

Chapter 5 *Stochastic Theories of Equity Value*

5.1 Foundations of Modern Finance

- A. Basics of Mean-Variance Portfolio Analysis
- B. Separation, the CAPM and the Market Model
- C. Decision Making Under Uncertainty

5.2 Ergodicity and Asset Pricing Theories

- A. A Brief History of Ergodic Theory
- B. Uncertainty and the Ergodic Hypothesis
- C. Ergodicity in Financial Economics

5.3 Bifurcation and Multi-modal Densities

- A. The Phenomenological Approach
- B. Transition Density Decomposition
- C. Bifurcation and the Quartic Exponential Distribution

Appendix: Preliminaries and Proofs

Modern Finance and Equity Valuation

In order to be implemented in practice, the equity valuation methodology proposed by modern Finance academics requires knowledge of parameters from *ex ante* return distributions. In particular, the parameters required for the capital asset pricing model (CAPM) are the means, variances and covariances of the *ex ante* return distributions for the equity securities and market index of interest. Empirical implementation of such valuation models involves use of *ex post* estimates of the relevant *ex ante* distribution parameters. Estimators are chosen to have desirable statistical properties such as unbiasedness and consistency. Yet, despite decades of effort, there is little evidence that *ex post* optimal portfolios have similar *ex ante* performance. A range of explanations and apologies have been provided for this unexpectedly poor performance. This chapter demonstrates that fundamental theoretical difficulties can be traced to the pervasive use of time reversible ergodic stochastic processes. This assumption permits the use of a single *ex post* sample path to estimate the parameters of the *ex ante* stationary distribution for the ensemble of possible future time paths. In contrast, time irreversible ergodic processes provide a viable explanation for some unexplained behaviour of equity prices. Such processes can account for: the erratic forecasting properties of parameter estimates from conventional models; excess volatility in stock prices relative to the underlying fundamentals; and, the anecdotal claims regarding *ex ante* valuation accuracy achieved using technical analysis.

5.1 Foundations of Modern Finance

A. Basics of Mean-Variance Portfolio Analysis

Many of the essential analytical tools used in mean-variance portfolio analysis can be found in the results for linear combinations of random variables from mathematical statistics. One basic result is the following, e.g., Freund (1971, p.195):

Theorem: Moments of Linear Combinations of Random Variables

If $X(1), X(2), \dots, X(N)$ are random variables and a_1, a_2, \dots, a_N are constants and $Y = a_1 X(1) + a_2 X(2) + \dots + a_N X(N)$ then:

$$E[Y] = \sum_{i=1}^N a_i E[X(i)]$$

$$\text{var}[Y] = \sum_{i=1}^N a_i^2 \text{var}[X(i)] + 2 \sum_{i>j} a_i a_j \text{cov}[X(i), X(j)]$$

where the double sum over $i > j$ extends over all values of i and j , from 1 to N , for which $i > j$.

Derivation of $\text{var}[Y]$ requires the observation that $\text{cov}[X(i), X(j)] = \text{cov}[X(j), X(i)]$. One immediate corollary is that if $\text{cov}[X(i), X(j)] = 0$ for all i and j where $i \neq j$, i.e., the random variables are all independent, and $a_1 = a_2 = \dots = a_N = 1/N$ then $\text{var}[Y]$ has the property:

$$\lim_{N \rightarrow \infty} \text{var}[Y] = \lim_{N \rightarrow \infty} \left\{ \sum_{i=1}^N a_i^2 \text{var}[X(i)] \right\} = 0$$

This result has applications in insurance where the random variables are policy payouts and the a_i are the fraction of the portfolio of policies attributable to policy i .

Extending these results to portfolios follows immediately from identifying the random variables as the returns on individual securities held in a given portfolio, i.e., let $X(i) = R_i$ for all i . The definition of the portfolio expected return follows:

Definition: The expected return on the portfolio $E[R_p]$ is the value weighted sum of the expected returns on the individual securities, the $E[R_i]$:

$$E[R_p] = \sum_{i=1}^k w_i E[R_i]$$

where k is the number of securities in the portfolio. To calculate the value weights, w_i :

$$w_i = \frac{\$A_i}{\sum_{i=1}^k A_i} \quad \text{where} \quad \sum_{i=1}^k w_i = 1$$

with $\$A_i$ being the dollar value invested in security i and the sum over all $\$A_i$ being the total amount of money invested in the portfolio

As a simple example, consider having \$1 million invested in a portfolio of 2 securities, and there is \$500,000 in each security, then each $w_i = .5$. As a slightly more complicated example consider the following problem: At the beginning of the year, Joe Investor owned four securities in the following amounts: A, 100 shares; B, 400 shares; C, 200 shares; D, 200 shares. The current prices of the securities are: A = \$12.50; B = \$17.50; C = \$25; and, D = \$50. In one year's time, Joe expects the prices to be: A = \$25; B = \$20; C = \$30; and D = \$55. What is the expected return on Joe's portfolio for the year? The solution to this problem is determined by calculating the total value invested as: $100 (12.50) + 400 (17.50) + 200 (25) + 200 (50) = \$23,250$. This permits the calculation of the value weights: $w_A = 1250/23250 = .054$; $w_B = .301$; $w_C = .215$; $w_D = .430$. The expected return on the portfolio can now be calculated as: $E[R_p] = .054(E[R_A]) + .301(E[R_B]) + .215(E[R_C]) + .430(E[R_D]) = .054 (1.00) + .301 (.143) + .215 (.2) + .430 (.1) = .183$ (18.3%).

The other key element in the mean-variance portfolio model is the standard deviation of portfolio returns. As with calculating the risk for individual securities, calculations are done for the variance and the standard deviation is determined by taking a square root. The standard deviation, as opposed to the variance, is of the appropriate measure of risk because it is in the same units as the expected returns. However, calculations are done using the variance.

Definition: The standard deviation of portfolio returns, σ_p is the square root of the variance of portfolio returns $var[R_p] \equiv \sigma_p^2$. Various **equivalent** forms for the portfolio variance formula are available:

$$\begin{aligned} var[R_p] &= \sigma_p^2 = \sum_{i=1}^k \sum_{j=1}^k w_i w_j \sigma_{ij} = \sum_{i=1}^k w_i^2 \sigma_i^2 + 2 \sum_{i>j}^k w_i w_j \sigma_{ij} \\ &= \sum_{i=1}^k w_i^2 \sigma_i^2 + 2 \sum_{j=1}^k \sum_{i=1, i>j}^k w_i w_j \sigma_{ij} = \sum_{i=1}^k w_i^2 \sigma_i^2 + 2 \sum_{i>j} \sum w_i w_j \sigma_{ij} \end{aligned}$$

where $cov[R_i, R_j] = \sigma_{ij}$. In the double sum expression, when $i=j$ the covariance is a variance. These expressions can be further manipulated by making further substitutions using the definition for ρ_{ij} , the correlation between R_i and R_j , i.e., $cov[R_i, R_j] = \sigma_{ij} \equiv \rho_{ij} \sigma_i \sigma_j$.

It is easiest to understand these results for the case where $k = 2$, when there are only two securities in the portfolio. In this case:

$$\begin{aligned}\sigma_p^2 &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_{12} \\ &= w_1^2 \sigma_1^2 + (1 - w_1)^2 \sigma_2^2 + 2 w_1 (1 - w_1) \rho_{12} \sigma_1 \sigma_2\end{aligned}$$

Similarly for 3 assets in the portfolio:

$$\begin{aligned}\sigma_p^2 &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + w_3^2 \sigma_3^2 \\ &+ 2 \{w_1 w_2 \sigma_{12} + w_1 w_3 \sigma_{13} + w_2 w_3 \sigma_{23}\}\end{aligned}$$

When there are k securities in the portfolio, the resulting portfolio variance will contain k variance terms and $\{k(k - 1)\}/2$ covariance terms. The substitutions using the definition for the correlation coefficient and the restriction that the sum of the value weights equals one is left as an exercise.

Having the formula for the variance of portfolio return permits the ready identification of an important special case of an optimum portfolio: the minimum variance portfolio, the portfolio that has the smallest risk in the set of all possible portfolios. This formula for this portfolio can be derived by minimizing $\text{var}[R_p]$ with respect to the choice variables, the value weights for each of the individual securities, subject to the restriction that the sum of the value weights be equal to one. In the simple case of the minimum variance portfolio for two securities, using the result that $w_1 + w_2 = 1$:

$$\begin{aligned}\sigma_p^2 &= w_1^2 \sigma_1^2 + (1 - w_1)^2 \sigma_2^2 + 2 w_1 (1 - w_1) \sigma_{12} \\ \frac{d\sigma_p^2}{dw_1} &= 2 w_1 \sigma_1^2 - 2 (1 - w_1) \sigma_2^2 + 2 (1 - 2w_1) \sigma_{12} = 0 \\ w_1^* &= \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2 \sigma_{12}}\end{aligned}$$

This result demonstrates that the minimum variance portfolio will be a combination of the two securities and not just be fully invested in the security with the lowest risk.

The intuition behind the portfolio diversification problem can be illustrated with the following artificial situation: assume that all the securities in the market have the same expected return of 10% and the same standard deviation of security return of 15% with the covariance between all security returns being .02. Construct an equally weighted portfolio containing N securities. While the expected return on this portfolio will be 10%, the variance of the equally weighted portfolio containing N securities will be:

$$\sigma_p^2 = \sum_{i=1}^N \frac{\sigma_i^2}{N^2} + 2 \sum_{i>j}^N \frac{\sigma_{ij}}{N^2} < .15$$

While the expected return is a linear combination of the individual security expected returns, the result same does not apply to the variance. This property of the variance for a linear combination of random variables is an essential ingredient in the portfolio optimization models.

Recall the result stated previously where, if the random variables are uncorrelated, then the variance of an equally weighted linear combination will go to zero as N goes to infinity. What happens to the portfolio expected return and standard deviation of this portfolio as N gets large? Because it is assumed that all standard deviations are the same, there are N equal terms in the first sum and, because the covariances have been assumed to be equal, there are $N(N - 1)$ terms in the second sum and the portfolio variance reduces to:

$$\sigma_p^2 = \frac{\sigma_i^2}{N} + N(N-1) \frac{\sigma_{ij}}{N^2} = \frac{\sigma_i^2}{N} + (1 - \frac{1}{N}) \sigma_{ij}$$

As $N \rightarrow \infty$, the first term goes to zero and the portfolio variance is reduced to the covariance. To get this result, N must be very large. Even for portfolios containing, say, 100 securities, there is still some contribution to variance from the σ_i . As noted, when the covariance between all available securities is zero (independent returns), the standard deviation of the portfolio will go to zero as N gets large.

This simple example provides a pedagogical basis for illustrating the gains to diversification. Examining the variance of the equally weighted portfolio described above, the (σ_i^2 / N) term applies to the specific risk associated with the individual securities, where the σ_i^2 for each of individual security are associated with individual firm specific risks. Because this component of portfolio variance goes to zero as the number of securities gets large, it is appropriately described as ***diversifiable risk***. The $(1 - (1/N)) \sigma_{ij}$ term is associated with the covariance between security returns. Because this source of portfolio risk does not go to zero as N goes to infinity, it is appropriately described as ***nondiversifiable risk***. It follows that the portfolio standard deviation can be decomposed into the sum of diversifiable risk and nondiversifiable risk. Hence, the risk associated with any portfolio of securities equals the sum of the diversifiable risk and nondiversifiable risk for that specific portfolio. ‘Efficiently diversified’ portfolios have eliminated diversifiable risk.

As securities are added to the equally weighted portfolio, the risk of the portfolio is reduced until the lower bound provided by non-diversifiable risk is reached. However, the amount of risk reduction decreases as the number of the securities in the portfolio increases to the point where there is no more firm specific risk that can be eliminated. The lower bound on portfolio risk, associated with the non-diversifiable risk, is due to the covariance between security returns. Modern portfolio theory expends considerable theoretical effort in developing the capital asset pricing model (CAPM) where individual security returns depend on a combination of the riskless interest rate and the covariance of the individual security return with the return on the market portfolio. In this model, it is the nondiversifiable risk, referred to as systematic risk, that is compensated with higher expected return. Hence, ***there is a tradeoff between systematic risk and expected return***. Because firm specific risk (unsystematic risk) can be eliminated in an efficiently diversified portfolio, the security market will not reward this source of risk with higher expected return.

The Optimization Model

The mean-variance portfolio optimization model is a central paradigm of modern Finance. The essence of the model is captured in the following quadratic optimization problem, e.g., Elton and Gruber (1995), Luenberger (1998):¹

$$\begin{aligned} \min_{\{w_i\}} \quad & \text{var}[R_p] = \sum_{i=1}^k \sum_{j=1}^k w_i w_j \sigma_{i,j} \\ \text{subject to:} \quad & E[R_p] = \sum_{i=1}^k w_i E[R_i] = \bar{c}_n \\ \text{for } \quad & \bar{c}_n \in \{c_0, c_1, c_2, \dots\} \quad \text{where: } c_0 = c_{mv} \text{ and: } \sum_{i=1}^k w_i = 1 \end{aligned}$$

where: k is the number of risky securities or assets available for investment; $E[R_i]$ is the (conditional) expected return on security or asset i ; $E[R_p]$ is the expected return on the portfolio; c_{mv} is the return on the minimum variance portfolio; and, $\text{var}[R_p]$ is the variance of portfolio return. In this model, the $\{w_i\}$ are the value weights, the fraction of the total value of the portfolio invested in each asset. Though it is conventional to develop the model under the assumptions of perfect capital markets (see BOX 5.1), it is possible, even desirable, to impose additional restrictions on the optimization problem. One such additional restriction is $w_i \geq 0$ for all i . This restriction prevents short selling of securities. Without this restriction, all or almost all securities will be held in some form, either long or short. With the short selling restriction, the resulting optimal portfolios will have many securities that have value weights equal to zero.

INSERT BOX 5.1

Perfect capital markets assumptions

The quadratic optimization problem is to determine the value weights for each security which minimize the variance of the return on the portfolio, subject to a target level of expected return. Because there is range of possible expected returns that can be chosen, ***the solution to the optimization problem will be a set of portfolios***, each with its own set of optimal weights. This set of optimal portfolios is typically referred to as the ***efficient frontier***. Other terms such as efficient set (Fama 1976), portfolio possibilities curve (Elton and Gruber 1995) and mean-variance efficient locus (Ingersoll 1987) are also used. There are a number of solution methodologies that can be used to solve quadratic optimization problems. A simple iterative method involves initially solving the minimum variance problem. The resulting optimal minimum variance weights are used to identify c_{mv} . This value is used to specify $c_1 = c_{mv} + \epsilon$, where $\epsilon (> 0)$ is specified according to the desired precision required in the solutions. Using c_1 it is now possible to solve the Lagrangian problem for the next portfolio along the frontier. This process continues for $c_2 = c_{mv} + 2\epsilon$, $c_3 = c_{mv} + 3\epsilon$ and so on until the desired efficient frontier is determined. With the short sales constraint, the maximum c_i is given by having all funds invested in the highest returning security. Without the short sales constraint, the efficient frontier can be extended indefinitely. Because the underlying problem is quadratic, there will be two ‘optimal’ solutions, one of which is ignored because it will have higher

risk for the same expected return. Hence, the efficient frontier only contains the portfolios with high return/lowest risk.

The number of variations that have emerged from this basic model is staggering.² Initially, implementation of the model was impeded by the large number of parameters required to make the model operational. In addition to the k individual asset returns, $E[R_i]$, there are k variances, σ_i^2 , and $\{k(k-1)\}/2$ covariances which have to be estimated from past data. Even if these parameters are available, the model is only capable of generating a set of mean-variance optimal portfolios, the efficient frontier. Additional structure is needed to select a specific portfolio from the set of optimal portfolios. Tobin (1958), Sharpe (1963, 1964) and others handled this problem by introducing a riskless asset. This permits the investor to form portfolios which combine the riskless asset with an efficient frontier portfolio. In this fashion, the investor is able to achieve the same level of expected return as that generated by an efficient frontier portfolio, again with a lower level of risk. Effectively, the addition of a riskless asset transforms the investment opportunity set from a convex function, the efficient frontier, to a set of linear functions, the capital allocation lines.

In general, where there are many possible securities available for inclusion in the portfolio, solution of the efficient set from the optimization problem is complicated. For purposes of illustration, it convenient to assume that there is only two risky securities. In this case it is possible to derive the efficient frontier directly, permitting basic concepts to be illustrated. So, assume you are considering creating a portfolio combining a stock fund composed of large stocks (S) and a bond fund (B). The statistics for these funds are: $E[R]$, large stock fund = 12%, bond fund = 5%; σ , large stock fund = 15%, bond fund = 8%. For ease of calculation, assume the correlation coefficient between the funds is zero. It is now possible to calculate $E[R_p]$ and σ_p for all possible portfolios, starting from 0% invested in the stock fund ($w_s = 0$) and going to 100% ($w_s = 1$), in increments of 20%. This produces:

w_s	$E[R_p]$	σ_p
0	5.0%	8.0%
.2	6.4%	7.07%
.4	7.8%	7.68%
.6	9.2%	9.55%
.8	10.6%	12.11%
1.0	12%	15%

Plotting these values in $\{E[R], \sigma\}$ space produces the efficient frontier. From these values, it is apparent that an investment of 100% in the bond fund ($w_s = 0$) does not make sense because portfolios with values such as $w_s = .2$ and $w_s = .4$ both provide a higher portfolio expected return with a lower level of portfolio risk. Recalling that the returns were assumed to be independent, the portfolio weights, $E[R_{mv}]$ and σ_{mv} for the minimum-variance portfolio are:

$$w_s = \frac{\sigma_b^2}{\sigma_s^2 + \sigma_b^2} = \frac{.08^2}{.15^2 + .08^2} = .2215 \quad \rightarrow \quad w_b = .7785$$

It follows that $E[R_{mv}] = .0655$ and $\sigma_{mv} = .0706$.

Suppose the number of securities available for selection in the portfolio included a third fund

composed of small stocks which has $E[R] = 15\%$ and $\sigma = .20$. How would you derive the efficient frontier for portfolios combining the three funds? Attempting to use the method of direct calculation that worked for the two security case is no longer possible. The two security case reduced to solving for one weight because it was possible to substitute out the other weight using the constraint that the sum of the weights equals one. Hence, there was no optimization problem to solve as there was only one effective weight. Barring trivial cases, when there are three or more securities the optimal weights have to be determined by solving the first order conditions of the optimization problem in order to identify how much of a given frontier portfolio is invested in each security. Ingersoll (1987) discusses the relevant solution procedure.

Capital Allocation Lines: Introducing a Riskless Asset

The efficient frontier specifies a set of portfolios that achieve optimal combinations of the available **risky** assets. In order to provide a practical guide to portfolio selection, some method is required to identify a specific portfolio from the efficient frontier. Observing that the efficient frontier is a convex relationship between $E[R]$ and σ for portfolios of risky assets, by introducing a **riskfree** asset it is possible to produce a linear set of available portfolios. In particular, when a riskfree asset is introduced a range of portfolios can be identified which are not available with risky assets alone. What is the riskfree asset? This depends on the investment horizon or the portfolio rebalancing period. The riskfree asset must be free of default risk, have no coupons to reinvest (this would create coupon reinvestment risk), have maturity equal to the investment horizon and be denominated in domestic currency. For US investors, it is typical to use the 3 month Tbill rate as the riskfree interest rate. Though it is conventional to proxy the riskfree asset with a 3 month US Treasury bill, some presentations, e.g., Damodaran (1994), argue that the US Treasury long term bond yield is appropriate. The problem of specifying the riskfree asset will be explored in more detail shortly.

