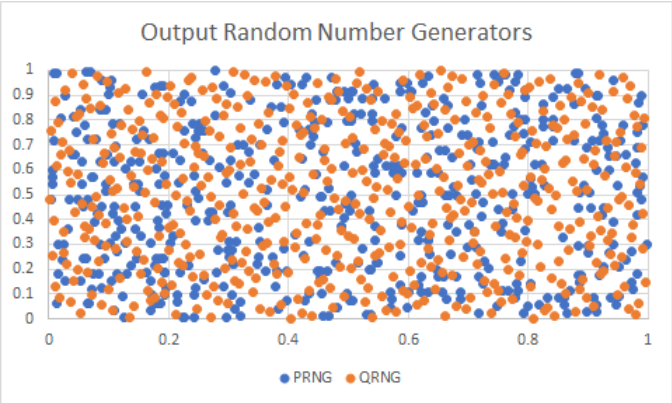

Figure 7.2: Comparison of QRNG and PRNG

ninths, then twenty-sevenths, etc. This results in the following sequence:

$$\left[0\,,1/3\,,2/3\,,1/9\,,4/9\,,7/9\,,2/9\,,5/9\,,7/9\,,1/27\,,10/27\,,19/27\,,4/27,\ldots\right] \quad (7.1.3)$$

Since the Halton sequence is defined on the unit line, it can be interpreted as random draws from a uniform distribution. Consecutive observations in the Halton sequence tend to be negatively correlated, so it generally provides better coverage than a pseudo-random generator. [1]

7.2 Simulating Non-Uniform Random Variables

Random numbers from non-uniform distributions are often based on an initial draw of a uniform random variable. One approach for deriving non-uniform random numbers is the inverse transform.

Definition: Inverse Transform Sampling Assume that X is a continuous random variable with a continuous distribution function, F, so that $Pr(X \leq x) =$

[1] Note that if you have two Halton sequences (based on different primes numbers) the first pairs in the two sequences will be highly correlated. If you need independent random pairs you must eliminate enough of the early pairs in the sequence to remove the correlation. Usually eliminating the first 10 or 15 pairs will suffice.

	CDF	Inverse
Cauchy	$1/2+(1/\pi)\arctan\ x$	$\tan(\pi U)$
Exponential	$1-e^{-x}$	$\log(1/U)$
Gumbel	$e^{-e^{-x}}$	$-\log\log(1/U)$
Logistic	$\frac{1}{1+e^{-x}}$	$-\log((1-U)/U)$
Pareto, $\alpha>0$,	$1-1/x^{\alpha}$	$1/U^{1/\alpha}$
Weibull, $\alpha>0$	$1-e^{-x^{\alpha}}$	$(\log(1/U))^{1/\alpha}$

Figure 7.3: Some Densities and Functions that are Explicitly Invertible

$F(x)$. Define the inverse of the cdf, F^{-1} as,

$$F^{-1}(u) = \inf\{x : F(x) = u, 0 < u < 1\} \tag{7.2.1}$$

If $U \sim U[0,1]$, then the random variable $F^{-1}(U)$ has the distribution F. To generate a random variable $X \sim F$, generate U from the Uniform distribution $U[0,1]$ and apply the transformation $x = F^{-1}(u)$.

If the inverse is easily computed, the inversion method is the easiest and fastest method of calculating a univariate random variable. Some useful transformations are shown in Figure 7.3.

Example 7.1. The Pareto Distribution
In this example the inverse transform method is used to sample 100 random observations from a Pareto distribution.[2] Figure 7.4 shows a QQ plot comparing the sample distribution to theoretical distribution. Given the relatively small sample size the fit seems most reasonable for values less than 2. Table **??** compares the 3 sample moments to the expected moments. The results support the QQ plot in that the sample variance and skew are quite different from the expected variance and skew.

	Theoretical	**Sample**
Mean	1.333	1.297
Variance	0.222	0.103
Skew	7.071	2.456

Table 7.1: Pareto Distribution Sample Moments vs. Theory

7.3 Importance Sampling

Importance sampling is a form of Monte Carlo integration. Suppose that we want to to estimate the expected value in equation 7.3.2 but we are unable to calculate

[2]See section 8.6.9 for the definition of a Pareto distribution.

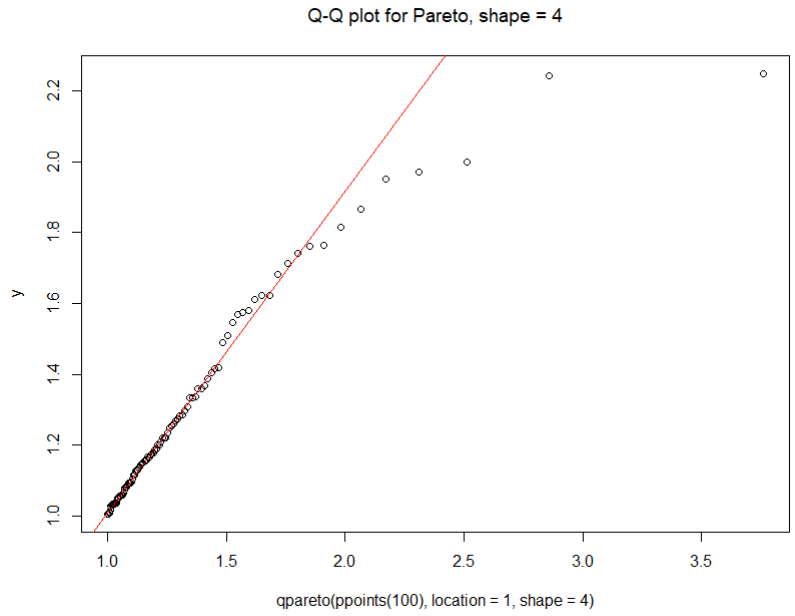

Figure 7.4: QQ Plot for Pareto Random Variables based on The Inverse Transform

the integral analytically.

$$E[g(\theta)] = \int g(\theta)p(\theta|x)d\theta \tag{7.3.1}$$

where $p(\theta|X)$ denotes the posterior density function.

Importance sampling (IS) introduces an auxiliary distribution $h(\theta)$ that can be sampled from, and rewrites the expected value in equation 7.3.2 as,

$$E[g(\theta)] = \int g(\theta)\frac{p(\theta|x)}{h(\theta)}h(\theta)d\theta \tag{7.3.2}$$

To implement IS, $\theta^i, i = 1, \ldots, N$ observations are drawn from $h(\theta)$. The sample is then used to create the following weights,

$$w_i = \sum_{i=1}^{N} \frac{p(\theta^i|x)}{h(\theta^i)} \tag{7.3.3}$$

Monte Carlo integration is then used to estimate the expected value of $g(\theta)$:

$$E[g(\theta)] = \sum_{i=1}^{N} w_i g(\theta^i) \tag{7.3.4}$$

Implementation of equation 7.3.4 does not account for the marginal distribution of the posterior, $p(\theta|x)$. That is, just using the weights as described does not estimate the normalizing constant in equation 7.3.5.

$$E[g(\theta)] = \int g(\theta)p(\theta|x)d\theta = \frac{\int g(\theta)p(x|\theta)p(\theta)d\theta}{\int p(x|\theta)d\theta} \tag{7.3.5}$$

The IS approximation for the entire posterior including the marginal density of y, can be estimated by normalizing the weights so that they sum to one.

Given a likelihood function $p(y|\theta, x)$ and a prior distribution, $p(\theta)$ IS can be used to sample from the posterior distributions as follows:

1. Draw N samples from the importance distribution, $h(\theta)$.

2. Compute the unnormalized weights
   $w^{*i} = \frac{p(x|\theta^i)p(\theta^i)}{h(\theta^i)}$

3. Compute normalized weights
   $w_i = \frac{w^{*i}}{\sum_{i=1}^{N} w^{*i}}$

4. The approximation of $E[g(\theta)]$ is
   $E[g(\theta)] \approx \sum_{i=1}^{N} w_i g(\theta^{)}$

Example 7.2. In financial risk analysis an important measure of loss is the expected value conditional on exceeding the quantile:

$$E[x|x < q] = \frac{\int_{-\infty}^{q} xf(x)dx}{\int_{-\infty}^{q} f(x)dx} \tag{7.3.6}$$

This measures the average size of the loss when it falls below the cutoff value. This measure is referred to as the expected shortfall, or expected tail loss, or Conditional VaR (CVaR).

For the standard normal distribution the expected loss is:

$$E[X|x < q] = \frac{-f(q)}{F(q)} \tag{7.3.7}$$

where f is the standard Normal pdf, and F is the standard Normal cdf.

To illustrate the IS algorithm, a random sample is drawn from the standard normal density, and the average value is estimated, for the subset of observations less than or equal to -2. Figure 7.5 shows the density for a sample of 1 million random draws from a standard Normal density. Figure 7.6 is the density for all values less than -2. The auxiliary distribution, shown in Figure 7.8 is Pareto with location and scale both equal to 2. It has a shape that is similar to 7.6, but the tail is much longer. [3] Using the IS method the $E[X|x < -2] = -2.364502$. Applying equation 7.3.7 provides a check of the IS result. The analytic solution yields $E[X|x < -2] =$.

7.4 The Perfect Sampler

Cassella(2001)

[3] When calculating the CVAR, the tail sample (Fig. 7.6) is multiplied by minus one.

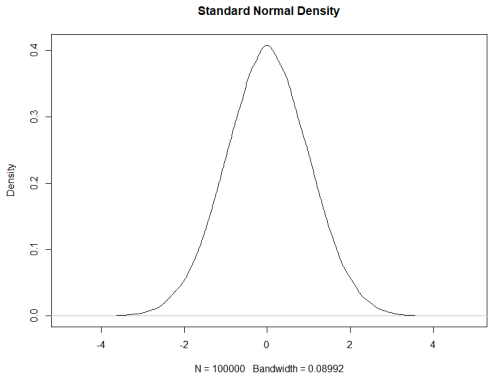

Figure 7.5: Std. Normal Density

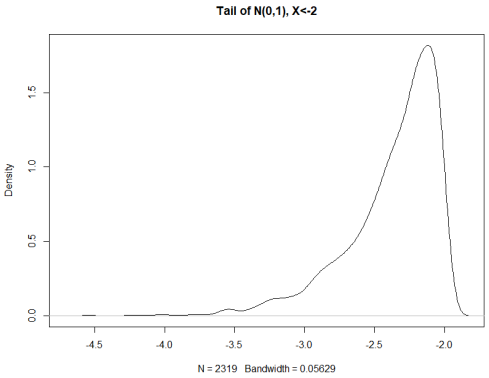

Figure 7.6: Tail Density

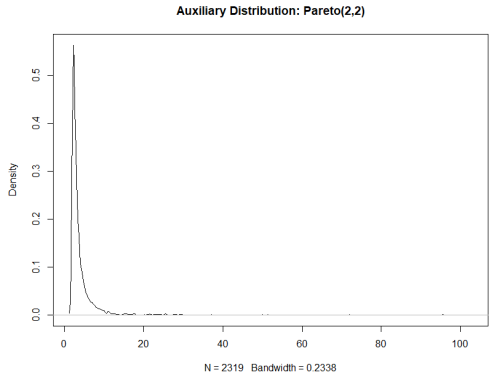

Figure 7.7: Auxiliary Distribution

7.5 Markov Processes

7.5.1 Ergodic Theorem

7.6 Metropolis - Hastings

7.6.1 Adaptive Metropolis-Hastings

Choosing an appropriate candidate density can be difficult when the one has very little information about the shape of the target distribution. There have been

7.7 Gibbs Sampler

Suppose that we have an estimate of a joint density $f(x, y_1, \ldots, y_p)$, and we are interested in obtaining characteristics of the marginal distribution,

$$f(x) = \int \cdots \int f(x, y_1, \ldots, y_p) dy_1 \ldots dy_p \tag{7.7.1}$$

One approach for obtaining $f(x)$ is to analytically solve the integral. An alternative approach is to use the Gibbs sampler.[4]

The Gibbs sampler is a procedure for finding the marginal density, f(x), without analytically solving the integral. Given a pair of random variables (X, Y) Gibbs generates a sample from f(x) by drawing a sequence of samples from the conditional distributions, $f(x|y)$ and $f(y|x)$.

