In time-series analysis, one is often interested in forecasting the future value of a random variable. Consider the first-order autoregression

\[ Y_t = \beta Y_{t-1} + U_t, \]

where \( U_t \) is a white-noise error with \( EU_t^2 = \sigma^2 \). The unconditional expectation of \( Y_t \) is zero while \( E(Y_t|Y_{t-1}) = \beta Y_{t-1} \), which clearly reveals the great gain in accuracy achieved with conditional forecasts. For such a model both the unconditional and conditional variances are constant

\[ \text{Var} (Y_t) = \frac{\sigma^2}{1-\beta} \quad \text{and} \quad \text{Var} (Y_t|Y_{t-1}) = \sigma^2. \]

One might expect better forecast intervals if additional information from the past were allowed to affect the forecast variance. The improved forecast intervals would arise under the idea that the mean and variance of the conditional distribution jointly evolve over time. Why might we need models of nonlinearities in finance? In analysis of asset returns, it is often the case that the conditional mean is unpredictable (at least the excess return above the riskless rate of interest) and so the interesting feature of the conditional distribution to model is the conditional variance. Accurate modeling of the conditional variance of an asset price (or return) is needed to value options on the asset. A related application is to correctly model a conditional covariance matrix for multiple assets to allow one to construct optimal portfolio weights.

Let \( Y_t \) be an asset return, such as the daily return on IBM. It is most often the case that we specify only that \( Y_t \) is a function of latent shocks \( \{\varepsilon_s\}_{s\leq t} \). Because the model is nonlinear, all structure is captured by the model and we assume that the shocks are iid. A general representation is

\[ Y_t = f (\varepsilon_t, \varepsilon_{t-1}, \ldots). \]

The general representation is often too vague to inform empirical work, so much empirical work is based on

\[ Y_t = g (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) + h (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \varepsilon_t, \]
often with the assumption that the shocks have mean zero and variance one to identify the conditional location and scale parameters as mean and variance. Such a model might arise if the volatility in the stock price of IBM reflects the changing structure of the computer market or the changing perception of risk in equities over time.

Observe that \( E_{t-1} Y_t = g (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \) and so the conditional variance is

\[
E_{t-1} (Y_t - E_{t-1} Y_t)^2 = h_t (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \equiv h_{t-1}.
\]

The simpler representation imposes substantial mathematical restrictions that one must consider in practice. While the first two conditional moments are allowed to vary, the time variation in all higher order conditional moments is inflexibly linked to the time variation in the conditional variance

\[
E_{t-1} (Y_t - E_{t-1} Y_t)^M = h_{t-1}^M E (\varepsilon_t^M).
\]

Models of Time-Varying Volatility

The quantity \( h_{t-1} \) captures time variation in the conditional distribution through the scale parameter. Because the concept of scale is less familiar to finance economists, it is much more common to refer to \( h_{t-1} \) (or \( h_{t-1}^2 \)) as the volatility. To link volatility to specific measurements of variation, one must make an assumption about \( \varepsilon_t \). If we assume \( E \varepsilon_t^2 = 1 \), then the volatility is the conditional variance, while if we assume that the interquartile range of \( \varepsilon_t \) is one, then the volatility is the conditional interquartile range.

To focus on the volatility parameter, we assume that \( Y_t \) is mean zero (such as the change in the asset price, rather than the price itself)

\[
Y_t = h_{t-1} \varepsilon_t.
\]

The unconditional variance of \( Y_t \) is

\[
EY_t^2 = Eh_{t-1}^2,
\]

so the variability of \( h_{t-1}^2 \) around its mean does not change the unconditional variance.

The variability of \( h_{t-1}^2 \) does, however, affect higher moments of the unconditional distribution of \( Y_t \). Even if the innovation \( \varepsilon_t \) is Gaussian, so the conditional
distribution of $Y_t$ is $N(0,h_t^2)$, the presence of time-varying volatility implies that
the unconditional distribution of $Y_t$ has thick tails (relative to scale). The kurtosis
of $Y_t$, defined as $K(Y_t) = \frac{EY_t^4}{(EY_t^2)^2}$, is

$$K(Y_t) = \frac{Eh_{t-1}^4E\varepsilon_{t-1}^4}{(Eh_{t-1}^2)^2} = 3\frac{Eh_{t-1}^4}{(Eh_{t-1}^2)^2} \geq 3,$$

where the last inequality is implied by Jensen’s inequality.