INSERT FIGURE 5.1.a
CAL with efficient frontier

The riskfree asset permits the creation of portfolios which combine the riskfree asset with a risky portfolio located on the efficient frontier (see Figure 5.1.a). In $(E[R], \sigma)$ space, the line connecting the riskfree rate with a point on the efficient frontier is referred to as a **capital allocation line** (CAL). (This terminology is used in Bodie, Kane and Marcus 1999). There are as many capital allocation lines as there are portfolios on the efficient frontier. Each point on a given capital allocation line defines a range of possible portfolios which combine the riskfree asset and an efficient portfolio composed of risky assets. Along a given capital allocation line connecting the riskfree rate, r , with an efficient portfolio X, the expected return ($E[R_R]$) and standard deviation (σ_R) of a portfolio combining the efficient frontier portfolio and the riskfree asset can be specified:

$$E[R_R] = w_r r + (1 - w_r) E[R_x]$$

$$\begin{aligned}\sigma_R^2 &= w_r^2 \sigma_r^2 + (1 - w_r)^2 \sigma_x^2 + 2 w_r (1 - w_r) \sigma_{rx} \\ &= (1 - w_r)^2 \sigma_x^2 \quad \rightarrow \quad \sigma_R = (1 - w_r) \sigma_x\end{aligned}$$

The result for the variance of the portfolio follows because the riskfree asset has no risk or covariance because it is risk free.

Some care has to be taken in interpreting the weight w_r . Though the notation w is used, this weight is not associated with the w_i weights for the various efficient frontier portfolios. The w_r weight is the fraction of the portfolio in the riskfree asset and $(1 - w_r)$ is the fraction held in the efficient portfolio. When $1 \geq w_r \geq 0$, this implies that the portfolio involves a positive investment in both the riskfree asset and the efficient portfolio. In Figure 3-a this condition is applicable to all the portfolios lying on the portion of the CAL between r and X . At r , $w_r = 1$ and at X on the efficient frontier $w_r = 0$. When $w_r \leq 0$, this implies that the riskfree asset is held short, i.e., the investor is borrowing at the riskless rate. This is equivalent to saying that, in addition to investment of the original capital, the investor has also borrowed money at the riskfree rate and has purchased additional units of the risky portfolio X . For example, if the investor has \$1 million of original capital to invest and $w_r = -.5$ ($w_x = 1.5$), then the investor has borrowed an additional \$500,000 and has used this money to purchase an additional \$500,000 of X . Portfolios where $w_r \leq 0$ lie to the right of X on the CAL. The key point to recognize is that the presence of the riskfree asset permits the investor to attain portfolios which are not available using risky assets alone.

The Capital Market Line and Market Equilibrium

To this point, the problem of picking an individual portfolio from the set of efficient frontier portfolios has not been solved. A convex set has been replaced by a set of CAL's, each of which has a theoretically infinite number of possible portfolios. What has been demonstrated is that, with a riskfree asset, it is possible to specify $(E[R], \sigma)$ tradeoffs that are unattainable with risky assets alone. In order to identify the best CAL and the appropriate portfolio to select on that CAL, it is conventional to introduce the mean-variance expected utility function: $EU[R] = E[R] - b \text{ var}[R]$. As depicted in Figure 5.1.b, the mean-variance EU function defines a preference ordering over the $(E[R], \sigma)$ space. Movements in a northwest direction indicate increasingly higher levels of expected utility. It follows that the CAL which is just tangent to efficient frontier will attain the highest level of expected utility and the tangency of that CAL with the highest EU curve will be the specific portfolio that maximizes EU . This importance of this particular CAL is recognized specifically by referring to it as the *capital market line* (CML).

INSERT FIGURE 5.1.b
EU map for Mean-Variance function

As it turns out, the slope of the capital allocation line for any efficient portfolio X will be of interest in identifying the properties of the capital market line. Observing that the rise of the CAL is $E[R_x] - r$ and the run from the origin is σ_x it follows that the slope of any CAL is $(E[R_x] - r)/\sigma_x$. Using this result, the equation of the capital allocation line for X becomes:

$$E[R] = r + \frac{E[R_x] - r}{\sigma_x} \sigma_R$$

Refer to Figure 5.1.b describing the indifference curves in $(E[R], \sigma)$ space for a risk averse investor. Because utility increases as the slope of the capital allocation line increases, the rational investor will achieve the maximum level of utility by selecting the capital allocation line with maximum slope, i.e., the CAL that is just tangent to the efficient set. Under perfect market assumptions, the CAL that is just tangent to the efficient set represents the highest level of utility. This **tangency portfolio** is the portfolio which represents the **market equilibrium**. With the additional theoretical apparatus provided by the capital asset pricing model (CAPM) it can be demonstrated that the tangency portfolio associated with the capital market line is the **market portfolio**.

To illustrate the process of identifying a specific portfolio, refer back to the stock/bond fund portfolio discussed previously. Assume that a riskfree asset is available with $r = 3\%$. By maximizing the slope of the CAL, the weights for the tangency portfolio are given to be $w_s = .561$ and $w_b = .439$. For these weights, $E[R_M] = .0893$ and $\sigma_M = .0912$ where M indicates the tangency portfolio. Observing that the slope of CML is $(.0893 - .03)/.0912 = .65$, the equation of the capital market line can be specified as: $E[R_R] = .03 + .65 \sigma_R$. Now, suppose the investor's indifference map is given by the expected utility function: $EU[R] = E[R] - \{3.25 \text{ var}[R]\}$. Recognizing that $\text{var}[R] = ((1 - w_r) \sigma_M)^2$ and $E[R_R] = w_r r + (1 - w_r) E[R_x]$, the process of maximizing $EU[R]$ gives $w_r = -.096$ as the optimal holding for the riskless asset. This implies that, for the specified mean-variance expected utility function, the optimal solution involves a solution on the CML to the right of the tangency with the efficient frontier. It can be verified that alternative values of the risk aversion parameter b give: $b = 5$, $w_r = .2875$; $b = 3$, $w_r = -.01786$; and, $b = 2$, $w_r = -.78125$.

Criticism of Mean-Variance Portfolio Analysis

While the mean-variance portfolio model has considerable theoretical appeal, there are a number of substantive problems that arise in implementing the model. One obvious problem concerns the large number of parameters that have to be estimated, e.g., Lummer and Riepe (1994). Even if this problem can be overcome, attempting to capture the gains, *out-of-sample*, has proved to be illusive, particularly when international assets are permitted to be part of the set of available securities. In practice, the use of *ex post* (in-sample) data to estimate the relevant parameters of the *ex ante* (out-of-sample) distribution creates numerous problems, not the least of which is instability in both the mean and variance-covariance parameter estimates. This is especially the case where expected returns are of interest. As pointed out by Eaker, et.al. (1991): "The problem with including returns in the portfolio selection decision is that such portfolios generally perform poorly in out-of-sample tests."

The mean-variance portfolio model is a central tenet of modern Finance. Much like another central tenet, the efficient markets hypothesis, enthusiasm for the mean-variance portfolio model within Finance has evolved considerably. In recent years the model has been subjected to substantial critical scrutiny, e.g., Fisher and Statman (1997). The first wave in the assault on the mean-variance approach can be attributed to Jorion (1985, p.265), which describes the problems emphatically in the context of internationally diversified portfolios:

Mean-variance analysis has serious shortcomings which are too often ignored ... Perhaps the most serious defect in the classical (portfolio) approach is the poor out-of-sample performance of the optimal portfolios. Performance measures always deteriorate substantially outside the sample period, and the supposedly optimal choice is sometimes dominated by a naive method....Another problem is the instability in the optimal portfolio: the proportions allocated to each asset are extremely sensitive to variations in expected returns, and adding a few observations may change the portfolio distribution completely. Also, optimal portfolios are not necessarily well diversified. Often a corner solution appears, where most of the investments are zero and large proportions are assigned to countries with relatively small capital markets and high average returns.

As it turns out, this attack is somewhat overstated. However, the basic point remains: *ex post* estimates of expected returns, based on arithmetic or weighted average estimators, can be unreliable estimates of future returns. Relative to estimates of variances and covariances, numerous empirical studies dating back to Jorion (1985) and Eun and Resnick (1988) demonstrate that estimates of expected returns are considerably more unstable over time.

Empirically, the parameter instability problem has a number of implications. For example, *ex ante* results concerning the return on a given portfolio may vary significantly from sample to sample. Jorion (1985) examines the out-of-sample performance of the two *ex post* optimal internationally diversified portfolios identified by Grubel (1968) and Levy and Sarnat (1974), together with two 'naive' portfolios, the equally weighted and market value weighted. As measured by the Sharpe ratio, Jorion found that over the next investment horizon, the *ex ante* performance of the two mean-variance efficient portfolios was inferior to the performance of the naive equally weighted portfolio. Jorion (1985) also provides evidence that, in estimating *ex post* returns, longer sampling windows, e.g., five years for monthly data, provides superior *ex ante* forecasting when compared with shorter sampling windows, e.g., 1 year of monthly data. The difficulty with longer sampling windows is that it takes a longer time interval for the estimates to react to changing market conditions.³

In addition to the length of the sampling window used to determine the relevant parameter inputs, the presence or absence of short-selling has been found to be fundamental in assessing the performance of mean-variance efficient portfolios. Even though the early studies implicitly assumed short selling was not permitted, at least since Jorion (1985) it has been recognized that odd results can be obtained when short selling is permitted. For example, Jorion reports results for the time series properties of the optimal weight on domestic assets in the *ex post* tangency portfolio. A considerable amount of short-selling is indicated at various times, as much as -2.4 times the total principal value of the portfolio at one point in 1978. For many types of investment situations, e.g., pension funds, life insurance companies, this amount of short selling would be unacceptable and unobtainable. Evidence on portfolio composition with short selling restrictions, e.g., Glen and Jorion (1993), indicates a dramatic narrowing of the number of assets held in the portfolio is likely, amplifying the concentration of a given portfolio in a small number of assets.

Somehow, proponents of the model believe that the out-of-sample prediction problems can be resolved by improving the estimation methods that are used. This still leaves the problem of identifying the appropriate portfolio from the set of mean-variance efficient portfolios. Following

Sharpe and others, the efficient frontier portfolio to be selected is that portfolio associated with the capital allocation line which is just tangent to the efficient frontier, i.e., the portfolio associated with the *capital market line*. This tangency portfolio can be determined by solving the following optimization problem, e.g., Eun and Resnick (1994):

$$\max_{\{w_i\}} \frac{E[R_p] - r}{\sigma_p} \quad \text{subject to:} \quad \sum_{i=1}^k w_i = 1$$

where r is, as before, the riskfree interest rate. On theoretical grounds, the tangency portfolio is the *ex ante* mean-variance-expected-utility optimizing risky portfolio. Even though the precise combination of riskless asset and risky tangency portfolio for any given investor requires specification of the relevant parameters for the investor's mean-variance expected utility function, the optimal risky portfolio has been determined.

Given the *ex post* estimates of the relevant means, variances and covariances, the optimality problem is solved and the resulting tangency portfolio will represent the optimal, *in-sample* portfolio. Whether this in-sample optimality translates into superior *out-of-sample* performance is an open question. The answer to this question becomes even more complex when foreign assets are admitted into the asset universe. In particular, the domestic currency return on a foreign asset depends on a combination two random variables: the return denominated in foreign currency terms; and, the change in the exchange rate. The correlation between foreign and domestic asset returns will tend to be lower than the correlations between domestic assets, making foreign assets excellent candidates for diversification. However, as illustrated in Goetzmann et al. (2005, p.20), the correlations of UK, US, French and German monthly equity market returns have changed dramatically over the available 1872-2000 sample period. For example, the correlation between US and French monthly equity returns for 1946-71 (1972-2000) period was -0.02 (.414). Using the same sub-samples, the US and UK correlation increased from .182 to .508.

In contrast to the modern period, the correlations between equity market returns investment over the 1872-1914 sample associated with the 'average investment trusts' confirms the empirical basis of the 'geographical distribution of risks'. As Lowenfeld (1909) observed:

The fact that whilst the world's trade is constantly expanding, the share of each separate nation in it is constantly altering is the fundamental principle of our *geographical method of equalizing risks*. For it follows that if an investor widely distributes his own capital over the earth's surface, local depression in one quarter will be counter-balanced by the local trade activity in another quarter. Further, it also follows, if an investor's capital is sufficiently large to enable him to purchase investments representative of every trading centre in the world, that the world's perpetual trade expansion will automatically increase the realisable capital value of his investments year by year.

Compared to an average correlation among major markets of .475 for 1972-2000, the correlation was .102 (.155) for the 1872-89 (1890-1914) sample. In the modern context, the higher geographical correlations make returns more dependent on global trade expansion.

Despite the considerable attention paid to the risk diversification benefits of foreign assets, it is the possibility of significantly higher *ex ante* and, in some cases, *ex post* returns than those on offer for domestic assets that drives the bulk of the demand for offshore equity securities. This creates real difficulties for the *ex ante* performance of mean variance optimal portfolios. In Eun and Resnick (1994), for example, the difficulties associated with estimating expected returns results in the minimum variance portfolio having generally superior *ex ante* performance compared to the *ex post* optimal tangency portfolio. The difficulties in using *ex post* estimates to proxy for *ex ante* expected returns is increased significantly for foreign securities compared to domestic securities due to the additional currency risk. The impact of currency risk is also observed in the home country bias observed in investment portfolios in different countries. Though currency risk can be hedged with currency derivatives, this is not common practice among individual investors. While this is partly due to having position sizes smaller than exchange trade minimums for derivative contracts, the amount to hedge is complicated by the amount of hedging that individual firms are also doing.

B. Separation, the CAPM and the Market Model

Two Fund Separation

The combination of mean-variance expected utility, perfect markets and the CML provides the basis for a version of the ***two fund separation property***: in market equilibrium, rational risk averse investors will hold portfolios which combine the riskfree asset with the tangency portfolio. The precise combination of the riskfree asset and the tangency portfolio will depend on the risk preferences of the individual investor. The CML result does not provide any information about how to determine the return on individual assets, or any portfolio of assets which is not efficient. The CML also does not provide specific information about the asset composition of the tangency portfolio. This information is provided by the capital asset pricing model (CAPM). If the CAPM is incorporated, then it can be shown that the tangency portfolio will be the market portfolio. In this case, the two fund separation property says that, in market equilibrium, ***rational risk averse investors will hold portfolios which combine the riskfree asset and the market portfolio***.

Two fund separation provides the theoretical basis for a persuasive and implementable investment strategy. This strategy requires a strong belief in efficient markets. If markets are efficient then the gains to individual security selection strategies, using either fundamental or technical analysis, will be illusory. The decision problem facing the rational investor is to determine what fraction of invested capital to hold in the risky market portfolio – effectively a fixed, open ended, equally weighted managed index fund – and what fraction to hold in the riskless asset. Investors with high levels of risk tolerance will leverage up, by borrowing at the riskfree rate, and purchase more of the market portfolio. Investors with moderate to low levels of risk tolerance will have positive investment weights for both the riskfree asset and the market portfolio. Though this perfect markets result requires some adjustment to account for market imperfections, e.g., differences between lending and borrowing rates, the basic intuition survives in tact. As pointed out by Roll (1978), the main practical ambiguities lies with the specification of the riskfree asset and the market portfolio.

The CAPM provides a method for determining the expected return, $E[R_i]$, for any asset i , not just for portfolios on the efficient frontier. The CAPM is an *ex ante* model that can be expressed as:

$$E[R_i] = r + \{E[R_m] - r\} \beta_i$$

where $E[R_m]$ is the expected return on the market portfolio and β_i is a measure of the *systematic risk* of asset i . In words, the CAPM can be expressed as: the expected return on asset i = risk free rate + systematic risk premium for asset i . A key variable in the CAPM is β which is specified as:

$$\beta_i = \frac{\text{cov}[R_i, R_m]}{\sigma_m^2} \equiv \frac{\sigma_{i m}}{\sigma_m^2}$$

where σ_m^2 is the variance of the return on the market portfolio.

Some examples of mechanical calculations that can be done with CAPM are: assume that the rate of return on the market $E[R_m] = .15$ and $r = .05$ and $\beta_i = 1.5$ then the expected return on asset i is $E[R_i] = .05 + (.15 - .05)(1.5) = .20$; assume that $E[R_i] = .1$, $E[R_m] = .105$ and $\beta_i = .9$ then the riskfree rate r is 5.5%; assume that $E[R_i] = .2$, $E[R_m] = .15$ and $r = .10$, then $\beta_i = 2$. Beta is applicable not only for individual securities but also for portfolio of securities. The following useful result can readily be derived: the beta of a portfolio, β_p is the value weighted sum of the individual betas (the β_i 's). This follows because the CAPM holds for any asset, including individual assets as well as portfolios of assets. Recognizing that the CAPM will hold for the efficient portfolios on the efficient frontier, the CAPM can be used to show that the tangency portfolio in the CML is the market portfolio.

To demonstrate this result, assume that the CAPM is true. If the tangency portfolio is the market portfolio, the CML provides the result:

$$E[R_R] = r + \frac{E[R_m] - r}{\sigma_m} \sigma_R$$

where m refers to the market portfolio. Using the result that $\sigma_R = (1 - w_r) \sigma_m$ this can be rewritten:

$$E[R_R] = r + \frac{E[R_m] - r}{\sigma_m} (1 - w_r) \sigma_m = r + \{E[R_m] - r\} (1 - w_r)$$

If the CAPM is true then it will hold for any portfolio along the CML. If the market portfolio is the tangency portfolio for the CML then $E[R] = w_r r + (1 - w_r) E[R_m]$ and evaluating β gives:

$$\begin{aligned} \beta &= \frac{\text{cov}[E[R], E[R_m]]}{\sigma_m^2} = \frac{\text{cov}[\{w_r r + (1 - w_r) E[R_m]\}, E[R_m]]}{\sigma_m^2} \\ &= \frac{(1 - w_r) \sigma_m^2}{\sigma_m^2} = (1 - w_r) \end{aligned}$$

Substituting this result back into the CML shows that the CAPM and CML are equivalent when the

tangency portfolio is the market portfolio.

INSERT Figure 5-c
Security Market Line and Equity Security Valuation

The relationship between the CAPM and the CML can be expressed in a linear form in $(E[R], \beta)$ space. This linear relationship is the **security market line** (SML). While the CAL and CML provide a linear relationship in $(E[R], \sigma)$ space between total risk, as measured by standard deviation, and expected return, the SML provides a linear relationship between systematic risk, as measured by β , and expected return. The equation for the SML can be derived by identifying two points on the line: the riskfree rate where $\beta_f = 0$ and $E[R_f] = r$ and the market portfolio where $\beta_m = 1$, $E[R_m] = E[R_m]$. It follows that the slope is $(E[R_m] - r)$. From this the equation for the SML can be stated: $E[R_i] = r + (E[R_m] - r) \beta_i$. Hence, the SML is the graphical representation of the CAPM. A useful pedagogical application of the SML is to describe whether a particular equity security is over or underpriced relative to its measure of systematic risk. As illustrated in Figure 5-c, points above the SML have an $E[R]$ that is higher than warranted for the associated β and, as a consequence, represent underpriced securities. The instability of alpha and beta estimates from the market model make this method of valuing equity securities inapplicable for vernacular Finance purposes.