A starting value $Y_0' = y_0'$ must be specified by the user, then the Gibbs sampler generates a sequence of random variables

$$X_0', Y_0', X_1', Y_1', \ldots, X_n', Y_n' \tag{7.7.2}$$

by drawing from the conditional distributions,

$$f(x_i'|Y_i' = y_i') \tag{7.7.3}$$

$$f(y_{i+1}'|X_i' = x_i') \tag{7.7.4}$$

Under general conditions, as $p \to \infty$ the distribution of X_p' converges to the true marginal distribution $f(x)$.

There are several possible approaches to sampling the marginal once the sequence has converged:

1. Sample the p^{th} or final value from many independent repetitions of the Gibbs sequence.

2. Generate one long Gibbs sequence and then extract every r^{th} observation. For r great enough the observations will be iid.

[4]This section is primarily based on the 1992 paper by Casella and George, [1]

3. Use all realizations of X_j' for $j > k$. The observations will not be iid, but the empirical series still converges to the true marginal density.

To understand how the Gibbs sampler works, consider the process of going from x_0 to x_1 in the case of two random variables x and y. Given x_0 the next step for the Gibbs sampler is to draw y_1. Given y_1 a value for x_1 can be drawn

This one step transition can be written as,

$$f(x_1|x_0) = \int f(x_1|y)f(y|x_0)dy \tag{7.7.5}$$

Similarly a k step transition matrix can be written as,

$$f(x_k|x_0) = \int f_{x_{k|k-1}}(x|t)f_{x_{k-1}|x_0}(t|x_0)dt \tag{7.7.6}$$

As $k \to \infty$ $f_{x_{k|k-1}}$ converges to the marginal distribution, $f(x)$ which is the stationary point of equation 7.7.6. The series of steps that define the sequence of a Gibbs sampler form a Markov chain. If the chain is recurent and does not have any absorbing states, the sequence will converge to a unique distribution which is the marginal distribution regardless of the starting point of the chain.

Example 7.3. In this example we use the Gibbs sampler to estimate the marginal density of the binomial variable in a beta-binomial posterior density. This example is a replication of an example in Casella and George [1]

One of the advantages to using the Gibbs sampler is that conditional distributions are often relatively easy to derive. Consider the case when the posterior distribution is the beta-binomial,

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-Y)^{n-x+\beta-1} \tag{7.7.7}$$

The marginal densities are,

$$f(x|y) \sim Bin(n, y) \tag{7.7.8}$$

$$f(y|x) \sim Beta(x+\alpha, n-x+\beta) \tag{7.7.9}$$

Note that is a discrete variable and y is continuous. The analytical solution for f(x) is:

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \tag{7.7.10}$$

for $x = 0, 1, \ldots, n$,

Following Cassela and George, two samples of size m=500 are drawn assuming, $n = 16$, $\alpha = 2$ and $\beta = 5$. One sample is based on the analytic solution and the other is based on the Gibbs sampler. The histograms in Figure 7.3 compare the sample two samples. The dark green area is the overlap between the two samples. There is no systematic difference between the Gibbs sampler and the analytic results.

Why does Gibbs work

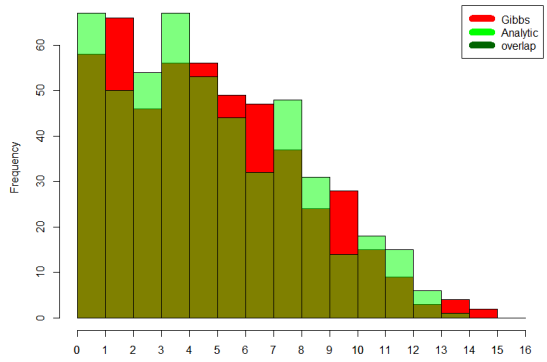

Figure 7.8: Beta-Binomial, Gibbs Sampler vs.Analytic Solution

7.8 Sequential Importance Sampling

7.9 Particle Filters

7.10 Exercises

1. Show the steps used to derive the Halton sequence in the illustration (see 7.1.3).

7.11 R Code for Examples

7.11.1 Example

```
library(EnvStats)
library(moments)
rp<-1/(runif(100,min=0,max=1)^(1/4))
#rp<-rp[rp<5]
plot(density(rp),main="Random Pareto Observations",col="Red",lwd=2.5)
mean(rp)
var(rp)
skewness(rp)
kurtosis(rp)
```

7.11.2 Example

```
library(EnvStats)
x<-rnorm(100000,0,1)
plot(density(x),main="Standard Normal Density")

xs<- -x[x< -2] #Flip the sign
plot(density(xs),main="Tail of N(0,1), X<-2")
set.seed(1240)
hs <- rpareto(length(xs),location=2,shape=2)  #sample from proposal density

plot(density(hs),main="Auxiliary Distribution: Pareto(2,2)")
wstar <- dnorm(hs,0,1)/dpareto(hs,location=2,shape=2)  # importance weights
w <- wstar/ sum(wstar)## Normalized weights
sum(-w*xs) #Approximation of expected shortfall

-dnorm(-2)/pnorm(-2) #Analytic approach
```

7.11.3 Example

```
library(VGAM)
a=2;b=4;n=16;k=10;size=500
analyticfx<-rbetabinom.ab(size,n,a,b) #//Draw 500 obs. From betabinomial distributi
nrep=1500
y<-matrix(0,nrep,1)
x<-matrix(0,nrep,1)
gibbsfx<-matrix(0,size,1)
y[1]<-0.1
x[1]<-rbinom(1,n,y[1])
for(i in 2:size) {  #//Run Gibbs Sampler for 1,500 repetitions.
y[i]<-rbeta(1,x[i-1]+a,n-x[i-1]+b)
x[i]<-rbinom(1,n,y[i])
}
```

```
gibbsfx<-x1[1001:1500] #marginal density f(x)

hist(gibbsfx, breaks=seq(from =0, to=16,by=1),freq=TRUE,col="red",main="",xlab="")
hist(analyticfx,add=T,col=rgb(0,1,0,0.5))
axis(side=1, at=seq(0,16,1),labels=seq(0,16,1))
legend("topright", c("Gibbs", "Analytic", "overlap"), col=c("red", "green","dark gr
```

Bibliography

[1] Casella, George and Edward I. George. "Explaining the Gibbs Sampler", The American Statistician, August 1992, Vol. 46, No. 3.

[2] Casella, George, Michael Lavine, and Christian P.Robert, "Explaining the Perfect Sampler", The American Statistician, 55:4, 299-305.

[3] Chib, Siddhartha and Edward Greenberg. "Understanding the Metropolis-Hastings Algorithm", The American Statistician, November 1995, Vol. 49, No. 4.

[4] Creal, Drew. "A Survey of Sequential Monte Carlo Methods for Economics and Finance", Econometric Reviews, 31(3):245–296, 2012.

[5] Devroye, Luc. "Non-Uniform Random Variate Generation", Springer-Verlag New York, 1986.

[6] Gelfand, Alan E. and Adrian F. M. Smith."Sampling-Based Approaches to Calculating Marginal Densities", Journal of the American Statistical Association June 1990,Vol. 85, No. 410, Theory and Methods.

[7] Geman, Stuart and Donald Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, NO. 6, November 1984.

[8] Geweke, John, "Bayesian Inference in Econometric Models Using Monte Carlo Integration", Econometrica, Vol. 57, No. 6 (Nov., 1989), pp. 1317-1339.

[9] Hastings, W.K., "Monte Carlo sampling methods using Markov chains and their applications", Biometrika (1970), 57, 1, p. 97.

[10] Robert, Christian and George Casella. "Monte Carlo Staistical Methods", Springer-Verlag New York, 2004.

[11] Sarkka:2013) Särkkä, S. Bayesian Filtering and Smoothing (Institute of Mathematical Statistics Textbooks). Cambridge: Cambridge University Press. (2013)

[12] Vihola, Matti. "Robust adaptive Metropolis algorithm with coerced acceptance rate", Statistics and Computing 22(5)"997-1008, 2012.

Chapter 8

Introduction to Bayesian Inference

This lecture is the first of a three lectures on Bayesian inference, Monte Carlo simulation, and applications in Finance. We begin by discussing the basic components of Bayesian statistical inference: the prior, the likelihood, and posterior distributions. The lecture ends with discuss model verification, and forecasting.

8.1 Bayes Theorem

Given two events A and B, the conditional probability of event A given event B is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{8.1.1}$$

The intersection of events A an B can be written in either of two ways,

$$P(A \cap B) = P(A|B)P(B)$$
$$P(A \cap B) = P(B|A)P(A)$$

Following Bayes we set these two equations equal to one another,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{8.1.2}$$

Equation 8.1.2 is Bayes theorem. It is a rule for calculating the conditional probability of event A given event B from the conditional probability of B given each A, and the unconditional probabilities of A and B. Bayes theorem can be generalized for multiple events, A_i for

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{i=k}(B|A_i)P(A_i)} \tag{8.1.3}$$

where $P(B) > 0$, and $i = 1, \ldots k$.

Equation 8.1.3 is a good starting point for understanding the components of Bayesian inference. $P(A_i)$ is a prior probability in the sense that it summarizes ones beliefs about the probability of event A_i before observing either event A_i

or event B. $P(B|A_i)$ summarizes the likelihood of event B if event A_i occurs. The denominator is the the marginal probability, $P(B)$. Note that is is calculated as the sum of the numerator quantities for all k events. It plays the role of a normalizing constant, ensuring that the conditional probabilities sum to one. The term on the left hand side of the equation, $P(A_i|B)$ is the posterior probability of event A_i after observing event B.

We can use Bayes Theorem for inference. Let y be a vector or matrix of data and let θ be a vector or matrix of parameters for a model that seeks to explain y. Using Bayes theorem we can write,

$$P(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{8.1.4}$$

where $p(\theta)$ is the prior distribution, $p(y|\theta)$ is the likelihood, and $P(\theta|y)$ is the posterior distribution. The numerator of the right had side of **??** is the un-normalized posterior. The normalizing constant is $P(y)$ is typically referred to as the marginal likelihood.

$$P(y) = \int_\theta p(\theta)p(y|\theta)d\theta \tag{8.1.5}$$

Equation 8.1.4 shows us that Bayesian inference is an updating procedure whereby prior beliefs of parameter values are updated based on observed data.

Definition: Kernel
If the probability density function of a random variable, y, can be written as $p(y) = kg(y)$, then $g(y)$, is called the kernel of the function. The term k, which is the portion of the equation that is not a function of y, is a numerical constant whose role is to ensure that $p(y)$ integrates to one. The kernel of the Binomial distribution is [1]

$$p(\theta) \propto \theta^y(1-\theta)^{n-y} \tag{8.1.6}$$

where N is the total number of trial and y is the number of successful trials. The kernel of the Beta distribution is

$$p(y) \propto y^{\alpha-1}(1-y)^{\beta-1} \tag{8.1.7}$$

and k is the Beta function.

