Financial Forecasting, Non-linear Time Series in

GLORIA GONZÁLEZ-RIVERA, TAE-HWY LEE
Department of Economics, University of California,
Riverside, USA

Article Outline

Glossary

Arbitrage pricing theory (APT) the expected return of an asset is a linear function of a set of factors.

Artificial neural network is a nonlinear flexible functional form, connecting inputs to outputs, being capable of approximating a measurable function to any desired level of accuracy provided that sufficient complexity (in terms of number of hidden units) is permitted.

Autoregressive conditional heteroskedasticity (ARCH) the variance of an asset returns is a linear function of the past squared surprises to the asset.

Bagging short for *b*ootstrap *agg*regat*ing*. Bagging is a method of smoothing the predictors' instability by averaging the predictors over bootstrap predictors and thus lowering the sensitivity of the predictors to training samples. A predictor is said to be unstable if perturbing the training sample can cause significant changes in the predictor.

Capital asset pricing model (CAPM) the expected return of an asset is a linear function of the covariance of the asset return with the return of the market portfolio.

Factor model a linear factor model summarizes the dimension of a large system of variables by a set of factors that are linear combinations of the original variables.

Financial forecasting prediction of prices, returns, direction, density or any other characteristic of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc.

Functional coefficient model a model with time-varying and state-dependent coefficients. The number of states can be infinite.

Linearity in mean the process $\{y_t\}$ is linear in mean conditional on X_t if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k .$$

Loss (cost) function When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c_{t+h}(Y_{t+h}, f_{t,h})$, and the function $c_{t+h}(\cdot)$ can change over t and the forecast horizon h.

Markov-switching model features parameters changing in different regimes, but in contrast with the threshold models the change is dictated by a non-observable state variable that is modelled as a hidden Markov chain.

Martingale property tomorrow's asset price is expected to be equal to today's price given some information set

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t .$$

Nonparametric regression is a data driven technique where a conditional moment of a random variable is specified as an unknown function of the data and estimated by means of a kernel or any other weighting scheme on the data.

Random field a scalar random field is defined as a function $m(\omega, x) : \Omega \times A \to R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$.

Sieves the sieves or approximating spaces are approximations to an unknown function, that are dense in the original function space. Sieves can be constructed using linear spans of power series, e. g., Fourier series, splines, or many other basis functions such as artificial neural network (ANN), and various polynomials (Hermite, Laguerre, etc.).

Smooth transition models threshold model with the indicator function replaced by a smooth monotonically increasing differentiable function such as a probability distribution function.

Threshold model a nonlinear model with time-varying coefficients specified by using an indicator which takes a non-zero value when a state variable falls on a specified partition of a set of states, and zero otherwise. The number of partitions is finite.

Varying cross-sectional rank (VCR) of asset i is the proportion of assets that have a return less than or equal to the return of firm i at time t

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t})$$

Volatility Volatility in financial economics is often measured by the conditional variance (e. g., ARCH) or the conditional range. It is important for any decision making under uncertainty such as portfolio allocation, option pricing, risk management.

Definition of the Subject

Financial Forecasting

Financial forecasting is concerned with the prediction of prices of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc. Though many agents in the economy, i. e. investors, money managers, investment banks, hedge funds, etc. are interested in the forecasting of financial prices per se, the importance of financial forecasting derives primarily from the role of financial markets within the macro economy. The development of financial instruments and financial institutions contribute to the growth and stability of the overall economy. Because of this interconnection between financial markets and the real economy, financial forecasting is also intimately linked to macroeconomic forecasting, which is concerned with the prediction of macroeconomic aggregates such as growth of the gross domestic product, consumption growth, inflation rates, commodities prices, etc. Financial forecasting and macroeconomic forecasting share many of the techniques and statistical models that will be explained in detail in this article.

In financial forecasting a major object of study is the return to a financial asset, mostly calculated as the continuously compounded return, i. e., $y_t = \log p_t - \log p_{t-1}$ where p_t is the price of the asset at time t. Nowadays financial forecasters use sophisticated techniques that combine the advances in modern finance theory, pioneered by Markowitz [113], with the advances in time series econometrics, in particular the development of nonlinear models for conditional moments and conditional quantiles of asset returns.

The aim of finance theory is to provide models for expected returns taking into account the uncertainty of the future asset payoffs. In general, financial models are concerned with investors' decisions under uncertainty. For instance the portfolio allocation problem deals with the allocation of wealth among different assets that carry different levels of risk. The implementation of these theories relies on econometric techniques that aim to estimate financial models and testing them against the data. Financial econometrics is the branch of econometrics that provides model-based statistical inference for financial variables, and therefore financial forecasting will provide their corresponding model-based predictions. However there are also econometric developments that inform the construction of ad hoc time series models that are valuable on describing the stylized facts of financial data.

Since returns $\{y_t\}$ are random variables, the aim of financial forecasting is to forecast conditional moments, quantiles, and eventually the conditional distribution of these variables. Most of the time our interest will be centered on expected returns and volatility as these two moments are crucial components on portfolio allocation problems, option valuation, and risk management, but it is also possible to forecast quantiles of a random variable, and therefore to forecast the expected probability density function. Density forecasting is the most complete forecast as it embeds all the information on the financial variable of interest. Financial forecasting is also concerned with other financial variables like durations between trades and directions of price changes. In these cases, it is also possible to construct conditional duration models and conditional probit models that are the basis for forecasting durations and market timing.

Critical to the understanding of the methodological development in financial forecasting is the statistical concept of *martingale*, which historically has its roots in the games of chance also associated with the beginnings of probability theory in the XVI century. Borrowing from the concept of fair game, financial prices are said to enjoy the *martingale property* if tomorrow's price is expected to be equal to today's price given some information set; in other words tomorrow's price has an equal chance to either move up or move down, and thus the best forecast must be the current price. The martingale property is written as

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t$$

where $\mathbb{E}$ is the expectation operator and the information set $\mathcal{F}_t \equiv \{p_t, p_{t-1}, p_{t-2}, \dots\}$ is the collection of past and current prices, though it may also include other variables known at time t such as volume. From a forecasting point of view, the martingale model implies that changes in financial prices $(p_{t+1} - p_t)$ are not predictable.

The most restrictive form of the martingale property, proposed by Bachelier [6] in his theory of speculation is the model (in logarithms)

$$\log p_{t+1} = \mu_t + \log p_t + \varepsilon_{t+1} ,$$

where $\mu_t = \mu$ is a constant drift and ε_{t+1} is an identically and independently distributed (i.i.d.) error that is assumed to be normally distributed with zero mean and constant variance σ^2. This model is also known as a random walk model. Since the return is the percentage change in prices, i. e. $y_t = \log p_t - \log p_{t-1}$, an equivalent model for asset returns is

$$y_{t+1} = \mu_t + \varepsilon_{t+1} .$$

Then, taking conditional expectations, we find that $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_t$. If the conditional mean return is not time-varying, $\mu_t = \mu$, then the returns are not forecastable based on past price information. In addition and given the assumptions on the error term, returns are independent and identically distributed random variables. These two properties, a constant drift and an i.i.d error term, are too restrictive and they rule out the possibility of any predictability in asset returns. A less restrictive and more plausible version is obtained when the i.i.d assumption is relaxed. The error term may be heteroscedastic so that returns have different (unconditional or conditional) variances and consequently they are not identically distributed, and/or the error term, though uncorrelated, may exhibit dependence in higher moments and in this case the returns are not independent random variables.

The advent of modern finance theory brings the notion of systematic risk, associated with return variances and covariances, into asset pricing. Though these theories were developed to explain the cross-sectional variability of financial returns, they also helped many years later with the construction of time series models for financial returns. Arguably, the two most important asset pricing models in modern finance theory are the Capital Asset Pricing Model (CAPM) proposed by Sharpe [137] and Lintner [103] and the Arbitrage Pricing Theory (APT) proposed by Ross [131]. Both models claim that the expected return to an asset is a linear function of risk; in CAPM risk is related to the covariance of the asset return with the return to the market portfolio, and in APT risk is measured as exposure to a set of factors, which may include the market portfolio among others. The original version of CAPM, based on the assumption of normally distributed returns, is written as

$$\mathbb{E}(y_i) = y_f + \beta_{im}\left[\mathbb{E}(y_m) - y_f\right] ,$$

where y_f is the risk-free rate, y_m is the return to the market portfolio, and β_{im} is the risk of asset i defined as

$$\beta_{im} = \frac{\operatorname{cov}(y_i, y_m)}{\operatorname{var}(y_m)} = \frac{\sigma_{im}}{\sigma_m^2} .$$

This model has a time series version known as the conditional CAPM [17] that it may be useful for forecasting purposes. For asset i and given an information set as $\mathcal{F}_t = \{y_{i,t}, y_{i,t-1}, \dots; y_{m,t}, y_{m,t-1}, \dots\}$, the expected return is a linear function of a time-varying beta

$$\mathbb{E}(y_{i,t+1}|\mathcal{F}_t) = y_f + \beta_{im,t}\left[\mathbb{E}(y_{m,t+1}|\mathcal{F}_t) - y_f\right]$$

where $\beta_{im,t} = \frac{\operatorname{cov}(y_{i,t+1}, y_{m,t+1}|\mathcal{F}_t)}{\operatorname{var}(y_{m,t+1}|\mathcal{F}_t)} = \frac{\sigma_{im,t}}{\sigma_{m,t}^2}$. From this type of models is evident that we need to model the conditional second moments of returns jointly with the conditional mean. A general finding of this type of models is that when there is high volatility, expected returns are high, and hence forecasting volatility becomes important for the forecasting of expected returns. In the same spirit, the APT models have also conditional versions that exploit the information contained in past returns. A K-factor APT model is written as

$$y_t = c + B' f_t + \varepsilon_t ,$$

where f_t is a $K \times 1$ vector of factors and B is a $K \times 1$ vector of sensitivities to the factors. If the factors have time-varying second moments, it is possible to specify an APT model with a factor structure in the time-varying covariance matrix of asset returns [48], which in turn can be exploited for forecasting purposes.

The conditional CAPM and conditional APT models are fine examples on how finance theory provides a base to specify time-series models for financial returns. However there are other time series specifications, more *ad hoc* in nature, that claim that financial prices are nonlinear functions – not necessarily related to time-varying second moments – of the information set and by that, they impose some departures from the martingale property. In this case it is possible to observe some predictability in asset prices. This is the subject of nonlinear financial forecasting. We begin with a precise definition of linearity versus nonlinearity.

Linearity and Nonlinearity

Lee, White, and Granger [99] are the first who precisely define the concept of "linearity". Let $\{Z_t\}$ be a stochastic process, and partition Z_t as $Z_t = (y_t\ X_t')'$, where (for simplicity) y_t is a scalar and X_t is a $k \times 1$ vector. X_t may (but need not necessarily) contain a constant and lagged values of y_t. LWG define that the process $\{y_t\}$ is *linear in mean conditional on* X_t if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k .$$

In the context of forecasting, Granger and Lee [71] define linearity as follows. Define $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$ being the optimum least squares h-step forecast of y_{t+h} made at time t. μ_{t+h} will generally be a nonlinear function of the contents of $\mathcal{F}_t$. Denote m_{t+h} the optimum *linear* forecast of y_{t+h} made at time t be the best forecast that is constrained to be a linear combination of the contents of $X_t \in \mathcal{F}_t$. Granger and Lee [71] define that $\{y_t\}$ is said to be *linear in conditional mean* if μ_{t+h} is linear in X_t, i. e., $\Pr\left[\mu_{t+h} = m_{t+h}\right] = 1$ for all t and for all h. Under this definition the focus is the conditional mean and thus a process exhibiting autoregressive conditional heteroskedasticity (ARCH) [44] may nevertheless exhibit linearity of this sort because ARCH does not refer to the conditional mean. This is appropriate whenever we are concerned with the adequacy of linear models for forecasting the conditional mean returns. See [161], Section 2, for a more rigorous treatment on the definitions of linearity and nonlinearity.

This definition may be extended with some caution to the concept of linearity in higher moments and quantiles, but the definition may depend on the focus or interest of the researcher. Let $\varepsilon_{t+h} = y_{t+h} - \mu_{t+h}$ and $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$. If we consider the ARCH and GARCH as linear models, we say $\{\sigma^2_{t+h}\}$ is linear in conditional variance if σ^2_{t+h} is a linear function of lagged ε^2_{t-j} and σ^2_{t-j} for some h or for all h. Alternatively, $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$ is said to be linear in conditional variance if σ^2_{t+h} is a linear function of $x_t \in \mathcal{F}_t$ for some h or for all h. Similarly, we may consider linearity in conditional quantiles. The issue of linearity versus nonlinearity is most relevant for the conditional mean. It is more relevant whether a certain specification is correct or incorrect (rather than linear or nonlinear) for higher order conditional moments or quantiles.

Introduction

There exists a nontrivial gap between martingale difference and serial uncorrelatedness. The former implies the latter, but not vice versa. Consider a stationary time series $\{y_t\}$. Often, serial dependence of $\{y_t\}$ is described by its autocorrelation function $\rho(j)$, or by its standardized spectral density

$$h(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \rho(j) \mathrm{e}^{-ij\omega}, \quad \omega \in [-\pi, \pi] .$$

Both $h(\omega)$ and $\rho(j)$ are the Fourier transform of each other, containing the same information of serial correlations of $\{y_t\}$. A problem with using $h(\omega)$ and $\rho(j)$ is that they cannot capture nonlinear time series that have zero autocorrelation but are not serially independent. Nonlinear MA and Bilinear series are good examples:

$$\begin{aligned} \text{Nonlinear MA}: \quad & Y_t = b e_{t-1} e_{t-2} + e_t , \\ \text{Bilinear}: \quad & Y_t = b e_{t-1} Y_{t-2} + e_t . \end{aligned}$$

These processes are serially uncorrelated, but they are predictable using the past information. Hong and Lee [86] note that the autocorrelation function, the variance ratios, and the power spectrum can easily miss these processes. Misleading conclusions in favor of the martingale hypothesis could be reached when these test statistics are insignificant. It is therefore important and interesting to explore whether there exists a gap between serial uncorrelatedness and martingale difference behavior for financial forecasting, and if so, whether the neglected nonlinearity in conditional mean can be explored to forecast financial asset returns.