The Capital Asset Pricing Model*

The Markowitz model of mean-variance portfolio optimization is concerned with the behavior of an individual investor selecting securities for inclusion in an optimal portfolio. Practical application of this model is complicated by the large number of parameters that have to be estimated and the associated complexity of the solutions as the number of securities is increased. The CAPM provided a theoretical mechanism for handling this problem. Using the CAPM, the problem of estimating the optimal weights for the individual assets in the tangency portfolio is replaced with a method of identifying the predetermined set of weights associated with the market portfolio. Though the notion of a 'market portfolio' is somewhat nebulous, in practice it has been interpreted to be a widely diversified value-weighted portfolio of common stocks such as the S&P 500, e.g., Damodaran (1994). This all raises the need to examine the derivation of the CAPM in more detail.

In the years since the CAPM was introduced by Sharpe (1964), Lintner (1965) and Mossin (1966), considerable effort has been given to extending and expanding the basic model. The basic CAPM, also referred to as the one factor or single index model, is derived under perfect markets assumptions. The process of extending and expanding has been largely concerned with relaxing these assumptions. Unlike the partial equilibrium approach of the mean-variance portfolio model, the CAPM is a general equilibrium model. It is this feature that permits the CAPM to go beyond the basic portfolio structure to make statements about the expected returns for individual assets. General equilibrium requires market clearing conditions for all assets to be satisfied. To accomplish this, the CAPM relies on the assumption the investors are homogeneous, possessing the same expectations about the means and variances of returns and the same investment horizon. All investors are assumed to have the same form of mean-variance expected utility function.

Much of the derivation of the CAPM follows the mean-variance optimization procedure. The

homogeneity assumption is invoked after the riskfree rate is introduced and the optimality of the tangency portfolio is established. Because investors are homogeneous and market clearing is required, it must be that the tangency portfolio is the market portfolio. Fama (1976, p.274-5) describes the logical argument:

a market equilibrium requires a market-clearing set of prices; a market equilibrium requires that, in aggregate, investors demand them in the proportions in which they are outstanding. Given the nature of the efficient set when there is risk-free borrowing and lending, this market-clearing condition means that a market equilibrium is not attained until the one tangency portfolio that all investors try to combine with risk-free borrowing or lending is a portfolio of all the positive variance securities in the market, where each security is weighted by the ratio of the total market value ... of all its outstanding units to the total market value of all outstanding units of securities. In short, a market equilibrium is not reached until the tangency portfolio ... is the value-weighted version of the market portfolio ... A market equilibrium – a set of security prices that clears the securities market and a value of (the risk-free rate) that clears the borrowing-lending market – requires that the tangency portfolio be the market-weighted version of the market portfolio.

This last step, the identification of the tangency portfolio with the market portfolio, follows immediately from the investor homogeneity assumption. In the derivation of the CAPM, this step is something of an afterthought.

The key parts of the CAPM relate to developing the relationship between the expected return on a given asset and the expected return on an efficient portfolio. This derivation requires one useful result associated with linear combinations of random variables:⁴

$$\begin{aligned} \text{var}[R_p] &= \sum_{i=1}^k \sum_{j=1}^k w_i w_j \sigma_{ij} = \sum_{i=1}^k w_i \left(\sum_{j=1}^k w_j \sigma_{ij} \right) \\ &= \sum_{i=1}^k w_i \text{cov}[R_i, R_p] \end{aligned}$$

Given this, the derivation of the CAPM proceeds by solving the Lagrangian arising from the mean-variance portfolio model:

$$\max_{\{w_i\}} L = \text{var}[R_p] - 2 \lambda_1 \left(\sum_{i=1}^k w_i E[R_i] - \bar{c}_n \right) - 2 \lambda_2 \left(\sum_{i=1}^k w_i - 1 \right)$$

This optimization problem will produce k first order conditions associated with the $\{w_i\}$ together with two additional first order conditions for the constraints to produce a system of $k+2$ equations.

For j th security, the first order condition provides:

$$\sum_{i=1}^k w_i \text{cov}[R_i, R_j] - \lambda_1 E[R_j] - \lambda_2 = 0$$

As the ordering of the securities is arbitrary, this result can be equated with the first order condition

for the first security to obtain:

$$\sum_{i=1}^k w_i \text{cov}[R_i, R_j] - \lambda_1 E[R_j] = \sum_{i=1}^k w_i \text{cov}[R_i, R_1] - \lambda_1 E[R_1]$$

The next step involves multiplying both sides by w_j and summing over j . This affects the right and left hand sides differently. The left hand side produces:

$$\sum_{j=1}^k w_j \sum_{i=1}^k w_i \text{cov}[R_i, R_j] - \lambda_1 \sum_{j=1}^k w_j E[R_j] = \text{var}[R_p] - \lambda_1 E[R_p]$$

The right hand side produces:

$$\sum_{j=1}^k w_j \left(\sum_{i=1}^k w_i \text{cov}[R_i, R_1] - \lambda_1 E[R_1] \right) = \sum_{i=1}^k w_i \text{cov}[R_i, R_1] - \lambda_1 E[R_1]$$

This result follows because there is no j on the right hand side and, as a result, the sum of the weights equals one and has no impact.

Observing that the weights apply to a mean-variance efficient portfolio, i.e., R_p is on the efficient frontier, manipulating the right and left hand sides produces the result:

$$E[R_1] - E[R_p] = \frac{1}{\lambda_1} \left(\sum_{i=1}^k w_i \text{cov}[R_i, R_1] - \text{var}[R_p] \right) = \frac{1}{\lambda_1} (\text{cov}[R_1, R_p] - \text{var}[R_p])$$

What remains is to determine λ , which is the Lagrange multiplier associated with the impact of changes in the target level of portfolio expected return on the variance of the portfolio. When there is a riskfree rate, the λ for the tangency portfolio can be determined as:

$$2 \lambda_1 = \frac{d \text{var}[R_p]}{d E[R_p]} = \frac{d \text{var}[R_p]}{d \sigma_p} \frac{d \sigma_p}{d E[R_p]} = 2 \sigma_p \left(\frac{\sigma_p}{E[R_p] - r} \right)$$

Substituting this λ result back into the prior equation and manipulating gives the CAPM.

Nothing in this derivation demonstrates that the tangency portfolio is the market portfolio. This result is obtained from the logical argument about market clearing with homogeneity of consumers. In a sense, two fund separation, where the rational investor holds combinations of the market portfolio and the riskless asset, is too strong a condition. A more appropriate result would be a partial equilibrium result where the rational investor holds combinations of a mean-variance efficient portfolio and the riskless asset. But this would require the mean-variance efficient portfolio to be determined, falling back to the problems associated with complexity, number of parameters to estimate, estimator forecasting error and the like. That the CAPM assumptions make proponents of modern Finance uncomfortable is implicit in the following quote from Elton and Gruber (1984, p.273):

the final test of a model is not how reasonable the assumptions behind it appear but how well the model describes reality ... many assumptions [may be] objectionable. Furthermore, the final model is so simple the reader may well wonder about its validity ... despite the stringent assumptions and the simplicity of the model, it does an amazingly good job of describing prices in capital markets.

The real world is sufficiently complex that to understand it and construct models of how it works, one must assume away those complexities that, hopefully have only a small (or no) affect on its behavior. As the physicist builds models of the movement of matter in a frictionless environment, the economist builds models where there are no institutional frictions to the movement of stock prices.

These words reflect the relationship between the philosophical foundation of modern Finance and the CAPM. While acceptable in an academic context, the claim about "amazingly good job of describing prices" is not supported in the world of vernacular Finance where forecasting future equity prices is of central importance.

The Market Model

Because the CAPM is an *ex ante* model it depends on expected returns and other unknown values, that are not directly observable or testable. The statistical representation of the CAPM which is testable is known as the **market model**. The market model is a bivariate regression model of the form:

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + e_{i,t}$$

where $R_{i,t}$ is the observed return on asset i at time t , $R_{m,t}$ is the observed return on the market portfolio at time t , α_i and β_i are statistical parameters to be estimated and $e_{i,t}$ is the asset specific error which is assumed to obey the statistical properties for ordinary least squares regression. For the market model the ordinary least squares (OLS) assumptions (in vector notation) are: $E[e_i] = 0$, the firm specific error has mean zero; $E[e_i R_m] = 0$, the firm specific risks are uncorrelated with the market return; $E[e_i e_j] = 0$ for i not equal to j , the firm specific risks for different securities are uncorrelated; in addition, it is assumed that the e_i are iid random variables. The additional assumption of normality of e_i facilitates hypothesis testing. With these assumptions, ordinary least squares can be used to estimate the coefficients α_i and β_i . The market model is sometimes referred to as the single index model, e.g., Elton and Gruber (1995).

Taking expected values for the market model gives for any t :

$$E[R_i] = \alpha_i + \beta_i E[R_m] \quad \text{because } E[e_i] = 0$$

If the CAPM is true, it follows that $\alpha_i = (1 - \beta_i) r$, (β_i has the same interpretation as $cov[R_i, R_m]/var[R_m]$). This interpretation for α depends on the riskless rate being a constant. While this assumption is correct in one-step-ahead decision making, it is problematic when estimating parameters in a time series. To account for changes in r over time, the market model is often expressed in **risk premium form**, by subtracting the observed riskless rate, in any given period, from

the observed returns:

$$R_{i,t} - r_t = \alpha_i + \beta_i [R_{m,t} - r_t] + u_{i,t}$$

where u_i is also a firm specific risk with ordinary least squares properties. In this form, taking expectations and assuming that the CAPM holds gives $\alpha_i = 0$ and β_i with the same interpretation.

Beta is a measure of systematic or **market risk**. It provides information on how the stock return reacted historically when the market portfolio changed. Beta estimates are reported at a number of websites, such as www.bloomberg.com. For $\beta_i > 1$ (high beta), when the return on the market portfolio changes, the return on the stock will tend to change by **more** than the return on the market portfolio. Stocks with higher than market betas are considered to be **aggressive**. When the market is expected to move up, shifting into stocks with high beta is indicated. In the US market, examples of high beta stock groups occur in industries such as trucking, consumer durables, construction and air transport. For $\beta_i < 1$ (low beta) when the return on the market portfolio changes, the return on the stock will tend to change by **less** than the return on the market portfolio. Stocks with lower than market betas are considered to be **defensive**. When the market is expected to move down, shifting into stocks with high beta is indicated. In the US market, examples of low beta stock groups occur in telephone stocks, utilities, breweries and food producers/distributors.

Alpha is an **asset specific** measure which indicates the **excess return** that the security earned beyond that warranted by the risk premium captured by the security's beta. For the market model expressed in risk premium form, positive alpha indicates that the stock outperformed the market, after adjusting for systematic risk. Negative alpha indicates that the stock underperformed the market, after adjusting for systematic risk. The market model provides a useful simplification for determining the betas and alphas for a portfolio from the alphas and betas of the individual securities:

$$E[R_p] = \sum_{i=1}^k w_i E[R_i] = \sum_{i=1}^k w_i \{ \alpha_i + \beta_i E[R_m] \}$$

$$\sum_{i=1}^k w_i \alpha_i + \sum_{i=1}^k \beta_i E[R_m] = \alpha_p + \beta_p E[R_m]$$

In other words, the alpha and beta for the portfolio are the value weighted sums of the individual portfolio alphas and betas.

The market model can also be used to demonstrate that portfolio diversification leads to the elimination of unsystematic or **firm specific** risk leaving only systematic risk as the determinant of portfolio variance. If the market model is true then the variance for individual security returns reduces to:

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{e_i}^2$$

Similarly, the covariance between the security returns becomes:

$$\sigma_{i,j} = \beta_i \beta_j \sigma_m^2$$

These results follow from the assumptions made in the market model. Substituting these results into the formula for the portfolio variance, σ_p^2 gives:

$$\begin{aligned}\sigma_p^2 &= \sum_{i=1}^k w_i^2 \beta_i^2 \sigma_m^2 + \sum_{i=1}^k w_i^2 \sigma_{e_i}^2 + \sum_{i>j} w_i w_j \beta_i \beta_j \sigma_m^2 \\ &= \sum_{i=1}^k \sum_{j=1}^k w_i w_j \beta_i \beta_j \sigma_m^2 + \sum_{i=1}^k w_i^2 \sigma_{e_i}^2\end{aligned}$$

Observing that the variance of the market portfolio is a common term in the double sum, it is possible to do some factoring:

$$\sigma_p^2 = \left\{ \sum_{i=1}^k w_i \beta_i \right\} \left\{ \sum_{j=1}^k w_j \beta_j \right\} \sigma_m^2 + \sum_{i=1}^k w_i^2 \sigma_{e_i}^2$$

Using the result that the beta of the portfolio is the value weighted sum of the individual security betas, the following simplification is available for the portfolio variance:

$$\sigma_p^2 = \beta_p^2 \sigma_m^2 + \sum_{i=1}^k w_i^2 \sigma_{e_i}^2$$

$$\sigma_p \rightarrow \beta_p \sigma_m \quad \text{as} \quad k \rightarrow \infty \quad \text{and} \quad w_i \Downarrow$$

The term involving β_p is the **systematic** or market related risk. It depends only on the composition of the portfolio and the variance of the return on the market. The second term involves only firm specific or **unsystematic** risks.

To show the impact of diversification on both systematic (market) risk and unsystematic (firm specific) risk, observe that the first term involving the beta of the portfolio is not much affected by increases in the number of securities in the portfolio. The second term involving the firm specific risks is directly affected by the number of securities in the portfolio. Take the case of an equally weighted portfolio where $w_i = 1/N$. In this case, as the number of the securities in the portfolio increases the firm specific risks are reduced at the rate of $(1/N^2)$, which is converging to zero quite rapidly. This follows from observing that when the firm specific risk has been eliminated then the last term on the rhs of the last equation is zero. Taking square roots and observing that the portfolio beta is the weighted sum of the individual security betas provides the required result. However, because there are a large number of securities in a portfolio, this does **not** mean that the firm specific risks are eliminated. Rather, the elimination of firm specific risk depends on reducing the value weights attached to each security as N increases. For example, if $w_i = .5$ for a particular security, and this value does not change as the number of securities in the portfolio increases, then the firm specific risk associated with that security is not eliminated as the number of securities in the portfolio increases.

5.2 Ergodicity and Asset Pricing Theories

A. A Brief History of Ergodic Theory

The Encyclopedia of Mathematics (2002) defines ergodic theory as the “metric theory of dynamical systems. The branch of the theory of dynamical systems that studies systems with an invariant measure and related problems.” This modern definition implicitly identifies the birth of ergodic theory with proofs of the mean ergodic theorem by von Neumann (1932) and the pointwise ergodic theorem by Birkhoff (1931). These early proofs have had significant impact in a wide range of modern subjects. For example, the notions of invariant measure and metric transitivity used in the proofs are fundamental to the measure theoretic foundation of modern probability theory (Doob 1953; Mackey 1974). Building on Kolmogorov (1933), a seminal contribution to probability theory, in the years immediately following it was recognized that the ergodic theorems generalize the strong law of large numbers. Similarly, the equality of ensemble and time averages – the essence of the mean ergodic theorem – is necessary to the concept of a strictly stationary stochastic process. Ergodic theory is the basis for the modern study of random dynamical systems, e.g., Arnold (1998). In mathematics, ergodic theory connects measure theory with the theory of transformation groups. This connection is important in motivating the generalization of harmonic analysis from the real line to locally compact groups.

From the perspective of modern mathematics, statistical physics or systems theory, Birkhoff (1931) and von Neumann (1932) are excellent starting points for a history of ergodic theory. Building on the ergodic theorems, subsequent developments in these and related fields have been dramatic. These contributions mark the solution to a problem in statistical mechanics and thermodynamics that was recognized sixty years earlier when Ludwig Boltzmann (1844-1906) introduced the ergodic hypothesis to permit the theoretical phase space average to be interchanged with the measurable time average. From the perspective of economics, the selection of the less formally correct and rigorous contributions of Boltzmann are a more auspicious beginning for a history of the ergodic hypothesis. Problems of interest in mathematics are generated by a range of subjects, such as physics, chemistry, engineering and biology. The formulation and solution of physical problems in, say, statistical mechanics will have mathematical features which are unnecessary in, say, economics. For example, in statistical mechanics, points in the phase space are often multi-dimensional functions representing the mechanical state of the system, hence the desirability of a group-theoretic interpretation of the ergodic hypothesis. From the perspective of both mainstream and Post Keynesian economics, such complications are largely irrelevant and an alternative history of ergodic theory that captures the etymology and basic physical interpretation is more revealing than a history that focuses on the relevance for mathematics. This arguably more revealing history begins with the formulation of the problems that von Neumann and Birkhoff were able to solve.

Mirowski (1989, esp. ch.5) establishes the importance of 19th century physics in the development of the neoclassical economic system advanced by Jevons, Walras and Menger during the marginalist revolution of the 1870's. As such, neoclassical economic theory inherited essential features of mid-19th century physics: deterministic rational mechanics; conservation of energy; and the non-atomistic continuum view of matter that inspired the energetics movement later in the 19th century.⁵ It was during the transition from rational to statistical mechanics during the last third of the century that Boltzmann made the contributions that led to the transformation of theoretical physics from the microscopic mechanistic models of Rudolf Clausius (1822-1888) and James Maxwell (1831-1879)

to the macroscopic probabilistic theories of Josiah Gibbs (1839-1903) and Albert Einstein (1879-1955).⁶ Coming largely after the start of the marginalist revolution in economics, this fundamental transformation in theoretical physics and mathematics had little impact on the progression of mainstream economic theory until the appearance of contributions on continuous time finance in the 1970's.⁷ The deterministic mechanics of the energistic model was well suited to the subsequent axiomatic formalization of neoclassical economic theory which culminated in the von Neumann and Morgenstern expected utility approach to modeling uncertainty and the Bourbaki inspired Arrow-Debreu general equilibrium theory, e.g., Davidson (2007), Weintraub (2002).

Having descended from the deterministic rational mechanics of mid-19th century physics, defining works of neoclassical economics, such as Hicks (1939) and Samuelson (1947), do not capture the probabilistic approach to modeling systems initially introduced by Boltzmann and further clarified by Gibbs.⁸ Mathematical problems raised by Boltzmann were subsequently solved using tools introduced in a string of later contributions by the likes of the Ehrenfests and Cantor in set theory, Gibbs and Einstein in physics, Lebesgue in measure theory, Kolmogorov in probability theory, Wiener and Levy in stochastic processes. Boltzmann was primarily concerned with problems in the kinetic theory of gases, formulating dynamic properties of the stationary Maxwell distribution – the velocity distribution of gas molecules in thermal equilibrium. Starting in 1871, Boltzmann took this analysis one step further to determine the evolution equation for the distribution function. The mathematical implications of this analysis still resonate in many subjects of the modern era. The etymology for “ergodic” begins with an 1884 paper by Boltzmann, though the initial insight to use probabilities to describe a gas system can be found as early as 1857 in a paper by Clausius and in the famous 1860 and 1867 papers by Maxwell.⁹

The Maxwell distribution is defined over the velocity of gas molecules and provides the probability for the relative number of molecules with velocities in a certain range. Using a mechanical model that involved molecular collision, Maxwell (1867) was able to demonstrate that, in thermal equilibrium, this distribution of molecular velocities was a ‘stationary’ distribution that would not change shape due to ongoing molecular collision. Boltzmann aimed to determine whether the Maxwell distribution would emerge in the limit whatever the initial state of the gas. In order to study the dynamics of the equilibrium distribution over time, Boltzmann introduced the probability distribution of the relative time a gas molecule has a velocity in a certain range while still retaining the notion of probability for velocities of a relative number of gas molecules. Under the ergodic hypothesis, the average behavior of the macroscopic gas system, which can objectively be measured over time, can be interchanged with the average value calculated from the ensemble of unobservable and highly complex microscopic molecular motions at a given point in time. In the words of Wiener (1939, p.1): “Both in the older Maxwell theory and in the later theory of Gibbs, it is necessary to make some sort of logical transition between the average behavior of all dynamical systems of a given family or ensemble, and the historical average of a single system.”