The kernel of the univariate Normal distribution is,

$$p(y) \propto exp\Big[(-\tau/2)(y-\mu)^2\Big] \tag{8.1.8}$$

where $\tau = precision = 1/variance$.

8.1.1 Specifying the Likelihood, $p(y|\theta)$

The likelihood function, $p(y|\theta)$ describes the plausibility of a data set, y, given a particular set of parameters, θ. For a set of n independent and identically distributed

[1] See section 8.6.3 for definitions of the distributions.

(iid) observations data points, $y_i, i = 1, \ldots, n$ the likelihood function is the product of the densities of the n data points:

$$L(y|\theta) = p(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta) \tag{8.1.9}$$

Example 8.1. The likelihood function for a single coin toss can be written as a Bernoulli,

$$p(y_i|\theta) = (\theta)^{y_i}(1-\theta)^{y_i} \tag{8.1.10}$$

In the case of n independent coin tosses the likelihood function is,

$$p(y_i|\theta) = \prod \theta^{y_i}(1-\theta)^{y_i} = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \tag{8.1.11}$$

Example 8.2. Consider a single observation y from a Normal distribution with a mean μ and variance σ^2. For a single observation the likelihood function is,

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{1}{2\sigma^2}(y_i-\mu)^2} \tag{8.1.12}$$

For a sequence of n iid observations of y_i the likelihood is,

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} \tag{8.1.13}$$

Now, consider a generalization of the univariate Normal distribution where the k-dimensional vector $Y = [y_1, y_2, \ldots, y_k]$ with mean μ and variance covariance matrix Σ. Denote this vector as,

$$X \sim N(\mu, \Sigma) \quad \text{where,} \quad \mu \sim kx1, \Sigma \sim kxk \tag{8.1.14}$$

The probability density function of Y is,

$$p(Y|\mu, \Sigma) = \frac{1}{2\pi^{k/2}}|\Sigma|^{-1/2} exp\Big[-\frac{1}{2}(y-\mu)'(y-\mu)\Big] \tag{8.1.15}$$

Example 8.3. The likelihood for a finite mixture distribution is,

$$L(Y|\pi_k, \mu_k, \Sigma_k) = \prod_i \sum_i \pi_k \phi(Y|\mu_k, \Sigma_k) \tag{8.1.16}$$

Where each k=1,...,K latent sub-distributions, and i = 1,...,N observations of y_i.

The finite mixture model can be applied to any distribution. Consider the following mixture of two bivariate Normal distributions. [2]

$$f(\mu, \mathbf{\Sigma}|\mathbf{Y}) = \pi\phi(\mu_\mathbf{1}, \mathbf{\Sigma_1}) + (\mathbf{1}-\pi)\phi(\mu_\mathbf{2}, \mathbf{\Sigma_2}) \tag{8.1.17}$$

$$L(\mu, \mathbf{\Sigma}|\mathbf{Y}) = \prod_{\mathbf{i=1}}^{\mathbf{N}} \Big[\pi\phi(\mu_\mathbf{1}, \mathbf{\Sigma_1}) + (\mathbf{1}-\pi)\phi(\mu_\mathbf{2}, \mathbf{\Sigma_2})\Big] \tag{8.1.18}$$

where, for $k = 1, 2$

$$\phi(y|\mu_k, \Sigma_k) = (2\pi)^{0.5}(|\Sigma_k|)^{-0.5} e^{-0.5(y_i-\mu_k)'\Sigma(y_i-\mu_k)} \tag{8.1.19}$$

[2] See section 8.6.8 for a definition of the bivariate Normal distribution

8.1.2 Specifying the Prior Distribution, $p(\theta)$

The explicit use of a prior distribution in Bayesian inference is a key point of distinction from classical regression analysis. The prior represents the modelers beliefs about the unobserved parameters prior to observing the data. The choice of a prior may be driven by a number of considerations.

1. In the absence of strong prior beliefs the modeler may choose a diffuse or weakly informative prior.

2. Alternatively, lacking strong prior beliefs the modeler may try a host of different priors in the context of a sensitivity analysis.

3. The modeler may select a particular prior because it is a conjugate prior for the likelihood function chosen for the analysis.

4. The problem under consideration may be best modeled with an hierarchical prior.

Conjugate Priors

If the product of the prior and likelihood results in a posterior distribution in the same family of distributions as the prior, then the prior is a conjugate prior.[3] Working with conjugate priors makes Bayesian inference easier in the sense that the functional form of the posterior distribution is known. [4] Table 8.1 contains a list of commonly used conjugate priors and their likelihoods.

Likelihood	**Conjugate Prior**
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Negative Binomial	Beta
Multinomial	Dirichlet
Normal with known variance, σ^2	Normal
Normal with known mean, μ	Inverse Gamma
Normal	Normal Inverse Gamma
Multivariate Normal	Normal-Wishart

Table 8.1: Likelihoods and Conjugate Priors

Example 8.4. The beta prior, binomial likelihood, and the beta posterior are an obvious example of the benefits of using a conjugate. Given a binomial likelihood and a prior that is $Beta(\alpha, \beta)$, the posterior $Beta(y + \alpha, n - y + \beta)$. This model is typically referred to as the beta-binomial model.

$$p(\theta) \propto \theta^{\alpha-1}(\theta)^{\beta-1} \tag{8.1.20}$$

[3] Two distributions are in the same family if they have the same form and different parameters.
[4] But conjugate priors are not required for Bayesian inference.

$$p(y|\theta) \propto \theta^y(1-\theta)^{n-y} \tag{8.1.21}$$

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \tag{8.1.22}$$

Example 8.5. The Normal Gamma prior is used in the specification of the Bayesian linear regression. Given a vector,Y, consisting of n IID observations, the likelihood function assuming Normality is shown in equation 8.1.25.

$$Y = (y1, y2, \ldots, y_n) \tag{8.1.23}$$

$$y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2) \tag{8.1.24}$$

$$L(\mu, \tau) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \tau^{\frac{1}{2}} exp\bigg[-\frac{1}{2}\tau(y_i - \mu)^2 \bigg] \tag{8.1.25}$$

where μ is the mean, σ^2 is the variance and $\tau = frac1\sigma^2$ is the precision. The priors for the mean and precision are,

$$\mu|\sigma^2 \sim N(\mu_0, n_0\sigma^2) \tag{8.1.26}$$

$$\tag{8.1.27}$$

8.1.3 The Posterior

The posterior distribution is the product of the likelihood function and the prior distribution. The result is a joint posterior distribution which can the be used evaluate the marginal distributions of each parameter. In instances where the posterior distribution does not have an analytic solution simulation is used for analysis.

Note the conceptual difference between Bayesian inference and maximum likelihood. ML selects the parameter set, θ, that maximizes the probability of observing the actual sample of data, y, that we observe. Bayesian inference updates the modelers prior beliefs regarding the distribution of the parameters based on the observed data. The output from the ML approach is a set of point estimates of the parameters. The output from the Bayesian approach is a joint distribution of the parameters. As we will see later on in this Section, the posterior can be viewed as a weighted average of the prior and likelihood.

A Beta-Binomial Example

Consider a coin tossing experiment, where $\theta = p(head)$ is distributed as $Beta(\alpha, \beta)$ and the sampling distribution of heads and tails is $Bin(n, \theta)$ where n is the total number of trials (heads + tails). We know from the discussion on conjugate priors that the posterior distribution is Beta. For this experiment the kernel of the posterior is,

$$p(\theta|y) \propto \theta^{h+\alpha-1}(1-\theta)^{t+\beta-1} \tag{8.1.28}$$

where h is the number of heads and t is the number of tails. Since the posterior is $Beta(h+\alpha, t+\beta)$ the expected value is,

$$E[\theta|x] = \frac{h+\alpha}{\alpha+\beta+n} \tag{8.1.29}$$

Let $w = \frac{\alpha+h}{\alpha+\beta+n}$, then the expected value can be written as,

$$E[\theta|x] = w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{h}{n} \tag{8.1.30}$$

The expected value of the posterior estimate of θ is a weighted average of the expected value of the prior distribution, and the average number of heads in the trial data. The weight, w, represents the relative size of the prior in terms of the number of trials. That is, α and β can be viewed as equivalent to the number of heads and tails in the prior information. As the sample size increases, the relative weight of the prior decreases, and the expected value of the posterior approaches the maximum likelihood estimate $\hat{\theta} = h/n$. Alternatively, as $\alpha \to 0$ and $\beta \to 0$ the effective sample size of the prior goes to zero and the expected value of the posterior mean approaches the MLE.

8.1.4 Normal-Gamma Inverse Model

8.2 Posterior Sampling

8.3 Model Verification

8.4 Forecasting

8.4.1 Non-informative Priors

8.5 Applications

8.5.1 Improving GDP Measurement

In this section we apply ffbs tp the estimation of mismeasurement in GDP as described by Arouba et.al (2015). The sample period is to growth rates in real GDP for the period through Q32018. We simulate for 60,000 iterations, and toss out the first 30,000.The results are shown in Figures 8.2 to 8.4.

```
> mean(outdat[m:n,1])
mu: 2.95856
> mean(outdat[m:n,2])
rho: 0.6139805
> mean(outdat[m:n,3])
SigGG: 7.828538
> mean(outdat[m:n,4])
SigEE: 4.642055
> mean(outdat[m:n,5])
SigII: 2.556732
> mean(outdat[m:n,6])
SigGE: 1.762951
```

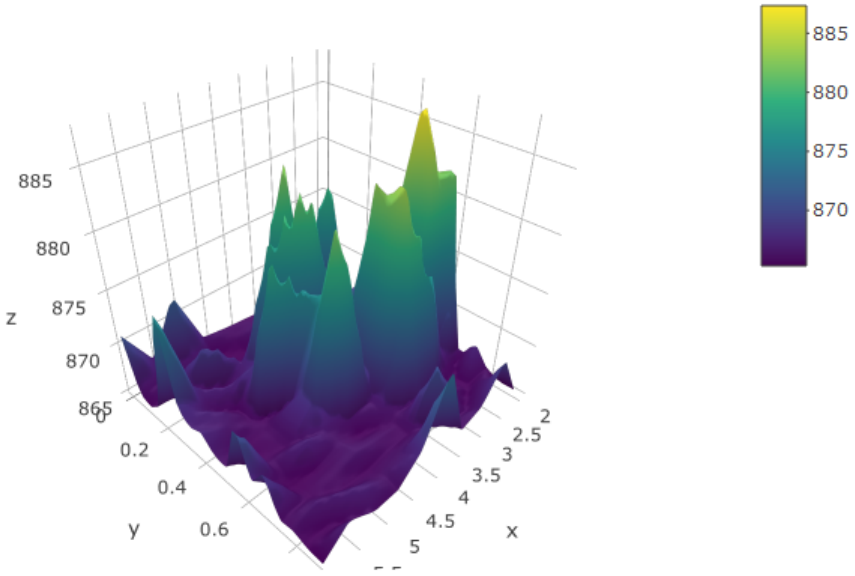

Figure 8.1: Log Likelihood Surface - Grid Search

```
> mean(outdat[m:n,7])
SigGI: -0.07106606
> mean(outdat[m:n,8])
SigEI: -0.7500625
>
```