There are two principle methods to estimate the volatility $h_{t-1}^2$. The oldest
method, which is coming back into play, is to extract an estimate directly from
the data. The realized volatility estimator is the cumulated sum of volatilities
over short intervals. For example, the realized estimator of daily volatility is
the cumulated sum of volatilities over each half hour on the day. (See Diebold’s
survey paper. For the question of what the appropriate interval is, Olsen &
Associates may have a paper on selecting the minimum allowable interval.)

The alternative approach is to specify a parametric model for volatility. Be-
cause we typically model an asset return as a function of past shocks, we do
not use the extensive body of results related to “classic” treatment in which het-
erskedasticity is introduced through an exogenous variable $X$ as $U_t = X_{t-1}V_t
$ with $V_t$ white noise. Rather, we want the conditional variance to depend on past
realizations of $Y_t$. Perhaps the most direct extension of the “classic” treatment is
simply to replace $X_{t-1}$ with $Y_{t-1}$ to yield the bilinear model $Y_t = Y_{t-1}U_t$
with $U_t$ white noise. For the bilinear model the conditional variance is $Y_{t-1}^2\sigma_V^2$, although
because $Y_t = \prod_{i=0}^\infty U_{t-i}$ the unconditional variance is 0 if $\sigma^2 < 1$ and $\infty$ if $\sigma^2 > 1$.
A closely related variant of the bilinear model, for which the unconditional vari-
ance is generally nonzero and finite, is the ARCH model in which $Y_{t-1}$ is replaced
with $h_{t-1}$ where

$$h_t^2 = \omega + \sum_{i=1}^P \alpha_i Y_{t-i}^2.$$

In practice, $P$ tends to be large. To avoid estimation of a large number of para-
eters, the Generalized Autoregressive Conditional Heteroskedasticity (GARCH)
model was developed (it is now the most widely used model)

$$h_t^2 = \omega + \beta (L) h_{t-1}^2 + \alpha (L) Y_t^2,$$

in which past squared innovations determine the magnitude of current volatility.
Such a model captures the temporal clustering evident in much financial data.
By analogy with ARMA models, the model is a GARCH($p,q$) model when the order of the polynomial $\beta(L)$ is $p$ and the order of the polynomial $\alpha(L)$ is $q$. For almost all applications $p = q = 1$, written as

$$h_t^2 = \omega + \beta h_{t-1}^2 + \alpha Y_t^2 = \omega + (\alpha + \beta) h_{t-1}^2 + \alpha (Y_t^2 - h_{t-1}^2).$$

In the second equality, the term $(Y_t^2 - h_{t-1}^2)$ has mean zero, conditional on time $t-1$ information, and can be thought of as the shock to volatility. The coefficient $\alpha$ measures the extent to which a volatility shock today feeds through into next period’s volatility, while $(\alpha + \beta)$ measures the rate at which the effect dies out over time.

To understand how to use the model to forecast squared innovations (recall the innovations themselves are mean zero), note that rewriting the equation for $h_t^2$ yields

$$Y_{t+1}^2 = \omega + (\alpha + \beta) Y_t^2 + (Y_{t+1}^2 - h_t^2) - \beta (Y_t^2 - h_{t-1}^2).$$

The representation makes it clear that the GARCH(1,1) model is an ARMA(1,1) model for squared innovations. Unlike a standard ARMA(1,1) model, however, the shocks are heteroskedastic

$$(Y_{t+1}^2 - h_t^2) = h_t^2 (\varepsilon_{t+1}^2 - 1).$$

As anticipated from our discussion above, $\alpha + \beta$ impacts the stationarity properties of the model. If $\alpha + \beta < 1$, then the unconditional expectation of $Y_{t+1}^2$ (that is, the unconditional variance of $Y_{t+1}$) is $\frac{\omega}{1 - \alpha - \beta}$. The conditional expectation of volatility $j$ periods ahead is

$$E_t (h_{t+j}^2) = (\alpha + \beta)^j \left( h_t^2 - \frac{\omega}{1 - \alpha - \beta} \right) + \frac{\omega}{1 - \alpha - \beta}.$$  

The multiperiod volatility forecast reverts to its unconditional mean at rate $(\alpha + \beta)$. The relation between single-period and multiperiod forecasts is the same as in a linear ARMA(1,1) model with autoregressive coefficient $(\alpha + \beta)$.