In the forthcoming sections, we will present, without being exhaustive, nonlinear time series models for financial returns, which are the basis for nonlinear forecasting. In Sect. "Nonlinear Forecasting Models for the Conditional Mean", we review nonlinear models for the conditional mean of returns. A general representation is $y_{t+1} = \mu(y_t, y_{t-1}, \dots) + \varepsilon_{t+1}$ with $\mu(\cdot)$ a nonlinear function of the information set. If $\mathbb{E}(y_{t+1}|y_t, y_{t-1}, \dots) = \mu(y_t, y_{t-1}, \dots)$, then there is a departure from the martingale hypothesis, and past price information will be relevant to predict tomorrow's return. In Sect. "Nonlinear Forecasting Models for the Conditional Variance", we review models for the conditional variance of returns. For instance, a model like $y_{t+1} = \mu + u_{t+1}\sigma_{t+1}$ with time-varying conditional variance $\sigma^2_{t+1} = \mathbb{E}((y_{t+1} - \mu)^2|\mathcal{F}_t)$ and i.i.d. error u_{t+1}, is still a martingale-difference for returns but it represents a departure from the independence assumption. The conditional mean return may not be predictable but the conditional variance of the return will be. In addition, as we have seen modeling time-varying variances and covariances will be very useful for the implementation of conditional CAPM and APT models.

Nonlinear Forecasting Models for the Conditional Mean

We consider models to forecast the expected price changes of financial assets and we restrict the loss function of the forecast error to be the mean squared forecast error (MSFE). Under this loss, the optimal forecast is $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$. Other loss functions may also be used but it will be necessary to forecast other aspects of the forecast density. For example, under a mean absolute error loss function the optimal forecast is the conditional median.

There is evidence for μ_{t+h} being time-varying. Simple linear autoregressive polynomials in lagged price changes are not sufficient to model μ_{t+h} and nonlinear specifications are needed. These can be classified into parametric and nonparametric. Examples of parametric models are autoregressive bilinear and threshold models. Examples of nonparametric models are artificial neural network, kernel and nearest neighbor regression models.

It will be impossible to have an exhaustive review of the many nonlinear specifications. However, as discussed in White [161] and Chen [25], some nonlinear models are universal approximators. For example, the sieves or approximating spaces are proven to approximate very well unknown functions and they can be constructed using linear spans of power series, Fourier series, splines, or many other basis functions such as artificial neural network (ANN), Hermite polynomials as used in e. g., [56] for modelling semi-nonparametric density, and Laguerre polynomials used in [119] for modelling the yield curve. Diebold and Li [36] and Huang, Lee, and Li [89] use the Nelson–Siegel model in forecasting yields and inflation.

We review parametric nonlinear models like threshold model, smooth transition model, Markov switching model, and random fields model; nonparametric models like local linear, local polynomial, local exponential, and functional coefficient models; and nonlinear models based on sieves like ANN and various polynomials approximations. For other nonlinear specifications we recommend some books on nonlinear time series models such as Fan and Yao [52], Gao [57], and Tsay [153]. We begin with a very simple nonlinear model.

A Simple Nonlinear Model with Dummy Variables

Goyal and Welch [66] forecast the equity premium on the S&P 500 index – index return minus T-bill rate – using many predictors such as stock-related variables (e. g., dividend-yield, earning-price ratio, book-to-market ratio, corporate issuing activity, etc.), interest-rate-related variables (e. g., treasury bills, long-term yield, corporate bond returns, inflation, investment to capital ratio), and ex ante consumption, wealth, income ratio (modified from [101]). They find that these predictors have better performance in bad times, such as the Great Depression (1930–33), the oil-shock period (1973–75), and the tech bubble-crash period (1999–2001). Also, they argue that it is reasonable to impose a lower bound (e. g., zero or 2%) on the equity premium because no investor is interested in (say) a negative premium.

Campbell and Thompson [23], inspired by the out-of-sample forecasting of Goyal and Welch [66], argue that if we impose some restrictions on the signs of the predictors' coefficients and excess return forecasts, some predictors can beat the historical average equity premium. Similarly to Goyal and Welch [66], they also use a rich set of forecasting variables – valuation ratios (e. g., dividend price ratio, earning price ratio, and book to market ratio), real return on equity, nominal interest rates and inflation, and equity share of new issues and consumption-wealth ratio. They impose two restrictions – the first one is to restrict the predictors' coefficients to have the theoretically expected sign and to set wrong-signed coefficients to zero, and the second one is to rule out a negative equity premium forecast. They show that the effectiveness of these theoretically-inspired restrictions almost always improve the out-of sample performance of the predictive regressions. This is an example where "shrinkage" works, that is to reduce the forecast error variance at the cost of a higher forecast bias but with an overall smaller mean squared forecast error (the sum of error variance and the forecast squared bias).

The results from Goyal and Welch [66] and Campbell and Thompson [23] support a simple form of nonlinearity that can be generalized to threshold models or time-varying coefficient models, which we consider next.

Threshold Models

Many financial and macroeconomic time series exhibit different characteristics over time depending upon the state of the economy. For instance, we observe bull and bear stock markets, high volatility versus low volatility periods, recessions versus expansions, credit crunch versus excess liquidity, etc. If these different regimes are present in economic time series data, econometric specifications should go beyond linear models as these assume that there is only a single structure or regime over time. Nonlinear time series specifications that allow for the possibility of different regimes, also known as state-dependent models, include several types of models: threshold, smooth transition, and regime-switching models.

Threshold autoregressive (TAR) models [148,149] assume that the dynamics of the process is explained by an autoregression in each of the n regimes dictated by a conditioning or threshold variable. For a process $\{y_t\}$, a general specification of a TAR model is

$$y_t = \sum_{j=1}^{n} \left[\phi_o^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} y_{t-i} + \varepsilon_t^{(j)} \right] \mathbf{1}(r_{j-1} < x_t \leq r_j).$$

There are n regimes, in each one there is an autoregressive process of order p_j with different autoregressive parameters $\phi_i^{(j)}$, the threshold variable is x_t with r_j thresholds and $r_o = -\infty$ and $r_n = +\infty$, and the error term is assumed i.i.d. with zero mean and different variance across regimes $\varepsilon_t^{(j)} \sim$ i.i.d. $\left(0, \sigma_j^2\right)$, or more generally $\varepsilon_t^{(j)}$ is assumed to be a martingale difference. When the threshold variable is the lagged dependent variable itself y_{t-d}, the model is known as self-exciting threshold autoregressive (SETAR) model. The SETAR model has been applied to the modelling of exchange rates, industrial production indexes, and gross national product (GNP) growth, among other economic data sets. The most popular specifications within economic time series tend to find two, at most three regimes. For instance, Boero and Marrocu [18] compare a two and three-regime SETAR models with a linear AR with GARCH disturbances for the euro exchange rates. On the overall forecasting sample, the linear model performs better than the SETAR models but there is some improvement in the predictive performance of the SETAR model when conditioning on the regime.

Smooth Transition Models

In the SETAR specification, the number of regimes is discrete and finite. It is also possible to model a *continuum* of regimes as in the Smooth Transition Autoregressive (STAR) models [144]. A typical specification is

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \left(\theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i} \right) F(y_{t-d}) + \varepsilon_t$$

where $F(y_{t-d})$ is the transition function that is continuous and in most cases is either a logistic function or an exponential,

$$F(y_{t-d}) = \left[1 + \exp\left(-\gamma \left(y_{t-d} - r\right)\right)\right]^{-1}$$
$$F(y_{t-d}) = 1 - \left[\exp\left(-\gamma \left(y_{t-d} - r\right)^2\right)\right]$$

This model can be understood as many autoregressive regimes dictated by the values of the function $F(y_{t-d})$, or alternatively as an autoregression where the autoregressive parameters change smoothly over time. When $F(y_{t-d})$ is logistic and $\gamma \to \infty$, the STAR model collapses to a threshold model SETAR with two regimes. One important characteristic of these models, SETAR and STAR, is that the process can be stationary within some regimes and non-stationary within others moving between explosive and contractionary stages.

Since the estimation of these models can be demanding, the first question to solve is whether the nonlinearity is granted by the data. A test for linearity is imperative before engaging in the estimation of nonlinear specifications. An LM test that has power against the two alternatives specifications SETAR and STAR is proposed by Luukkonen et al. [110] and it consists of running two regressions: under the null hypothesis of linearity, a linear autoregression of order p is estimated in order to calculate the sum of squared residuals, SSE_0; the second is an auxiliary regression

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \sum_{j=1}^{p} \psi_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p} \sum_{j=1}^{p} \zeta_{ij} y_{t-i} y_{t-j}^2 + \sum_{i=1}^{p} \sum_{j=1}^{p} \xi_{ij} y_{t-i} y_{t-j}^3 + u_t$$

from which we calculate the sum of squared residuals, SSE_1. The test is constructed as $\chi^2 = T(SSE_0 - SSE_1)/SSE_0$ that under the null hypothesis of linearity is chi-squared distributed with $p(p+1)/2 + 2p^2$ degrees of freedom. There are other tests in the literature, for instance Hansen [80] proposes a likelihood ratio test that has a non-standard distribution, which is approximated by implementing a bootstrap procedure. Tsay [151] proposes a test based on arranged regressions with respect to the increasing order of the threshold variable and by doing this the testing problem is transformed into a change-point problem.

If linearity is rejected, we proceed with the estimation of the nonlinear specification. In the case of the SETAR model, if we fix the values of the delay parameter d and the thresholds r_j, the model reduces to n linear regressions for which least squares estimation is straightforward.

Tsay [151] proposes a conditional least squares (CLS) estimator. For simplicity of exposition suppose that there are two regimes in the data and the model to estimate is

$$y_t = \left[\phi_o^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} y_{t-i}\right] \mathbf{1}(y_{t-d} \le r) + \left[\phi_o^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} y_{t-i}\right] \mathbf{1}(y_{t-d} > r) + \varepsilon_t$$

Since r and d are fixed, we can apply least squares estimation to the model and to obtain the LS estimates for the parameters ϕ_i's. With the LS residual $\hat{\varepsilon}_t$, we obtain the total sum of squares $S(r,d) = \sum_t \hat{\varepsilon}_t^2$. The CLS estimates of r and d are obtained from $(\hat{r}, \hat{d}) = \arg\min S(r,d)$.

For the STAR model, it is also necessary to specify a priori the functional form of $F(y_{t-d})$. Teräsvirta [144] proposes a modeling cycle consisting of three stages: specification, estimation, and evaluation. In general, the specification stage consists of sequence of null hypothesis to be tested within a linearized version of the STAR model. Parameter estimation is carried out by nonlinear least squares or maximum likelihood. The evaluation stage mainly consists of testing for no error autocorrelation, no remaining nonlinearity, and parameter constancy, among other tests.

Teräsvirta and Anderson [146] find strong nonlinearity in the industrial production indexes of most of the OECD countries. The preferred model is the logistic STAR with two regimes, recessions and expansions. The dynamics in each regime are country dependent. For instance, in USA they find that the economy tends to move from recessions into expansions very aggressively but it will take a large negative shock to move rapidly from an expansion into a recession. Other references for applications of these models to financial series are found in [28,73,94].

For forecasting with STAR models, see Lundbergh and Teräsvirta [109]. It is easy to construct the one-step-ahead forecast but the multi-step-ahead forecast is a complex problem. For instance, for the 2-regime threshold model, the one-step-ahead forecast is constructed as the conditional mean of the process given some information set

$$\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) = \left[\phi_o^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} y_{t+1-i}\right] \mathbf{1}(y_{t+1-d} \le r) + \left[\phi_o^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} y_{t+1-i}\right] \mathbf{1}(y_{t+1-d} > r)$$

provided that $y_{t+1-i}, y_{t+1-d} \in \mathcal{F}_t$. However, a multi-step-ahead forecast will be a function of variables that being dated at a future date do not belong to the information set; in this case the solution requires the use of numerical integration techniques or simulation/bootstrap procedures. See Granger and Teräsvirta [72], Chapter 9, and Teräsvirta [145] for more details on numerical methods for multi-step forecasts.

Markov-Switching Models

A Markov-switching (MS) model [76,77] also features changes in regime, but in contrast with the SETAR models the change is dictated by a non-observable state variable that is modelled as a Markov chain. For instance, a first order autoregressive Markov switching model is specified as

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \varepsilon_t$$

where $s_t = 1, 2, \dots, N$ is the unobserved state variable that is modelled as an N-state Markov chain with transition probabilities $p_{ij} = P(s_t = j|s_{t-1} = i)$, and $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$ or more generally ε_t is a martingale difference. Conditioning in a given state and an information set $\mathcal{F}_t$, the process $\{y_t\}$ is linear but unconditionally the process is nonlinear. The conditional forecast is $\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t; \theta) = c_j + \phi_j y_t$ and the unconditional forecast based on observable variables is the sum of the conditional forecasts for each state weighted by the probability of being in that state,

$$\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) = \sum_{j=1}^{N} P(s_{t+1} = j|\mathcal{F}_t;\theta)\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t;\theta)\,.$$

The parameter vector $\theta = (c_1 \dots c_N, \phi_1 \dots \phi_N, \sigma^2)'$ as well as the transition probabilities p_{ij} can be estimated by maximum likelihood.