B. Uncertainty and the Ergodic Hypothesis

In aiming to achieve a scientific approach, positivists are fundamentally concerned with the quantification, measurement and empirical verification of hypotheses. As a key assumption in the application of statistical methodology to time series data, ergodicity lies at the philosophical core of

modern Finance. This statement captures the thrust of the strong criticisms that Davidson (1991, p.132-3) and others make about the economic foundations of modern Finance: "Acceptance of the presumption of an ergodic economic environment is often rationalized by the necessity of developing economics as an empirically based science. Indeed, Samuelson has made the acceptance of *the 'ergodic hypothesis'* the sine qua non of the scientific method in economics." In Finance, ergodicity plays a fundamental role in converting *ex post* logical relationships, such as the CAPM or Markowitz mean-variance diversification models, into *ex ante* prescriptions for investment strategy. It is an essential component of the efficient markets hypothesis and is the driving force behind the fascination with the risk-return tradeoff and the equity risk premia, e.g., Mehra and Prescott (1985), Kocherlakota (1996), Constantinides (2002).

Formal Definitions

Ergodicity is a property of stochastic processes. Formally, a stochastic process can be defined:

Definition: Let $\{X(t)\}$ be a family of random variables indexed by the linear (index) set \mathfrak{X} , where $t \in \mathfrak{X}$. Then $\{X(t)\}$ is said to be a **stochastic process**.

In Finance, the terms stochastic (random) process and time series are often used interchangeably, though it is possible for the index set to refer to some linear variable other than time. Following Karlin and Taylor (1975, p.32), "a stochastic process may be considered as well defined once its state space, index parameter and family of joint distributions are prescribed." Similar approaches can be found in other sources, e.g., Dhrymes (1974, p.383): "The probability characteristics of a stochastic process $\{X(t)\}$ are completely specified if we determine the joint density function of a finite number of members of the family of random variables comprising the process."

Heuristically, the theory of stochastic processes describes the behavior of random variables, the X 's, over time, $t \in \mathfrak{X}$. Conventionally, **a random variable** is a function that maps from a prespecified domain, or sample space, to some portion of the real line, \mathfrak{R}^1 . In the theory of stochastic processes, a single realization of X defines a sample path starting at, say, $X(0)$ and ending at $X(T)$. When the distribution of X is continuous, there are an infinite number of such possible sample paths. In order to make reference to individual sample paths, it is necessary to further introduce another indexing variable for X , $\xi \in \Xi$. This index allows individual samples paths or 'states' of X to be identified. It follows that $X(\xi, T)$ would refer to the time $t=T$ observation from a single sample path in $\{X(t)\}$ that starts at $X(0)$ and ends at $X(T)$ and $\{X(\xi, T)\}$ would be the set of all $X(\xi, T)$ at $t=T$. The $\xi \in \Xi$ index makes it possible to define the operation of summing over the ξ at any time $t=T$. Such operations are relevant to identifying the properties of one of the joint distributions of the $\{X(t)\}$ at a single point in time.

In certain financial applications, e.g., where X refers to a security price, X takes values only on the positive, half line. In this case as well as when the X values are allowed to assume any value along the real line, it is conventional to assume that there is a zero probability of X being equal to plus or minus infinity. When t is fixed at a given point, $X(t)$ has the conventional interpretation of a random variable, with associated (one-dimensional) probability density function. In contrast, the ergodicity assumption is concerned with using the $X(\xi, t)$, $X(\xi, t+1)$, $X(\xi, t+2)$... $X(\xi, T)$ observations from a

single sample path to estimate the parameters of the joint distributions defining the $\{X(t)\}$. Specification of the stochastic process for X requires **specification of the joint density functions** that relate X 's at different points in time: the joint densities provide a probabilistic specification of how X evolves over time. This potentially complicated mapping can involve various combinations of discrete or continuous observations on X and t .

In many empirical applications of stochastic process theory, the objective is to rationalize how to use past and present observations on $X(t)$ ($t \leq 0$) to predict future values ($t > 0$). A classical example of this type of reasoning in Finance is: "Stock returns will outperform bond returns in the long run". Based on past realizations of the time series of returns on stocks and bonds, a prediction is made about the future path for returns. The task of prediction is difficult because the past and present $X(t)$ represent only one realization of the process, i.e., **there is only one observed sample path**. Yet, for any given $t \in \mathcal{T}$ the joint probability densities can be used to specify an infinite number of future possible paths for $X(t)$. Theoretically, an *assumption* is required to permit the statistics for the joint probability densities, i.e., the means, variances and other parameters, to be calculated from a single realization of the process. The requisite assumption invokes some form of ergodicity for the stochastic process.

To visualize how an ergodicity assumption works, choose a given starting value for a stochastic process, $X(0)$. From this starting point, the (continuous) joint probability distributions of the stochastic process define an infinite number of possible future paths for $X(t)$. Between $t = 0$ and $t = T$ each of these paths will start at $X(0)$ and reach some point $X(\xi, T)$ at time T . It is now possible to take a 'large number' of the points for these paths at T and calculate an **arithmetic mean** of the $\{X(\xi, T)\}$. Setting N to be a large number this gives:

$$\bar{X}[N, T] = \frac{1}{N} \sum_{\xi=1}^N X(\xi, T)$$

The set of X defined by the ξ is referred to as the **ensemble** of time paths. Ergodic theorems are concerned with the conditions under which $M(T)$, the arithmetic average calculated from an individual time path from $t=1$ to $t=T$, converges to the same limit (mean value) as the ensemble average taken at T . More precisely, for t measured discretely:

$$M(T) = \frac{1}{T} \sum_{t=1}^T X(t) = \frac{1}{T} \sum_{t=1}^T X(t \mid \xi = a) \leftrightarrow \bar{X}[N, T]$$

Being concerned with the convergence properties of the arithmetic mean, ergodic theorems are closely related to the strong and weak laws of large numbers, e.g., Feller (1957, ch X).

Laws of Large Numbers

An important convergence property of the arithmetic mean of the ensemble of time paths is given by the strong law of large numbers. Under certain conditions, such as stationarity of the stochastic process, the process is ergodic and the strong law also applies to time averages. More precisely, if the $\{X(\xi, T)\}$ are independently and identically distributed (iid) with mean $|\mu| < \infty$, the **strong law of large numbers** for the ensemble average states:

$$Pr \left\{ \lim_{N \rightarrow \infty} \bar{X}[N, T] = \mu \right\} = 1$$

where μ is the population mean of $\{X(\xi, T)\}$, i.e., $\mu = E[X(\xi, T)]$. In words, the strong law states that, for a random sample of iid $\{X(\xi, T)\}$ observations, the sample mean will converge to the population mean with probability 1. This is purely a convergence property of the mean, no restriction is imposed on the variance or higher moments. Because $\{X(\xi, T)\}$ is iid, it follows that $\mu = E[X(\xi, T)]$.

The weak law of large numbers is so-called because it deals with convergence in probability. A process which converges with probability one will also converge in probability, but not conversely. Applied to the ensemble averages, the **weak law of large numbers** requires:

$$\lim_{N \rightarrow \infty} Pr \left\{ |\bar{X}[N, T] - \mu| > \epsilon \right\} = 0$$

where, for large enough N , ϵ can be chosen to be an arbitrarily small positive number. A key result, due to Khinchine (Khintchine), is that if $\{X(\xi, T)\}$ or, more generally, $\{X(\xi, t)\}$ is a sequence of independently, identically distributed random variables with a finite mean μ , then this sequence will obey the weak law. The difference between the strong and weak law relates to the type of convergence which is imposed. By imposing convergence with probability one, the strong law applies to the properties of the arithmetic average as N increases to the limit. In using convergence in probability, the weak law only applies to the arithmetic average at the limit.

In modern presentations, the strong and weak laws apply to the properties of the arithmetic mean. Where additional conditions are imposed on the variance and, possibly, higher moments, then attention shifts to the central limit theorems which provide information not only about the mean but also the distribution of the sequence. In particular, by imposing additional restrictions to those required for the strong law, the central limit theorem can be used to estimate the size of the discrepancy between the arithmetic average and the population mean.¹⁰ This is accomplished by demonstrating that the distribution of the arithmetic average is asymptotically normal. The central limit theorem is a development on **Chebyshev's inequality** which states:

$$Pr \left\{ |X(\xi, T) - \mu| \geq \theta \right\} \leq \frac{\sigma^2}{\theta^2}$$

where $\sigma^2 = E[(X(\xi, t) - \mu)^2]$ and θ is a given constant. In this form, Chebyshev's inequality provides a relationship between the variance of a distribution and the probability for the size of observed deviations from the mean.

Feller (1966, p.219) observes: "Chebyshev's inequality must be regarded as a theoretical tool rather than a practical method of estimation. Its importance is due to its universality, but no statement of great generality can be expected to yield sharp results in individual cases." The use of the variance to specify Chebyshev's inequality is an essential component of the result. However, if it assumed that the random variable $X(\xi, T)$ is strictly positive, as is the case where X refers to a security price, then it is possible to derive a form of the inequality that does not involve the variance, i.e.:

$$Pr \left\{ X(\xi, T) \geq \alpha \right\} \leq \frac{E[X(T)]}{\alpha}$$

where $\alpha > 0$ is a given constant. This form of Chebyshev's inequality illustrates the extensions that are possible where X can be restricted to be positive.

The central limit theorem goes well beyond Chebyshev's inequality to make a precise statement about the form of the probability distribution which, in turn, can be used to provide a practical estimate of the size of the deviation of the arithmetic average from the mean ($|\mu| < \infty$) of the distribution, in terms of the distribution's standard deviation ($0 < \sigma < \infty$). More precisely, at any arbitrary time $t=T$:

Central Limit Theorem

Let $\{X(\xi, T)\}$ be a sequence of independently, identically distributed random variables with $\mu = E[X(\xi, T)]$ and $\sigma^2 = E[(X(\xi, T) - \mu)^2]$. Then for every fixed β at time T :

$$\lim_{N \rightarrow \infty} Pr \left\{ \sqrt{N} \frac{\bar{X}[N, T] - \mu}{\sigma} < \beta \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta} e^{-\frac{u^2}{2}} du = \Phi[\beta]$$

where $\Phi[\cdot]$ is the standard normal distribution.

This basic result has been generalized in a number of different ways, e.g., to stable processes that do not have a finite variance. The central limit theorem forms the basis of classical parametric tests of empirical hypotheses. As such, the central limit theorem is a key element in the statistical analysis of stochastic processes.

Stationarity and Ergodicity

With this background, it is now possible to proceed to the key logical step that has been identified as the *sine qua non* of the scientific method in modern Finance: the ergodicity theorem. Much as with the laws of large numbers and the central limit theorem, there are a number of possible variations of the ergodic hypothesis that depend on different assumptions. In comparison to the results which have already been presented, the ergodic theorems are something of a hybrid. As used in Finance, the theorems require assumptions about the stationarity of the stochastic process which, at the least, imposes conditions on the covariance function relating $X(t)$ with $X(t+i)$ for all t and $t+i$ defined by the time index set \mathfrak{X} . However, as the ultimate objective is to identify conditions under which time averages equal ensemble averages, a correspondence is usually drawn between the ergodic theorems and the strong and weak laws, e.g., Karlin and Taylor (1975, p.474-89). As such, results about ergodicity rely on assumptions about the stationarity of the stochastic process.

Two definitions for stationarity are usually presented: *strict stationarity* and *covariance stationarity*. Strict stationarity applies to the joint distributions of the stochastic process:

Definition: A stochastic process $\{X(t)\}$ is a **strictly stationary** process if, for any positive integer k ,

for all t to $t+k$ and $t+i$ to $t+i+k$ in the time index set \mathfrak{J} , the joint distribution of $\{X(t), X(t+1), X(t+2) \dots X(t+k)\}$ has the same joint distribution as $\{X(t+i), X(t+i+1), X(t+i+2) \dots X(t+i+k)\}$.

It is possible for a strictly stationary process to have no finite moments, e.g., a strictly stationary Cauchy process. Strict stationarity could be considered to be a strong assumption because it imposes requirements on the joint distributions when, for many results, all that is required is restrictions on the first two moments of the distribution. With this in mind, the definition for a covariance stationary (*weakly stationary*) process follows:

Definition: A stochastic process $\{X(t)\}$ is a **covariance stationary** process if the second moment $E[X(t)^2]$ (variance) is finite, the mean $E[X(t)] = \mu$ is constant and the temporal covariance $E[(X(t) - \mu)(X(s) - \mu)] = E[(X(t+i) - \mu)(X(s+i) - \mu)]$ depends only on the time difference $t - s$.

In the same fashion that a strictly stationary process may not satisfy covariance stationarity because the variance (and possibly the mean) are not finite as $T \rightarrow \infty$, it is also possible for a covariance stationary process to not be strictly stationary. In the special case where the joint distribution of the stochastic process is Gaussian, then covariance stationarity and strict stationarity have the same meaning. The covariance stationary process leads naturally to the definition of the covariance function, $C[k]$, that defines the temporal covariance $E[(X(t) - \mu)(X(t-k) - \mu)]$ for lag k .

This considerable background is now sufficient to state the conditions under which a single realization of a stochastic process $\{X(0), X(\xi_1), \dots, X(\xi_T)\}$ can be used to estimate the constant mean value of the joint distributions. Two types of ergodic theorems are available, one type which applies to covariance stationary processes and ‘corresponds’ to the weak law and one type which applies to strictly stationary processes and ‘corresponds’ to the strong law. The weak law variation takes the form:

Mean-Square Ergodicity Theorem¹¹

Suppose $\{X(t)\}$ is a covariance stationary process with covariance function $C[k]$. Then:

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{k=0}^{M-1} C[k] = 0$$

if and only if:

$$\lim_{T \rightarrow \infty} E[(M(T) - \mu)^2] = 0$$

Because the process is assumed to be covariance stationary with $\mu = E[X(t)]$, the convergence in quadratic mean part of the theorem relates to the limit of the variance of $M(T)$. The first condition relates to the convergence of covariance between $M(T)$ and $X(0)$. It follows that the mean square ergodic theorem says that the variance of $M(T)$ will go to zero in the limit if, and only if, the covariance between $M(T)$ and any arbitrary starting point $X(0)$ also goes to zero in the limit.

The connection of this theorem to the weak law is facilitated by observing that ***convergence in quadratic mean implies convergence in probability*** (but not the converse). In terms of quadratic mean convergence, the weak law applies when the elements of the sequence $\{X(t)\}$ are asymptotically uncorrelated.¹² In this vein, the mean-square ergodic theorem requires that as the lag (k) increases in the covariance function between $X(t)$ and $X(t-k)$, the covariance function goes to zero. If this condition is satisfied, then a single realization (time path) of a covariance stationary process can be used to estimate the mean of the ensemble of time paths, provided that the observed time path has a large enough number of observations. Though not easy to prove, this result is intuitive. The action, so to speak, is in the assumption of stationarity.

Casual inspection of the weak and strong laws, as well as the condition for mean square ergodicity reveals the dependence of these results on taking the limit as N or T goes to infinity. Hence, even accepting that the stochastic process satisfies the conditions needed for stationarity, ***a time path of "a sufficiently long duration"*** (Karlin and Taylor 1975, p.475) is still needed. In Finance applications this requirement can create complications, e.g., the longer is the time path the greater the possibility that the fundamentals driving the stochastic process will change due, say, to substantive regulatory changes or evolution of investor sentiments and so on. If the time path is not sufficiently long enough, then the distribution governing the outcomes will be subject to short run influences, such as the picking of an $X(0)$ which is too high or low relative to μ or to the possible impact of boundary conditions. More formally, for 'short' time paths the observed distribution will be a combination of the ergodic distribution and a sequence of transient terms, e.g., Linetsky (2005). Only if the process is allowed to run for a sufficiently long duration will the stochastic process dampen out the possible transients and permit the ergodic distribution to determine the properties of the arithmetic average.

Key elements in the specification of the strong and weak laws of large numbers and the related central limit theorem are the properties of independence and identical distribution. The statement of the strong law given above is only applicable if the $\{X(t)\}$ are independently and identically distributed. With some difficulty, it is possible to generalize this result to the case where the random sample uses $\{X(t)\}$ that are only independently distributed, where values of X observed at times up to and including T will not necessarily have constant mean, e.g., Dhrymes (1974, p.102). This generalization requires a $\mu(t)$, the mean at time t , to be introduced. Invoking ergodicity, the strong law can now be stated:

$$Pr \left\{ \lim_{T \rightarrow \infty} \frac{\sum_{t=1}^T |X(\xi, t) - \mu(t)|}{T} = 0 \right\} = 1$$

It is in this form that the strong and weak laws, the central limit theorem and related results are typically applied in applications using regression analysis. The operative random variable is the error term in a linear regression equation: $y = W\beta + u$, where y is a $T \times 1$ vector containing a time series of observations on the dependent variable of interest, W is a $T \times (k+1)$ matrix of the time series for k independent variables and a constant, β is a $(k+1) \times 1$ parameter vector to be estimated and u is a $T \times 1$ vector of unobserved error terms that is assumed to be strictly stationary with $E[u] = 0$. To see the

connection to the independence form of the strong law, let $y(t) = X(\xi, t)$ and $W(t)\beta = \mu(t)$. It follows that the law of large numbers applies to $u(t)$.

Shackle and Non-Additive Probability

One point where difficulties arise with this interpretation of an ergodic process occurs when the stationary densities are not unimodal. This can occur when the time dependence in the mean is non-linear. Stochastic theories of equity value based on time reversible ergodic processes imply unimodal stationary densities. In the case of path dependent bifurcating stochastic processes, such as the double well or pitchfork bifurcation processes, the time of the bifurcation point along any particular path is not known at $X(t=0)$. The analytical stability of the Lyapunov exponent along trajectories at a stochastic bifurcation point (Baxendale 1999) suggests trending behavior once the bifurcation point is breached stochastically. In an equity valuation context, introducing nonlinear time dependence in the mean of the *ex ante* stationary density can be used to resolve the *ex ante* / *ex post* dilemma confronting empirical implementation of modern Finance theories. In addition, stochastic trending behavior around bifurcation points provides a theoretical justification for the profitability of some technical analysis methods. The key insight is to recognize that the parameters of the *ex ante* stationary density, which determines current pricing, will almost certainly not correspond to parameter estimates based on the *ex post* realization of a single sample path.