If $\alpha + \beta = 1$, the conditional expectation of volatility $j$ periods ahead is instead

$$E_t (h_{t+j}^2) = h_t^2 + j \omega.$$  

The model has a unit autoregressive root so that today’s volatility affects forecasts of future volatility into the indefinite future. It is therefore known as the integrated GARCH, or IGARCH(1,1) model.
The IGARCH(1,1) model looks like a linear random walk with drift $\omega$. However, Nelson (1990) shows that this analogy must be treated with caution. A linear random walk is nonstationary in two senses. First, it has no stationary distribution, hence the process is not strictly stationary. Second, it has no unconditional first or second moments, hence it is not covariance stationary. In the IGARCH(1,1) model, in contrast, $h_t^2$ is strictly stationary even though its stationary distribution generally lacks unconditional moments. Thus the IGARCH(1,1) model is strictly stationary but not generally covariance stationary.

To understand the point, consider the case in which $\omega = 0$. Here $E_t (h_{t+j}^2) = h_t^2$, so volatility is a martingale. At the same time, volatility remains bounded because it cannot become negative. The martingale convergence theorem states that a bounded martingale must converge; in this case, the only value to which it can converge is zero. The stationary distribution for $h_t^2$ is then a degenerate point mass at zero, and this implies that the stationary distribution for $Y_{t+1}$ is also degenerate at zero. In this case the stationary distributions for $h_t^2$ and $Y_{t+1}$ have moments, but they are all trivially zero.

When $\omega > 0$, Nelson shows that there exists a nondegenerate stationary distribution for $h_t^2$. But this distribution does not have finite mean (hence, it does not have finite higher order moments). This implies that $Y_{t+1}$ has a stationary distribution with a zero mean but with tails that are so thick that second moments do not exist. (These conditions continue to hold with $\alpha + \beta > 1$ but with $E [\log (\beta + \alpha \epsilon_t^2)] < 0$.)

The GARCH model assumes that it is the magnitude of the innovation, rather than the sign of the innovation, which affects volatility. Yet many have noticed an asymmetry in stock market data (at least for the US): Negative innovations to stock prices tend to increase volatility more than positive innovations of the same magnitude. Perhaps the most general way in which this is captured is through the specification (note the move to conditional standard deviations)

$$h_t = \omega + \beta h_{t-1} + \alpha h_{t-1} f (\epsilon_t),$$

where

$$f (\epsilon_t) = |\epsilon_t - b| - c (\epsilon_t - b).$$

The parameter $b$ measures the shift, so that the impact of a shock on volatility depends on the distance between the shock and $b$. The parameter $c$ measures the tilt, that is the differential impact on volatility of two shocks of equal magnitude but different sign. To ensure that volatility estimates are positive, it must be the
case that \( f(\varepsilon_t) \) is positive, which in turn requires \(|c| \leq 1\). The attached *news impact curves* show the effect of different values of \( b \) and \( c \). Empirical research seems to have concluded \( b = 0 \), so search for asymmetry tends to focus on the exponential GARCH (EGARCH) model

\[
\log h_t = \omega + \beta \log h_{t-1} + \alpha (|\varepsilon_t| - c\varepsilon_t).
\]

This model is appealing because no parameter restrictions are required to ensure that the conditional variance is always positive. Also, if \( \alpha + \beta = 1 \), \( h_t \) is neither covariance stationary nor strictly stationary, avoiding the unusual statistical properties of the IGARCH(1,1) model. One drawback, multiperiod ahead forecasts of future variances are harder to calculate; no closed-form expressions like those for the GARCH model are available.

**Estimation**

Estimation of any of the above models is accomplished through the method of maximum likelihood. Let \( \theta \) be the vector of model parameters and define

\[
\varepsilon_{t+1}(\theta) = \frac{Y_{t+1}}{h_t(\theta)},
\]

where \( \varepsilon_{t+1}(\theta_0) \) is iid with density \( g(\cdot) \). In much work, it is assumed that \( g(\cdot) \) is standard Gaussian, so the quasi-log likelihood of \( Y_{t+1} \) is

\[
L(\theta; Y_{t+1} = y_{t+1}) = \ln g \left( \frac{y_{t+1}}{h_t(\theta)} \right) - \frac{1}{2} \ln \left[ h_t^2(\theta) \right],
\]

where the last term is the Jacobian term that appears because we observe \( Y_{t+1} \) and not \( \frac{Y_{t+1}}{h_t(\theta)} \). To form the QMLE we need an initial \( h_0^2(\theta) \) (for the log-likelihood of the first observation). Consistency of the QMLE is straightforward to establish but asymptotic normality is not. The difficulty is that the regularity conditions are hard to verify for GARCH processes. There are results for GARCH(1,1) models, but few other results are available.