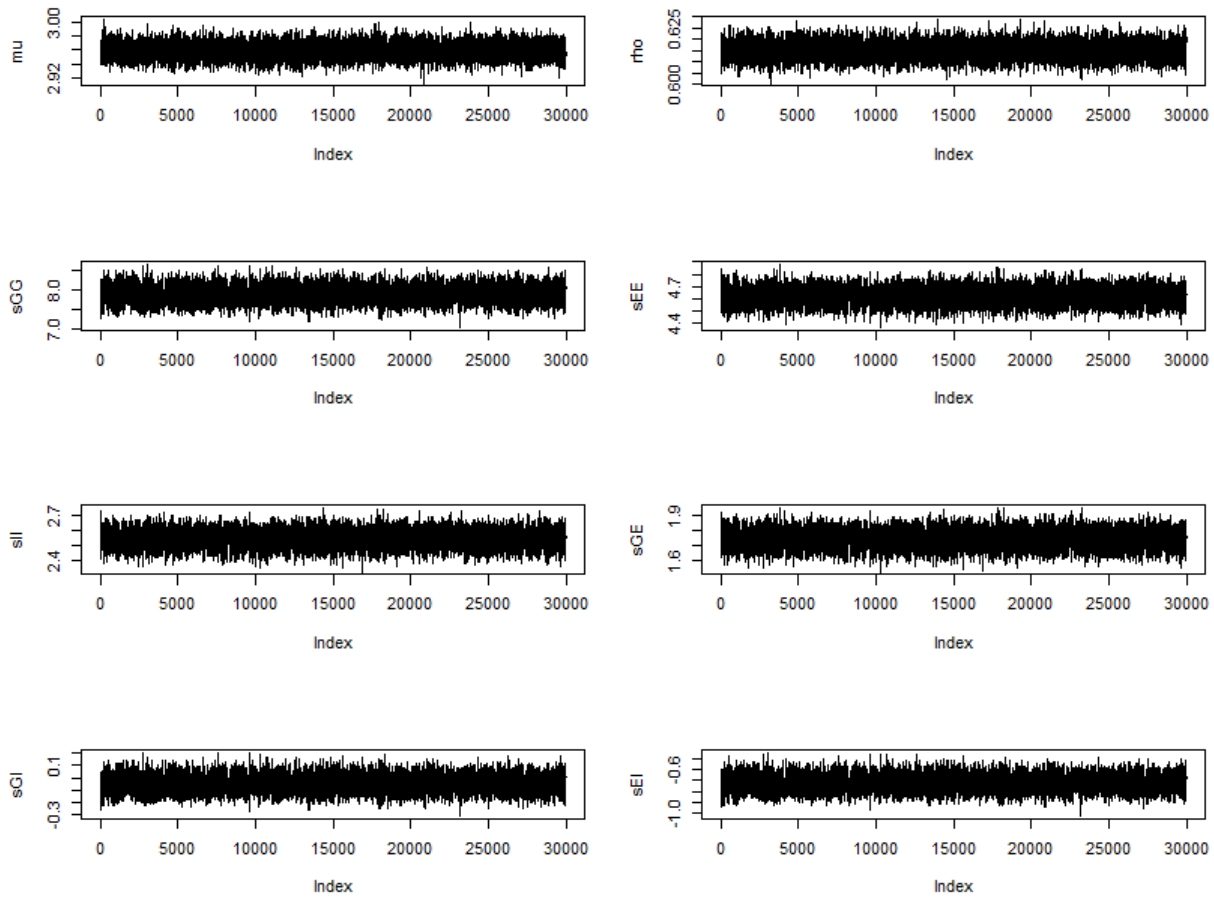

Figure 8.2: Metropolis-Hastings Parameter Sequences

8.6 Distributions

In this section we define several of the distributions that we will be working with throughout the lectures on Bayesian inference.

8.6.1 Bernoulli Distribution

A Bernoulli random variable y is a discrete random variable which takes the value 1 with probability θ, and 0 with probability $1 - \theta$. The probability mass function is,

$$p(y|\theta) = \theta^{y_i}(1-\theta)^{1-y_i} \tag{8.6.1}$$

. The mean and variance are defined as

$$E[y] = \theta \tag{8.6.2}$$

$$Var[y] = \theta(1-\theta) \tag{8.6.3}$$

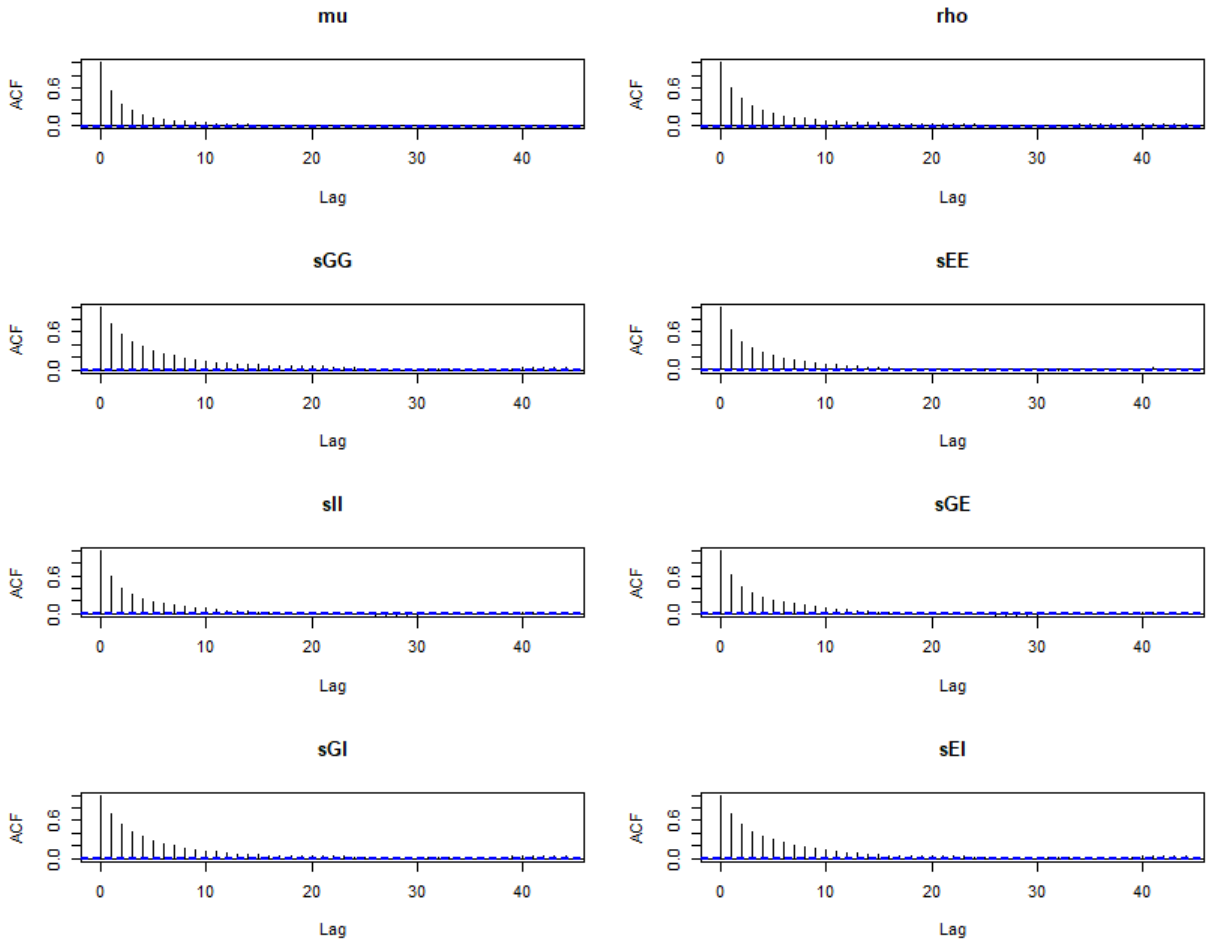

Figure 8.3: Metropolis-Hastings ACF of Parameter Draws

8.6.2 Binomial Distribution

If $x_1, x_2, \ldots x_n$ are iid Bernoulli with success probability θ then

$$\sum_{k=1}^{n} y_i \sim Binomial(n, \theta) \tag{8.6.4}$$

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta^{n-y}) \tag{8.6.5}$$

$$\text{where,} \quad \binom{n}{y} = \frac{n!}{y!(n-y)!} \tag{8.6.6}$$

where n=number of trials, and y=number of successes. The mean and variance are defined as

$$E[y] = n\theta \tag{8.6.7}$$

$$Var[y] = n\theta(1-\theta) \tag{8.6.8}$$

The binomial is the discrete probability distribution of the number of successes in a sequence of n independent Bernoulli experiments each with a probability θ of success. When n=1, the Binomial is Bernoulli.

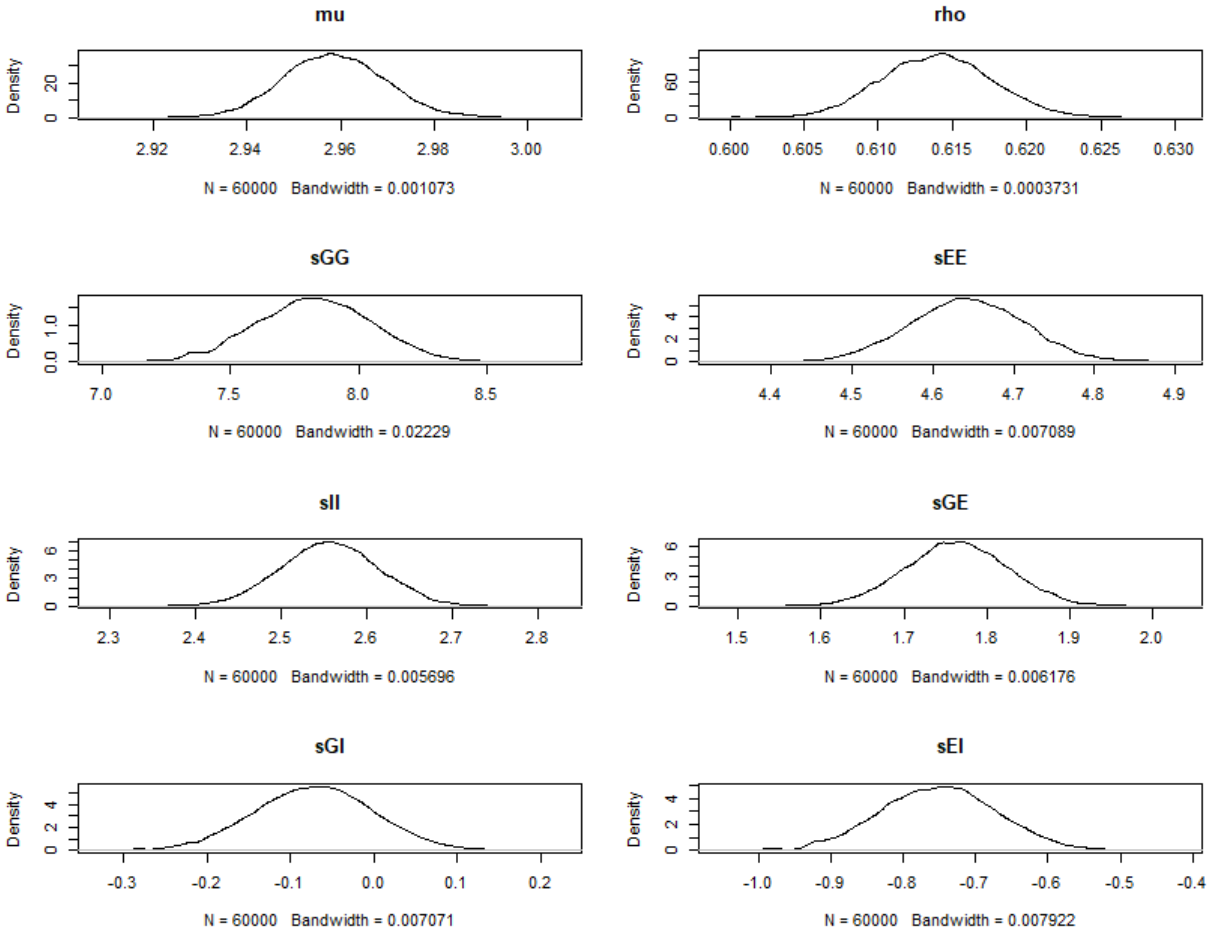

Figure 8.4: Metropolis-Hastings Parameter Density

8.6.3 Beta Distribution

The Beta is a family of continuous probability distributions defined on the interval $[0, 1]$. The random variable $y \sim Beta(\alpha, \beta)$ if $0 < y < 1$ and

$$p(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \tag{8.6.9}$$

$$\text{where,} \quad \Gamma(\alpha) = (\alpha - 1)!. \tag{8.6.10}$$

Where $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalizing constant. It is defined as,

$$B[\alpha, \beta] = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy \tag{8.6.11}$$

$B[\alpha, \beta]$ is called the Beta function. As illustrated in Figure 8.6, the beta density can take on many shapes depending on the value of the two shape parameters *alpha* and β.