Confronted with the difficulties raised by uncertainty in forming expectations, Shackle proposed “non-additive” measures of uncertainty. “After Shackle wrote on this subject, mainly in the 1950s, Probability Theory has come to consider classical probabilities as a special case of monotone measures of uncertainty, of which Choquet capacities are the best known representatives” (Fioretti 2009, p.284; Klir 2006). Following Shafer (1976), evidence theory uses Choquet capacities of infinite order to specify the monotone uncertainty measures. The context is cognitive, “it does not take a gambler as its prototypical subject, but a judge or a detective ... managers making investments [are] ... more akin to a judge or a detective looking for cues than to a gambler looking for luck.” A gambler is able to fully enumerate the possibilities that can occur. “On the contrary, judges and detectives know that unexpected proves and testimonies may open up unexpected possibilities” (Fioretti, *ibid.*). Beyond this, a number of alternative, not completely compatible, approaches to the handling of uncertainty are possible.

Confronted with the shortcomings of probability theory as a method of handling uncertainty, Shackle proposed “potential surprise”, and the associated concepts of ‘focus-loss’ and ‘focus-gain’, as a replacement. This concept was a non-additive, cardinal measure that depended on the “certainty of wrongness”. As Ford (1993, p.696) observes: “Shackle raised fundamental issues concerned with the individual’s perception and measurement of uncertainty: some of those basic issues await resolution.” Though the analytical methods Shackle proposed have not gained much traction, “in spite of its conceptual problems and the criticisms it has received, the theory can be utilised to explain many types of economic behaviour under uncertainty” (Ford, *ibid.*). Consider the following from Shackle (1952, p.76):

Investment can be depressed by either an increase in focus-losses or a decrease in focus-gains, and the effect in *both* these cases will seem to be immediate, though in fact the

downward movement following an event which is going to increase focus-losses will be due at first to the mere 'standstill' effect of the surprise event. This asymmetry seems at least partly to explain why the downturn of investment and employment after a boom is usually more abrupt and rapid than their upturn after a slump.

Altering the context slightly from aggregate investment to equity investment, the Shackle 'potential surprise' approach appears capable of explaining the empirical observation of a more violent downdraft of prices with a bear market compared to the gradual upswing of a bull market.

Ford (1993, p.688) describes the *leit motif* for Shackle as 'time, expectations and uncertainty'. Regarding time, Shackle shared much in common with the philosopher Henri Bergson where time evolves creatively not deterministically (Shackle 1958, p.23-4):

In classical dynamics of the physicist time is merely and purely a mathematical variable. The essence of his scheme of thought is the fully abstract idea of function, the idea of some working model or coded procedure which, applied to any particular and specified value of set of values of one or more independent variables, generates a value of a dependent variable. For the independent variable in a mental construction of this kind, *time* is a misnomer ... The solution to the differential equation, if it can be found, is complete in an instantaneous and timeless sense.

This timelessness ... abolishes the distinction between past and future. The physicist has, within the stated limits of his problem, complete, perfect and indisputable *knowledge* of where his particle will be at any instant; the very nature of human consciousness ... depends ... upon *ignorance* of the future ... upon the necessity to live in one moment at a time.

From this observation Shackle is able to make two strong conclusions (Ford 1993, p.690-1): "it was impossible to construct a dynamic model of an economic system except for one period at a time"; and, "the formal mathematical models of growth and of business cycles which were based on such mechanisms as difference equations were otiose, have no meaning, being necessarily built on mechanical, non-expectational time."

Shackle provides a number of avenues to justify certain types of stochastic behaviour exhibited by equity prices that confound conventional methods. One of these avenues leads to the Post Keynesian critique of the ergodicity hypothesis in mainstream economics. The critique recognizes that substantive difficulties arise in economic models, such as the rational expectations model, when it is not possible to use an individual sample time path of the stochastic process to estimate the mean value of the stationary distribution for the ensemble of future time paths, e.g., Davidson (1988, p.332): "If economic observations are generated in nonergodic circumstances, then the calculation of either time and/or space averages based on past data can not be expected to provide a statistically reliable estimate of either (1) the current space average or (2) any time or space averages that will be observed over future calendar time". This leads to the conclusion (Davidson 1988, p.333): "*in a nonergodic world, the future is uncertain in the sense that history and current events can not provide a reliable statistical guide to knowledge about future outcomes!*" Significantly, the possibility that some plausible ergodic processes can also create statistical havoc, such as having objectively indeterminate and non-comparable probabilities for future realizations, goes

unrecognized.

Speaking about the debate over the connection between Babylonian thought and Post Keynesianism, Dow (2005, p.390) observes: “Difficulties have arisen, as they so often do, as a result of different understandings of language.” This observation is particularly telling in reference to “the axiom of a generally nonergodic world” (Dow 2005, p.386). This axiom involves defining the world in negative terms: the world is not ergodic. For example, Davidson (1991, p.133) observes: “Whenever economists talk about ‘structural breaks’ or ‘changes in regime’, they are implicitly admitting that the economy is, at least at that point of time, not operating under the ergodic presumption that the past objective probabilities will continue to govern future events.” There is less clarity on precisely what type of process would apply. At times it appears that there is no admissible process: “in a world where economic observations need not be generated by any stochastic process, *uncertainty about the future can be defined in terms of the absence of governing ergodic processes*” (Davidson 1988, p.332). However, at other times it appears that nonergodic means that ergodic processes are not sufficiently accurate representations of economic processes: “In the real world, some economic processes may be ergodic, at least for short sub-periods of calendar time, while others are not.”

C. Ergodicity in Economics and Financial Economics

The ergodic hypothesis associated with the statistical mechanics of the kinetic gas model is distinct from the various concepts of ergodicity encountered in modern economics; if only because the complex microscopic interactions of individual gas molecules have to obey the second law of thermodynamics which has no corresponding concept in economics.¹³ Despite differences in physical interpretation, there are two fundamental difficulties associated with the ergodic hypothesis in Boltzmann’s statistical mechanics – reversibility and recurrence – that have a rough similarity to notions arising in the Post Keynesian critique of mainstream economics, e.g., Davidson (1996). These difficulties are compounded by different applications of the ergodic hypothesis arising in mainstream economics. In econometrics, ergodicity is necessary for both strict and covariance stationarity of a stochastic process. In addition, the economic models being tested may also have ergodic restrictions, e.g., the models employ rational expectations or Markovian dynamics. In financial economics, ergodicity is required for the stochastic differential equations used in option pricing models. Examples in other economic subjects are readily available, e.g., whenever a ‘mean-reverting’ process is used.

Even though the formal solutions proposed were inadequate by standards of modern mathematics, the thermodynamic model introduced by Boltzmann to explain the dynamic properties of the Maxwell distribution is a pedagogically useful starting point to develop the implications of ergodicity in economics. To be sure, von Neumann (1932) and Birkhoff (1931) correctly specify ergodicity using Lebesgue integration – an essential analytical tool unavailable to Boltzmann – but the analysis is too complex to be of much value to all but the most mathematically specialized economists. The physical intuition of the kinetic gas model is lost in the generality of the results. Using Boltzmann as a starting point, the large number of mechanical and complex molecular collisions could correspond to the large number of microscopic, atomistic competitors and consumers interacting to determine the macroscopic market price.¹⁴ In this context, it is variables

such as the asset price or the interest rate or the exchange rate, or some combination, that is being measured over time and ergodicity would be associated with the properties of the transition density generating the macroscopic variables. Ergodicity can fail for a number of reasons and there is value in determining the source of the failure. In addition, certain ergodic processes do exhibit behavior that have features with a decidedly Post Keynesian flavor.

Halmos (1949, p.1017) is a helpful starting point to sort out the differing notions of ergodicity that can arise in range of subjects: “The ergodic theorem is a statement about a space, a function and a transformation”. In mathematical terms, ergodicity or ‘metric transitivity’ is a property of ‘indecomposable’, measure preserving transformations. Because the transformation acts on points in the space, there is a fundamental connection to the method of measuring relationships such as distance or volume in the space. In von Neumann (1932) and Birkhoff (1931), this is accomplished using the notion of Lebesgue measure: the admissible functions are either integrable (Birkhoff) or square integrable (von Neumann). In contrast to, say, statistical mechanics where spaces and functions account for the complex physical interaction of large numbers of particles, economics can often specify the space in a mathematically convenient fashion. For example, in the case where there is a single random variable, then the space is “superfluous” (Mackey 1974, p.182) as the random variable is completely described by the distribution. Multiple random variables can be handled by assuming the random variables are discrete with finite state spaces. In effect, conditions for an ‘invariant measure’ can often be assumed in economics in order to focus attention on “finding and studying the invariant measures” (Arnold 1998, p.22) where, in the terminology of econometrics, the invariant measure corresponds to the stationary distribution or likelihood function.

The mean ergodic theorem of von Neumann (1932) provides an essential connection to the ergodicity hypothesis in econometrics. It is well known that, in the Hilbert and Banach spaces common to econometric work, the mean ergodic theorem corresponds to the strong law of large numbers. In statistical applications where strictly stationary distributions are assumed, the relevant ergodic transformation, U^* , is the unit shift operator: $U^* \Psi[x(t)] = \Psi[U^* x(t)] = \Psi[x(t+1)]$; $[(U^*)^k] \Psi[x(t)] = \Psi[x(t+k)]$; and $\{(U^*)^{-k}\} \Psi[x(t)] = \Psi[x(t-k)]$ with k being an integer and $\Psi[x]$ the strictly stationary distribution for x that in the strictly stationary case is replicated at each t . Significantly, this reversible transformation is independent of initial time and state. Because this transformation imposes strict stationarity on $\Psi[x]$, U^* will only work for certain ergodic processes. In effect, the ergodic requirement that the transformation be measure preserving is weaker than the strict stationarity of the stochastic process required for U^* . Post Keynesians accurately identify the implications of the reversible ergodic transformation U^* , e.g., “In an economic world governed entirely by ergodic processes ... economic relationships among variables are timeless, or ahistoric in the sense that the future is merely a statistical reflection of the past” (Davidson 1991, p.331). The Post Keynesian critique argues that the real world distribution for $x(t)$ cannot be assumed to be sufficiently similar to those for $x(t+k)$ or $x(t-k)$ making the ergodic transformation U^* unacceptable.

The axiom of a generally nonergodic world is fundamental to Post Keynesian analysis, e.g., Dow (2005, p.386). In this axiom, a nonergodic process $\{x(t); t = 1, 2, 3 \dots T\}$ is defined such that the time series of a single observed sample path up to $t=j$ can not reliably be used to estimate the parameters – especially the mean – of the distribution for $x(T)$ where $T > j$. This definition directs attention to the ergodic unit shift transformation: $U^* \Psi[x(t)] = \Psi[U^* x(t)] = \Psi[x(t+1)]$.¹⁵ Because this transformation is measure preserving, the replication property of strictly stationary distributions is

captured.¹⁶ Recognizing the possibility that there may be other ergodic transformations than the unit shift, a connection between nonstationarity and nonergodicity is recognized: “Nonstationarity is a sufficient, but not a necessary condition, for nonergodicity” (Davidson 1991, p.332). Limit cycles are given as an example of a stationary stochastic process that is nonergodic.¹⁷ In turn, nonstationarity is identified with path dependence and structural breaks. Beyond this point, the discussion is less clear. The possibility that certain ergodic transformations, more complicated than the unit shift transformation, are consistent with structural breaks is unrecognized. The reasons that limit cycles can be stationary but not ergodic is unexplored.¹⁸ There is the distinct impression that Post Keynesians maintain the class of ‘real world’ processes can not be formally specified because such processes are nonstationary.

The Post Keynesian critique is rooted in the insights of *The General Theory*. These insights include recognizing the distinction between fundamental uncertainty and objective probability for wealth accumulation and liquidity preference. It is unlikely that Keynes gained much exposure to Birkhoff (1931) and von Neumann (1932) or the substantive extensions and applications that appeared in the years following these contributions. As a consequence, the definition of ergodic theory in Post Keynesian thought lacks formal precision, e.g., “There is no reason to presume that structures will remain stable; the economic system is nonergodic” (Dow 2005, p.387). Ergodic theory is implicitly seen as another piece of the mathematical formalism inspired by Hilbert and Bourbaki and captured in the Arrow-Debreu general equilibrium model of mainstream economics. Yet, the 19th century statistical mechanics that inspired ergodic theory is grounded in real world problems; in particular, Birkhoff (1931) and von Neumann (1932) formally solved the Boltzmann problem of incorporating dynamic phase transitions – from gas to liquid and from liquid to solid – where the mechanical model governing the ergodic process changes abruptly. Though there are a variety of ergodic transformations that incorporate such possibilities, the Post Keynesian critique only considers transformations that do not allow for such possibilities.

5.3 Bifurcation and Multi-modal Densities

A. *The Phenomenological Approach*

The distributional implications of boundary restrictions, derived by modeling the random variable as a diffusion process subject to reflecting barriers, have been studied for many years, e.g., Feller (1952,1954). The diffusion process framework is useful because it imposes a functional structure that is sufficient for known partial differential equation (PDE) solution procedures to be used to derive the relevant transition probability densities. Wong (1964) demonstrated that with appropriate specification of parameters in the PDE, the transition densities for popular stationary distributions such as the exponential, uniform, and normal distributions can be derived using SL methods. Following Karlin and Taylor (1981), the transition probability density function U is associated with the random (economic) variable x at time t ($U = U[x, t | x_0]$) that follows a regular, time homogeneous diffusion process with a state space that is either a possibly infinite open interval $I_o = (a, b; -\infty \leq a < b \leq \infty)$, a finite closed interval $I_c = [a, b; -\infty < a < b < +\infty]$, or the specific interval $I_s = [0 = a < b = \infty]$.¹⁹ Assuming that U is twice continuously differentiable in x and once in t and vanishes outside the relevant interval, then U obeys the forward equation (e.g., Gihhman and

Skorohod 1972, p.102-4):

$$\frac{\partial^2}{\partial x^2} \{ B[x] U \} - \frac{\partial}{\partial x} \{ A[x] U \} = \frac{\partial U}{\partial t} \quad (1)$$

where: $B[x]$ ($= \frac{1}{2} \sigma^2[x] > 0$) is the one half the infinitesimal variance and $A[x]$ the infinitesimal drift of the process. $B[x]$ is assumed to be twice and $A[x]$ once continuously differentiable in x . Being time homogeneous, this formulation permits state, but not time, variation in the drift and variance parameters.

The specific problem of deriving the transition probability density for a diffusion process starting at an interior point $x_0 > 0$ with constant parameters $A[x] = \mu$ (≤ 0) and $B[x] = \frac{1}{2} \sigma^2$ subject to a regular, fixed lower reflecting barrier at $x = 0$ is well known (e.g., Cox and Miller 1965). Because the process can reach but not pass below the barrier this imposes a restriction on the density to integrate to 1 over the specific interval $I_s = [0, \infty)$ or the open interval $I_o = (0 < a < b < \infty)$, depending on singularities at $x=0$ in $B[x]$ arising from, say, natural boundary restrictions. Differentiating with respect to time the condition that the density integrate to 1 over the state space then switching the order of integration and differentiation, permits the forward equation to be substituted for the time derivative. Letting $U = U[x, t | x_0] = U[x, t]$ for ease of notation, evaluating the remaining integral gives the reflecting boundary condition:

$$\frac{\partial}{\partial x} \{ B[x] U[x, t] \} \big|_{x=0} - A[0] U[0, t] = 0 \quad (2)$$

In effect, reflecting barriers can be represented as first derivative restrictions at the boundaries, in this case a lower boundary at $x = 0$. The drift term is required in the boundary condition to ensure conservation of probability. When the drift is zero, (2) reduces to the ‘flux zero’ condition.

More generally, if the diffusion process is subject to upper and lower reflecting boundaries that are regular and fixed ($-\infty < a < b < \infty$), the “Sturm-Liouville problem” involves solving (1) subject to the separated boundary conditions:²⁰

$$\frac{\partial}{\partial x} \{ B[x] U[x, t] \} \big|_{x=a} - A[a] U[a, t] = 0 \quad (3)$$

$$\frac{\partial}{\partial x} \{ B[x] U[x, t] \} \big|_{x=b} - A[b] U[b, t] = 0 \quad (4)$$

And the initial condition:

$$U[x, 0] = f[x_0] \quad \text{where:} \quad \int_a^b f[x_0] = 1 \quad (5)$$

and $f[x_0]$ is the continuous density function associated with x_0 where $a \leq x_0 \leq b$. When the initial starting value, x_0 , is known with certainty, the initial condition becomes the Dirac delta function: $U[x, 0] = \delta[x - x_0]$ and the resulting solution for U is referred to as the ‘principal solution’.

Recognizing time homogeneity of the process eliminates the need to explicitly consider the location of t_0 , for ease of notation it is assumed that $t_0 = 0$. In practice, solving (1) combined with (3)-(5) requires a and b to be specified. While a and b have ready interpretations in physical applications, e.g., the heat flow in an insulated bar, determining these values in economic applications can be more challenging. Some situations, such as the determination of the distribution of an exchange rate subject to control bands, are relatively straight forward. Other situations, such profit distributions with arbitrage boundaries or output distributions subject to production possibility frontiers, may require the basic SL framework to be adapted to the specifics of the modeling situation.

In general, solving the forward equation (1) for U subject to (3), (4) and some admissible form of (5) is difficult, e.g., Feller (1952), Risken (1989). In such circumstances, it is expedient to restrict the problem specification to permit closed form solutions for the transition density to be obtained. Wong (1964) provides an illustration of this approach. The PDE (1) is reduced to an ODE by only considering the non-trivial stationary distributions arising from the Pearson system. More precisely, the processes considered obey:

$$\lim_{t \rightarrow \infty} U[x, t | x_0] = \int_a^b f[x_0] U[x, t | x_0] dx_0 = \Psi[x]$$

where only the principal solution ($f[x_0] = \delta[x - x_0]$) is considered. Restrictions on the stationary distributions $\Psi[x]$ are constructed by imposing the fundamental ODE condition for the unimodal Pearson system of distributions:

$$\frac{d\Psi[x]}{dx} = \frac{e_1 x + e_0}{d_2 x^2 + d_1 x + d_0} \Psi[x]$$

The transition probability density U can then be reconstructed by working back from a specific closed form for the stationary distribution using known results for the solution of specific forms of the forward equation. In this procedure, the d_0, d_1, d_2, e_0 and e_1 in the Pearson ODE are used to specify the relevant parameters in (1). The U for important distributions that fall within the Pearson system, such as the normal, beta, central t , and exponential, can be derived by this method.

The solution procedure employed by Wong (1964) depends crucially on restricting the PDE problem sufficiently to apply classical S-L techniques. Using S-L methods, various studies have generalized the set of solutions for U to cases where the stationary distribution is not a member of the Pearson system or U is otherwise unknown, e.g., Linetsky (2005). While the conventional method is to employ an eigenfunction expansion solution, Veerstraeten (2004) demonstrates that a more revealing solution is provided for the special case where $B[x]$ and $A[x]$ are constants if the Green's function is used to solve the S-L problem²¹. In order to employ the separation of variables technique used in solving S-L problems, (1) has to be transformed into the canonical form of the forward equation. To do this, the following important function has to be introduced:²²

$$r[x] = B[x] \exp \left[- \int_a^x \frac{A[s]}{B[s]} ds \right]$$

Using this function, the forward equation can be rewritten in the form (see Appendix):

$$\frac{1}{r[x]} \frac{\partial}{\partial x} \left\{ p[x] \frac{\partial U}{\partial x} \right\} + q[x] U = \frac{\partial U}{\partial t} \quad (6)$$

$$\text{where: } p[x] = B[x] r[x] \quad q[x] = \frac{\partial^2 B}{\partial x^2} - \frac{\partial A}{\partial x}$$

Equation (6) is the canonical form of equation (1). The S-L problem now involves solving (6) subject to appropriate initial and boundary conditions.