**Conditional Nonnormality**

The standardized residuals from the volatility models should be Gaussian. Yet in most applications, the standardized residuals are leptokurtic. In an effort to overcome this problem, researchers have specified QML estimators based on distributions other than the Gaussian. See Newey and Steigerwald for an analysis of this case. (One can also use GMM to estimate the model, by assuming
that squared returns, less the appropriate function of the observed variables, are orthogonal to observed variables.)

The real issue is that the inliers are caused by overly high estimates of the conditional variance. See Steigerwald microstructure for this case.

**Stochastic-Volatility Models**

Another response is to assume that there is a random variable conditional upon which returns are normal, but that this variable—which is termed stochastic volatility—is latent. A simple example of a stochastic-volatility model is

\[
Y_t = \varepsilon_t e^{\alpha t}, \quad \alpha_t = \phi \alpha_{t-1} + v_t,
\]

where \(\varepsilon_t \sim N(0, \sigma^2_{\varepsilon})\), \(v_t \sim N(0, \sigma^2_v)\) and we assume that \(\varepsilon_t\) and \(v_t\) are serially uncorrelated and independent of each other. In essence \(e^{\alpha t}\) is the volatility parameter which is no longer measurable with respect to period \(t-1\) information. The model implies

\[
\log Y_t^2 = \alpha_t + \log \varepsilon_t^2, \quad \alpha_t = \phi \alpha_{t-1} + v_t.
\]

This is in linear state-space form except that the first equation has an error with a \(\log \chi^2\) distribution instead of a Gaussian distribution. To understand the specification, note that if \(\varepsilon_t\) is close to zero (an inlier), \(\log \varepsilon_t^2\) is a large negative outlier. Estimation of such models is complicated, but Bayesian methods have been shown to work well.

**Multivariate Models**

One of the most exciting avenues for research is to take these models to full scale analysis of portfolios to construct conditional covariance matrices. Let

\[
h_{ii,t;v} = \text{Var}(Y_{i,t+1})\]

and let

\[
h_{ij,t;v} = \text{Cov}(Y_{i,t+1}, Y_{j,t+1}).
\]

To demonstrate the logic of the models, we focus on the case in which we analyze two assets.

The most obvious way forward is to generalize the GARCH(1,1) model as

\[
h_{11,t;v} = \omega_1 + \beta_1 h_{11,t-1} + \beta_2 h_{12,t-1} + \beta_3 h_{22,t-1} + \alpha_1 Y^2_{1,t-1} + \alpha_2 Y_{1,t-1} Y_{2,t-1} + \alpha_3 Y^2_{2,t-1}.
\]

As similar equations are needed for \(h_{12,t;v}\) and \(h_{22,t;v}\) there are 21 parameters! In fact, the number of parameters grows with \(N^4\), so this model is not feasible.

Two goals then emerge. First, how to reduce the number of parameters. Second, how to ensure that the conditional covariance matrix is positive definite. The first approach is to assume that each conditional covariance is really a univariate process. For the model at hand

\[
h_{11,t;v} = \omega_1 + \beta_1 h_{11,t-1} + \alpha_1 Y^2_{1,t-1}.
\]
With similar equations for $h_{12,t}^2$ and $h_{22,t}^2$ there are now only 9 parameters (the number of parameters grows with $N^2$) and the conditional covariance matrix is positive definite if the matrices of parameters $[\omega_{ij}]$, $[\beta_{ij}]$ and $[\alpha_{ij}]$ are all positive definite.

A closely related alternative model defines the conditional variances as univariate GARCH(1,1) processes, but defines the conditional covariance term as

$$h_{12,t}^2 = \rho_{ij} \sqrt{h_{11,t}^2 h_{22,t}^2}.$$ 

Again, the number of parameters grows with $N^2$ and the conditional covariance matrix is positive definite if the correlations $\rho_{ij}$ make up a well defined correlation matrix and all parameters are positive.

To understand the differences in the models, consider what happens to the conditional variance after large shocks of opposite signs hit the two assets. For the first model, a positive $\alpha_{ij}$ implies that the negative cross product $Y_{1,t-1} Y_{2,t-1}$ lowers the conditional variance. In the remaining models, the conditional variance is unaffected.

**Nonparametric Analysis**

For other problems in finance, we may wish to estimate the conditional distribution nonparametrically. The most commonly used nonparametric estimators are smoothing estimators, in which observational errors are (hopefully) reduced by smoothing the data in sophisticated ways. Kernel regression, orthogonal series expansion, projection pursuit, nearest-neighbor estimators, average derivative estimators, splines and artificial neural networks are all examples of smoothing.