MS models have been applying to the modeling of foreign exchange rates with mixed success. Engel and Hamilton [43] fit a two-state MS for the Dollar and find that there are long swings and by that they reject the random walk behavior in the exchange rate. Marsh [114] estimates a two-state MS for the Deutschemark, the Pound Sterling, and the Japanese Yen. Though the model approximates the characteristics of the data well, the forecasting performance is poor when measured by the profit/losses generated by a set of trading rules based on the predictions of the MS model. On the contrary, Dueker and Neely [40] find that for the same exchange rate a MS model with three states variables – in the scale factor of the variance of a Student-t error, in the kurtosis of the error, and in the expected return – produces out-of-sample excess returns that are slightly superior to those generated by common trading rules. For stock returns, there is evidence that MS models perform relatively well on describing two states in the mean (high/low returns) and two states in the variance (stable/volatile periods) of returns [111]. In addition, Perez-Quiros and Timmermann [124] propose that the error term should be modelled as a mixture of Gaussian and Student-t distributions to capture the outliers commonly found in stock returns. This model provides some gains in predictive accuracy mainly for small firms returns. For interest rates in USA, Germany, and United Kingdom, Ang and Bekaert [5] find that a two-state MS model that incorporates information on international short rate and on term spread is able to predict better than an univariate MS model. Additionally they find that in USA the classification of regimes correlates well with the business cycles.

SETAR, STAR, and MS models are successful specifications to approximate the characteristics of financial and macroeconomic data. However, good in-sample performance does not imply necessarily a good out-of-sample performance, mainly when compared to simple linear ARMA models. The success of nonlinear models depends on how prominent the nonlinearity is in the data. We should not expect a nonlinear model to perform better than a linear model when the contribution of the nonlinearity to the overall specification of the model is very small. As it is argued in Granger and Teräsvirta [72], the prediction errors generated by a nonlinear model will be smaller only when the nonlinear feature modelled in-sample is also present in the forecasting sample.

A State Dependent Mixture Model Based on Cross-sectional Ranks

In the previous section, we have dealt with nonlinear time series models that only incorporate time series information. González-Rivera, Lee, and Mishra [63] propose a nonlinear model that combines time series with cross sectional information. They propose the modelling of expected returns based on the joint dynamics of a sharp jump in the cross-sectional rank and the realized returns. They analyze the marginal probability distribution of a jump in the cross-sectional rank within the context of a duration model, and the probability of the asset return conditional on a jump specifying different dynamics depending on whether or not a jump has taken place. The resulting model for expected returns is a mixture of normal distributions weighted by the probability of jumping.

Let $y_{i,t}$ be the return of firm i at time t, and $\{y_{i,t}\}_{i=1}^{M}$ be the collection of asset returns of the M firms that constitute the *market* at time t. For each time t, the asset returns are ordered from the smallest to the largest, and define $z_{i,t}$, the *Varying Cross-sectional Rank* (VCR) of firm i within the market, as the proportion of firms that have a return less than or equal to the return of firm i. We write

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t}) \,, \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and for M large, $z_{i,t} \in (0, 1]$. Since the rank is a highly dependent variable, it is assumed that small movements in the asset ranking will not contain significant information and that most likely large movements in ranking will be the result of news in the overall market and/or of news concerning a particular asset. Focusing on large rank movements, we define, at time t, a sharp jump as a binary variable that takes the value one when there is a minimum (upward or downward) movement of 0.5 in the ranking of asset i, and zero otherwise:

$$J_{i,t} \equiv \mathbf{1}(|z_{i,t} - z_{i,t-1}| \geq 0.5) \,. \tag{2}$$

A jump of this magnitude brings the asset return above or below the median of the cross-sectional distribution of returns. Note that this notion of jumps differs from the more traditional meaning of the word in the context of continuous-time modelling of the univariate return process. A jump in the cross-sectional rank implicitly depends on numerous univariate return processes.

The analytical problem now consists in modeling the joint distribution of the return $y_{i,t}$ and the jump $J_{i,t}$, i. e. $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1})$ where $\mathcal{F}_{t-1}$ is the information set up to time $t-1$. Since $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1}) = f_1(J_{i,t}|\mathcal{F}_{t-1}) f_2(y_{i,t}|J_{i,t}, \mathcal{F}_{t-1})$, the analysis focuses first on the modelling of the marginal distribution of the jump, and subsequently on the modelling of the conditional distribution of the return.

Since $J_{i,t}$ is a Bernoulli variable, the marginal distribution of the jump is $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1 - p_{i,t})^{(1-J_{i,t})}$ where $p_{i,t} \equiv \Pr(J_{i,t} = 1|\mathcal{F}_{t-1})$ is the conditional probability of a jump in the cross-sectional ranks. The modelling of $p_{i,t}$ is performed within the context of a dynamic duration model specified in calendar time as in Hamilton and Jordà [79]. The calendar time approach is necessary because asset returns are reported in calendar time (days, weeks, etc.) and it has the advantage of incorporating any other available information also reported in calendar time.

It is easy to see that the probability of jumping and duration must have an inverse relationship. If the probability of jumping is high, the expected duration must be short, and vice versa. Let $\Psi_{N(t)}$ be the expected duration. The expected duration until the next jump in the cross-sectional rank is given by $\Psi_{N(t)} = \sum_{j=1}^{\infty} j(1-p_t)^{j-1} p_t = p_t^{-1}$. Note that $\sum_{j=0}^{\infty}(1-p_t)^j = p_t^{-1}$. Differentiating with respect to p_t yields $\sum_{j=0}^{\infty} -j(1-p_t)^{j-1} = -p_t^{-2}$. Multiplying by $-p_t$ gives $\sum_{j=0}^{\infty} j(1-p_t)^{j-1} p_t = p_t^{-1}$ and thus $\sum_{j=1}^{\infty} j(1-p_t)^{j-1} p_t = p_t^{-1}$. Consequently, to model $p_{i,t}$, it suffices to model the expected duration and compute its inverse. Following Hamilton and Jordà [79], an autoregressive conditional hazard (ACH) model is specified. The ACH model is a calendar-time version of the autoregressive conditional duration (ACD) of Engle and Russell [49]. In both ACD and ACH models, the expected duration is a linear function of lag durations. However as the ACD model is set up in event time, there are some difficulties on how to introduce information that arrives between events. This is not the case in the ACH model because the set-up is in calendar time. In the ACD model, the forecasting object is the expected time between events; in the ACH model, the objective is to forecast the probability that the event will happen tomorrow given the information known up to today. A general ACH model is specified as

$$\Psi_{N(t)} = \sum_{j=1}^{m} \alpha_j D_{N(t)-j} + \sum_{j=1}^{r} \beta_j \Psi_{N(t)-j} \,. \tag{3}$$

Since p_t is a probability, it must be bounded between zero and one. This implies that the conditional duration must have a lower bound of one. Furthermore, working in calendar time it is possible to incorporate information that becomes available between jumps and can affect the probability of a jump in future periods. The conditional hazard rate is specified as

$$p_t = [\Psi_{N(t-1)} + \delta' X_{t-1}]^{-1} \,, \tag{4}$$

where X_{t-1} is a vector of relevant calendar time variables such as past VCRs and past returns. This completes the marginal distribution of the jump $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1-p_{i,t})^{(1-J_{i,t})}$.

On modelling $f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$, it is assumed that the return to asset i may behave differently depending upon the occurrence of a jump. The modelling of two potential different states (whether a jump has occurred or not) will permit to differentiate whether the conditional expected return is driven by active or/and passive movements in the asset ranking in conjunction with its own return dynamics. A priori, different dynamics are possible in these two states. A general specification is

$$f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) = \begin{cases} N(\mu_{1,t}, \sigma_{1,t}^2) & \text{if } J_t = 1 \\ N(\mu_{0,t}, \sigma_{0,t}^2) & \text{if } J_t = 0 \end{cases} , \tag{5}$$

where $\mu_{j,t}$ is the conditional mean and $\sigma_{j,t}^2$ the conditional variance in each state ($j = 1, 0$). Whether these two states are present in the data is an empirical question and it should be answered through statistical testing.

Combining the models for the marginal density of the jump and the conditional density of the returns, the estimation can be conducted with maximum likelihood techniques. For a sample $\{y_t, J_t\}_{t=1}^{T}$, the joint log-likelihood function is

$$\begin{aligned} &\sum_{t=1}^{T} \ln f(y_t, J_t|\mathcal{F}_{t-1}; \theta) \\ &= \sum_{t=1}^{T} \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1) + \sum_{t=1}^{T} \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) \,. \end{aligned}$$

Let us call $\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1)$ and $\mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$. The maximization of the joint log-likelihood function can be achieved by maximizing $\mathcal{L}_1(\theta_1)$ and $\mathcal{L}_2(\theta_2)$ separately without loss of efficiency by assuming that the parameter vectors θ_1 and θ_2 are "variation free" in the sense of Engle et al. [45].

The log-likelihood function $\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \ln f_1(J_t| \mathcal{F}_{t-1}; \theta_1)$ is

$$\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \left[J_t \ln p_t(\theta_1) + (1-J_t) \ln(1-p_t(\theta_1)) \right] , \tag{6}$$

where θ_1 includes all parameters in the conditional duration model.

The log-likelihood function $\mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln f_2(y_t| J_t, \mathcal{F}_{t-1}; \theta_2)$ is

$$\begin{aligned} \mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln \Bigg[&\frac{J_t}{\sqrt{2\pi\sigma_{1,t}^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_t - \mu_{1,t}}{\sigma_{1,t}} \right)^2 \right\} \\ &+ \frac{1-J_t}{\sqrt{2\pi\sigma_{0,t}^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_t - \mu_{0,t}}{\sigma_{0,t}} \right)^2 \right\} \Bigg] , \end{aligned}$$

where θ_2 includes all parameters in the conditional means and conditional variances under both regimes.

If the two proposed states are granted in the data, the marginal density function of the asset return must be a mixture of two normal density functions where the mixture weights are given by the probability of jumping p_t:

$$\begin{aligned} g(y_t|\mathcal{F}_{t-1};\theta) &\equiv \sum_{J_t=0}^{1} f(y_t, J_t|\mathcal{F}_{t-1};\theta) \\ &= \sum_{J_t=0}^{1} f_1(J_t|\mathcal{F}_{t-1};\theta_1) f_2(y_t|J_t,\mathcal{F}_{t-1};\theta_2) \\ &= p_t \cdot f_2(y_t|J_t=1,\mathcal{F}_{t-1};\theta_2) \\ &\quad + (1-p_t)\cdot f_2(y_t|J_t=0,\mathcal{F}_{t-1};\theta_2), \end{aligned} \tag{7}$$

as $f_1(J_t|\mathcal{F}_{t-1};\theta_1) = p_t^{J_t}(1-p_t)^{(1-J_t)}$. Therefore, the one-step ahead forecast of the return is

$$\begin{aligned} &\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) \\ &= \int y_{t+1}\cdot g(y_{t+1}|\mathcal{F}_t;\theta)\mathrm{d}y_{t+1} \\ &= p_{t+1}(\theta_1)\cdot \mu_{1,t+1}(\theta_2) + (1-p_{t+1}(\theta_1))\cdot \mu_{0,t+1}(\theta_2)\,. \end{aligned} \tag{8}$$

The expected return is a function of the probability of jumping p_t, which is a nonlinear function of the information set as shown in (4). Hence the expected returns are nonlinear functions of the information set, even in a simple case where $\mu_{1,t}$ and $\mu_{0,t}$ are linear.

This model was estimated for the returns of the constituents of the SP500 index from 1990 to 2000, and its performance was assessed in an out-of-sample exercise from 2001 to 2005 within the context of several trading strategies. Based on the one-step-ahead forecast of the mixture model, a proposed trading strategy called VCR-Mixture Trading Rule is shown to be a superior rule because of its ability to generate large risk-adjusted mean returns when compared to other technical and model-based trading rules. The VCR-Mixture Trading Rule is implemented by computing for each firm in the SP500 index the one-step ahead forecast of the return as in (8). Based on the forecasted returns $\{\hat{y}_{i,t+1}(\hat{\theta}_t)\}_{t=R}^{T-1}$, the investor predicts the VCR of all assets in relation to the overall market, that is,

$$\hat{z}_{i,t+1} = M^{-1}\sum_{j=1}^{M}\mathbf{1}(\hat{y}_{j,t+1} \le \hat{y}_{i,t+1}), \qquad t = R,\ldots,T-1\,, \tag{9}$$

and buys the top K performing assets if their forecasted return is above the risk-free rate. In every subsequent out-of-sample period ($t = R,\ldots,T-1$), the investor revises her portfolio, selling the assets that fall out of the top performers and buying the ones that rise to the top, and she computes the one-period portfolio return

$$\pi_{t+1} = K^{-1}\sum_{j=1}^{M} y_{j,t+1}\cdot \mathbf{1}\left(\hat{z}_{j,t+1} \ge z_{t+1}^{K}\right), \qquad t = R,\ldots,T-1\,, \tag{10}$$

where z_{t+1}^K is the cutoff cross-sectional rank to select the K best performing stocks such that $\sum_{j=1}^{M}\mathbf{1}\left(\hat{z}_{j,t+1} \ge z_{t+1}^{K}\right) = K$. In the analysis of González-Rivera, Lee, and Mishra [63] a portfolio is formed with the top 1% ($K = 5$ stocks) performers in the SP500 index. Every asset in the portfolio is weighted equally. The evaluation criterion is to compute the "mean trading return" over the forecasting period

$$MTR = P^{-1}\sum_{t=R}^{T-1}\pi_{t+1}\,.$$

It is also possible to correct MTR according to the level of risk of the chosen portfolio. For instance, the traditional Sharpe ratio will provide the excess return per unit of risk measured by the standard deviation of the selected portfolio

$$SR = P^{-1}\sum_{t=R}^{T-1}\frac{(\pi_{t+1} - r_{f,t+1})}{\sigma_{t+1}^{\pi}(\hat{\theta}_t)}\,,$$

where $r_{f,t+1}$ is the risk free rate. The VCR-Mixture Trading Rule produces a weekly MTR of 0.243% (63.295% cumulative return over 260 weeks), equivalent to a yearly compounded return of 13.45%, that is significantly more than the next most favorable rule, which is the Buy-and-Hold-the-Market Trading Rule with a weekly mean return of −0.019%, equivalent to a yearly return of −1.00%. To assess the return-risk trade off, we implement the Sharpe ratio. The largest SR (mean return per unit of standard deviation) is provided by the VCR-Mixture rule with a weekly return of 0.151% (8.11% yearly compounded return per unit of standard deviation), which is lower than the mean return provided by the same rule under the MTR criterion, but still a dominant return when compared to the mean returns provided by the Buy-and-Hold-the-Market Trading Rule.