Because the methods for solving the S-L problem are ODE-based, some method of eliminating the time derivative in (1) is required. The eigenfunction expansion approach applies separation of variables, permitting (6) to be specified as:

$$U[x,t] = e^{-\lambda t} \varphi[x] \quad (7)$$

Where $\varphi[x]$ must satisfy the ODE:

$$\frac{1}{r[x]} \frac{d}{dx} \left[p[x] \frac{d\varphi}{dx} \right] + [q[x] + \lambda] \varphi[x] = 0 \quad (1')$$

Transforming the boundary conditions involves substitution of (7) into (3) and (4) and solving to get:

$$\frac{d}{dx} \{ B[x] \varphi[x] \} \big|_{x=a} - A[a] \varphi[a] = 0 \quad (3')$$

$$\frac{d}{dx} \{ B[x] \varphi[x] \} \big|_{x=b} - A[b] \varphi[b] = 0 \quad (4')$$

Significant analytical advantages are obtained by making the S-L problem ‘regular’ which involves assuming: $[a,b]$ is a closed interval with $r[x]$, $p[x]$ and $q[x]$ being real valued and $p[x]$ having a continuous derivative on $[a,b]$; and, $r[x] > 0$, $p[x] > 0$ at every point in $[a,b]$. ‘Singular’ S-L problems arise where these conditions are violated due to, say, an infinite state space or a vanishing coefficient in the interval $[a,b]$. The separated boundary conditions (3) and (4) ensure the problem is self-adjoint (Berg and McGregor 1966, p.91).

The S-L problem of solving (6) subject to the initial and boundary conditions admits a solution only for certain critical values of λ , the eigenvalues. Further, since equation (1) is linear in U , the general solution for (7) is given by a linear combination of solutions in the form of eigenfunction expansions. Details of these results can be found in Hille (1969, ch. 8), Boyce & Di Prima (2001) and Birkhoff and Rota (1989, ch. 10). When the S-L problem is self-adjoint and regular the

solutions for the transition probability density can be summarized in the following:

Proposition I:

The regular, self-adjoint Sturm-Liouville problem has an infinite sequence of real eigenvalues, $0 = \lambda_0 < \lambda_1 < \lambda_2 < \dots$ with:

$$\lim_{n \rightarrow \infty} \lambda_n = \infty$$

To each eigenvalue there corresponds a unique eigenfunction $\varphi_n \equiv \varphi_n[x]$. Normalization of the eigenfunctions produces:

$$\psi_n[x] = \left[\int_a^b r[x] \varphi_n^2 dx \right]^{-1/2} \varphi_n$$

The $\psi_n[x]$ eigenfunctions form a complete orthonormal system in $L_2[a,b]$. The unique solution in $L_2[a,b]$ to (1), subject to the boundary conditions (3)-(4) and initial condition (5) is, in general form:

$$U[x,t] = \sum_{n=0}^{\infty} c_n \psi_n[x] e^{-\lambda_n t} \quad (8)$$

$$\text{where: } c_n = \int_a^b r[x] f[x_0] \psi_n[x] dx$$

This Proposition provides the general solution to the regular, self-adjoint S-L problem of deriving U when the process is subject to reflecting barriers. The Proposition demonstrates that having a discrete spectrum permits a representation for the transition probability density in the summation form of (8).²³ However, while useful, (8) is not immediately revealing because time is allowed to vary over $[0, \infty]$. The issue of decomposing U into time dependent and time independent components is addressed in the following section.

B. Transition Density Decomposition

By providing an appropriate foundation, Proposition I facilitates the derivation of the general form of U for the regular, self-adjoint S-L problem. This section demonstrates that for this problem U can be decomposed into two components: a limiting equilibrium stationary density which is independent of time and the initial condition; and, a power series of transient terms that are time, boundary and initial condition dependent but with zero net density. In many econometric applications, the assumption of stationarity permits the $\Psi[x]$ distribution to be used directly as the likelihood function. This implicitly assumes that only the limiting behavior of U as $t \rightarrow \infty$ is relevant. The impact of the transient component is ignored. In a sampling context, this can be rationalized by standardizing the

variables and assuming the transient components will average out to leave only the asymptotic behavior of a stationary process. Using the S-L approach, theoretical results on the U 's associated with different types of boundary restrictions can be derived and the implications for, say, testing theory can be formulated by examining the shape and *iid* behavior of the relevant distributions and proposing appropriate adjustment factors for confidence intervals.

Being in the form of an eigenfunction expansion, (8) cannot be readily applied to the types of closed form distributions typically encountered in econometrics. Further simplification is required. This leads to the following result:²⁴

Proposition II: Density Decomposition

Under the conditions required for Proposition I, the transition probability density function for x at time t (U) can be expressed as the sum of a stationary limiting equilibrium distribution that is linearly independent²⁵ of the boundaries and a power series of transient terms that are boundary and initial condition dependent:

$$U[x, t | x_0] = \Psi[x] + T[x, t | x_0] \quad (9)$$

$$\text{where: } \Psi[x] = \frac{r[x]^{-1}}{\int_a^b r[x]^{-1} dx} \quad (10)$$

Using the specifications of λ_n , c_n , and ψ_n from Proposition I, the properties of $T[x, t]$ are defined as:

$$T[x, t | x_0] = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} \psi_n[x] = \frac{1}{r[x]} \sum_{n=1}^{\infty} e^{-\lambda_n t} \psi_n[x] \psi_n[x_0] \quad (11)$$

$$\text{with: } \int_a^b T[x, t | x_0] dx = 0 \quad \text{and} \quad \lim_{t \rightarrow \infty} T[x, t | x_0] = 0$$

Proposition II permits (9) to be combined with appropriately specified (10)-(11) to analyze the distributional implications of reflecting barriers. The distributional impact of the boundary restrictions enter through $T[x, t]$ ($= T[x, t | x_0]$). From the restriction on $T[x, t]$ in (11), the total mass of the transient term is zero. The transient acts to redistribute the mass of the stationary distribution, thereby causing a change in shape. The specific degree and type of alteration depends on the relevant assumptions made about the parameters and initial functional forms. A key feature of the Proposition is that (11) is in the form of a discrete spectrum.²⁶ Because the power series given in (11) involves powers of $\exp[-\lambda_n]$, from Proposition I it follows that for given t the terms in the sum will decrease as $n \rightarrow \infty$. This property and the discrete spectrum significantly simplifies the calculation of the transient $T[x, t]$ in practical applications.

To see the implications of Proposition II, consider the variety of boundary independent stationary

densities $\Psi[x]$ generated by appropriate choices of $A[x]$ and $B[x]$. A range of results are available in Wong (1964), Borodin and Salminen (2002), Veerstraeten (2004) and Linetsky (2005). A benchmark solution is given by the Brownian motion, constant coefficient case where $A[x] = \mu$ ($\neq 0$) and $B[x] = \frac{1}{2}\sigma^2$. Evaluating (10) gives the solution as (e.g., Veerstraeten 2004) :

$$\Psi[x] = \frac{2\mu}{\sigma^2} \frac{\exp\left\{\frac{2\mu}{\sigma^2} x\right\}}{\exp\left\{\frac{2\mu}{\sigma^2} b\right\} - \exp\left\{\frac{2\mu}{\sigma^2} a\right\}} = \frac{2\mu}{\sigma^2} \frac{\exp\left\{\frac{2\mu}{\sigma^2} (x - a)\right\}}{\exp\left\{\frac{2\mu}{\sigma^2} (b - a)\right\} - 1} \quad (12)$$

There is no convention for a specific closed form to use for expressing this case. For example, Linetsky (2005) simplifies this solution by setting $\sigma^2 = 1$ and $a = 0$. In either form, $\Psi[x]$ is a scaled exponential density. If $\mu = 0$, the exponential density reduces to a uniform density: $\Psi[x] = 1/(b - a)$. The uniform stationary density is intuitive: if the reflecting boundaries are constant and the process has no drift then as $t \rightarrow \infty$ each point in the state space will be equally likely. It follows that the exponential stationary distributions are a consequence of the sample paths drifting to the upper ($A[x] > 0$) or the lower ($A[x] < 0$) boundary and ‘bouncing off’. These solutions can be contrasted with Wong (1964) where the stationary exponential density $\Psi[x] = \exp[-x]$ corresponding to the Pearson system $\{d\Psi/dx\} = -\Psi[x]$ is used but the specific interval $I_s = [0, \infty)$ is required due to the density having to integrate to one over the state space.

The simplicity of the closed form stationary density component, $\Psi[x]$, of the transition density in the Brownian motion, constant parameter case does not carry over to the transient component. Following Borodin and Salminen (2002, p.121-2), the simplest constant parameter solution selects the principal solution (delta function initial condition), sets the drift to zero (uniform stationary), $B[x] = 1$ ($\sigma = \sqrt{2}$), and $I_c = [0 \leq x \leq 1]$ ($\Psi[x] = 1$). This produces the power series of transient terms which using (11) defines the eigenvalues ($\lambda_n = n^2 \pi^2$) and eigenfunctions ($\psi_n = \sqrt{2} \cos n \pi x$), i.e.:

$$T[x, t | x_0] = 2 \left(\sum_{n=1}^{\infty} \exp[-n^2 \pi^2 t] \cos[n \pi x] \cos[n \pi x_0] \right)$$

This ‘simplest’ solution can be used to illustrate the implications of altering the specification of the S-L problem. In particular, the drift zero, principal solution with $I_c = [0 \leq x \leq L]$ ($\Psi[x] = 1/L$) and $B[x] = \frac{1}{2} \sigma^2$ produces eigenvalues ($\lambda_n = (n^2 \pi^2 \sigma^2) / 2L^2$), eigenfunctions ($\psi_n = \sqrt{2/L} \cos [(n \pi x) / L]$) and the solution:

$$T[x, t | x_0] = \frac{2}{L} \left(\sum_{n=1}^{\infty} \exp\left[-\frac{n^2 \pi^2 \sigma^2}{2L^2} t\right] \cos\left[\frac{n \pi}{L} x\right] \cos\left[\frac{n \pi}{L} x_0\right] \right)$$

Both the interval length and dispersion value act to scale the simple solution. This formulation permits the impact on $T[x, t]$ of increasing the interval length for given t to be assessed. This result

can also be used to illustrate the analytical significance of having a process with zero drift.

To see the importance of drift specification, consider the principal solution where $\Psi[x]$ is given by (12) with $I_c = [a, b]$, $A[x] = \mu (\neq 0)$ and $B[x] = \frac{1}{2}\sigma^2$. The U for this case has been derived in the context of exchange rates distributions with target rate bands (Svensson 1991; de Jong 1994). Formal treatments of the Brownian motion, constant parameter solution are available in Linetsky (2005) and, using the alternative Green's function approach, in Veerstraeten (2004):

$$T[x,t] = \frac{\exp\left[\frac{\mu}{\sigma^2}(x - x_0)\right]}{(b - a)} \sum_{n=1}^{\infty} \exp[-\lambda_n t] \frac{\sigma^2 \pi^2}{\lambda_n (b-a)^2} \left[n \cos\left[n\pi \frac{x - a}{b - a}\right] + \frac{\mu (b - a)}{\sigma^2 \pi} \sin\left[n\pi \frac{x - a}{b - a}\right] \right] \\ \left[n \cos\left[n\pi \frac{x_0 - a}{b - a}\right] + \frac{\mu (b - a)}{\sigma^2 \pi} \sin\left[n\pi \frac{x_0 - a}{b - a}\right] \right]$$

where the eigenvalues are $\lambda_n = (\mu / 2\sigma^2) + ((\sigma^2 \pi^2 n^2) / 2(b - a)^2)$ and the eigenfunctions retain both the sin and cos terms from the general solution. It is apparent that processes possessing a non-zero drift pose increased analytical complications associated with solving variable coefficient PDE's. This substantial increase in the complexity of the solution for the transient component in the constant coefficient case does not bode well for finding ready to implement solutions in more complicated cases.

This intuition about increased complexity is confirmed by Linetsky (2005) where the Sturm-Liouville problem is solved for the U associated with an Ornstein-Uhlenbeck (OU) process. In this case, the drift is state dependent $A[x] = \kappa (\chi - x)$ with $\kappa > 0$ and χ the long run mean of x ($b > \chi > a$). The infinitesimal variance is constant with $B[x] = \frac{1}{2}\sigma^2$. Evaluating (10) for these values gives the solution of the stationary distribution as (e.g., Linetsky 2005, p.447):

$$\Psi[x] = \frac{\sqrt{2\kappa}}{\sigma} \frac{n[z]}{N[\beta] - N[\alpha]} \\ z = \frac{\sqrt{2\kappa}}{\sigma} (x - \chi) \quad \alpha = -\frac{\sqrt{2\kappa}}{\sigma} (\chi - a) \quad \beta = \frac{\sqrt{2\kappa}}{\sigma} (b - \chi)$$

where $n[\cdot]$ and $N[\cdot]$ are the standard normal density and the cumulative standard normal distribution function, respectively. The process of determining the eigenfunctions is decidedly more complicated (Linetsky 2005, p.447-9), involving functions not commonly encountered in econometrics. More precisely, changing variables to transform the forward equation into Weber-Hermite form permits solutions involving Weber-Hermite parabolic cylinder functions, which are related to Kummer confluent hypergeometric functions available in standard software packages, e.g., Mathematica.²⁷ The solutions require derivatives of the Kummer functions to be evaluated numerically leading to solutions involving digamma functions. The worked solution for the eigenfunction expansion of U in this case is available in Linetsky (2005, p.449).

Beyond Regular Boundaries

Section II demonstrates that, despite having the theoretical advantage of a discrete spectrum, imposing regular reflecting barriers on the state space for the forward equation quickly leads to analytical complexity in actually deriving the eigenfunction expansion for the transition probability density. These disadvantages need to be tempered by considering the alternatives to imposing reflecting boundaries. Consider the well known solution (e.g., Cox and Miller 1965, p.209) for U involving a constant coefficient standard normal variate $Y(t) = (x - x_0 - \mu t) / \sigma$ over the unbounded state space $I_o = (-\infty \leq x \leq \infty)$. In this case the forward equation (1) reduces to: $\frac{1}{2}\{\partial^2 U / \partial Y^2\} = \partial U / \partial t$. By evaluating these derivatives, it can be verified that the principal solution for U is:

$$U[x, t | x_0] = \frac{1}{\sigma \sqrt{2\pi t}} \exp \left[-\frac{(x - x_0 - \mu t)^2}{2\sigma^2 t} \right]$$

and as $t \rightarrow -\infty$ or $t \rightarrow +\infty$ then $U \rightarrow 0$ and the stochastic process does not possess a non-trivial stationary distribution. In effect, if the process runs long enough then U will evolve to where there is no discernible probability associated with starting from x_0 and achieving a given point x .²⁸ The absence of a stationary distribution raises a number of practical problems, e.g., unit roots. Imposing regular reflecting boundaries is a certain method of obtaining an stationary distribution and a discrete spectrum (Hansen and Schienkman 1998, p.13). Alternative methods, such as specifying the process to admit natural boundaries where the parameters of the diffusion are zero within the state space, can give rise to continuous spectrum and raise significant analytical complexities. At least since Feller (1952), the search for useful solutions, including those for singular diffusion problems, has produced a number of specific cases of interest. However, without the analytical certainty of the S-L framework, analysis proceeds on a case by case basis.

One possible method of obtaining a stationary distribution without imposing both upper and lower boundaries is to impose only a lower (upper) reflecting barrier and construct the stochastic process such that positive (negative) infinity is non-attracting, e.g., Linetsky (2005); Ait-Sahalia (1999). This can be achieved by using an OU drift term. In contrast, Cox and Miller (1964, p.223-5) use the Brownian motion, constant coefficient forward equation with $x_0 > 0$, $A[x] = \mu < 0$ and $B[x] = \frac{1}{2}\sigma^2$ subject to the lower reflecting barrier at $x = 0$ given in (2) to solve for both the U and the stationary density. The principal solution is solved using the ‘method of images’ to obtain:

$$U[x, t | x_0] = \frac{1}{\sigma \sqrt{2\pi t}} \left\{ \exp \left(-\frac{(x - x_0 - \mu t)^2}{2\sigma^2 t} \right) + \exp \left(-\frac{4x_0\mu t - (x - x_0 - \mu t)^2}{2\sigma^2 t} \right) \right\} \\ + \frac{1}{\sigma \sqrt{2\pi t}} \left\{ \frac{2\mu}{\sigma^2} \exp \left(\frac{2\mu x}{\sigma^2} \right) \left(1 - N \left[\frac{x + x_0 + \mu t}{\sigma \sqrt{t}} \right] \right) \right\}$$

where $N[x]$ is again the cumulative standard normal distribution function. Observing that $A[x] = \mu > 0$ again produces $U \rightarrow 0$ as $t \rightarrow +\infty$, the stationary density for $A[x] = \mu < 0$ follows:

$$\psi[x] = \frac{2|\mu|}{\sigma^2} \exp\left(-\frac{2|\mu|x}{\sigma^2}\right)$$

Though x_0 does not enter the solution, combined with the location of the boundary at $x = 0$, it does implicitly impose the restriction $x > 0$. From Proposition II, $T[x, t | x_0]$ can be determined as $U[x, t | x_0] - \Psi[x]$.

Following Linetsky (2005), Veerstraeten (2004) and others, the analytical procedure used in section II to determine U involved specifying the parameters of the forward equation and the boundary conditions and then solving for $\Psi[x]$ and $T[x, t]$. Wong (1964) uses a different approach, initially selecting a stationary distribution and then solving for U using the restrictions of the Pearson system to specify the forward equation. In this approach, the functional form of the desired stationary distribution determines the appropriate boundary conditions. While application of this approach has been limited to the restricted class of distributions associated with the Pearson system, it is expedient when a known stationary distribution, such as the standard normal distribution, is of interest. More precisely, let:

$$\Psi[x] = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right], \quad I_o = (-\infty < x < \infty)$$

In this case, the boundaries of the state space are non-attracting and not regular. Solving the Pearson equation gives: $d\Psi[x]/dx = -x \Psi[x]$ and a forward equation of the OU form:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial}{\partial x} xU = \frac{\partial U}{\partial t}$$

The principal solution for this unrestricted equation is:

$$U[x, t | x_0] = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] \sum_{n=0}^{\infty} \frac{\exp[-nt]}{n!} H_n[x_0] H_n[x]$$

where $H_n[x]$ are the Hermite polynomials, e.g., Kendall and Stuart (1963, p.155), and the solution for the (discrete spectrum) $T[x, t]$ is given by taking the sum from $n = 1$. Following Wong (1964, p.268) Mehler's formula can be used to express the solution for U as:

$$U[x, t | x_0] = \frac{1}{\sqrt{2\pi(1 - e^{-2t})}} \exp\left[\frac{-(x - x_0 e^{-t})^2}{2(1 - e^{-2t})}\right]$$

Given this, as $t \rightarrow -\infty$ then $U \rightarrow 0$ and as $t \rightarrow +\infty$, U achieves the standard normal ergodic distribution.