To understand the motivation for smoothing, suppose we wish to estimate the relation between two variables $Y_t$ and $X_t$, which satisfy

$$Y_t = m(X_t) + U_t, \quad t = 1, \ldots, n,$$

where $m(\cdot)$ is a fixed but unknown nonlinear function and $\{U_t\}$ is a zero-mean iid process. Consider estimating $m(\cdot)$ at a particular date $t_0$ for which $X_{t_0} = x_0$. Suppose that for this one observation $X_{t_0}$ we can obtain repeated independent observations of the variable $Y_{t_0}$, say $Y_{t_0}^1 = y_1, \ldots, Y_{t_0}^T = y_T$. Then a natural estimator of the function $m(\cdot)$ at the point $x_0$ is

$$\hat{m}(x_0) = \frac{1}{T} \sum_{i=1}^{T} y_i = m(x_0) + \frac{1}{T} \sum_{i=1}^{T} u_i.$$
By the Law of Large Numbers, the second term vanishes as $T$ grows.

Of course, if $\{Y_t\}$ is a time series, we do not have the luxury of repeated observations. If the function $m(\cdot)$ is sufficiently smooth, then for observations of $X_t$ near $x_0$ the corresponding value of $Y_t$ should be close to $m(x_0)$. From this logic one can conclude that a natural estimator is a weighted average of the $Y_t$'s, where the weights decline as the $X_t$'s get further from $x_0$. This weighted average procedure is the essence of smoothing. More formally, a smoothing estimator of $m(x)$ is

$$\hat{m}(x) = \frac{1}{n} \sum_{t=1}^{n} w_{t,n}(x) Y_t,$$

where the weights depend on the distance between $X_t$ and $x$. Selection of the weight function is crucial; if we select too large a neighborhood, the weighted average will be too smooth and will not exhibit the nonlinearities of $m(x)$ (we will introduce bias); if we select a neighborhood that is too small, the weighted average will be too variable, reflecting noise as well as variation in $m(x)$ (we will have large variance). The weights are chosen carefully to balance the bias against the variance.

An important smoothing technique is kernel regression. The weight function $w_{t,n}(x)$ is constructed from a probability density function $K(x)$, also called a kernel. For the most basic kernel

$$K(x) > 0, \quad \int K(u) \, du = 1.$$  

Despite the fact that $K(x)$ is a probability density function, it plays no probabilistic role in the subsequent analysis. The kernel is simply a convenient method for computing a weighted average. In particular, use of $K(x)$ does not imply that $X$ is distributed as $K(x)$, which would be a parametric assumption.

By rescaling the kernel with respect to a variable $h > 0$, we can change the spread of the kernel by varying $h$. The rescaled kernel is defined as

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right), \quad \int K_h(u) \, du = 1.$$  

To understand the import of rescaling, consider the following two kernels. The uniform kernel, which we define without loss of generality over $u \in [0, 1]$, gives equal weight to each element in its range. Let $h = 2$, so that

$$K_h(u) = \frac{1}{2} K\left(\frac{u}{h}\right) = \frac{1}{2}, \quad u \in [0, 1].$$
We see that for the uniform kernel, introduction of $h$ does not alter the fact that the weights are uniform over the range, but does alter the magnitude of the weight. For a Gaussian kernel, in which $K(u)$ represents a standard Gaussian density, the range is the entire real line and the weights decline as we move from the mean. For $h > 0$,

$$K_h(u) = \frac{1}{h} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{u^2}{h^2}},$$

so $h$ is the standard deviation and thus controls the spread of the weight function.

The weight function is defined as

$$w_{t,n}(x) = \frac{K_h(x - X_t)}{g_h(x)},$$

where

$$g_h(x) = \frac{1}{n} \sum_{t=1}^{n} K_h(x - X_t).$$

The numerator varies with distance a given observation is from $x$. The denominator, which does not vary over $t$, is the average over all the weights and ensures that $\sum_{t=1}^{n} w_{t,n}(x) = n$. If $h$ is small, then almost all the weight is concentrated in a small neighborhood around $x$. Controlling the degree of smoothing amounts to adjusting the parameter $h$, termed the bandwidth parameter. With this weight function obtained from a kernel, the resultant estimator of $m(x)$ is the Nadaraya-Watson kernel estimator

$$\hat{m}_h(x) = \frac{1}{n} \sum_{t=1}^{n} w_{t,n}(x) Y_t = \frac{\sum_{t=1}^{n} K_h(x - X_t) Y_t}{\sum_{t=1}^{n} K_h(x - X_t)}.$$

Under certain regularity conditions on the shape of the kernel and the magnitudes and behavior of the weights as the sample size grows, it may be shown that $\hat{m}_h(x)$ converges to $m(x)$ in several ways.