Random Fields

Hamilton [78] proposed a flexible parametric regression model where the conditional mean has a linear parametric component and a potential nonlinear component represented by an isotropic Gaussian random field. The model has a nonparametric flavor because no functional form is assumed but, nevertheless, the estimation is fully parametric.

A scalar random field is defined as a function $m(\omega, x) : \Omega \times A \rightarrow R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$. A random field is also denoted as $m(x)$. If $m(x)$ is a system of random variables with finite dimensional Gaussian distributions, then the scalar random field is said to be Gaussian and it is completely determined by its mean function $\mu(x) = \mathbb{E}\left[m(x)\right]$ and its covariance function with typical element $C(x,z) = \mathbb{E}\left[(m(x) - \mu(x))(m(z) - \mu(z))\right]$ for any $x, z \in A$. The random field is said to be homogeneous or stationary if $\mu(x) = \mu$ and the covariance function depends only on the difference vector $x - z$ and we should write $C(x,z) = C(x - z)$. Furthermore, the random field is said to be isotropic if the covariance function depends on $d(x,z)$, where $d(\cdot)$ is a scalar measure of distance. In this situation we write $C(x,z) = C(d(x,z))$.

The specification suggested by Hamilton [78] can be represented as

$$y_t = \beta_0 + x_t'\beta_1 + \lambda m(g \odot x_t) + \epsilon_t \,, \tag{11}$$

for $y_t \in R$ and $x_t \in R^k$, both stationary and ergodic processes. The conditional mean has a linear component given by $\beta_0 + x_t'\beta_1$ and a nonlinear component given by $\lambda m(g \odot x_t)$, where $m(z)$, for any choice of z, represents a realization of a Gaussian and homogenous random field with a moving average representation; x_t could be predetermined or exogenous and is independent of $m(\cdot)$, and ϵ_t is a sequence of independent and identically distributed $N(0, \sigma^2)$ variates independent of both $m(\cdot)$ and x_t as well as of lagged values of x_t. The scalar parameter λ represents the contribution of the nonlinear part to the conditional mean, the vector $g \in R^k_{0,+}$ drives the curvature of the conditional mean, and the symbol $\odot$ denotes element-by-element multiplication.

Let H_k be the covariance (correlation) function of the random field $m(\cdot)$ with typical element defined as $H_k(x,z) = \mathbb{E}\left[m(x)m(z)\right]$. Hamilton [78] proved that the covariance function depends solely upon the Euclidean distance between x and z, rendering the random field isotropic. For any x and $z \in R^k$, the correlation between $m(x)$ and $m(z)$ is given by the ratio of the volume of the overlap of k-dimensional unit spheroids centered at x and z to the volume of a single k-dimensional unit spheroid. If the Euclidean distance between x and z is greater than two, the correlation between $m(x)$ and $m(z)$ will be equal to zero. The general expression of the correlation function is

$$H_k(h) = \begin{cases} G_{k-1}(h,1)/G_{k-1}(0,1) & \text{if } h \leq 1 \\ 0 & \text{if } h > 1 \end{cases} , \tag{12}$$

$$G_k(h,r) = \int_h^r (r^2 - w^2)^{k/2} \mathrm{d}w \,,$$

where $h \equiv \frac{1}{2} d_{L_2}(x,z)$, and $d_{L_2}(x,z) \equiv \left[(x-z)'(x-z)\right]^{1/2}$ is the Euclidean distance between x and z.

Within the specification (11), Dahl and González-Rivera [33] provided alternative representations of the random field that permit the construction of Lagrange multiplier tests for neglected nonlinearity, which circumvent the problem of unidentified nuisance parameters under the null of linearity and, at the same time, they are robust to the specification of the covariance function associated with the random field. They modified the Hamilton framework in two directions. First, the random field is specified in the L_1 norm instead of the L_2 norm, and secondly they considered random fields that may not have a simple moving average representation. The advantage of the L_1 norm, which is exploited in the testing problem, is that this distance measure is a linear function of the nuisance parameters, in contrast to the L_2 norm which is a nonlinear function. Logically, Dahl and González-Rivera proceeded in an opposite fashion to Hamilton. Whereas Hamilton first proposed a moving average representation of the random field, and secondly, he derived its corresponding covariance function, Dahl and González-Rivera first proposed a covariance function, and secondly they inquire whether there is a random field associated with it. The proposed covariance function is

$$C_k(h^*) = \begin{cases} (1 - h^*)^{2k} & \text{if } h^* \leq 1 \\ 0 & \text{if } h^* > 1 \end{cases} , \tag{13}$$

where $h^* \equiv \frac{1}{2} d_{L_1}(x,z) = \frac{1}{2}|x - z|'1$. The function (13) is a permissible covariance, that is, it satisfies the positive semidefiniteness condition, which is $q'C_k q \geq 0$ for all $q \neq 0_T$. Furthermore, there is a random field associated with it according to the Khinchin's theorem (1934) and Bochner's theorem (1959). The basic argument is that the class of functions which are covariance functions of homogenous random fields coincides with the class of positive semidefinite functions. Hence, (13) being a positive semidefinite function must be the covariance function of a homogenous random field.

The estimation of these models is carried out by maximum likelihood. From model (11), we can write $y \sim N(X\beta, \lambda^2 C_k + \sigma^2 I_T)$ where $y = (y_1, y_2, \ldots, y_T)'$, $X_1 = (x_1', x_2', \ldots, x_T')'$, $X = (\mathbf{1} : X_1)$, $\beta = (\beta_0, \beta_1')'$, $\epsilon =$

$(\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$ and σ^2 is the variance of ϵ_t. C_k is a generic covariance function associated with the random field, which could be equal to the Hamilton spherical covariance function in (12), or to the covariance in (13). The log-likelihood function corresponding to this model is

$$\ell(\beta, \lambda^2, g, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\lambda^2 C_k + \sigma^2 I_T| \\ -\frac{1}{2}(y - X\beta)'(\lambda^2 C_k + \sigma^2 I_T)^{-1}(y - X\beta). \tag{14}$$

The flexible regression model has been applied successfully to detect nonlinearity in the quarterly growth rate of the US real GNP [34] and in the Industrial Production Index of sixteen OECD countries [33]. This technology is able to mimic the characteristics of the actual US business cycle. The cycle is dissected according to measures of duration, amplitude, cumulation and excess cumulation of the contraction and expansion phases. In contrast to Harding and Pagan [82] who find that nonlinear models are not uniformly superior to linear ones, the flexible regression model represents a clear improvement over linear models, and it seems to capture just the right shape of the expansion phase as opposed to Hamilton [76] and Durland and McCurdy [41] models, which tend to overestimate the cumulation measure in the expansion phase. It is found that the expansion phase must have at least two subphases: an aggressive early expansion after the trough, and a moderate/slow late expansion before the peak implying the existence of an inflexion point that we date approximately around one-third into the duration of the expansion phase. This shape lends support to parametric models of the growth rate that allow for three regimes [136], as opposed to models with just two regimes (contractions and expansions). For the Industrial Production Index, testing for nonlinearity within the flexible regression framework brings similar conclusions to those in Teräsvirta and Anderson [146], who propose parametric STAR models for industrial production data. However, the tests proposed in Dahl and González-Rivera [33], which have superior performance to detect smooth transition dynamics, seem to indicate that linearity cannot be rejected in the industrial production indexes of Japan, Austria, Belgium and Sweden as opposed to the findings of Teräsvirta and Anderson.

Nonlinear Factor Models

For the last ten years forecasting using a data-rich environment has been one of the most researched topic in economics and finance, see [140,141]. In this literature, factor models are used to reduce the dimension of the data but mostly they are linear models. Bai and Ng (BN) [7] introduce a nonlinear factor model with a quadratic principal component model as a special case. First consider a simple factor model

$$x_{it} = \lambda_i' F_t + e_{it}\,. \tag{15}$$

By the method of principal component, the elements of $\mathbf{f}_t$ are linear combinations of elements of $\mathbf{x}_t$. The factors are estimated by minimizing the sum of squared residuals of the linear model, $x_{it} = \lambda_i F_t + e_{it}$.

The factor model in (15) assumes a linear link function between the predictor $\mathbf{x}_t$ and the latent factors F_t. BN consider a more flexible approach by a nonlinear link function $g(\cdot)$ such that

$$g(x_{it}) = \phi_i' J_t + v_{it}\,,$$

where J_t are the common factors, and ϕ_i is the vector of factor loadings. BN consider $g(x_{it})$ to be x_{it} augmented by some or all of the unique cross-products of the elements of $\{x_{it}\}_{i=1}^N$. The second-order factor model is then $x_{it}^* = \phi_i' J_t + v_{it}$ where x_{it}^* is an $N^* \times 1$ vector. Estimation of J_t then proceeds by the usual method of principal components. BN consider $x_{it}^* = \{x_{it}\, x_{it}^2\}_{i=1}^N$ with $N^* = 2N$, which they call the SPC (squared principal components).

Once the factors are estimated, the forecasting equation for y_{t+h} would be

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + \varepsilon_t\,.$$

The forecasting equation remains linear whatever the link function g is. An alternative way of capturing nonlinearity is to augment the forecasting equation to include functions of the factors

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + a(\hat{F}_t) + \varepsilon_t\,,$$

where $a(\cdot)$ is nonlinear. A simple case when $a(\cdot)$ is quadratic is referred to as PC2 (squared factors) in BN.

BN note that the PC2 is conceptually distinct from SPC. While the PC2 forecasting model allows the volatility of factors estimated by linear principal components to have predictive power for y, the SPC model allows the factors to be possibly nonlinear functions of the predictors while maintaining a linear relation between the factors and y. Ludvigson and Ng [108] found that the square of the first factor estimated from a set of financial factors (i. e., volatility of the first factor) is significant in the regression model for the mean excess returns. In contrast, factors estimated from the second moment of data (i. e., volatility factors) are much weaker predictors of excess returns.

Artificial Neural Network Models

Consider an augmented single hidden layer feedforward neural network model $f(x_t, \theta)$ in which the network output y_t is determined given input x_t as

$$\begin{aligned} y_t &= f(x_t, \theta) + \varepsilon_t \\ &= x_t\beta + \sum_{j=1}^{q} \delta_j \psi(x_t\gamma_j) + \varepsilon_t \end{aligned}$$

where $\theta = (\beta'\gamma'\delta')'$, β is a conformable column vector of connection strength from the input layer to the output layer; γ_j is a conformable column vector of connection strength from the input layer to the hidden units, $j = 1, \ldots, q$; δ_j is a (scalar) connection strength from the hidden unit j to the output unit, $j = 1, \ldots, q$; and ψ is a squashing function (e. g., the logistic squasher) or a radial basis function. Input units x send signals to intermediate hidden units, then each of hidden unit produces an activation ψ that then sends signals toward the output unit. The integer q denotes the number of hidden units added to the affine (linear) network. When $q = 0$, we have a two layer *affine* network $y_t = x_t\beta + \varepsilon_t$. Hornick, Stinchcombe and White [88] show that neural network is a nonlinear flexible functional form being capable of approximating any Borel measurable function to any desired level of accuracy provided sufficiently many hidden units are available. Stinchcombe and White [138] show that this result holds for any $\psi(\cdot)$ belonging to the class of "generically comprehensively revealing" functions. These functions are "comprehensively revealing" in the sense that they can reveal arbitrary model misspecifications $\mathbb{E}(y_t|x_t) \neq f(x_t, \theta^*)$ with non-zero probability and they are "generic" in the sense that almost any choice for γ will reveal the misspecification.