The ergodic normal distribution is an example where a discrete spectrum is obtained without imposing boundaries on the state space. Another example is given by Wong (1964, p.268-9) where the stochastic process has a state space $I_s = [0 \leq x < \infty)$ and a discrete spectrum involving Laguerre

polynomials with a stationary density of the form:

$$\Psi[x] = \frac{x^\alpha}{\Gamma[\alpha + 1]} e^{-x} = \frac{1}{\Gamma[\alpha + 1]} \exp\{\alpha \ln[x] - x\}$$

and forward equation:²⁹

$$\frac{\partial^2}{\partial x^2}[xU] - \frac{\partial}{\partial x}[(\alpha + 1 - x)U] = \frac{\partial U}{\partial t}$$

where the gamma function $\Gamma[\alpha + 1]$ has $\alpha > -1$. This process is significant in having x dependence on the infinitesimal variance and a solution for U involving Laguerre polynomials that can be solved in closed form. Linetsky (2005) provides results for affine diffusion processes where the coefficients of the forward equation are given by $B[x] = \frac{1}{2} \sigma \sqrt{x - \ell}$ with the shift parameter $\ell < 0$ and $A[x] = \kappa(\chi - x)$ with the same parameter restrictions as for the OU process of section II. When subjected to a lower reflecting barrier (because ∞ is non-attracting) the affine diffusion also has a discrete spectrum. However, a closed form solution is unavailable.³⁰

Generalized Pearson Systems

The results in Wong (1964), Linetsky (2005), Veerstraeten (2004) and related studies apply directly to the transition probability densities associated with the unimodal Pearson system. Generalizing this approach to allow more flexibility in the shape of the stationary distribution can be achieved using a higher order exponential density, e.g., Fisher (1921), Cobb et al. (1983), Crauel and Flandoli (1998). Increasing the degree of the polynomial in the exponential comes at the expense of introducing additional parameters resulting in a substantial increase in the analytical complexity of the transition density spectrum. As a consequence, the generalized Pearson distributions typically defy a closed form solution for the transition densities. However, at least since Elliott (1955), it has been recognized that the solution of the associated regular S-L problem will still have a discrete spectrum, even if the specific form of the eigenfunctions and eigenvalues in $T[x, t | x_0]$ are not precisely determined (Horsthemke and Lefever 1984, sec. 6.7). Inferences about transient stochastic behavior can be obtained by examining the solution of the deterministic non-linear dynamics. In this process, attention initially focuses on the properties of the higher order exponential distributions.

To this end, assume that the stationary distribution is a fourth degree or “general quartic” exponential:

$$\Psi[x] = K \exp[-\Phi[x]] = K \exp[-(\beta_4 x^4 + \beta_3 x^3 + \beta_2 x^2 + \beta_1 x)]$$

where: K is a constant determined such that the density integrates to one; and, $\beta_4 > 0$.³¹ Following Fisher (1921), the class of distributions associated with the general quartic exponential admits both unimodal and bimodal densities and nests the standard normal as a limiting case where $\beta_4 = \beta_3 = \beta_1 = 0$ and $\beta_2 = \frac{1}{2}$ with $K = 1/(\sqrt{2\pi})$. The generalized Pearson ODE restriction for the quartic exponential takes the form:

$$\frac{d\Psi[x]}{dx} = \frac{e_3 x^3 + e_2 x^2 + e_1 x + e_0}{g[x]} \Psi[x]$$

In this case, the generalization occurs because the degree of the ‘shape polynomial’ in the numerator has been increased from one to three. It is possible to further generalize to a stationary distribution with a k (> 4) degree exponential. With correct selection of parameters, the quartic exponential density is sufficient to capture the implications of stationary bimodality; a higher degree polynomial is needed if the possible number of stationary modes is greater than two.

The implications of generalizing the Pearson system by increasing the degree of the exponential is apparent from the ODE restriction. In the Pearson system, $g[x] = d_0 + d_1 x + d_2 x^2$ is a polynomial of degree at most two in x that depends on the particular specification of the stationary or transition density desired. The classification of Pearson system of stationary distributions into the various Types I-VII follows the specification of the degree of the polynomial $g[x]$ (Johnson and Kotz 1970, p.9-14). Extending this approach to the generalized Pearson system, when $g[x]$ is a constant there is a family \mathbb{N} of distributions that include the standard normal and the quartic exponential densities, instead of a single distribution defined by the standard normal that is the limiting case of all Pearson distributions types. Permitting g to be a linear transformation of the form $d_1 x$ restricts the admissible $x \in \{0 < x < \infty\}$. In particular, a stationary distribution of the form:

$$\psi[x] = K_G \exp[-g_0 \ln[x] + g_1 x + g_2 x^2 + g_3 x^3]$$

produces the family \mathbb{G} of gamma densities that nest the Pearson Type III. Allowing $g[x] = d_2 x^2$ produces the inverse gamma family nesting the Pearson Type V; and, setting $d_1 = d_2 = 1$ and $g[x] = x(1-x)$ produces the beta family nesting the Pearson Type I. Each of these families requires an appropriate version of the exponential stationary distribution to correspond with the desired $g[x]$ in the generalized Pearson ODE (Cobb et al. 1983).

In specifying the generalized Pearson system, the additional complications introduced by the higher degree polynomial in the numerator of the ODE augments the concern with different solutions of the quadratic polynomial in the denominator that arises with the Pearson system. To avoid complicated solutions for the generalized Pearson ODE involving ratios of polynomials in x , it is expedient to focus attention on the non-linearity in the drift and away from state dependence of the infinitesimal variance. In effect, enhancing precision in the estimation of distributional shape comes at the expense of incorporating state dependence in the variance. This induces a fundamental shift in the conceptual approach to modelling random behavior using diffusion processes. Consider the problem of modeling the drift and diffusion parameters for the short term interest rate process, e.g., Stanton (1997). Following Ait-Sahalia (1996) the preferred approach to empirically determining these parameters has been to fit a flexible, nonlinear functional form for each parameter, such as:

$$A[x;\theta] = \theta_0 + \theta_1 x + \theta_2 x^2 + \frac{\theta_3}{x} \quad B[x;\beta] = \beta_1 x + x^{\beta_2}$$

The generalized Pearson ODE suggests that such ‘flexibility’ is misleading. Consistent with this observation, Chapman and Pearson (2000) argue against the flexible, non-linear function form approach to capturing nonlinearity in the drift of short-term interest rates due to multicollinearity

between the drift and diffusion coefficients. Similarly, Hurn and Lindsay (2002) address the multicollinearity problem by employing orthogonal Legendre polynomials and find estimation of the non-linear drift function depends crucially on “specification of the drift in terms of orthogonal constituents” (p.563). Hence, permitting state dependence of both the drift and volatility imposes significant restrictions on the parameters of the stationary distribution.

Despite being recognized as early as Fisher (1921) as the class of distributions for which the efficiency of the method of moments coincides with maximum likelihood, generalized Pearson distributions such as the quartic exponential density have been mostly ignored in econometrics in favor of processes, such as affine diffusions, that feature state dependence of the infinitesimal variance. Where diffusions from this class are used, as in the “double- well” diffusion process in Ait-Sahalia (1999):

$$dX(t) = (X(t) - X(t)^3) dt + dW(t)$$

where $dW(t)$ is a Weiner process, the parametric flexibility needed to fit the non-linearity in the drift is ignored.³² While the double-well process does have a symmetric bimodal stationary density, the *a priori* restrictions on the non-linear drift term are apparent from the specification of the generalized Pearson ODE. The analytical advantage of setting the infinitesimal variance to a constant is to enhance fitting of the shape polynomial for the stationary distribution. The restrictions imposed by the double well process produce a quartic exponential distribution that is bimodal and symmetric about zero. The parameter restrictions imposed are too severe to be representative of actual economic time series.

C. Bifurcation and the Quartic Exponential Distribution

The stationary distribution of the double well process is a special case of the symmetric quartic exponential distribution:

$$\Psi[y] = K_S \exp[-\{\beta_2 (x - \mu)^2 + \beta_4 (x - \mu)^4\}] \quad \text{where } \beta_4 \geq 0$$

where μ is the population mean and the symmetry restriction requires $\beta_1 = \beta_3 = 0$. To see why the condition on β_1 is needed, consider change of origin $X = Y - \{\beta_3 / 4 \beta_4\}$ to remove the cubic term from the general quartic exponential (Matz 1978, p.480):

$$\Psi[y] = K_Q \exp[-\{\kappa (y - \mu_y) + \alpha (y - \mu_y)^2 + \gamma (y - \mu_y)^4\}] \quad \text{where } \gamma \geq 0$$

The substitution of y for x indicates the change of origin which produces the following relations between coefficients for the general and specific cases:

$$\kappa = \frac{8\beta_1\beta_4^2 - 4\beta_2\beta_3\beta_4 + \beta_3^3}{8\beta_4^2} \quad \alpha = \frac{8\beta_2\beta_4 - 3\beta_3^2}{8\beta_4} \quad \gamma = \beta_4$$

The symmetry restriction $\kappa = 0$ can only be satisfied if both β_3 and $\beta_1 = 0$. Given the symmetry restriction, the double well process further requires $-\alpha = \gamma = \sigma = 1$. Solving for the modes of $\Psi[y]$ gives $\pm \sqrt{\{|\alpha| / (2\gamma)\}}$ which reduces to ± 1 for the double well process, as in Ait-Sahlia (1999, Figure 6B, p.1385).

INSERT FIGURE 5.3.a HERE
Family of Stationary Densities

As illustrated in Figure 5.3.a, the selection of a_i in the stationary density $\Psi_i[x] = K_Q \exp\{-(.25 x^4 - .5 x^2 - a_i x)\}$ defines a family of general quartic exponential densities, where a_i is the selected value of κ for that specific density.³³ The coefficient restrictions on the parameters α and γ dictate that these values cannot be determined arbitrarily. For example, given that β_4 is set at .25, then for $a_i = 0$, it follows that $\alpha = \beta_2 = 0.5$. ‘Slicing across’ the surface in Figure 5.3.a at $a_i = 0$ reveals a stationary distribution that is equal to the double well density. Continuing to slice across as a_i increases in size, the bimodal density becomes progressively more asymmetrically concentrated in positive x values. Though the location of the modes does not change, the amount of density between the modes and around the negative mode falls. Similarly, as a_i decreases in size the bimodal density becomes more asymmetrically concentrated in positive x values. While the stationary density is bimodal over $a_i \in \{-1, 1\}$, for $|a_i|$ large enough the density becomes so asymmetric that only a unimodal density appears. For the general quartic, asymmetry arises as the amount of the density surrounding each mode (the sub-density) changes with a_i . In this, the individual sub-densities have a symmetric shape. To introduce asymmetry in the sub-densities, the reflecting boundaries at a and b that bound the state space for the regular S-L problem can be used to introduce positive asymmetry in the lower sub-density and negative asymmetry in the upper sub-density.

Solving the forward equation to obtain a closed form for the transition density of a diffusion process with a quartic exponential stationary distribution is confounded by the presence of the cubic non-linearity in the numerator of the generalized Pearson ODE and in the forward equation term:

$$\frac{\partial}{\partial x} \{ A[x] U[x, t] \}$$

Except in the special generalized Pearson cases such as the family \mathfrak{G} of gamma densities, also permitting state variation in $B[x]$ renders the forward equation for higher order exponential densities unsolvable in closed form. To obtain information about $T[x, t | x_0]$, attention focuses on solving the non-linear dynamics of the deterministic equation associated with the drift term. For the symmetric quartic exponential, these deterministic dynamics are described by the pitchfork bifurcation ODE:

$$\frac{dx}{dt} = -x^3 + \rho_1 x + \rho_0$$

where ρ_0 and ρ_1 are the ‘normal’ and ‘splitting’ control variables, respectively (e.g., Cobb 1978). While ρ_0 has significant information in a stochastic context, this is not usually the case in the deterministic problem so $\rho_0 = 0$ is assumed. Given this, for $\rho_1 \leq 0$, there is one real equilibrium ($\{dx/dt\} = 0$) solution to this ODE at $x = 0$ where “all initial conditions converge to the same final point

exponentially fast with time” (Crauel and Flandoli 1998, p.260). For $\rho_1 > 0$, the solution bifurcates into three equilibrium solutions $x = \{ 0, \pm\sqrt{\rho_1} \}$, one unstable and two stable. In this case, the state space is split into two physically distinct regions (at $x = 0$) with the degree of splitting controlled by the size of ρ_1 . Even for initial conditions that are ‘close’, the equilibrium achieved will depend on the sign of the initial condition.

It is well known that the introduction of randomness to the pitchfork ODE changes the properties of the equilibrium solution, e.g., (Arnold 1998, sec.9.2). It is no longer necessary that the state space for the principal solution be determined by the location of the initial condition relative to the bifurcation point. The possibility for randomness to cause some paths to cross over the bifurcation point depends on the size of σ which measures the non-linear signal to white noise ratio. Of the different approaches to introducing randomness (e.g., multiplicative noise), the simplest approach to converting from a deterministic to a stochastic context is to add a Weiner process to the ODE. Augmenting the diffusion equation to allow for σ to control the relative impact of non-linear drift versus random noise produces the “pitchfork bifurcation with additive noise” (Arnold 1998, p.475) which in symmetric form is:

$$dX(t) = (\rho_1 X(t) - X(t)^3) + \sigma dW(t)$$

While capable of sustaining the common approach in econometrics based on a one-to-one correspondence between invariant Markov forward measures and stationary distributions, the dynamics of the pitchfork process captured by $T[x, t | x_0]$ have been “forgotten” (Arnold 1998, p.473).

Bifurcation and the *Ex Ante* / *Ex Post* Dilemma

The significance of the *ex ante* / *ex post* distinction was recognized as early as Myrdal (1939), though only in connection with the more general problem of determining aggregate savings and investment equilibrium (Meacci 2009). As Shackle (1967, 242-3) observes:

it is the level of incomes which *moves in search* of an equilibrium between (designed *ex ante*) saving and (designed *ex ante*) investment ... when there is a disparity, a disequilibrium, between the two *ex ante* quantities, there will almost inevitably follow one period later, that is, so soon as this disparity is revealed, *ex post*, a set of decisions by business men to change designed general output, and thus aggregate income.

For purposes of equity valuation, the *ex ante* / *ex post* distinction applies to the use of past data to forecast future values, especially for equity prices and those variables that drive equity prices. The single *ex post* observed sample path used to estimate parameters of the transition density is the outcome of decisions made using *ex ante* distributions that can differ substantively from the *ex post* distribution. In particular, if the future time paths for the ensemble of sample paths are of long enough duration, there will almost certainly be differences in the *ex ante* and *ex post* distributions.

That *ex ante* and *ex post* distributions are equivalent in stochastic theories of equity value associated with modern Finance is due to the usually unstated assumption of time reversible ergodic processes. Despite the importance of the assumption of ergodicity, a sufficiently accurate definition

of an ergodic process has not been provided. More precisely, if the stock price observations are generated by time irreversible ergodic processes, then the calculation of time averages based on a sufficiently long enough set of past data can not be expected to provide a statistically reliable estimate of any state space averages that will be observed in a sufficiently distant future calendar time. In the case of a time irreversible bifurcating process, this follows because the *ex ante* stationary distributions are bimodal. The precise location and density associated with each mode is subjectively determined and depends on the anticipated time to bifurcation. In other words, the type of reversible ergodic stochastic processes commonly encountered in mainstream academic Finance are less capable of representing the *ex ante* stochastic behavior of equity prices than irreversible, bifurcating ergodic processes.

Significantly, a type of fundamental uncertainty is inherent in bifurcating processes as illustrated in the selection of a_i in Figure 5.3.a. A semantic connection can be established between the subjective uncertainty about encountering a future bifurcation point and, say, the possible collapse of an asset price bubble due to a change in Keynesian convention about market valuations. Examining the quartic exponential stationary distribution associated with a bifurcating ergodic process, it is apparent that this distribution nests the Gaussian distribution as a special case. In this sense, the bifurcating process represents a stochastic generalization of the mainstream stochastic theory of equity price behavior. By construction, ergodic bifurcating processes become unstable due to the mean being non-linear in time, with degree and timing of the instability being uncertain at $t = 0$. Following Shackle, once the surprise generated instability has been assimilated the ‘standstill effect’ ends with the formation of new *ex ante* distributions. While the *ex post* process at any point in time is given, the *ex ante* distributions can vary depending on changes subjective perceptions and the like.

A specification of the uncertainty critique based on decomposing the transition probability density of an ergodic one-dimensional diffusion process subject to regular upper and lower reflecting barriers is seemingly inconsistent with the spirit of Keynes (1936), Shackle (1958), or Davidson (2003, 2007). In contrast to Shackle, the use of diffusion processes involves additive probability and the phenomenological approach involves the solution of differential equations. Uncertainty appears, not with some portion of the total probability that is not allocated, but rather with the determination of the amount of the total ergodic density allocated to subdensities associated with the different modes. Post Keynesian properties of the ergodic process – where use of time path for a single *ex post* realization to estimate the parameters of the *ex ante* ensemble of future time paths is questioned – can be captured by using bimodal or multi-modal stationary densities such as the quartic exponential. As such, estimates of means and variances have less meaning than identifying the ‘best’ or ‘worst’ that can happen associated with the modes of the *ex ante* distribution.

The use of bifurcating stochastic processes involves a number of fundamental differences with conventional time reversible processes. Because multi-modality implies a mean process that is non-linear in time, the variation in the *ex ante* process originates with changes in mean values, not changes in variance and covariance. At least since Shiller (1981), the ‘excess’ variation in stock prices compared to variation in the underlying fundamentals has puzzled academics. This ‘excess’ variation can be explained using: *ex ante* processes with mean values that are non-linear in time; and, observing that the *ex post* sample path is only a single realization of the ensemble of possible *ex ante* paths between $t=0$ and $t=T$. In order to efficiently recover information about the distribution of the

mean value process, only in special cases is it feasible to introduce time variation in the volatility parameter. This follows because the generalized Pearson conditions have to be satisfied in order to recover information about the functional form of the underlying stationary distribution from the solution to the forward equation.

Following Ford (1993, p.690-1), Shackle has a number of theoretical insights about decision making under uncertainty applicable to the interpretation of time irreversible processes: it is not possible “to construct a dynamic model of an economic system except for one period at a time”; “at the outset of the ensuing period market participants would form new expectations, not necessarily linked to anything which had gone before”; and, economics cannot be a predictive science as it is built on concepts of “mechanical, non-expectational time”. This leads to the following interpretation of the decision problem: at a particular $t=0$ with $X(0)$, the two modes for the ensemble of *ex ante* time paths reflect the possibilities of continued prosperity or market collapse. This is consistent with Shackle’s theory that individuals concentrated on the ‘best’ and ‘worst’ that can happen in making decisions under uncertainty (Ford 1993, p.696). Observing the a_i evolves over time to capture change in subjective expectations of prosperity or collapse, changes in a_i corresponds to Shackle’s ‘degrees of potential surprise’.