- Beta(1/2,1/2) is u-shaped.

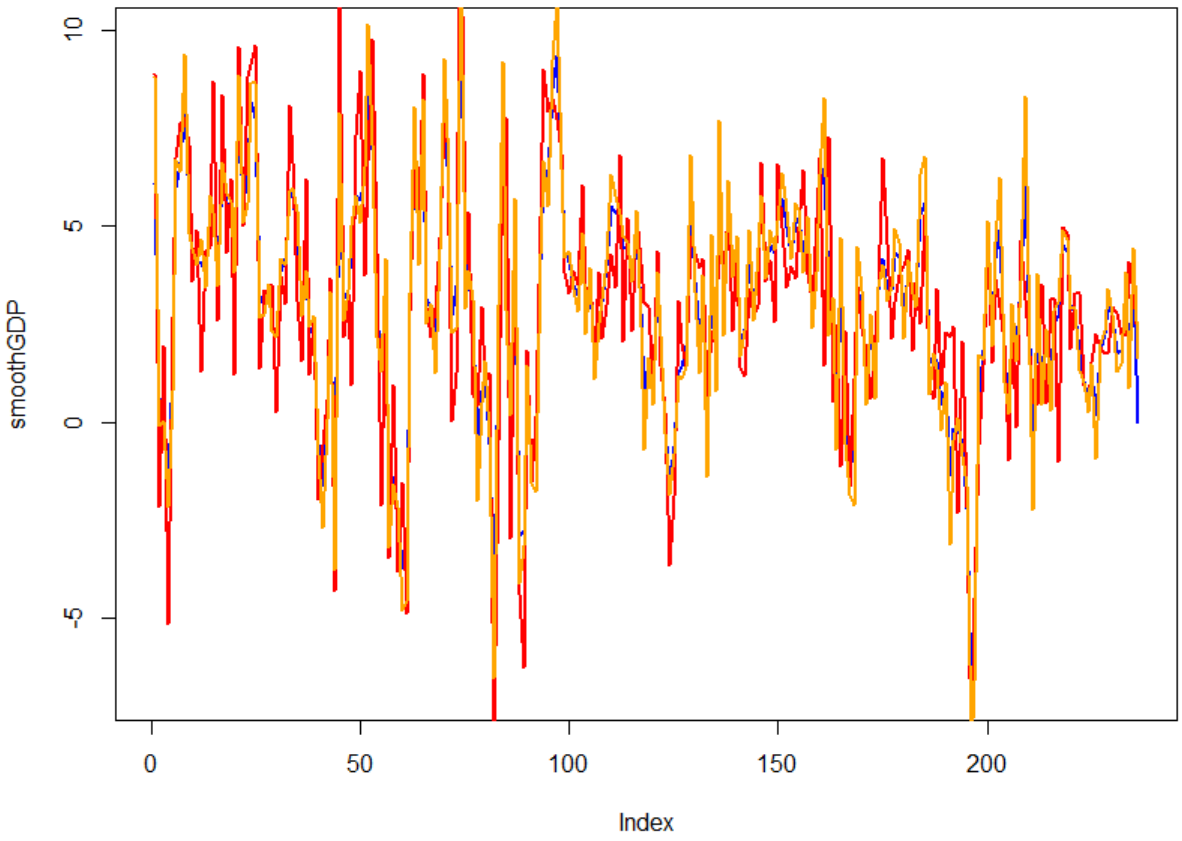

Figure 8.5: GDE,GDI and smoothed GDP

- Beta(1,1) is the uniform distribution
- When $\alpha = \beta$ the distribution is symmetrical about 1/2.

The Beta distribution is a common choice for a prior distribution because it can produce a wide variety of shapes. It is a conjugate prior for the Bernoulli distribution, meaning that the product of the Bernoulli likelihood and a Beta prior yields a posterior with a Beta distribution . It is a conjugate prior in the sense that the prior and posterior are in the same family.

Note that a $Binomial(y, n)$ random variable is $Beta(y + 1, n - y + 1)$.

8.6.4 Cauchy Distribution

The probability density function of the Cauchy distribution is (eq. 8.6.12):

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)\right]} \tag{8.6.12}$$

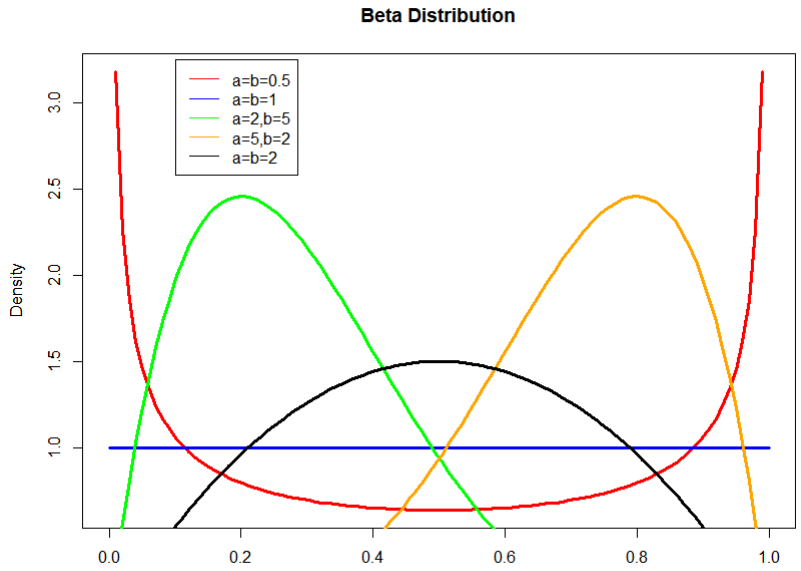

Figure 8.6: The Beta Distribution for Different Parameter Values

where x_0 is a location parameter, and γ is a scale parameter. The Cauchy is an interesting distribution in that the mean and all of the higher moments are undefined. That is, the value of the integral for the expected value of Cauchy random variable are undefined. The median and mode are defined, and are equal to the location parameter x_0. The $student-t$ distribution with one degree of freedom is Cauchy. The ratio of two independent standard Normal random variables is Cauchy.

8.6.5 Gamma Distribution

The Gamma is typically used to model waiting time. i.e. time until failure The density function for the Gamma distribution is:

$$p(y) = \frac{y^{\alpha-1}e^{-\beta y}}{\Gamma(a)\beta^{-\alpha}} \tag{8.6.13}$$

The mean and variance are:

$$E[y] = \frac{\alpha}{\beta} \tag{8.6.14}$$

$$Var[y] = \frac{\alpha}{\beta^2} \tag{8.6.15}$$

where $y > 0$, $\alpha > 0$, and $\beta > 0$. The gamma family is a generalization of the exponential family which is Gamma with $\alpha = 1$. The χ^2 distribution is gamma $\alpha = \nu/2$ and $\alpha = 1/2$ where ν is positive parameter, usually an integer. Figure 8.7 illustrates the variety of shapes that the Gamma family can assume. Note that unlike the beta distribution, the range of the gamma distribution is unlimited for positive values.

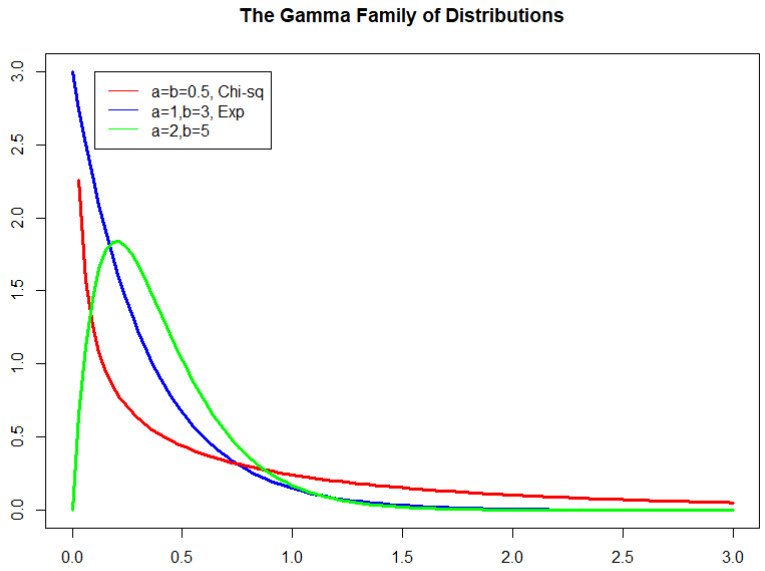

Figure 8.7: The Gamma Distribution for Different Parameter Values

8.6.6 Negative Binomial Distribution

Given a sequence of Bernoulli trials, if we observe the sequence until a certain number of successes, k, occurs, then the distribution of k will be negative binomial. The density is:

$$p(y=k) = \binom{k+r-1}{k}\theta^k(1-\theta)^k \tag{8.6.16}$$

Where k = number of successes and r = number of failures.

8.6.7 Normal Distribution

The probability density function for the Normal distribution is,

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{1}{2\sigma^2}(y_i-\mu)^2} \tag{8.6.17}$$

where $E[y] = \mu$ and $var(y) = \sigma^2$

8.6.8 Bivariate Normal

Let X_1 and X_2 be bivariate normal random variables, The bivariate density can be written as,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$$

where μ_1 and μ_2 are the mean of X_1 and X_2, respectively. $\Sigma_{i,j}$, $i, j = 1, 2$ are the variance and covariance of X_1 and X_2. One of the properties of the bivariate Normal distribution is that given the joint density of X_1 and X_2 there is a closed form solution for the conditional densities $X_1 \mid X_2$ and $X_2 \mid X_1$. The converse also holds, meaning that given the conditional densities we can write the joint density. The conditional of $X_1 \mid X_2$ is defined as,

$$X_1 \mid X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{8.6.18}$$

Figure 8.8 shows a bivariate Normal density with zero correlation between X1 and X2.