**Bandwidth Selection**

There is a vast literature on optimal selection of the bandwidth parameter. On an intuitive level, one can think of selection $h = 3s_X$, or perhaps $1s_X$ with $s_X$ the sample standard deviation of $\{X_t\}$. Such a selection is naturally motivated by the Gaussian kernel, in which the smoothing parameter is related to the standard deviation of $x$. (Note, we do not assume the kernel is the density for $X_t$, so it is not an implication that $h$ is the standard deviation of $X_t$.) The most common
bandwidth selection method is (least-squares) cross validation, popular because of its robustness and asymptotic optimality. With cross validation, the bandwidth is chosen to minimize a weighted-average squared error of the kernel estimator. To determine the error, we evaluate $m(\cdot)$ at each point in the sample

$$m_{h,j}(X_j) = \frac{1}{n} \sum_{t \neq j} w_{t,n}(X_j) Y_t.$$ 

As the expression makes clear, observation $X_j$ is not used to estimate the function, which is the essence of cross validation. The cross-validation function, that is the weighted-average squared error of the kernel estimator, is

$$CV(h) = \frac{1}{n} \sum_{t=1}^{n} [Y_t - m_{h,t}(X_t)]^2 \delta(X_t),$$

where $\delta(X_t)$ is a nonnegative weight function that is required to reduce boundary effects. The term cross validation arises because $CV(h)$ validates the success of the kernel estimator in fitting $\{Y_t\}$ across $n$ subsamples $\{X_t, Y_t\}_{t \neq j}$, each with one observation omitted. The optimal bandwidth minimizes $CV(h)$.

The preceding analysis is for a scalar quantity. With multivariate quantities, the problem becomes more complex. One path forward is the average derivative estimator, available when $m(X_{i1}, \ldots, X_{iK})$ is a function of a single index $m(X_{i1}^{(1)})$. See page 504 of Campbell et al. to begin this analysis.

**Estimation of State-Price Densities**

One of the fundamental problems in finance is to estimate the density that prices assets. The density, which is derived from equilibrium theory, embodies the value of Arrow-Debreu securities. These securities pay $1 in one specific state of nature and nothing otherwise and form the building blocks of understanding economic equilibrium in an uncertain environment.

To begin our analysis, consider the pricing of assets by agents with different risk preferences. Consider the equilibrium price $P_t$ of a security at date $t$ with a single liquidating payoff $Y(C_T)$ at date $T$ that is a function of aggregate consumption $C_T$. A standard dynamic exchange economy yields the equilibrium price (derived by Lucas in 1978)

$$P_t = E_t \left[ Y(C_T) M_{t,T} \right], \quad M_{t,T} = \delta^{T-t} \frac{U^{(1)}(C_T)}{U^{(1)}(C_t)}.$$
The quantity $M_{t,T}$ is the marginal rate of substitution between dates $t$ and $T$, where $\delta$ is the rate of time preference. If we assume that the conditional distribution of future consumption has a density representation $f_t(\cdot)$, then

$$E_t [Y(C_T)M_{t,T}] = \int Y(C_T) M_{t,T} f_t(C_T) \, dC_T$$

$$= e^{-r_{t,T}(T-t)} \int Y(C_T) f_t^*(C_T) \, dC_T$$

$$= e^{-r_{t,T}(T-t)} E_t^* [Y(C_T)],$$

where

$$f_t^*(C_T) = \frac{M_{t,T} f_t(C_T)}{\int M_{t,T} f_t(C_T) \, dC_T},$$

in which $r_{t,T}$ is the risk-free return, the continuously compounded net rate of return between $t$ and $T$ of an asset promising one unit of consumption at $T$. The asset price is expressed as the discounted expected payoff, in which $e^{-r_{t,T}(T-t)}$ is the discount and the expected payoff is evaluated with $f_t^*(C_T)$ (termed the state-price density or SPD) rather than with $f_t(C_T)$. (The function $f_t^*(C_T)$ is also known as the risk-neutral pricing density or the equivalent martingale measure.)