Appendix: Preliminaries and Proofs

Preliminaries on solving the Forward Equation:

Due to the widespread application in a wide range of subjects, textbook presentations of the Sturm-Liouville problem possess subtle differences that require some clarification to be applicable to the formulation used in this paper. In particular, to derive the canonical form (6) of the Fokker-Planck equation (1) observe that evaluating the derivatives in (1) gives:

$$B[x] \frac{\partial^2 U}{\partial x^2} + \left[2 \frac{\partial B}{\partial x} - A[x] \right] \frac{\partial U}{\partial x} + \left[\frac{\partial^2 B}{\partial x^2} - \frac{\partial A}{\partial x} \right] U = \frac{\partial U}{\partial t}$$

This can be rewritten as:

$$\frac{1}{r[x]} \frac{\partial}{\partial x} \left[P[x] \frac{\partial U}{\partial x} \right] + Q[x] U = \frac{\partial U}{\partial t}$$

where:

$$P[x] = B[x] r[x] \quad \frac{1}{r[x]} \frac{\partial P}{\partial x} = 2 \frac{\partial B}{\partial x} - A[x] \quad Q[x] = \frac{\partial^2 B}{\partial x^2} - \frac{\partial A}{\partial x}$$

It follows that:

$$\frac{\partial B}{\partial x} = \frac{1}{r[x]} \frac{\partial P}{\partial x} - \frac{1}{r^2} \frac{\partial r}{\partial x} P[x] = 2 \frac{\partial B}{\partial x} - A[x] - \frac{B[x]}{r[x]} \frac{\partial r}{\partial x}$$

This provides the solution for the key function $r[x]$:

$$\frac{1}{r[x]} \frac{\partial r}{\partial x} - \frac{1}{B[x]} \frac{\partial B}{\partial x} = - \frac{A[x]}{B[x]} \quad \rightarrow \quad \ln[r] - \ln[k] = - \int^x \frac{A[s]}{B[s]} ds$$

$$r[x] = B[x] \exp \left[- \int^x \frac{A[s]}{B[s]} ds \right]$$

This $r[x]$ function is used to construct the scale and speed densities commonly found in presentations of solutions to the forward equation, e.g., Karlin and Taylor (1981), Linetsky (2005).

Another specification of the forward equation that is of importance is found in Wong (1964, eq.6-7):

$$\frac{d}{dx} \left[B[x] \rho[x] \frac{d\theta}{dx} \right] + \lambda \rho[x] \theta[x] = 0 \quad \text{with b.c.} \quad B[x] \rho[x] \frac{d\theta}{dx} = 0$$

This formulation occurs after separating variables, say with $U[x] = g[x] h[t]$. Substituting this result into (1) gives:

$$\frac{\partial^2}{\partial x^2}[B g h] - \frac{\partial}{\partial x}[A g h] = g[x] \frac{\partial h}{\partial t}$$

Using the separation of variables substitution $(1/h)\{\partial h / \partial t\} = -\lambda$ and redefining $g[x] = \rho\theta$ gives:

$$\frac{d}{dx} \left[\frac{d}{dx} B g - A g \right] = -\lambda g = \frac{d}{dx} \left[\frac{d}{dx} B[x] \rho[x] \theta[x] - A[x] \rho[x] \theta[x] \right] = -\lambda \rho \theta$$

Evaluating the derivative inside the bracket and using the condition $\{d/dx\} [B\rho] - A\rho = 0$ to specify admissible ρ gives:

$$\frac{d}{dx} \left[\theta \frac{d}{dx} (B\rho) + B\rho \frac{d}{dx} \theta - A\rho\theta \right] = \frac{d}{dx} \left[B\rho \frac{d}{dx} \theta \right] = -\lambda \rho[x] \theta[x]$$

which is equation (6) in Wong (1964). The condition used to define ρ is then used to identify the specification of $B[x]$ and $A[x]$ from the Pearson system. The associated boundary condition follows from observing the $\rho[x]$ will be the ergodic density and making appropriate substitutions into the boundary condition:

$$\frac{\partial}{\partial x} \{B[x] f[t] \rho[x] \theta[x]\} - A[x] f[t] \rho[x] \theta[x] = 0 \quad \rightarrow \quad \frac{d}{dx} [B \rho \theta] - A \rho \theta = 0$$

Evaluating the derivative and taking values at the lower (or upper) boundary gives:

$$\begin{aligned} & B[a]\rho[a] \frac{d\theta[a]}{dx} + \theta[a] \frac{dB\rho}{dx} - A[a]\rho[a]\theta[a] = 0 \\ & = B[a]\rho[a] \frac{d\theta[a]}{dx} + \theta[a] \left[\frac{dB[a]\rho[a]}{dx} - A[a]\rho[a] \right] \end{aligned}$$

Observing the expression in the last bracket is the original condition with the ergodic density serving as U gives the boundary condition stated in Wong (1964, eq.7).

Proof of Proposition II:

(a) ψ_n has exactly n zeroes in $[a,b]$

Hille (1969, p.398, Theorem 8.3.3) and Birkhoff and Rota (1989, p.320, Theorem 5) shows that the eigenfunctions of the Sturm-Liouville system (1') with (3'), (4') and (5) have exactly n zeroes in the interval (a,b) . More precisely, since it assumed that $r > 0$, the eigenfunction ψ_n corresponding to the

n th eigenvalue has exactly n zeroes in (a,b) .

$$(b) \text{ For } \psi_n = 0, \quad \int_a^b \psi_n[x] \, dx = 0$$

Proof:

For $\psi_n = 0$ the following applies:

$$\begin{aligned} \psi_n &= \frac{1}{\lambda_n} \frac{d}{dx} \left\{ \frac{d}{dx} [B[x] \psi_n] - A[x] \psi_n \right\} \\ \therefore \int_a^b \psi_n[x] \, dx &= \frac{1}{\lambda_n} \left\{ \frac{d}{dx} [B[x] \psi_n] \Big|_{x=b} - A[b] \psi_n[b] - \frac{d}{dx} [B[x] \psi_n] \Big|_{x=a} + B[a] \psi_n[a] \right\} = 0 \end{aligned}$$

Since each $\psi_n[x]$ satisfies the boundary conditions (B.2).

(c) For some k , $\lambda_k = 0$.

Proof:

From Proposition 1:

$$U[x,t] = \sum_{k=0}^{\infty} c_k e^{-\lambda_k t} \psi_k[x]$$

Since $\int_a^b U[x,t] \, dx = 1$ then:

$$1 = \sum_{k=0}^{\infty} c_k e^{-\lambda_k t} \int_a^b \psi_k[x] \, dx$$

But from part (b) this will = 0 (which is a contradiction) unless $\lambda_k = 0$ for some k .

(d) $\lambda_0 = 0$

Proof:

From part (a), $\psi_0[x]$ has no zeroes in (a,b) . Therefore, either $\int_a^b \psi_0[x] \, dx > 0$ or $\int_a^b \psi_0[x] \, dx < 0$.

It follows from part (b) that $\lambda_0 = 0$.

(e) $\lambda_n > 0$ for $n \neq 0$. This follows from the strict inequality conditions provided in Proposition 1 and in part (d).

(f) Obtaining the solution for $T[x]$ in Proposition 2.

From part (d) it follows: $\frac{d}{dx} \left\{ [B[x] \psi_0[x,t]]_x - A[x] \psi_0[x,t] \right\} = 0$

Integrating this equation from a to x and using the boundary condition gives:

$$[B[x] \psi_0[x,t]]_x - A[x] \psi_0[x,t] = 0$$

This equation can be solved for ψ_0 to get:

$$\psi_0 = A [B[x]]^{-1} \exp \left[\int_a^x \frac{A[s]}{B[s]} ds \right] = C [r[x]]^{-1} \quad \text{where: } C = \text{constant}$$

Therefore:

$$\psi_0[x] = \left[\int_a^b r[x] C^2 [r[x]]^{-2} dx \right]^{-1/2} C [r[x]]^{-1} = \frac{[r[x]]^{-1}}{\left[\int_a^b r[x]^{-1} dx \right]^{1/2}}$$

Using the definition in Proposition I and observing that the integral of $f[x]$ over the state space is one it follows:

$$c_0 = \int_a^b \left\{ f[x] r[x] \frac{r[x]^{-1}}{\left[\int_a^b r[x] dx \right]^{1/2}} \right\} dx = \frac{1}{\left[\int_a^b r[x] dx \right]^{1/2}}$$

$$\therefore c_0 \psi[x] = \frac{r[x]^{-1}}{\left[\int_a^b [r[x]]^{-1} dx \right]}$$

(g) The Proof of Proposition 2 now follows from parts (f), (e) and (b).

NOTES

- 1 The σ_{ij} term is interpreted as being a covariance when $i \neq j$ and as a variance when $i = j$. Because the basic optimization problem is quadratic, it follows that the optimal solutions will take the form of an ellipse or a parabola. Consider the case where the $\{w_i\}$ are restricted to be non-negative, then the solution will be an ellipse. At any given target level of expected return, there will be two values of σ which solve the optimization problem. In evaluating the solutions, it is conventional to ignore the optimal solution which has the higher level of σ and consider only the portfolios which have the lowest σ .
2. Markowitz (1999) reviews the historical development of the model.
3. Such is the reason for using moving sampling windows instead of using all the data available. For example, 100 years of monthly data produces estimates of the arithmetic average which would not be affected by an additional observation. Hence, the optimal weights would not change over time.
4. The general approach in the following discussion is adapted from Fama (1976).
5. In rational mechanics, once the initial positions of the particles of interest, e.g., molecules, are known, the mechanical model fully determines the future evolution of the system. This scientific and philosophical approach is often referred to as Laplacian determinism.
6. Boltzmann and Max Planck were vociferous opponents of energetics. The debate over energetics was part of a larger intellectual debate concerning determinism and reversibility. Jevons (1877, p.738-9) reflects the entrenched determinist position of the marginalists: "We may safely accept as a satisfactory scientific hypothesis the doctrine so grandly put forth by Laplace, who asserted that a perfect knowledge of the universe, as it existed at any given moment, would give a perfect knowledge of what was to happen thenceforth and for ever after. Scientific inference is impossible, unless we may regard the present as the outcome of what is past, and the cause of what is to come. To the view of perfect intelligence nothing is uncertain." What Boltzmann, Planck and others had observed in statistical physics was that, even though the behavior of one or two molecules can be completely determined, it is not possible to generalize these mechanics to the describe the macroscopic motion of molecules in large, complex systems, e.g., Brush (1983, esp. ch.II).
7. This ignores developments in econometrics that commenced in the 1950's. These developments were concentrated on discrete time models that featured additive errors with strictly stationary distributions. In other words, probabilistic implications were incorporated by solving a deterministic model and then adding an error to the postulated theoretical relationship. This static probabilistic approach to modeling uncertainty has difficulty determining the non-linear dynamics that are captured by models associated with statistical mechanics.

8. As such, Boltzmann was part of the larger: “Second Scientific Revolution, associated with the theories of Darwin, Maxwell, Planck, Einstein, Heisenberg and Schrödinger, (which) substituted a world of process and chance whose ultimate philosophical meaning still remains obscure” (Brush 1983, p.79). This revolution superseded the: “First Scientific Revolution, dominated by the physical astronomy of Copernicus, Kepler, Galileo, and Newton, ... in which all changes are cyclic and all motions are in principle determined by causal laws.” As such, the irreversibility and indeterminism of the Second Scientific Revolution replaces the reversibility and determinism of the First.

9. There are many interesting sources on these points which provide citations for the historical papers that are being discussed. Cercignani (1998, p.146-50) discusses the role of Maxwell and Boltzmann in the development of the ergodic hypothesis. Maxwell (1879) is identified as “perhaps the strongest statement in favour of the ergodic hypothesis”. Brush (1974) has a detailed account of the development of the ergodic hypothesis. Gallavotti (1995) traces the etymology of “ergodic” to the ‘ergode’ in an 1884 paper by Boltzmann. More precisely, an ergode is shorthand for ‘ergomonode’ which is a ‘monode with given energy’ where a ‘monode’ can be either a single stationary distribution taken as an ensemble or a collection of such stationary distributions with some defined parameterization. The specific use is clear from the context. Boltzmann proved that an ergode is an equilibrium ensemble and, as such, provides a mechanical model consistent with the second law of thermodynamics. It is generally recognized that the modern usage of ‘the ergodic hypothesis’ originates with Ehrenfest and Ehrenfest (1912).

10. Reference to ‘the’ central limit theorem is somewhat misplaced as there are numerous varieties of central limit theorems which vary according to the initial assumptions imposed, e.g., Feller (1966, ch. VIII). Reference to the central limit theorem is to the general result which establishes the conditions under which sums of independent random variables are asymptotically normally distributed.

11. See Karlin and Taylor (1975, p.476) for a proof of this theorem. Because a covariance stationary process has a constant mean, the theorem says that the time variance of the stochastic process will converge to zero if and only if the covariance between elements of the process goes to zero as the time distance between the elements increases.

12. If the elements are correlated, it is possible to specify the correlation between elements, form correlation adjusted differences and then apply the weak law to these differences. For example, assume that the elements have a first order correlation ρ where $X(t) = \rho X(t-1) + u(t)$. The weak law can be applied to $(X(t) - \rho X(t-1))$.

13. The second law of thermodynamics is the universal law of increasing entropy – a measure of the randomness of molecular motion and the loss of energy to do work. First recognized in the early 19th century, the second law maintains that the entropy of an isolated system, not in equilibrium, will necessarily tend to increase over time. Entropy approaches a maximum value at thermal equilibrium. A number of attempts have been made to apply the entropy of information to problems in economics, with mixed success. In addition to the second law, physics now recognizes the zeroth law of thermodynamics that “any system approaches an equilibrium state” (Reed and Simon 1980, p.54).

The implications of the second law for theories in economics was explored by N. Georgescu-Roegen The Entropy Law and the Economic Process (1971).

14. This interpretation of the microscopic collisions differs from Davidson (1988, p.332): “If there is only one actual economy, and we do not possess, never have possessed and conceptually never will possess an *ensemble* of economics worlds, then even a definition of probability distribution functions is questionable.” In this context, points in the phase space at time t represent individual realizations of different macroscopic outcomes for the economic system at t . This interpretation of the ensembles is closer to Gibbs than Maxwell. Precisely how to interpret the ensembles in an economic context has not been closely examined. One exception is Nicola (1997).

15. That the unit shift transformation implies stationarity is apparent from the definition: A stochastic process is strictly stationary if all its finite-dimensional probability densities are invariant against time shifts. In other words, strict stationarity requires time homogeneity. This transformation is applicable to gambling-type situations where the probability distribution of the outcome, e.g., head or tail, is the same for all replications. This requires the distribution of the initial $x(0)$ to be identical to the distribution for $x(t)$ and permits objective relative frequencies to be used to estimate parameters such as the mean.

16. There are ergodic theorems for transformations that are not measure preserving, e.g., Hurewicz’s ergodic theorem (Halmos 1949, p.1017). In this case, weighted averages replace equally weighted averages and the weighted sums are shown to converge to a finite limit. Such transformations are not considered here and, as a consequence, all ergodic transformations considered are measure preserving.

17. This position is adopted in other sources, e.g., Dunn (2001, p.573): “One must be careful not to conflate the concepts of stationarity and ergodicity. A stochastic process is stationary if the estimates of times averages do not vary with the period under observation. Since some stationary processes are nonergodic, that is, limit cycles, nonstationarity is not necessary for nonergodicity. But since all nonstationary processes are nonergodic nonstationarity is a sufficient condition.”

18. The failure of limit cycles to be ergodic can be generalized to the case where ‘dynamical systems possessing periodic orbits are never ergodic’ (Wightman 1971, p.20). This is because a periodic system retains information about the initial condition as t goes to infinity, violating the ergodicity requirement. Stationarity in this case, which holds only for certain parameter values, is sometimes referred to as periodic stationarity to distinguish this case from the conventional strictly stationary case where processes are required to be ergodic.

19. A diffusion process is ‘regular’ if starting from any point in the state space I , any other point in I can be reached with positive probability (Karlin and Taylor 1981, p.158). This condition is distinct from other definitions of regular that will be introduced: ‘regular boundary conditions’ and ‘regular S-L problem’.

20. The classification of boundary conditions is typically an important issue in the study of solutions to the forward equation, e.g., Berg and McGregor (1966). Important types of boundaries include: regular; exit; entrance; and natural. Also important in boundary classification are: the properties of attainable and unattainable; whether the boundary is attracting or non-attracting; and whether the boundary is reflecting or absorbing. In the present context, only regular, attainable, reflecting boundaries are being considered in Sec. II with a few specific extensions to other types of boundaries being incorporated in Sec. III. In general, the specification of boundary conditions is essential in determining whether a given PDE is self-adjoint. The presence of the drift term in the boundary condition is required to ensure that the density integrates to one or, in the terminology of Feller (1952), that the boundary condition be norm preserving.

21. In Veerstraeten (2004), the use of Green's functions is implemented by using a transformation that achieves the PDE form: $g U_{xx} = U_t$ where the subscript denotes partial differentiation. A Laplace transform is then used to eliminate the time derivative. It is well known that using Laplace transforms to determine closed form solutions is usually restricted to the constant coefficient case because, without constant coefficients, the solution to the transform would involve another differential equation and nothing substantive is achieved by doing the transform. Hence, while Veerstraeten (2004) produces an insightful solution, more general cases require a different solution procedure if the Green's function solution is used to determine the transition probability density.

22. Following Karlin and Taylor (1991, p. 194-5) and Linetsky (2005, p.437) $r[x]$ can be interpreted as $B[x]$ times the 'scale density'.

23. Birkhoff and Rota (1989, p.337) demonstrate that the regular S-L problem has a spectrum that is always discrete and have eigenfunctions that are (trivially) square-integrable. These eigenfunctions will be orthogonal with respect to the weight function $r[x]$.

24. In the following. Proposition, $\Psi[x]$ is proportional to the "speed density" given in Karlin and Taylor (1981, p. 195).

25. This excludes the affect of the normalizing constant: $\int_a^b r[x]^{-1} dx$.

26. Hansen et al. (1998, p.12-3) recognize the importance of having solutions with a discrete spectrum and provide a sufficient general condition required for this result: 'finite first moment with the stationary density in natural scale'. This condition will always apply where there are reflecting barriers. The well-known result that a discrete spectrum is possible with certain singular diffusion problems arising with natural boundaries is also identified.

27. Whittaker and Watson (1963) is a useful source on Kummer and other transcendental functions.

28. More precisely, the probability $U[x, \infty | x_0]$ is associated with the set of time paths that start from x_0 and achieve an ending in the given volume element dx as $t \rightarrow \infty$.

29. The drift coefficient follows from observing $d \ln[\Psi] / dx = (\alpha - x)/x$ where the drift is specified as $A[x] - (dB[x]/dx) = (\alpha - x) \rightarrow A[x] = (\alpha - x) + 1$ (Cobb 1978).

30. Following standard convention, a closed-form solution is available if, and only if, at least one solution can be expressed in terms of a bounded number of well-known functions. These well-known functions are defined to be the elementary functions, including the error function, gamma function and the general hypergeometric functions. Solutions which involve infinite series, limits, and continued fractions are not consistent with closed forms.

31. In what follows, except where otherwise stated, it is assumed that $\sigma = 1$. Hence, the condition that K be a constant such that the density integrates to one incorporates $\sigma = 1$ assumption. Allowing $\sigma \neq 1$ will alter either the value of K or the β 's from that stated.

32. Ait-Sahalia (1999) also considers a diffusion with non-linear drift ($\alpha_0 + \alpha_1 X(t) + \alpha_2 X(t)^2 + \alpha_3 X(t)^{-1}$) and state dependent infinitesimal variance ($\sigma X(t)^p$). This complicated process could be readily transformed into the family \mathfrak{G} by setting $p=1$, and changing $1/X$ to $\ln[X]$ in the drift. Conceptual advantage can be gained by adding a cubic term in the drift, e.g., (Cobb 1981, p.76).

33. A number of simplifications were used to produce the 3D image in Figure 5.3.a: x has been centered about μ ; and, $\sigma = K_Q = 1$. Changing these values will impact the specific size of the parameter values for a given x but will not change the appearance of the density plots.