We build an artificial neural network (ANN) model based on a test for neglected nonlinearity likely to have power against a range of alternatives. See White [158] and Lee, White, and Granger [99] on the neural network test and its comparison with other specification tests. The neural network test is based on a test function $h(x_t)$ chosen as the activations of 'phantom' hidden units $\psi(x_t\Gamma_j)$, $j = 1, \ldots, q$, where Γ_j are random column vectors independent of x_t. That is,

$$\mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*|\Gamma_j] = \mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*] = 0 \quad j = 1, \ldots, q, \tag{16}$$

under H_0, so that

$$\mathbb{E}(\Psi_t\varepsilon_t^*) = 0\,, \tag{17}$$

where $\Psi_t = (\psi(x_t\Gamma_1), \ldots, \psi(x_t\Gamma_q))'$ is a phantom hidden unit activation vector. Evidence of correlation of ε_t^* with Ψ_t is evidence against the null hypothesis that y_t is linear in mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations $\psi(x_t\Gamma_j)$ would permit an improvement in network performance. Thus the tests are based on sample correlation of affine network errors with phantom hidden unit activations,

$$n^{-1}\sum_{t=1}^{n} \Psi_t\hat{\varepsilon}_t = n^{-1}\sum_{t=1}^{n} \Psi_t(y_t - x_t\hat{\beta})\,. \tag{18}$$

Under suitable regularity conditions it follows from the central limit theorem that $n^{-1/2}\sum_{t=1}^{n} \Psi_t\hat{\varepsilon}_t \xrightarrow{d} N(0, W^*)$ as $n \to \infty$, and if one has a consistent estimator for its asymptotic covariance matrix, say $\hat{W}_n$, then an asymptotic chi-square statistic can be formed as

$$\left(n^{-1/2}\sum_{t=1}^{n} \Psi_t\hat{\varepsilon}_t\right)' \hat{W}_n^{-1} \left(n^{-1/2}\sum_{t=1}^{n} \Psi_t\hat{\varepsilon}_t\right) \xrightarrow{d} \chi^2(q)\,. \tag{19}$$

Elements of Ψ_t tend to be collinear with X_t and with themselves. Thus LWG conduct a test on $q^* < q$ principal components of Ψ_t not collinear with x_t, denoted Ψ_t^*. This test is to determine whether or not there exists some advantage to be gained by adding hidden units to the affine network. We can estimate $\hat{W}_n$ robust to the conditional heteroskedasticity, or we may use with the empirical null distribution of the statistic computed by a bootstrap procedure that is robust to the conditional heteroskedasticity, e. g., wild bootstrap.

Estimation of an ANN model may be tedious and sometimes results in unreliable estimates. Recently, White [161] proposes a simple algorithm called QuickNet, a form of "relaxed greedy algorithm" because QuickNet searches for a single best additional hidden unit based on a sequence of OLS regressions, that may be analogous to the least angular regressions (LARS) of Efron, Hastie, Johnstone, and Tibshirani [42]. The simplicity of the QuickNet algorithm achieves the benefits of using a forecasting model that is nonlinear in the predictors while mitigating the other computational challenges to the use of nonlinear forecasting methods. See White [161], Section 5, for more details on QuickNet, and for other issues of controlling for overfit and the selection of the random parameter vectors Γ_j independent of x_t.

Campbell, Lo, and MacKinlay [22], Section 12.4, provide a review of these models. White [161] reviews the research frontier in ANN models. Trippi and Turban [150] review the applications of ANNs to finance and investment.

Functional Coefficient Models

A functional coefficient model is introduced by Cai, Fan, and Yao [24] (CFY), with time-varying and state-dependent coefficients. It can be viewed as a special case of Priestley's [127] state-dependent model, but it includes the models of Tong [149], Chen and Tsay [26] and regime-switching models as special cases. Let $\{(y_t, s_t)'\}_{t=1}^n$ be a stationary process, where y_t and s_t are scalar variables. Also let $X_t \equiv (1, y_{t-1}, \ldots, y_{t-d})'$. We assume

$$\mathbb{E}(y_t|\mathcal{F}_{t-1}) = a_0(s_t) + \sum_{j=1}^{d} a_j(s_t) y_{t-j} ,$$

where the $\{a_j(s_t)\}$ are the autoregressive coefficients depending on s_t, which may be chosen as a function of X_t or something else. Intuitively, the functional coefficient model is an AR process with time-varying autoregressive coefficients. The coefficient functions $\{a_j(s_t)\}$ can be estimated by local linear regression. At each point s, we approximate $a_j(s_t)$ locally by a linear function $a_j(s_t) \approx a_j + b_j(s_t - s)$, $j = 0, 1, \ldots, d$, for s_t near s, where a_j and b_j are constants. The local linear estimator at point s is then given by $\hat{a}_j(s) = \hat{a}_j$, where $\{(\hat{a}_j, \hat{b}_j)\}_{j=0}^d$ minimizes the sum of local weighted squares $\sum_{t=1}^n [y_t - \mathbb{E}(y_t|\mathcal{F}_{t-1})]^2 K_h(s_t - s)$, with $K_h(\cdot) \equiv K(\cdot/h)/h$ for a given kernel function $K(\cdot)$ and bandwidth $h \equiv h_n \to 0$ as $n \to \infty$. CFY [24], p. 944, suggest to select h using a modified multi-fold "leave-one-out-type" cross-validation based on MSFE.

It is important to choose an appropriate smooth variable s_t. Knowledge on data or economic theory may be helpful. When no prior information is available, s_t may be chosen as a function of explanatory vector X_t or using such data-driven methods as AIC and cross-validation. See Fan, Yao and Cai [52] for further discussion on the choice of s_t. For exchange rate changes, Hong and Lee [85] choose s_t as the difference between the exchange rate at time $t-1$ and the moving average of the most recent L periods of exchange rates at time $t-1$. The moving average is a proxy for the local trend at time $t-1$. Intuitively, this choice of s_t is expected to reveal useful information on the direction of changes.

To justify the use of the functional coefficient model, CFY [24] suggest a goodness-of-fit test for an AR(d) model against a functional coefficient model. The null hypothesis of AR(d) can be stated as

$$\mathbb{H}_0 : a_j(s_t) = \beta_j, \quad j = 0, 1, \ldots, d ,$$

where β_j is the autoregressive coefficient in AR(d). Under $\mathbb{H}_0$, $\{y_t\}$ is linear in mean conditional on X_t. Under the alternative to $\mathbb{H}_0$, the autoregressive coefficients depend on s_t and the AR(d) model suffers from "neglected nonlinearity". To test $\mathbb{H}_0$, CFY compares the residual sum of squares (RSS) under $\mathbb{H}_0$

$$RSS_0 \equiv \sum_{t=1}^{n} \hat{\varepsilon}_t^2 = \sum_{t=1}^{n} \left[Y_t - \hat{\beta}_0 - \sum_{j=1}^{d} \hat{\beta}_j Y_{t-j} \right]^2$$

with the RSS under the alternative

$$RSS_1 \equiv \sum_{t=1}^{n} \tilde{\varepsilon}_t^2 = \sum_{t=1}^{n} \left[Y_t - \hat{a}_0(s_t) - \sum_{j=1}^{d} \hat{a}_j(s_t) Y_{t-j} \right]^2 .$$

The test statistic is $T_n = (RSS_0 - RSS_1)/RSS_1$. We reject $\mathbb{H}_0$ for large values of T_n. CFY suggest the following bootstrap method to obtain the p-value of T_n: (i) generate the bootstrap residuals $\{\varepsilon_t^b\}_{t=1}^n$ from the centered residuals $\tilde{\varepsilon}_t - \bar{\varepsilon}$ where $\bar{\varepsilon} \equiv n^{-1} \sum_{t=1}^n \tilde{\varepsilon}_t$ and define $y_t^b \equiv X_t' \hat{\beta} + \varepsilon_t^b$, where $\hat{\beta}$ is the OLS estimator for AR(d); (ii) calculate the bootstrap statistic T_n^b using the bootstrap sample $\{y_t^b, X_t', s_t\}_{t=1}^n$; (iii) repeat steps (i) and (ii) B times ($b = 1, \ldots, B$) and approximate the bootstrap p-value of T_n by $B^{-1} \sum_{b=1}^B \mathbf{1}(T_n^b \geq T_n)$. See Hong and Lee [85] for empirical application of the functional coefficient model to forecasting foreign exchange rates.

Nonparametric Regression

Let $\{y_t, x_t\}$, $t = 1, \ldots, n$, be stochastic processes, where y_t is a scalar and $x_t = (x_{t1}, \ldots, x_{tk})$ is a $1 \times k$ vector which may contain the lagged values of y_t. Consider the regression model

$$y_t = m(x_t) + u_t$$

where $m(x_t) = \mathbb{E}\left(y_t|x_t\right)$ is the true but unknown regression function and u_t is the error term such that $\mathbb{E}(u_t|x_t) = 0$.

If $m(x_t) = g(x_t, \delta)$ is a correctly specified family of parametric regression functions then $y_t = g(x_t, \delta) + u_t$ is a correct model and, in this case, one can construct a consistent least squares (LS) estimator of $m(x_t)$ given by $g(x_t, \hat{\delta})$, where $\hat{\delta}$ is the LS estimator of the parameter δ.

In general, if the parametric regression $g(x_t, \delta)$ is incorrect or the form of $m(x_t)$ is unknown then $g(x_t, \hat{\delta})$ may not be a consistent estimator of $m(x_t)$. For this case, an alternative approach to estimate the unknown $m(x_t)$ is to use the consistent nonparametric kernel regression estimator which is essentially a local constant LS (LCLS) estimator. To obtain this estimator take a Taylor series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= m(x) + e_t \end{aligned}$$

where $e_t = (x_t - x)m^{(1)}(x) + \frac{1}{2}(x_t - x)^2 m^{(2)}(x) + \cdots + u_t$ and $m^{(s)}(x)$ represents the sth derivative of $m(x)$ at $x_t = x$. The LCLS estimator can then be derived by minimizing

$$\sum_{t=1}^{n} e_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - m(x))^2 K_{tx}$$

with respect to constant $m(x)$, where $K_{tx} = K\left(\frac{x_t - x}{h}\right)$ is a decreasing function of the distances of the regressor vector x_t from the point $x = (x_1, \dots, x_k)$, and $h \to 0$ as $n \to \infty$ is the window width (smoothing parameter) which determines how rapidly the weights decrease as the distance of x_t from x increases. The LCLS estimator so estimated is

$$\hat{m}(x) = \frac{\sum_{t=1}^{n} y_t K_{tx}}{\sum_{t=1}^{n} K_{tx}} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\,\mathbf{i}'\mathbf{K}(x)\mathbf{y}$$

where $\mathbf{K}(x)$ is the $n \times n$ diagonal matrix with the diagonal elements K_{tx} $(t = 1, \dots, n)$, $\mathbf{i}$ is an $n \times 1$ column vector of unit elements, and $\mathbf{y}$ is an $n \times 1$ vector with elements y_t $(t = 1, \dots, n)$. The estimator $\hat{m}(x)$ is due to Nadaraya [118] and Watson [155] (NW) who derived this in an alternative way. Generally $\hat{m}(x)$ is calculated at the data points x_t, in which case we can write the leave-one out estimator as

$$\hat{m}(x) = \frac{\sum_{t'=1, t' \neq t}^{n} y_{t'} K_{t't}}{\sum_{t'=1, t' \neq t}^{n} K_{t't}}\,,$$

where $K_{t't} = K\frac{x_{t'} - x_t}{h}$. The assumption that $h \to 0$ as $n \to \infty$ gives $x_t - x = O(h) \to 0$ and hence $\mathbb{E}e_t \to 0$ as $n \to \infty$. Thus the estimator $\hat{m}(x)$ will be consistent under certain smoothing conditions on h, K, and $m(x)$. In small samples however $\mathbb{E}e_t \neq 0$ so $\hat{m}(x)$ will be a biased estimator, see [122] for details on asymptotic and small sample properties.

An estimator which has a better small sample bias and hence the mean square error (MSE) behavior is the local linear LS (LLLS) estimator. In the LLLS estimator we take a first order Taylor-Series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t = m(x) + (x_t - x)m^{(1)}(x) + v_t \\ &= \alpha(x) + x_t\beta(x) + v_t \\ &= X_t\delta(x) + v_t \end{aligned}$$

where $X_t = (1\ x_t)$ and $\delta(x) = [\alpha(x)\ \beta(x)']'$ with $\alpha(x) = m(x) - x\beta(x)$ and $\beta(x) = m^{(1)}(x)$. The LLLS estimator of $\delta(x)$ is then obtained by minimizing

$$\sum_{t=1}^{n} v_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - X_t\delta(x))^2 K_{tx}$$

sand it is given by

$$\tilde{\delta}(x) = (\mathbf{X}'\mathbf{K}(x)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(x)\mathbf{y}\,. \tag{20}$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix with the tth row X_t $(t = 1, \dots, n)$.

The LLLS estimator of $\alpha(x)$ and $\beta(x)$ can be calculated as $\tilde{\alpha}(x) = (1\ 0)\tilde{\delta}(x)$ and $\tilde{\beta}(x) = (0\ 1)\tilde{\delta}(x)$. This gives

$$\tilde{m}(x) = (1\ x)\tilde{\delta}(x) = \tilde{\alpha}(x) + x\tilde{\beta}(x)\,.$$

Obviously when $X = \mathbf{i}$, $\tilde{\delta}(x)$ reduces to the NW's LCLS estimator $\hat{m}(x)$. An extension of the LLLS is the local polynomial LS (LPLS) estimators, see [50].

In fact one can obtain the local estimators of a general nonlinear model $g(x_t, \delta)$ by minimizing

$$\sum_{t=1}^{n} [y_t - g(x_t, \delta(x))]^2 K_{tx}$$

with respect to $\delta(x)$. For $g(x_t, \delta(x)) = X_t\delta(x)$ we get the LLLS in (20). Further when $h = \infty$, $K_{tx} = K(0)$ is a constant so that the minimization of $K(0)\sum[y_t - g(x_t, \delta(x))]^2$ is the same as the minimization of $\sum[y_t - g(x_t, \delta)]^2$, that is the local LS becomes the global LS estimator $\hat{\delta}$.