The expression provides a fundamental insight. If $f_t(\cdot) = f_t^*(\cdot)$, then asset prices are a martingale. In general $f_t(\cdot) \neq f_t^*(\cdot)$ and so rational determination of asset prices does not imply that asset prices are a martingale. LeRoy first made this point in 1973, when he noted that if agents are risk-neutral, and so care only about expected consumption, prices will follow a martingale while if agents are risk-averse, and so care about the variance of consumption as well, prices need not follow a martingale. Observe that if individuals are risk neutral, then $U(C_T)$ is a linear function of $C_T$ (as individuals care only about expected $C_T$), hence $M_{t,T} = \delta^{T-t}$ (because $U^{(1)}(C_t)$ is constant for all values of consumption) and

$$f_t^*(C_T) = \frac{\delta^{T-t} f_t(C_T)}{\int \delta^{T-t} f_t(C_T) \, dC_T} = f_t(C_T).$$

To work in the opposite direction, and so understand how the weight function arises in $f_t^*(C_T)$, it is easiest to consider a model in which $C_T$ takes only two values. We do not do this now. (Discuss with LeRoy, note if there are only two values, then the densities discussed above must be replaced with distributions.)

Arrow-Debreu securities are not traded and so their prices cannot be directly observed. Their prices can be inferred, however, from the assets for which we do
observe prices. The goal of estimating the SPD is closely linked to determining the data-generating process for asset prices and to inferring the preferences of a representative agent in an equilibrium model of asset prices. In fact, any two of the following implies the third: (1) the representative agent’s preferences; (2) asset price dynamics; and (3) the SPD. It is the relation between the three that lends great weight to certain analysis. LeRoy (and Shiller) demonstrated that, in a world with risk-neutral investors, asset prices are too variable in relation to expected future dividends. As the introduction of risk-averse investors could lead to greater variation in the present discounted sum, the restriction to risk-neutral investors blunted the impact of the analysis. Mehra and Prescott demonstrated that with risk-averse investors, the estimated premium on equities implied a risk aversion parameter of 50 (which in turn implies that individuals would prefer a certain loss of $1000 to a gamble in which they would lose $0 or $1001 with equal probability). The potential weakness of Mehra and Prescott is that they allowed for only two states for $C_T$. Hansen and Jaganathann allowed for multiple states for $C_T$ and, while unable to construct a point estimate for risk aversion were able to construct bounds. Their bounds support the Mehra and Prescott conclusion.

These results present a fundamental question for finance. The empirical estimates from asset data are not in accord with the implications of equilibrium theory. As the equilibrium theory is based on axioms from economic analysis, such a conclusion is troubling. One may question the axioms on which the equilibrium theory rests. Few are willing. Rather, one possible answer is to propose that the restrictions in the economy require a corner solution, generating a different answer than the one provided above. Cochrane & Campbell and Constantinides and Mehra have such solutions. They push very hard on the rationality of agents. An alternative line of attack is to say that the empirical estimates are biased. When one observes that over horizons of 20 years, equities are in fact less risky than Treasury bonds, one wonders what is the makeup of individuals who price equities as more risky? This leads to behavioral finance, or more appropriately to the study of how individuals deal with and understand risk. (Papers by Toobey and Cosmides.)

With these caveats in mind, we return to analysis of the SPD. Under certain regularity conditions, we may express $f^*$ as an explicit function of $t$ and $T$ so that a single SPD $f^* (C_T; t, T)$ may be used to price an asset at any date $t$ with a single liquidating payoff $Y (C_T)$ at any future date $T \geq t$ (see page 509 of Campbell, Lo,
MacKinlay to understand multiple payoffs and infinite horizons)

\[ P_t = e^{-r_t,\tau(T-t)} \int Y(C_T) f^*(C_T; t, T) dC_T. \]

If the asset is European call option, then

\[ Y(C_T) = \max [C_T - X, 0] \text{ with price } G_t \]

(or call-pricing function)

\[ G_t = e^{-r_t,\tau(T-t)} \int \max [C_T - X, 0] f^*(C_T; t, T) dC_T. \]

When dealing with option prices, as opposed to general asset prices, the relation between the price and the SPD is even closer. The second derivative of the call-pricing function with respect to the strike price \( X \) must equal the SPD

\[ \frac{\partial^2 G_t}{\partial X^2} = e^{-r_t,\tau(T-t)} f^*. \]

Implicit in every option pricing formula is the SPD \( f^* \).