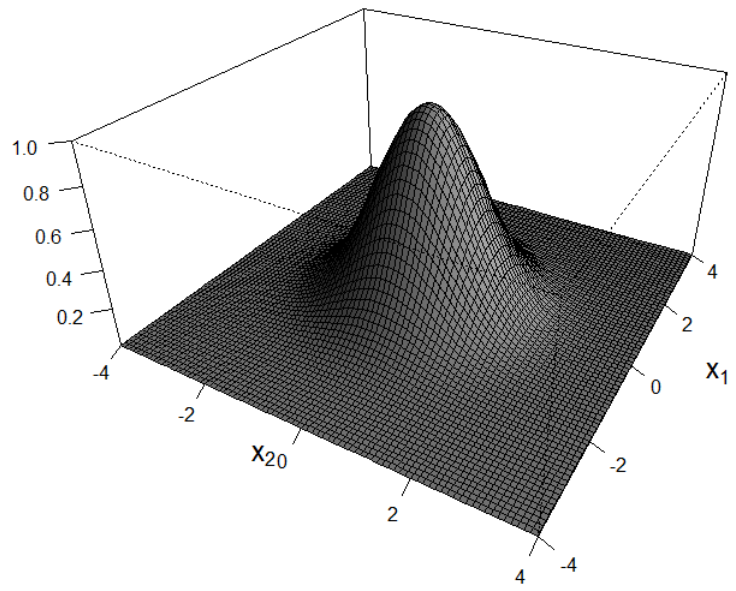

Figure 8.8: Bivariate Normal Density

8.6.9 Pareto

The Pareto density function is,

$$f(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^2 & \text{if} x_m \geq x_m \\ 0 & \text{if} x < x_m \end{cases}$$

where x_m is the minimum possible value of x. x_m is the scale parameter, and α is the shape parameter. Figure 8.9 shows the Pareto density function with $x_m = 1$, and $\alpha = 3$.

$$E(x) = \begin{cases} \frac{\alpha x_m}{x^{\alpha+1}} & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \leq 1 \end{cases}$$

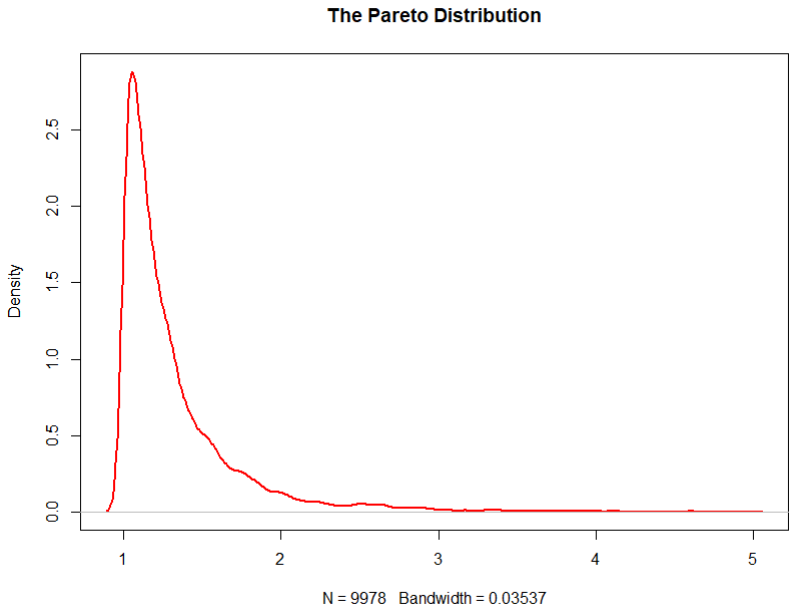

Figure 8.9: Pareto Probability Density Function

$$Var(x) = \begin{cases} \left(\frac{\alpha x_m}{x^{\alpha-1}}\right)^2 \frac{\alpha}{\alpha-2} & \text{if } \alpha > 2 \\ \infty & \text{if } \alpha \in (1,2] \end{cases}$$

In general the raw moments are defined as,

$$\mu' = \begin{cases} \frac{\alpha x_m^n}{x^{\alpha-n}} \frac{\alpha}{\alpha-2} & \text{if } \alpha > n \\ \infty & \text{if } \alpha \leq n \end{cases}$$

8.6.10 Scaled Inverse Chi-square

The density function for the scaled inverse chi-square is,

$$p(y) = \frac{(\nu/2)^{-\nu/s}}{\Gamma(\nu/2)} s^\nu y^{-(\nu/2+1)} e^{-\nu s^2/(2y)} \tag{8.6.19}$$

where ν is degrees of freedom, s is scale, and $y > 0$. Figure 8.10 shows the scaled inverse chi-square distribution for for sets of parameters. The positive range for y makes this a good distribution for modeling variances.

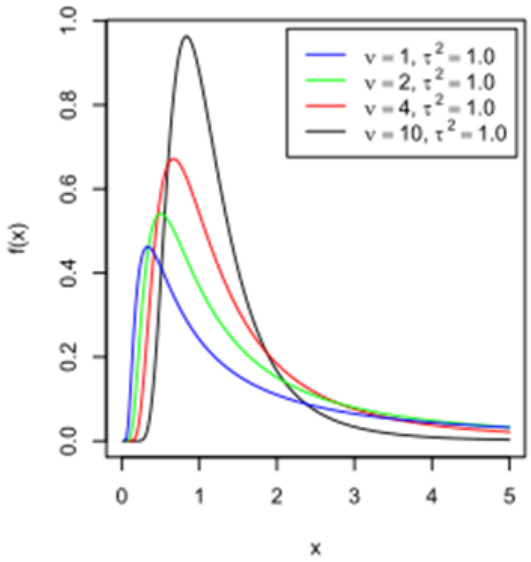

Figure 8.10: The Scaled Inverse Chi-Square for Different Parameter Values

Bibliography

[1] Aruoba, S. Borağan, Francis X. Diebold, Jeremy Nalewaik, Frank Schorfheide, Dongho Song, "Improving GDP measurement: A measurement-error perspective", Journal of Econometrics, 191 (2016), 384-397.

1

Chapter 9

Spectral Analysis

9.1 An Introduction to Spectral Analysis

Spectral analysis is a widely used method for analyzing time series data. It basically involves describing a time series, x_t, by comparing it to sets of sine and cosine curves. By summing up sine and cosine curves with different amplitudes we can create an artificial time series that resembles the series we observe. The first step in spectral analysis is transform a data series from time domain to frequency domain. This is done using the Fourier transform.

9.2 Fourier Transform

Fourier showed in 1807 that any periodic function can be rewritten as a weighted sum of sines and cosines of different frequencies. The basic building block used in the Fourier transform is the periodic process:

$$x = Asin(\omega x + \phi)$$

where ω is the frequency and ϕ is the phase. Figure 9.1 shows two sine curves that differ by $\pi/2$. The amplitude of the curve is the height to the peak. The phase is the difference in the peaks of the two curves.A cycle is defined as one complete period of a sine or cosine function defined over a time interval of length 2π.

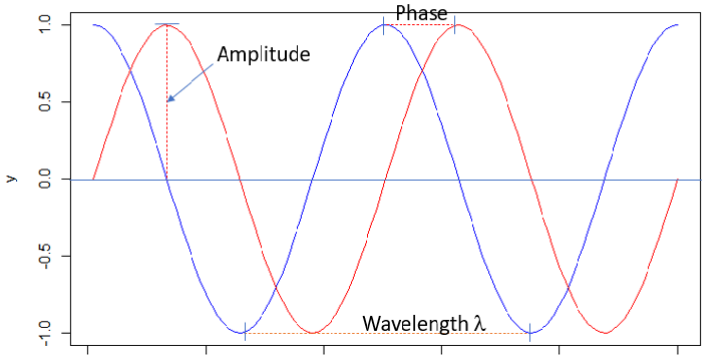

Figure 9.1: Two sine curves: $y = sin(x)$ and $y = sin(x + \pi/2)$

Given a time series $f(t)$ we can transform it to frequency domain $F(\omega)$ using the continuous Fourier transform. The Fourier transform of a value that is a function of time returns a value that is a function of frequency, ω. This equation is called the synthesis equation.

$$F(\omega) = \int_{\omega=-\infty}^{+\infty} f(t)e^{-2\pi i\omega t}d\omega \tag{9.2.1}$$

$$e^{2\pi i\omega t} = cos(2\pi\omega t) + isin(2\pi\omega t) \tag{9.2.2}$$

where i is the imaginary component of a complex number. For every ω from 0 to infinity, $F(\omega)$ contains the amplitude A and phase ϕ of the corresponding periodic process, $Asin(\omega x + \phi)$.

Let's spend some time trying to understand these equations. First, note that even if $f(t)$ real the Fourier transform will probably be complex. A complex number z, can be written as the sum of real part x, and imaginary parts, y:

$$z = x + iy$$

In addition we can make use of Euler's identity which states that,

$$e^{i\theta} = cos\theta + isin\theta$$

Finally note the polar form of a complex number is a more common representation.

$$z = |z|e^{i\theta}$$

where $|z|$ is the magnitude,

$$|z| = |x + iy| = \sqrt{x^2 + y^2}$$

Using Euler's identity the polar form can be written in real and imaginary parts. We see that the real part is a cosine function and the imaginary part is a sine function.

$$x = R[|z|e^{i\theta}] = |z|cos\theta$$
$$y = Im[|z|e^{i\theta}] = |z|sin\theta$$

9.2.1 Spectral Density

The spectral density is the Fourier transform of the autocovariance function. If a time series x_t has an autocovariance, γ satisfying

$$\sum_{k=-\infty}^{+\infty} |\gamma(k)| < \infty$$

then the spectral density is defined as

$$f(\omega) = \sum_{k=-\infty}^{+\infty} \gamma(k)e^{-2\pi\omega k}$$
$$\text{for } -\infty < \omega < +\infty$$

Properties

1. $f(\omega) = |\sum_{k=-\infty}^{+\infty} \gamma(k)e^{-2\pi\omega k}| < \infty$

2. f is periodic with a period of one which means that we can restrict the domain of f to $-1/2 \leq \omega \leq +1/2$.

3. $f(\omega) = f(-\omega)$

4. $f(\omega) \geq o$

5. $\gamma(k) = \int_{-1/2}^{+1/2} e^{2\pi i\omega k} f(\omega) d\omega$

White Noise Process
Let e_t be a white noise process. The variance $\gamma(0) = \sigma_e^2$ and $\gamma(k) = 0$ for $k \neq 0$. The spectral density is:

$$f(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k)e^{-2\pi i\omega k} = \gamma(0) = \sigma_e^t$$

The spectral density is constant across all frequencies.
AR(1) Process
Now, let's consider the spectral density of an AR(1).

$$x_t = \phi_1 x_{t-1} + e_t$$
$$\gamma(k) = \sigma_e^2 \frac{\phi^k}{(1-\phi^k)}$$
$$f(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k)e^{-2\pi i\omega k} = \frac{\sigma_e^2}{1-2\phi cos(2\pi\omega)} + \phi^k$$

- If $\phi > 0$ the process has positive autocorrelation. It is dominated by low frequency components. This is illustrated in Figure 9.2 where the time series and periodogram for $x_t = 0.7 * x_{t-1} + e_t$ are shown. The periodogram is negtively correlated with the frequency.