The LLLS estimator in (20) can also be interpreted as the estimator of the functional coefficient (varying coefficient) linear regression model

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= X_t\delta(x_t) + u_t \end{aligned}$$

where $\delta(x_t)$ is approximated locally by a constant $\delta(x_t) \simeq \delta(x)$. The minimization of $\sum u_t^2 K_{tx}$ with respect to $\delta(x)$ then gives the LLLS estimator in (20), which can be interpreted as the LC varying coefficient estimator. An extension of this is to consider the linear approximation $\delta(x_t) \simeq \delta(x) + D(x)(x_t - x)'$ where $D(x) = \frac{\partial\delta(x_t)}{\partial x_t'}$ evaluated at $x_t = x$. In this case

$$\begin{aligned} y_t &= m(x_t) + u_t = X_t\delta(x_t) + u_t \\ &\simeq X_t\delta(x) + X_t D(x)(x_t - x)' + u_t \\ &= X_t\delta(x) + [(x_t - x) \otimes X_t]vecD(x) + u_t \\ &= X_t^x\delta^x(x) + u_t \end{aligned}$$

where $X_t^x = [X_t \;\; (x_t - x) \otimes X_t]$ and $\delta^x(x) = [\delta(x)' \;\; (vecD(x))']'$. The LL varying coefficient estimator of $\delta^x(x)$ can then be obtained by minimizing

$$\sum_{t=1}^{n} [y_t - X_t^x \delta^x(x)]^2 K_{tx}$$

with respect to $\delta^x(x)$ as

$$\dot{\delta}^x(x) = (\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{X}^x)^{-1}\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{y} \,. \tag{21}$$

From this $\dot{\delta}(x) = (\mathbf{I} \;\, 0)\dot{\delta}^x(x)$, and hence

$$\dot{m}(x) = (1 \;\, x \;\, 0)\dot{\delta}^x(x) = (1 \;\, x)\dot{\delta}(x) \,.$$

The above idea can be extended to the situations where $\xi_t = (x_t \;\, z_t)$ such that

$$\mathbb{E}(y_t|\xi_t) = m(\xi_t) = m(x_t, z_t) = X_t\delta(z_t) \,,$$

where the coefficients are varying with respect to only a subset of ξ_t; z_t is $1 \times l$ and ξ_t is $1 \times p$, $p = k + l$. Examples of these include functional coefficient autoregressive models of Chen and Tsay [26] and CFY [24], random coefficient models of Raj and Ullah [128], smooth transition autoregressive models of Granger and Teräsvirta [72], and threshold autoregressive models of Tong [149].

To estimate $\delta(z_t)$ we can again do a local constant approximation $\delta(z_t) \simeq \delta(z)$ and then minimize $\sum[y_t - X_t\delta(z)]^2 K_{tz}$ with respect to $\delta(z)$, where $K_{tz} = K(\frac{z_t - z}{h})$. This gives the LC varying coefficient estimator

$$\tilde{\delta}(z) = (\mathbf{X}'\mathbf{K}(z)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(z)\mathbf{y} \tag{22}$$

where $\mathbf{K}(z)$ is a diagonal matrix of $K_{tz}, t = 1, \ldots, n$. When $z = x$, (22) reduces to the LLLS estimator $\tilde{\delta}(x)$ in (20).

CFY [24] consider a local linear approximation $\delta(z_t) \simeq \delta(z) + D(z)(z_t - z)'$. The LL varying coefficient estimator of CFY is then obtained by minimizing

$$\begin{aligned}
&\sum_{t=1}^{n} [y_t - X_t\delta(z_t)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n} [y_t - X_t\delta(z) - [(z_t - z) \otimes X_t]vecD(z)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n} [y_t - X_t^z\delta^z(z)]^2 K_{tz}
\end{aligned}$$

with respect to $\delta^z(z) = [\delta(z)' \;\; (vecD(z))']'$ where $X_t^z = [X_t \;\; (z_t - z) \otimes X_t]$. This gives

$$\ddot{\delta}^z(z) = (\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{X}^z)^{-1}\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{y} \,, \tag{23}$$

and $\ddot{\delta}(z) = (\mathbf{I} \;\, 0)\ddot{\delta}^z(z)$. Hence

$$\ddot{m}(\xi) = (1 \;\, x \;\, 0)\ddot{\delta}^z(z) = (1 \;\, x)\ddot{\delta}(z) \,.$$

For the asymptotic properties of these varying coefficient estimators, see CFY [24]. When $z = x$, (23) reduces to the LL varying coefficient estimator $\dot{\delta}^x(x)$ in (21). See Lee and Ullah [98] for more discussion of these models and issues of testing nonlinearity.

Regime Switching Autoregressive Model Between Unit Root and Stationary Root

To avoid the usual dichotomy between unit-root nonstationarity and stationarity, we may consider models that permit two regimes of unit root nonstationarity and stationarity.

One model is the Innovation Regime-Switching (IRS) model of Kuan, Huang, and Tsay [96]. Intuitively, it may be implausible to believe that all random shocks exert only one effect (permanent or transitory) on future financial asset prices in a long time span. This intuition underpins the models that allow for breaks, stochastic unit root, or regime switching. As an alternative, Kuan, Huang, and Tsay [96] propose the IRS model that permits the random shock in each period to be permanent or transitory, depending on a switching mechanism, and hence admits distinct dynamics (unit-root nonstationarity or stationarity) in different periods. Under the IRS framework, standard unit-root models and stationarity models are just two extreme cases. By applying the IRS model to real exchange rate, they circumvent the difficulties arising from unit-root (or stationarity) testing. They allow the data to speak for themselves, rather than putting them in the straitjacket of unit-root nonstationarity or stationarity. Huang and Kuan [90] re-examine long-run PPP based on the IRS model and their empirical study on US/UK real exchange rates shows that there are both temporary and permanent influences on the real exchange rate such that approximately 42% of the shocks in the long run are more likely to have a permanent effect. They also found that transitory shocks dominate in the fixed-rate regimes, yet permanent shocks play a more important role during the floating regimes. Thus, the long-run PPP is rejected due to the presence of a significant amount of permanent shocks, but there are still long periods of time in which the deviations from long-run PPP are only transitory.

Another model is a threshold unit root (TUR) model or threshold integrated moving average (TIMA) model of Gonzalo and Martíneza [65]. Based on this model they examine whether large and small shocks have different long-run effects, as well as whether one of them is purely transitory. They develop a new nonlinear permanent – transitory decomposition, that is applied to US stock prices to analyze the quality of the stock market.

Comparison of these two models with the linear autoregressive model with a unit root or a stationary AR model for the out-of-sample forecasting remains to be examined empirically.

Bagging Nonlinear Forecasts

To improve on unstable forecasts, bootstrap aggregating or bagging is introduced by Breiman [19]. Lee and Yang [100] show how bagging works for binary and quantile predictions. Lee and Yang [100] attributed part of the success of the bagging predictors to the small sample estimation uncertainties. Therefore, a question that may arise is that whether the good performance of bagging predictors critically depends on algorithms we employ in nonlinear estimation.

They find that bagging improves the forecasting performance of predictors on highly nonlinear regression models – e. g., artificial neural network models, especially when the sample size is limited. It is usually hard to choose the number of hidden nodes and the number of inputs (or lags), and to estimate the large number of parameters in an ANN model. Therefore, a neural network model generate poor predictions in a small sample. In such cases, bagging can do a valuable job to improve the forecasting performance as shown in [100], confirming the result of Breiman [20]. A bagging predictor is a combined predictor formed over a set of training sets to smooth out the “instability” caused by parameter estimation uncertainty and model uncertainty. A predictor is said to be “unstable” if a small change in the training set will lead to a significant change in the predictor [20].

As bagging would be valuable in nonlinear forecasting, in this section, we will show how a bagging predictor may improve the predicting performance of its underlying predictor. Let

$$\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t} \quad (t = R, \dots, T)$$

be a training set at time t and let $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ be a forecast of Y_{t+1} or of the binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$ using this training set $\mathcal{D}_t$ and the explanatory variable vector $\mathbf{X}_t$. The optimal forecast $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ for Y_{t+1} will be the conditional mean of Y_{t+1} given $\mathbf{X}_t$ under the squared error loss function, or the conditional quantile of Y_{t+1} on $\mathbf{X}_t$ if the loss is a tick function. Below we also consider the binary forecast for $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$.

Suppose each training set $\mathcal{D}_t$ consists of R observations generated from the underlying probability distribution $\mathbf{P}$. The forecast $\{\varphi(\mathbf{X}_t, \mathcal{D}_t)\}_{t=R}^{T}$ can be improved if more training sets were able to be generated from $\mathbf{P}$ and the forecast can be formed from averaging the multiple forecasts obtained from the multiple training sets. Ideally, if $\mathbf{P}$ were known and multiple training sets $\mathcal{D}_t^{(j)}$ $(j = 1, \dots, J)$ may be drawn from $\mathbf{P}$, an ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ can be constructed by the weighted averaging of $\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)})$ over j, i. e.,

$$\varphi_A(\mathbf{X}_t) \equiv \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t) \equiv \sum_{j=1}^{J} w_{j,t}\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)}) ,$$

where $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ denotes the expectation over $\mathbf{P}$, $w_{j,t}$ is the weight function with $\sum_{j=1}^{J} w_{j,t} = 1$, and the subscript A in φ_A denotes “aggregation”.

Lee and Yang [100] show that the ensemble aggregating predictor $\varphi_A(X_t)$ has not a larger expected loss than the original predictor $\varphi(X_t, \mathcal{D}_t)$. For any convex loss function $c(\cdot)$ on the forecast error z_{t+1}, we will have

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) \geq \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1})),$$

where $\mathbb{E}_{\mathcal{D}_t}(z_{t+1})$ is the aggregating forecast error, and $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}\{\mathbb{E}_{\mathcal{D}_t}(\cdot)|X_t\}]$ denotes the expectation $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ taken over $\mathbf{P}$ (i. e., averaging over the multiple training sets generated from $\mathbf{P}$), then taking an expectation of Y_{t+1} conditioning on X_t, and then taking an expectation of X_t. Similarly we define the notation $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}(\cdot)|X_t]$. Therefore, the aggregating predictor will always have no larger expected cost than the original predictor for a convex loss function $\varphi(X_t, D_t)$. The examples of the convex loss function includes the squared error loss and a tick (or check) loss $\rho_\alpha(z) \equiv [\alpha - \mathbf{1}(z < 0)]z$.

How much this aggregating predictor can improve depends on the distance between $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1})$ and $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. We can define this distance by $\Delta \equiv \mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) - \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. Therefore, the effectiveness of the aggregating predictor depends on the *convexity* of the cost function. The more convex is the cost function, the more effective this aggregating predictor can be. If the loss function is the squared error loss, then it can be shown that $\Delta = \mathbb{V}_{\mathcal{D}_t}\left[\varphi(\mathbf{X}_t, \mathcal{D}_t)\right]$ is the variance of the predictor, which measures the “instability” of the predictor. See Lee and Yang [100], Proposition 1, and Breiman [20]. If the loss is the tick function, the effectiveness of bagging is also different for different quantile predictions: bagging works better for tail-quantile predictions than for mid-quantile predictions.

In practice, however, $\mathbf{P}$ is not known. In that case we may estimate $\mathbf{P}$ by its empirical distribution, $\hat{\boldsymbol{P}}(\mathcal{D}_t)$, for a given $\mathcal{D}_t$. Then, from the empirical distribution $\hat{\boldsymbol{P}}(\mathcal{D}_t)$, multiple training sets may be drawn by the bootstrap method. Bagging predictors, $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$, can then be computed by taking weighted average of the predictors trained over a set of bootstrap training sets. More specifically, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be obtained in the following steps:

1. Given a training set of data at time t, $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^t$, construct the jth bootstrap sample $\mathcal{D}_t^{*(j)} \equiv \{(Y_s^{*(j)}, \mathbf{X}_{s-1}^{*(j)})\}_{s=t-R+1}^t$, $j = 1, \ldots, J$, according to the empirical distribution of $\hat{\boldsymbol{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$.
2. Train the model (estimate parameters) from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
3. Compute the bootstrap predictor $\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})$ from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
4. Finally, for mean and quantile forecast, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by averaging over J bootstrap predictors

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \sum_{j=1}^{J} \hat{w}_{j,t} \varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) \, ;$$

and for binary forecast, the bagging binary predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by majority voting over J bootstrap predictors:

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \mathbf{1}\left(\sum_{j=1}^{J} \hat{w}_{j,t} \varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) > 1/2 \right)$$

with $\sum_{j=1}^{J} \hat{w}_{j,t} = 1$ in both cases.

One concern of applying bagging to time series is whether a bootstrap can provide a sound simulation sample for dependent data, for which the bootstrap is required to be consistent. It has been shown that some bootstrap procedure (such as moving block bootstrap) can provide consistent densities for moment estimators and quantile estimators. See, e. g., Fitzenberger [54].

Nonlinear Forecasting Models for the Conditional Variance

Nonlinear Parametric Models for Volatility

Volatility models are of paramount importance in financial economics. Issues such as portfolio allocation, option pricing, risk management, and generally any decision making under uncertainty rely on the understanding and forecasting of volatility. This is one of the most active ares of research in time series econometrics. Important surveys as in Bollerslev, Chou, and Kroner [15], Bera and Higgins [13], Bollerslev, Engle, and Nelson [16], Poon and Granger [125], and Bauwens, Laurent, and Rombouts [12] attest to the variety of issues in volatility research. The motivation for the introduction of the first generation of volatility models namely the ARCH models of Engle [44] was to account for clusters of activity and fat-tail behavior of financial data. Subsequent models accounted for more complex issues. Among others and without being exclusive, we should mention issues related to asymmetric responses of volatility to news, probability distribution of the standardized innovations, i.i.d. behavior of the standardized innovation, persistence of the volatility process, linkages with continuous time models, intraday data and unevenly spaced observations, seasonality and noise in intraday data. The consequence of this research agenda has been a vast array of specifications for the volatility process.