Parametric option pricing formulae form a large family, whose most famous member is the Black-Scholes formula. Recent work has concentrated on non-parametric estimation of \( f^* \). Given observed call option prices \( \{G_i, X_i, \tau_i\} \), with \( \tau_i = T_i - t_i \), prices of the underlying asset \( \{P_i\} \) and the riskless rate of interest \( \{r_{\tau_i}\} \), the nonparametric estimator of the call-pricing function is

\[ \hat{G}(P, X, \tau, r_{\tau}) = \hat{E}(G|P, X, \tau, r_{\tau}). \]

Because there are four conditioning arguments, a multivariate kernel is needed and is formed as the product of four univariate kernels (each of which has its own bandwidth). The kernel regression estimator is

\[ \hat{G}(p, x, \tau, r_{\tau}) = \frac{\sum_{i=1}^{n} K(i) C_i}{\sum_{i=1}^{n} K(i)}, \]

where

\[ K(i) = \sum_{i=1}^{n} K_{\tau}(p - P_i) K_{x}(x - X_i) K_{\tau}(\tau - \tau_i) K_{r_{\tau}}(r_{\tau} - r_{\tau_i}). \]

From \( \hat{G} \), one can estimate the option's delta (the derivative of \( G \) with respect to \( P \)) as

\[ \hat{\Delta}(P, X, \tau, r_{\tau}) = \frac{\partial \hat{G}}{\partial P}. \]
and the SPD

\[ \hat{f}^* (P_T | P, \tau, r_\tau) = e^{r_\tau \tau} \left[ \frac{\partial^2 \hat{G}}{\partial X^2} \right]_{X = P_T} . \]

With the SPD estimator, one can price any derivative security with characteristic \( \theta \) and payoff function \( Y(\theta, P_T) \) as

\[ G(P, \theta, \tau, r_\tau) = e^{-r_\tau \tau} \int_0^\infty Y(\theta, P_T) \hat{f}^* (P_T | P, \tau, r_\tau) dP_T. \]

Some empirical success has been reported in capturing features of asset prices that parametric formulae have difficulty matching.

**Artificial Neural Networks**

An alternative to nonparametric regression that has recently received much attention is the **artificial neural network**. Artificial neural networks may be viewed as a nonparametric technique, but they are here treated separately because of their early origins in biology (specifically the physiology of nerve cells).

**Multilayer Perceptrons**

Perhaps the simplest example of an artificial neural network is the **binary threshold model** in which an output variable \( Y \) taking on only the values zero and one is nonlinearly related to a collection of \( J \) input variables \( X_j, j = 1, \ldots, J \) in the following way

\[ Y = g \left( \sum_{j=1}^{J} \beta_j X_j - \mu \right) \]

\[ g(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases} . \]

Each input is weighted by the coefficient \( \beta_j \), called the connection strength, and the summed across all inputs. If this weighted sum exceeds the threshold \( \mu \), then the artificial neuron is switched on or activated via the activation function \( g(\cdot) \). Without loss of generality, we set \( \mu = 0 \) because a nonzero threshold is captured by setting the first input \( X_1 = 1 \), in which case the negative of this input’s connection strength \(-\beta_1\) is the threshold. The form of \( g(\cdot) \) is that of an
indicator variable, also known as the *Heaviside* function. An alternative, with a smooth transition path, is the logistic function

\[ g(v) = \frac{1}{1 + e^{-v}}. \]

Perhaps the most important extension of the binary threshold model is the introduction of a hidden layer between the input (layer) and the output (layer). Specifically, let

\[
Y = h \left( \sum_{k=1}^{K} \alpha_k g (\beta'_k X) \right)
\]

\[ \beta_k = [\beta_{k1} \beta_{k2} \cdots \beta_{kJ}]', \quad X = [X_1 X_2 \cdots X_J]', \]

where \( h(\cdot) \) is another arbitrary nonlinear activation function. Here the inputs are connected to multiple hidden units (indexed by \( k \)), and at each hidden unit they are weighted (differently) and transformed by the same activation function \( g(\cdot) \). The output of each hidden unit is weighted yet again, this time by the \( \alpha'_k \)’s, and summed and transformed by a second activation function \( h(\cdot) \). Such a model is an example of a multilayer perceptron, with a single hidden layer in this case although more can clearly be added.

Even the single hidden-layer MLP possesses the universal approximation property: It can approximate any nonlinear function to an arbitrary degree of accuracy with a suitable number of hidden units. The parameters, the scalars \( \alpha_k \) and the vectors \( \beta_k, k = 1, \ldots, K \) are obtained by minimizing the sum of squared deviations between the output and the network

\[
\sum_t \left[ Y_t - \sum_k \alpha_k g (\beta'_k X) \right]^2.
\]

The main advantage of this method over nonparametric regression is that the functions can be quite simple, often \( h(\cdot) \) is the identity function and \( g(\cdot) \) is the logistic function, which leads naturally to parallel processing.