- if $\phi < 0$ the spectrum is dominated by high frequency components. This is illustrated in Figure 9.3. In this example the process is $x_t = -0.7x_{t-1} + e_t$. The frequenct is positively correlated with the periodogram

MA(1) Process
The spectral density for the MA(1) process $x_t = e_t + \phi e_{t-1}$ is:

$$f(\omega) = \frac{\sigma_e^2(1+2\phi cos\omega+\phi^2)}{2\pi}$$

If $\phi > 0$ the spectral density is maximal at ω=0. If $\phi < 0$ the spectral desnity is minimal at $\omega = 0$.
Periodogram
The periodogram is the sample analog of the spectral density.

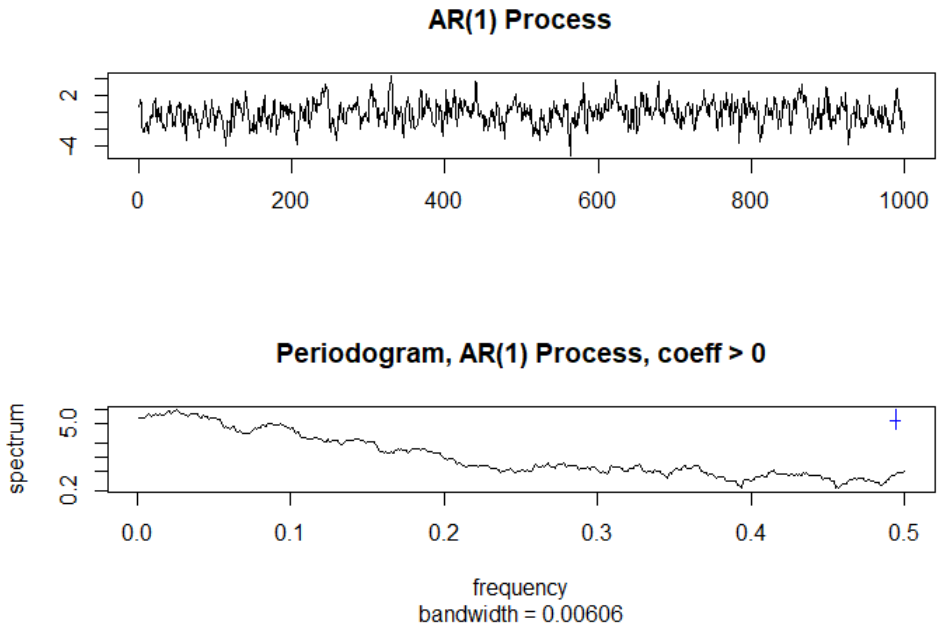

Figure 9.2: Periodogram of AR(1) with positive coefficient

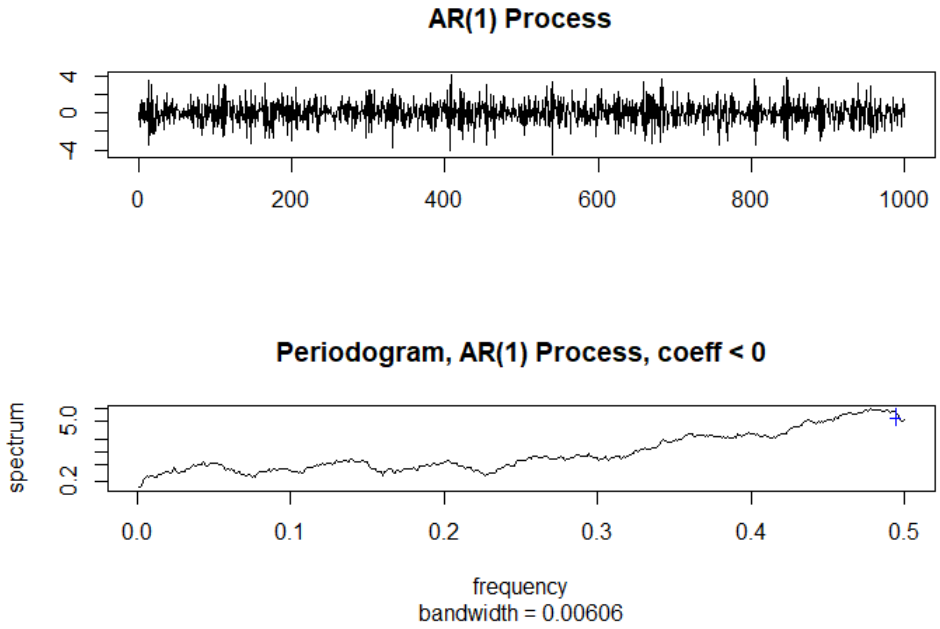

Figure 9.3: Periodogram of AR(1) with negative coefficient

Chapter 10

Wavelets

10.1 Introduction

This lecture introduces the wavelt transform. Wavelet transforms or filters are a mathematical technique that have been used in economics and finance for about 20 years, but have yet to become a mainstream econometric tool. It's unfortunate as wavelets can offer important insight into the dynamics of a time series.

The wavelet transform can be applied in both continuous and discrete form. The continuous wavelet transform (CWT) can be used to examine the power or variance of a single series across time and frequency, or the coherence of two series across time and frequency. The discrete wavelet transform (DWT) is a selective sample of the CWT at the scale level which results in a set of scale level time series.

Figure 10.2 contains annual returns for the S & P 500 from 1928 - 2016. The series shows a relatively high degree of variation, and it is somewhat difficult to discern the periods with the greatest volatility. Now, let's compare the time series of returns to wavelet power of the series which

is an example of a continuous wavelet transform. The wavelet power function (WPF) displays the energy of a times series. In Figure 10.2 the horizontal axis is in units of time, and the vertical axis shows periodicity measured in years. Dark blue indicates low oscillations and deep red indicates very high oscillations. The WPF shows that the strongest oscillation in market returns are the range of 1 to 8 years and occurred at the start of the Great Recession in 1929 to 1930. Market oscillations during the Great Repression (2008-2009) was mild in comparison. There is also a period of relatively high oscillations in the mid-1970's at a periodicity of 4 years.

Figure 10.3 show the multiresolution of the series of returns for the S & P 500. The return series has been transformed into 4 time series, each representing a different scale.

This chapter describes the wavelet transform and presents examples along with code. The first section describes the continuous wavelet transform, and the second section describes the discrete wavelet transform. The third section contains applications which include an example of wavelet coherence using the CWT, and calculation of the wavelet CAPM beta using the discrete wavelet transform (DWT).

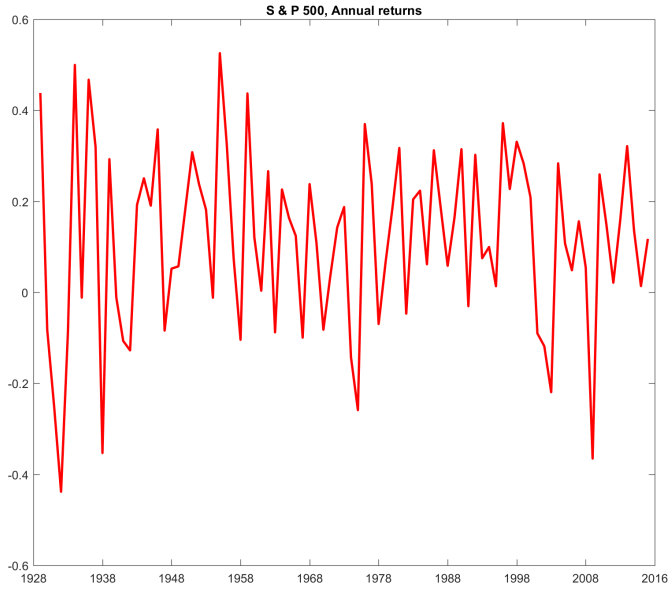

Figure 10.1: S & P 500 Returns, Annual Returns

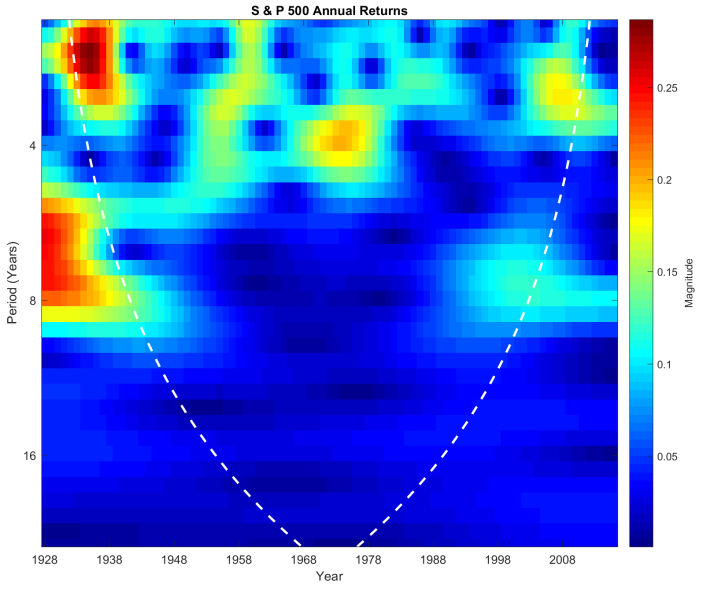

Figure 10.2: Wavelet Power Spectrum - S & P 500 Returns

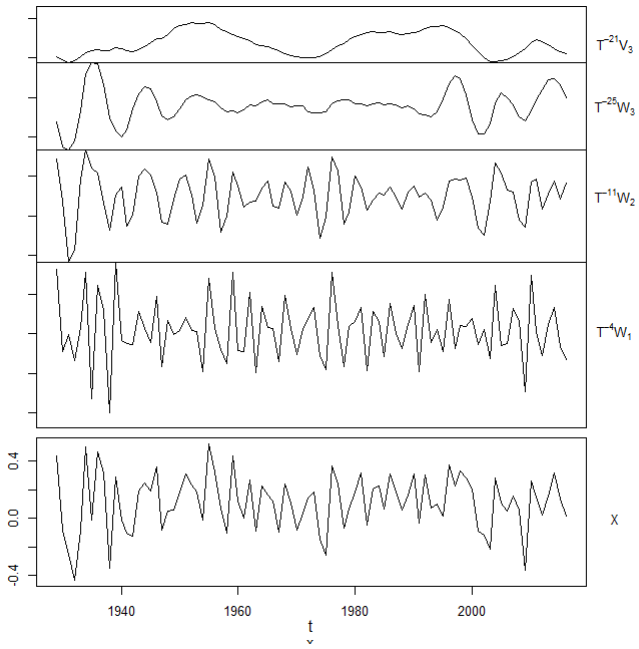

Figure 10.3: Multiresolution Analysis - S & P 500 Returns

10.2 The Discrete Wavelet Filter

It is intuitive to think of the wavelet transform in terms of a filter on the time series X_t. Let

$$\mathbf{h_l} = [h_0, h_1, ..., h_{L-1})$$

be a discrete finite length wavelet filter with the following properties:

$$\sum_{l=0}^{L-1} h_l = 0, \ (1)$$

$$\sum_{l=0}^{L-1} h_l^2 = 1, \ (2) \text{ unit energy}$$

$$\sum_{l=0}^{L-1} h_l h_{l+2n} = 0, \ (3) \text{ orthogonality to even shifts}$$

Condition (3), which states that the wavelet filters are orthogonal to even shifts, is needed to construct the orthonormal matrix that defines the DWT.