Suppose that the return series $\{y_t\}_{t=1}^{T+1}$ of a financial asset follows the stochastic process $y_{t+1} = \mu_{t+1} + \varepsilon_{t+1}$, where $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_{t+1}(\theta)$ and $\mathbb{E}(\varepsilon_{t+1}^2|\mathcal{F}_t) = \sigma_{t+1}^2(\theta)$ given the information set $\mathcal{F}_t$ (σ-field) at time t. Let $z_{t+1} \equiv \varepsilon_{t+1}/\sigma_{t+1}$ have the conditional normal distribution with zero conditional mean and unit conditional variance. Volatility models can be classified in three categories: MA family, ARCH family, and stochastic volatility (SV) family.

The simplest method to forecast volatility is to calculate a historical moving average variance, denoted as MA(m), or an exponential weighted moving average (EWMA):

MA(m)	$\sigma_t^2 = \frac{1}{m}\sum_{j=1}^m (y_{t-j} - \hat{\mu}_t^m)^2, \quad \hat{\mu}_t^m = \frac{1}{m}\sum_{j=1}^m y_{t-j}$
EWMA	$\sigma_t^2 = (1-\lambda)\sum_{j=1}^{t-1} \lambda^{j-1}(y_{t-j} - \hat{\mu}_t)^2,$ $\hat{\mu}_t = \frac{1}{t-1}\sum_{j=1}^{t-1} y_{t-j}$

In the EWMA specification, a common practice is to fix the λ parameter, for instance $\lambda = 0.94$ [129]. For these two MA family models, there are not parameters to estimate.

Second, the ARCH family is very extensive with many variations on the original model ARCH(p) of Engle [44]. Some representative models are: GARCH model of Bollerslev [14]; Threshold GARCH (T-GARCH) of Glosten et al. [60]; Exponential GARCH (E-GARCH) of Nelson [120]; quadratic GARCH models (Q-GARCH) as in Sentana [135]; Absolute GARCH (ABS-GARCH) of

Taylor [143] and Schwert [134] and Smooth Transition GARCH (ST-GARCH) of González-Rivera [61].

ARCH(p)	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2$
GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2$
I-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2,\ \alpha + \beta = 1$
T-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 \mathbf{1}(\varepsilon_{t-1} \geq 0)$
ST-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 F(\varepsilon_{t-1}, \delta)$ with $F(\varepsilon_{t-1}, \delta) = [1 + \exp(\delta\varepsilon_{t-1})]^{-1} - 0.5$
E-GARCH	$\ln\sigma_t^2 = \omega + \beta \ln\sigma_{t-1}^2 + \alpha[\lvert z_{t-1}\rvert - cz_{t-1}]$
Q-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(\varepsilon_{t-1} + \gamma)^2$
ABS-GARCH	$\sigma_t = \omega + \beta\sigma_{t-1} + \alpha\lvert\varepsilon_{t-1}\rvert$

The EWMA specification can be viewed as an integrated GARCH model with $\omega = 0, \alpha = \lambda$, and $\beta = 1 - \lambda$. In the T-GARCH model, the parameter γ allows for possible asymmetric effects of positive and negative innovations. In Q-GARCH models, the parameter γ measures the extent of the asymmetry in the news impact curve. For the ST-GARCH model, the parameter γ measures the asymmetric effect of positive and negative shocks, and the parameter $\delta > 0$ measures the smoothness of the transition between regimes, with a higher value of δ making ST-GARCH closer to T-GARCH.

Third, the stationary SV model of Taylor [143] with η_t is i.i.d. N $(0, \sigma_\eta^2)$ and ξ_t is i.i.d. N$(0, \pi^2/2)$ is a representative member of the SV family.

SV	$\sigma_t^2 = \exp(0.5h_t), \quad \ln(y_t^2) = -1.27 + h_t + \xi_t,$ $h_t = \gamma + \phi h_{t-1} + \eta_t.$

With so many models, the natural question becomes which one to choose. There is not a universal answer to this question. The best model depends upon the objectives of the user. Thus, given an objective function, we search for the model(s) with the best predictive ability controlling for possible biases due to "data snooping" [105]. To compare the relative performance of volatility models, it is customary to choose either a statistical loss function or an economic loss function.

The preferred statistical loss functions are based on moments of forecast errors (mean-error, mean-squared error, mean absolute error, etc.). The best model will minimize a function of the forecast errors. The volatility forecast is often compared to a measure of realized volatility. With financial data, the common practice has been to take squared returns as a measure of realized volatility. However, this practice is questionable. Andersen and Bollerslev [2] argued that this measure is a noisy estimate, and proposed the use of the intra-day (at each five minutes interval) squared returns to calculate the daily realized volatility. This measure requires intra-day data, which is subject to the variation introduced by the bid-ask spread and the irregular spacing of the price quotes.

Some other authors have evaluated the performance of volatility models with criteria based on economic loss functions. For example, West, Edison, and Cho [157] considered the problem of portfolio allocation based on models that maximize the utility function of the investor. Engle, Kane, and Noh [46] and Noh, Engle, and Kane [121] considered different volatility forecasts to maximize the trading profits in buying/selling options. Lopez [107] considered probability scoring rules that were tailored to a forecast user's decision problem and confirmed that the choice of loss function directly affected the forecast evaluation of different models. Brooks and Persand [21] evaluated volatility forecasting in a financial risk management setting in terms of Value-at-Risk (VaR). The common feature to these branches of the volatility literature is that none of these has controlled for forecast dependence across models and the inherent biases due to data-snooping.

Controlling for model dependence [160], González-Rivera, Lee, and Mishra [62] evaluate fifteen volatility models for the daily returns to the SP500 index according to their out-of-sample forecasting ability. The forecast evaluation is based, among others, on two economic loss functions: an option pricing formula and a utility function; and a statistical loss function: a goodness-of-fit based on a Value-at-Risk (VaR) calculation. For option pricing, volatility is the only component that is not observable and it needs to be estimated. The loss function assess the difference between the actual price of a call option and the estimated price, which is a function of the estimated volatility of the stock. The second economic loss function refers to the problem of wealth allocation. An investor wishes to maximize her utility allocating wealth between a risky asset and a risk-free asset. The loss function assesses the performance of the volatility estimates according to the level of utility they generate. The statistical function based on the goodness-of-fit of a VaR calculation is important for risk management. The main objective of VaR is to calculate extreme losses within a given probability of occurrence, and the estimation of the volatility is central to the VaR measure. The preferred models depend very strongly upon the loss function chosen by the user. González-Rivera, Lee, and Mishra [62] find that, for option pricing, simple models such as the exponential weighted moving average (EWMA) proposed by Riskmetrics [64] performed as well as any GARCH model. For an utility loss function, an asymmetric quadratic GARCH model is the most preferred. For VaR calculations, a stochastic volatility model dominates all other models.

Nonparametric Models for Volatility

Ziegelmann [163] considers the kernel smoothing techniques that free the traditional parametric volatility estimators from the constraints related to their specific models. He applies the nonparametric local 'exponential' estimator to estimate conditional volatility functions, ensuring its nonnegativity. Its asymptotic properties are established and compared with those for the local linear estimator for the volatility model of Fan and Yao [51]. Long, Su, and Ullah [106] extend this idea to semiparametric multivariate GARCH and show that there may exist substantial out-of-sample forecasting gain over the parametric models. This gain accounts for the presence of nonlinearity in the conditional variance-covariance that is neglected in parametric linear models.

Forecasting Volatility Using High Frequency Data

Using high-frequency data, quadratic variation may be estimated using realized volatility (RV). Andersen, Bollerslev, Diebold, and Labys [3] and Barndorff-Nielsen and Shephard [11] establish that RV, defined as the sum of squared intraday returns of small intervals, is an asymptotically unbiased estimator of the unobserved quadratic variation as the interval length approaches zero. Besides the use of high frequency information in volatility estimation, volatility forecasting using high frequency information has been addressed as well. In an application to volatility prediction, Ghysels, Santa-Clara, and Valkanov [58] investigate the predictive power of various regressors (lagged realized volatility, squared return, realized power, and daily range) for future volatility forecasting. They find that the best predictor is realized power (sum of absolute intraday returns), and more interestingly, direct use of intraday squared returns in mixed data sampling (MIDAS) regressions does not necessarily lead to better volatility forecasts.

Andersen, Bollerslev, Diebold, and Labys [4] represent another approach to forecasting volatility using RV. The model they propose is a fractional integrated AR model: ARFI(5, d) for logarithmic RV's obtained from foreign exchange rates data of 30-minute frequency and demonstrate the superior predictive power of their model.

Alternatively, Corsi [32] proposes the heterogeneous autoregressive (HAR) model of RV, which is able to reproduce long memory. McAleer and Medeiros [115] propose a new model that is a multiple regime smooth transition (ST) extension of the HAR model, which is specifically designed to model the behavior of the volatility inherent in financial time series. The model is able to describe simultaneously long memory as well as sign and size asymmetries. They apply the model to several Dow Jones Industrial Average index stocks using transaction level data from the Trades and Quotes database that covers ten years of data, and find strong support for long memory and both sign and size asymmetries. Furthermore, they show that the multiple regime smooth transition HAR model, when combined with the linear HAR model, is flexible for the purpose of forecasting volatility.

Forecasting Beyond Mean and Variance

In the previous section, we have surveyed the major developments in nonlinear time series, mainly modeling the conditional mean and the conditional variance of financial returns. However it is not clear yet that any of those nonlinear models may generate profits after accounting for various market frictions and transactions costs. Therefore, some research efforts have been directed to investigate other aspects of the conditional density of returns such as higher moments, quantiles, directions, intervals, and the density itself. In this section, we provide a brief survey on forecasting these other features.

Forecasting Quantiles

The optimal forecast of a time series model depends on the specification of the loss function. A symmetric quadratic loss function is the most prevalent in applications due to its simplicity. Under symmetric quadratic loss, the optimal forecast is simply the conditional mean. An asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67].

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \ldots, Z_t\}$ of size R or the whole past sample $\{Z_1, \ldots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^{T}$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^{*}$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) ,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e^*_{t+1} = Y_{t+1} - f^*_{t,1} .$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f^*_{t,1}) \mathrm{d}F_t(y) = 0 .$$

When we interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f^*_{t,1}) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1}) | I_t \right)$$

Based on the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1})$, the condition for forecast optimality is:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t \right) = 0 \quad a.s. ,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

To see this we consider the following two examples. First, when the loss function is the squared error loss

$$c(Y_{t+1} - f_{t,1}) = (Y_{t+1} - f_{t,1})^2 ,$$

the generalized forecast error will be $g_{t+1} \equiv \frac{\partial}{\partial f_t} c(Y_{t+1} - f^*_{t,1}) = -2e^*_{t+1}$ and thus $\mathbb{E}\left(e^*_{t+1} | I_t \right) = 0$ $a.s.$, which implies that the optimal forecast

$$f^*_{t,1} = \mathbb{E}\left(Y_{t+1} | I_t \right)$$

is the conditional mean. Next, when the loss is the check function, $c(e) = \left[\alpha - \mathbf{1}(e < 0) \right] \cdot e \equiv \rho_\alpha(e_{t+1})$, the optimal forecast $f_{t,1}$, for given $\alpha \in (0, 1)$, minimizing

$$\min_{f_{t,1}} \mathbb{E}\left[c(Y_{t+1} - f_{t,1}) | I_t \right]$$

can be shown to satisfy

$$\mathbb{E}\left[\alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1}) | I_t \right] = 0 \quad a.s.$$

Hence, $g_{t+1} \equiv \alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})$ is the generalized forecast error. Therefore,

$$\alpha = \mathbb{E}\left[\mathbf{1}(Y_{t+1} < f^*_{t,1}) | I_t \right] = \Pr(Y_{t+1} \leq f^*_{t,1} | I_t) ,$$

and the optimal forecast $f^*_{t,1} = q^\alpha\left(Y_{t+1} | I_t \right) \equiv q^\alpha_t$ is the conditional α-quantile.

Forecasting conditional quantiles are of paramount importance for risk management, which nowdays is key activity in financial institutions due to the increasing financial fragility in emerging markets and the extensive use of derivative products over the last decade. A risk measurement methodology called Value-at-Risk (VaR) has received a great attention from both regulatory and academic fronts. During a short span of time, numerous papers have studied various aspects of the VaR methodology. Bao, Lee, and Saltoglu [8] examine the relative out-of-sample predictive performance of various VaR models.

An interesting VaR model is the CaViaR (conditional autoregressive Value-at-Risk) model suggested by Engle and Manganelli [47]. They estimate the VaR from a quantile regression rather than inverting a conditional distribution. The idea is similar to the GARCH modeling in that VaR is modeled autoregressively

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + h(x_t | \theta) ,$$

where $x_t \in \mathcal{F}_{t-1}$, θ is a parameter vector, and $h(\cdot)$ is a function to explain the VaR model. Depending on the specification of $h(\cdot)$, the CaViaR model may be

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + a_2 |r_{t-1}| ,$$

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + a_2 |r_{t-1}| + a_3 |r_{t-1}| \cdot \mathbf{1}(r_{t-1} < 0),$$

where the second model allow nonlinearity (asymmetry) similarly to the asymmetric GARCH models.

Bao, Lee, and Saltoglu [8] compare various VaR models. Their results show that the CaViaR quantile regression models of Engle and Manganelli [47] have shown some success in predicting the VaR risk measure for various periods of time, and it is generally more stable than the models that invert a distribution function.

Forecasting Directions

It is well known that, while financial returns $\{Y_t\}$ may not be predictable, their variance, sign, and quantiles may be predictable. Christofferson and Diebold [27] show that binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$, where $\mathbf{1}(\cdot)$ takes the value of 1 if the statement in the parenthesis is true, and 0 otherwise, is predictable when some conditional moments are time varying, Hong and Lee [86], Hong and

Chung [85], Linton and Whang [104], Lee and Yang [100] among many others find some evidence that the directions of stock returns and foreign exchange rate changes are predictable.