Conditions (2) and (3) can be expressed in terms of the squared gain function which shows the relationship between wavelet filters and the discrete Fourier transform (DFT):

$$\mathcal{H}(f) + \mathcal{H}(f + 2) = 2$$

where $\mathcal{H}(f)$ is the the squared gain function for the transfer function $H(f)$, defined as:

$$H(f) \equiv \sum_{t=-\infty}^{\infty} h_l e^{-i2\pi fl} = \sum_{l=0}^{L-1} h_l e^{-i2\pi fl}$$

This last equation expresses the wavelet filter as a high pass filter (DFT). (Why?)

Finally, the orthogonality and unit energy conditions for the wavelet can be expressed in terms the inverse DFT:

$$\sum_{l=-\infty}^{\infty} h_l h_{l+2n} = \int_{-1/2}^{1/2} e^{i2\pi fn} df = \begin{cases} 1, & n = 0 \\ 0, & n = ..., -2, -2, 1, 2, ... \end{cases}$$

The wavelet coefficients associated with the unit scale are found by circularly filtering $X_t : t = 0, ..., N-1$ with h_t and keeping every other value, where $N \equiv 2^J$. The result is as follows:

$$2^{1/2}\tilde{W}_{1,t} \equiv \sum_{l=0}^{L-1} h_l X_{t-lmodN},\ \text{t=0,...,N-1.}$$

The wavelet coefficients for unit scale are defined as

$$\mathbf{W_{1,t}} \equiv 2^{1/2}\tilde{W}_{1,t} \equiv \sum_{l=0}^{L-1} h_l X_{2t+1-lmodN}, t = 0, ..., \tfrac{N}{2} - 1.$$

The first subscript, which is one above, refers to the scale. The scale is defined as $\tau_j = 2^{j-1}$.[1]

10.3 The Discrete Wavelet Transform (DWT)

We can define the discrete wavelet transform in matrix notation as follows: Let $\mathbf{W} = \mathcal{W}\mathbf{X}$, where $\mathbf{W}$ is a column vector of length $N = J^N$ containing wavelet coefficients, and $\mathcal{W}$ is an NxN matrix that defines the DWT and has the property $\mathcal{W}^T\mathcal{W} = I_N$.

The vector of wavelet coefficients, $\mathbf{w}$, can be written as J+1 vectors,

$$\mathbf{W} = [\mathbf{W_1}, \mathbf{W_2}, ..., \mathbf{W_J}, \mathbf{V_J}]^\mathbf{T}$$

where $\mathbf{W_j}$ is vector of wavelet coefficients of length $N/2^j$, and V_J is a vector of scaling coefficients of length $N/2^j$.

The matrix $\mathcal{W}$ is made up of wavelet and scaling coefficients arranged by row. The first $N/2$ rows of $\mathbf{W}$, denoted $\mathbf{W_1}$, are populated with the elements of $\mathbf{W_{1,t}}$.

Since $\mathbf{W}_1 = \mathcal{W}_1\mathbf{X}$ where $\mathcal{W}_1$ is an $N/2 \times N$ matrix consisting of the first $N/2$ rows of $\mathcal{W}$, the definition of $\mathbf{W}_1$ implies a definition of $\mathcal{W}_1$.

Let

$$\mathbf{h_1} = [h_{1,N-1}, h_{1,N-2}, ..., h_{1,1}, h_{1,0}]^T$$

[1] The procedure of taking every other value of the filter output is referred to as downsampling by two.

be the vector of unit scale wavelet filter coefficients in reverse order. The wavelet coefficients come from an orthonormal wavelet family of length L with all values in the range $L < t < N$ set to zero.

Next, shift the coefficients in $\mathbf{h_1}$ by factors of two, so for example

$$\mathbf{h_1^{(2)}} = [h_{1,1}, h_{1,0}, ..., h_{1,3}, h_{1,2}]^T$$
$$\mathbf{h_1^{(4)}} = [h_{1,3}, h_{1,2}, ..., h_{1,5}, h_{1,4}]^T$$

Do this shift to create N/2 shifted versions of $\mathbf{h_1}$. Define the $N/2$ x N matrix W_1 as the matrix of the shifted versions of $\mathbf{h_1}$, so that

$$\mathcal{W}_1 = [\mathbf{h_1^{(2)}}, \mathbf{h_1^{(4)}}, ..., \mathbf{h_1^{(N/2-1)}}, \mathbf{h_1}]^\mathbf{T}$$

Next, define $\mathbf{h_2}$ as the vector of scale 2 wavelet filter coefficient in a similar manner to the definition of $\mathbf{h_1}$, and construct $\mathcal{W}_2$ by circularly shifting by factors of four. Repeat this procedure to construct $\mathcal{W}_j$ by circularly shifting by factors of 2^j. The matrix v_j is defined a a column vector whose elements are all equal to $1/\sqrt{(N)}$. The N x N matrix $\mathcal{W}$ is as follows:

$$\mathcal{W}^T = [\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_J, \mathcal{V}_J]$$

The DWT of $\mathbf{X}$ is an orthonormal transform, which implies that $\mathbf{X} = \mathcal{W}^T\mathbf{w}$ and $||\mathbf{w}||^2 = ||\mathbf{X}||^2$. The energy of the time series, $\mathbf{X}$ (measured by the squared norm), is equal to the energy of the wavelet coefficients. W_n^2 is a measure of the energy attributable to the nth DWT coefficient.

The wavelet transform is composed of a father wavelet and a set of mother wavelets. Given a function Φ, the father wavelet for the discrete transform is defined as:

$$\Phi_{J,k} 2^{-\frac{J}{2}} \Phi \frac{t-2^J * k}{2^J} \quad (2)$$

$$\int \Phi(t)dt = 1$$

The mother wavelets, also in discrete form, are defined as:

$$\Psi_{j,k}\, 2^{-\frac{j}{2}} \Psi \frac{t-2^j * k}{2^j}, j = 1, ..., J \quad (3)$$

$$\int \Psi(t)dt = 0$$

Where J is the number of scales or levels,$2^J$is a scale factor and k is the time domain index.

The father and mother wavelets are each indexed by both scale and time. It is precisely this dual indexing that makes wavelet analysis appealing since as a time series, f(t), is represented as a linear combination of wavelet functions that are localized in space and time.

The scale parameter is inversely proportional to frequency.[2] The father and mother wavelet functions may also be represented as filters. In this alternative representation the father wavelet is a low pass filter, and the mother wavelets are high pass filters.[3] We can use the wavelet functions to transform a time series, f(t), into a series of wavelet coefficients,

$$S_{J,k} = \int f(t)\Phi_{J,k} \quad (4)$$

and,

$$d_{j,k} = \int f(t)\Psi_{j,k} \text{ j=1,...,J} \quad (5)$$

Where $S_{J,k}$ are the coefficients for the father wavelet at the maximal scale, 2^J, and the $d_{j,k}$, are the coefficients of the mother wavelets at the scales from 1 to 2^J. The $d_{j,k}$ are referred to as the detailed coefficients and the $s_{J,k}$ are referred to as the smooth coefficients. Applying the transforms results in a time series of length k of smooth coefficients at the maximal scale J, and J time series of detailed coefficients each of length k. If there are 6 scales, the frequency of the first scale is associated with the interval [1/4,1/2], and the frequency of scale 6 is associated with the interval [1/128, 1/64]. For a monthly time series decomposing into six scales (D1-D6) corresponds to periods 2-4, 4-8, 8-16, 16-32, 32-64, and 64-128 months. The smooth component (S6) captures the trend of the original series. The high frequency component is associated with the shortest scale D1, while the low frequency component is associated with the longest scale D6.

Given the smooth and detailed coefficients, a time series f(t) can be represented in decomposed form, known as the multi-resolution analysis of f(t), as follows:

$$f(t) = \Sigma_k S_{J,k}\, \Phi_{J,k}\,(t) + \Sigma_k d_J, k\Psi_{J,k}\,(t) + ... + \Sigma_k d_{j,k}\, \Psi_{j,k}\,(t) + ... + \Sigma_k d_{1,k}\, \Psi_{1,k}\,(t) \quad (6)$$

Or, using summary notation,

$$f(t) = S_J + D_J + D_{J-1} + ... + D_1$$

The discrete wavelet transform decomposes a time series into orthogonal signal components at different scales. S_j is a smooth signal, and each D_j is a signal of higher detail. The number of coefficients differs by scale. If the length of the data series is n, and divisible by 2^J, there are $n/2^j$ $d_{j,k}$ coefficients at scale j=1,...,J-1. At the coarsest scale there are $n/2^J$ $d_{J,k}$ and $s_{J,k}$ coefficients. The wavelet variance at each scale is captured as the wavelet power of each scale.

[2]See Gencay, et al. 2010, pp. 99-103 for a complete discussion.
[3]See Ramsey (2002).

10.4 The Continuous Wavelet Transform

The continuous wavelet transform (CWT) is also useful for gaining insight into the time-scale characteristics of a time series. The CWT is defined as,

$$W(\lambda,t) = \int_{-\infty}^{+\infty} \Psi_{\lambda,t}(u)x(u)du \quad (7)$$
$$\text{where, } \Psi_{\lambda,t}(u) \equiv \frac{1}{\sqrt{\lambda}}\Psi\left(\frac{u-t}{\lambda}\right)$$

As noted by Ramsey, the main difference between the CWT and DWT is that the CWT considers continuous variations in the scale (λ) and time components (t). The discrete wavelet transform can be derived independently of the CWT, but it can also be viewed as a critical sampling of the CWT with $\lambda = 2^{-j}$ and $t = k2^{-j}$.

The wavelet power spectrum which measures the local variance of a time series at different scales is defined as $|W(\lambda,t)^2|$, and aids our analysis in terms of understanding how periodic components evolve over time when applied to the market, as well as, the eleven sectors examined in our analysis. A clear advantage that the CWT has over the discrete transform is that it produces a powerful visual for detecting time-scale patterns. The wavelet power spectrum is helpful for understanding how the power varies with the scaling of the wavelet. But we also need to understand how periodic components evolve jointly over time. The Fourier coherency identifies frequency bands where two time series are related, while the wavelet coherency identifies both frequency bands and time intervals when time series are related. The wavelet coherence of two series, x and y, is a measure of co-movement across time and scale based on the CWT. To define it we need the definition of two other measures, the cross wavelet transform (XWT) and the cross wavelet power (XWP). The XWT is defined as

$$W_{xy} = W_x(\lambda,t)W_{y*}(\lambda,t) \quad (8)$$

The XWP is the defined as the absolute value of the XWT, $|W_{xy}(\lambda,t)|$. It measures the local covariance of x and y at different time scales. The XWP identifies areas in time-scale space where the two series have high common power. In addition to identifying the common power of two time series, we are also interested in identifying areas of co-movement in time-scale space, even if the cross wavelet power is low. A measure of co-movement, the wavelet coherence, is defined as:

$$R^2(\lambda,t) = \frac{|S(S^{-1}Wxy(\lambda,t))|^2}{S(S^{-1}|W_x(\lambda,t)|^2)*S(S^{-1}|W_{xy}(\lambda,t)|^2)} \quad (9)$$

Where S is a smoothing operator in time and scale, and $0 \leq R^2(\lambda,t) \geq 1$. The wavelet coherence is similar to the correlation coefficient, and is typically interpreted as a localized correlation in time-scale space.

Chapter 11

Extreme Value Theory

Chapter 12

Rational Bubbles

Chapter 13

Empirical Options Pricing

13.1 Black-Scholes Model

13.2 Heston Model

13.3 Ghysels, Garcia, Renault

13.4 Gourieoux, Jasiak

13.5 Bates

Chapter 14

Mixture Models

14.1 Introduction

14.2 The EM Algorithm

14.3 Bayesian Mixture Inference