Lee and Yang [100] also show that forecasting quantiles and forecasting binary (directional) forecasts are related, in that the former may lead to the latter. As noted by Powell [126], using the fact that for any monotonic function $h(\cdot)$, $q_t^\alpha(h(Y_{t+1})|\mathbf{X}_t) = h(q_t^\alpha(Y_{t+1}|\mathbf{X}_t))$, which follows immediately from observing that $\Pr(Y_{t+1} < y|\mathbf{X}_t) = \Pr[h(Y_{t+1}) < h(y)|\mathbf{X}_t]$, and noting that the indicator function is monotonic, $q_t^\alpha(G_{t+1}|\mathbf{X}_t) = q_t^\alpha(\mathbf{1}(Y_{t+1} > 0)|\mathbf{X}_t) = \mathbf{1}(q_t^\alpha(Y_{t+1}|\mathbf{X}_t) > 0)$. Therefore, predictability of conditional quantiles of financial returns may imply predictability of conditional direction.

Probability Forecasts

Diebold and Rudebush [38] consider the probability forecasts for the turning points of the business cycle. They measure the accuracy of predicted probabilities, that is the average distance between the predicted probabilities and observed realization (as measured by a zero-one dummy variable). Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time $t\,(R \le t < T + 1)$, to estimate a model. We then have time series of $P = T - R + 1$ probability forecasts $\{p_{t+1}\}_{t=R}^T$ where p_t is the predicted probability of the occurrence of an event (e. g., business cycle turning point) in the next period $t + 1$. Let $\{d_{t+1}\}_{t=R}^T$ be the corresponding realization with $d_t = 1$ if a business cycle turning point (or any defined event) occurs in period t and $d_t = 0$ otherwise. The loss function analogous to the squared error is the Brier's score based on quadratic probability score (QPS):

$$QPS = P^{-1} \sum_{t=R}^{T} 2(p_t - d_t)^2 \,.$$

The QPS ranges from 0 to 2, with 0 for perfect accuracy. As noted by Diebold and Rudebush [38], the use of the symmetric loss function may not be appropriate as a forecaster may be penalized more heavily for missing a call (making a type II error) than for signaling a false alarm (making a type I error). Another loss function is given by the log probability score (LPS)

$$LPS = -P^{-1} \sum_{t=R}^{T} \ln\left(p_t^{d_t}(1 - p_t)^{(1-d_t)}\right) ,$$

which is similar to the loss for the interval forecast. A large mistake is penalized more heavily under LPS than under QPS. More loss functions are discussed in Diebold and Rudebush [38].

Another loss function useful in this context is the Kuipers score (KS), which is defined by

$$KS = \text{Hit Rate} - \text{False Alarm Rate} ,$$

where Hit Rate is the fraction of the bad events that were correctly predicted as good events (power, or 1− probability of type II error), and False Alarm Rate is the fraction of good events that had been incorrectly predicted as bad events (probability of type I error).

Forecasting Interval

Suppose Y_t is a stationary series. Let the one-period ahead conditional interval forecast made at time t from a model be denoted as

$$J_{t,1}(\alpha) = (L_{t,1}(\alpha), U_{t,1}(\alpha)), \quad t = R, \ldots, T ,$$

where $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ are the lower and upper limits of the ex ante interval forecast for time $t + 1$ made at time t with the coverage probability α. Define the indicator variable $X_{t+1}(\alpha) = \mathbf{1}[Y_{t+1} \in J_{t,1}(\alpha)]$. The sequence $\{X_{t+1}(\alpha)\}_{t=R}^T$ is i.i.d. Bernoulli (α). The optimal interval forecast would satisfy $\mathbb{E}(X_{t+1}(\alpha)|I_t) = \alpha$, so that $\{X_{t+1}(\alpha) - \alpha\}$ will be an MD. A better model has a larger expected Bernoulli log-likelihood

$$\mathbb{E}\alpha^{X_{t+1}(\alpha)}(1 - \alpha)^{[1-X_{t+1}(\alpha)]} \,.$$

Hence, we can choose a model for interval forecasts with the largest out-of-sample mean of the predictive log-likelihood, which is defined by

$$P^{-1} \sum_{t=R}^{T} \ln\left(\alpha^{x_{t+1}(\alpha)}(1 - \alpha)^{[1-x_{t+1}(\alpha)]}\right) .$$

Evaluation of Nonlinear Forecasts

In order to evaluate the possible superior predictive ability of nonlinear models, we need to compare competing models in terms of a certain loss function. The literature has recently been exploding on this issue. Examples are Granger and Newbold [69], Diebold and Mariano [37], West [156], White [160], Hansen [81], Romano and Wolf [130], Giacomini and White [59], etc. In different perspective, to test the optimality of a given model, Patton and Timmermann [123] examine various testable properties that should hold for an optimal forecast.

Loss Functions

The loss function (or cost function) is a crucial ingredient for the evaluation of nonlinear forecasts. When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c(Y_{t+h}, f_{t,h})$. The loss function can depend on the time of prediction and so it can be $c_{t+h}(Y_{t+h}, f_{t,h})$. If the loss function is not changing with time and does not depend on the value of the variable Y_{t+h}, the loss can be written simply as a function of the error only, $c_{t+h}(Y_{t+h}, f_{t,h}) = c(e_{t+h})$.

Granger [67] discusses the following required properties for a loss function: (i) $c(0) = 0$ (no error and no loss), (ii) $\min_e c(e) = 0$, so that $c(e) \geq 0$, and (iii) $c(e)$ is monotonically nondecreasing as e moves away from zero so that $c(e_1) \geq c(e_2)$ if $e_1 > e_2 > 0$ and if $e_1 < e_2 < 0$.

When $c_1(e), c_2(e)$ are both loss functions, Granger [67] shows that further examples of loss functions can be generated: $c(e) = ac_1(e) + bc_2(e)$, $a \geq 0, b \geq 0$ will be a loss function. $c(e) = c_1(e)^a c_2(e)^b$, $a > 0, b > 0$ will be a loss function. $c(e) = 1(e > 0)c_1(e) + 1(e < 0)c_2(e)$ will be a loss function. If $h(\cdot)$ is a positive monotonic nondecreasing function with $h(0)$ finite, then $c(e) = h(c_1(e)) - h(0)$ is a loss function.

Granger [68] notes that an expected loss (a risk measure) of financial return Y_{t+1} that has a conditional predictive distribution $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$ with $\mathbf{X}_t \in I_t$ may be written as

$$\mathbb{E}c(e) = A_1 \int_0^\infty |y-f|^p dF_t(y) + A_2 \int_{-\infty}^0 |y-f|^p \mathrm{d}F_t(y),$$

with A_1, A_2 both > 0 and some $\theta > 0$. Considering the symmetric case $A_1 = A_2$, one has a class of volatility measures $V_p = \mathbb{E}\left[|y - f|^p\right]$, which includes the variance with $p = 2$, and mean absolute deviation with $p = 1$.

Ding, Granger, and Engle [39] study the time series and distributional properties of these measures empirically and show that the absolute deviations are found to have some particular properties such as the longest memory. Granger remarks that given that the financial returns are known to come from a long tail distribution, $p = 1$ may be more preferable.

Another problem raised by Granger is how to choose optimal L_p-norm in empirical works, to minimize $\mathbb{E}[|\varepsilon_t|^p]$ for some p to estimate the regression model $Y_t = \mathbb{E}(Y_t|X_t; \beta) + \varepsilon_t$. As the asymptotic covariance matrix of $\hat{\beta}$ depends on p, the most appropriate value of p can be chosen to minimize the covariance matrix. In particular, Granger [68] refers to a trio of papers [84,116,117] who find that the optimal $p = 1$ from Laplace and Cauchy distribution, $p = 2$ for Gaussian and $p = \infty$ (min/max estimator) for a rectangular distribution. Granger [68] also notes that in terms of the kurtosis κ, Harter [84] suggests to use $p = 1$ for $\kappa > 3.8$; $p = 2$ for $2.2 \leq \kappa \leq 3.8$; and $p = 3$ for $\kappa < 2.2$. In finance, the kurtosis of returns can be thought of as being well over 4 and so $p = 1$ is preferred.

Forecast Optimality

Optimal forecast of a time series model extensively depends on the specification of the loss function. Symmetric quadratic loss function is the most prevalent in applications due to its simplicity. The optimal forecast under quadratic loss is simply the conditional mean, but an asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67], as the expected loss function if formulated with the expectation taken with respect to the conditional distribution. Specification of the loss function defines the model under consideration.

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1\,(\equiv R + P)$ observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \ldots, Z_t\}$ of size R or the whole past sample $\{Z_1, \ldots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^T$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^*$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) \,,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e_{t+1}^* = Y_{t+1} - f_{t,1}^*.$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f^*_{t,1}) \mathrm{d}F_t(y) = 0 \, .$$

When we may interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f^*_{t,1}) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1}) | I_t \right)$$

the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1})$, forms the condition of forecast optimality:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t \right) = 0 \quad a.s.,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

Forecast Evaluation of Nonlinear Transformations

Granger [67] note that it is implausible to use the same loss function for forecasting Y_{t+h} and for forecasting $h_{t+1} = h(Y_{t+h})$ where $h(\cdot)$ is some function, such as the log or the square, if one is interested in forecasting volatility. Suppose the loss functions $c_1(\cdot), c_2(\cdot)$ are used for forecasting Y_{t+h} and for forecasting $h(Y_{t+h})$, respectively. Let $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c_1(e_{t+1})$, for which the optimal forecast $f^*_{t,1}$ will be chosen from $\min_{f_{t,1}} \int_{-\infty}^{\infty} c_1(y - f_{t,1}) \mathrm{d}F_t(y)$, where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$. Let $\varepsilon_{t+1} \equiv h_{t+1} - h_{t,1}$ will result in a cost of $c_2(\varepsilon_{t+1})$, for which the optimal forecast $h^*_{t,1}$ will be chosen from $\min_{h_{t,1}} \int_{-\infty}^{\infty} c_2(h - h_{t,1}) \mathrm{d}H_t(h)$, where $H_t(h) \equiv \Pr(h_{t+1} \leq h | I_t)$. Then the optimal forecasts for Y and h would respectively satisfy

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c_1(y - f^*_{t,1}) \mathrm{d}F_t(y) = 0 \, ,$$
$$\int_{-\infty}^{\infty} \frac{\partial}{\partial h_{t,1}} c_2(h - h^*_{t,1}) \mathrm{d}H_t(h) = 0 \, .$$

It is easy to see that the optimality condition for $f^*_{t,1}$ does not imply the optimality condition for $h^*_{t,1}$ in general. Under some strong conditions on the functional forms of the transformation $h(\cdot)$ and of the two loss functions $c_1(\cdot), c_2(\cdot)$, the above two conditions may coincide. Granger [67] remarks that it would be strange behavior to use the same loss function for Y and $h(Y)$. We leave this for further analysis in a future research.

Density Forecast Evaluation

Most of the classical finance theories, such as asset pricing, portfolio selection and option valuation, aim to model the surrounding uncertainty via a parametric distribution function. For example, extracting information about market participants' expectations from option prices can be considered another form of density forecasting exercise [92]. Moreover, there has also been increasing interest in evaluating forecasting models of inflation, unemployment and output in terms of density forecasts [29]. While evaluating each density forecast model has become versatile since Diebold et al. [35], there has been much less effort in comparing alternative density forecast models.

Given the recent empirical evidence on volatility clustering and asymmetry and fat-tailedness in financial return series, relative adequacy of a given model among alternative models would be useful measure of evaluating forecast models. Deciding on which distribution and/or volatility specification to use for a particular asset is a common task even for finance practitioners. For example, despite the existence of many volatility specifications, a consensus on which model is most appropriate has yet to be reached. As argued in Poon and Granger [125], most of the (volatility) forecasting studies do not produce very conclusive results because only a subset of alternative models are compared, with a potential bias towards the method developed by the authors. Poon and Granger [125] argue that lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task. They wrote (p. 507), "… it seems clear that one form of study that is included is conducted just to support a viewpoint that a particular method is useful. It might not have been submitted for publication if the required result had not been reached. This is one of the obvious weaknesses of a comparison such as this; the papers being prepared for different reasons, use different data sets, many kinds of assets, various intervals between readings, and a variety of evaluation techniques".

Following Diebold et al. [35], it has become common practice to evaluate the adequacy of a forecast model based on the probability integral transform (PIT) of the process with respect to the model's density forecast. If the density forecast model is correctly specified, the PIT follows an i.i.d. uniform distribution on the unit interval and, equivalently, its inverse normal transform follows an i.i.d. normal distribution. We can therefore evaluate a density forecast model by examining the departure of the transformed PIT from this property (i.i.d. and normality). The departure can be quantified by the Kullback-Leibler [97] information criterion, or KLIC, which is the expected logarithmic value of the likelihood ratio (LR) of the transformed PIT and the i.i.d. normal variate. Thus the LR statistic measures the distance of a candidate model to the unknown true model.

Consider a financial return series $\{y_t\}_{t=1}^{T}$. This observed data on a univariate series is a realization of a stochastic process $\mathbf{Y}^T \equiv \{Y_\tau : \Omega \to \mathbb{R}, \tau = 1, 2, \ldots, T\}$ on a complete probability space $(\Omega, \mathcal{F}_T, P_0^T)$, where $\Omega = \mathbb{R}^T \equiv \times_{\tau=1}^{T} \mathbb{R}$ and $\mathcal{F}_T = \mathcal{B}(\mathbb{R}^T)$ is the Borel σ-field generated by the open sets of $\mathbb{R}^T$, and the *joint* probability measure $P_0^T(B) \equiv P_0[\mathbf{Y}^T \in B]$, $B \in \mathcal{B}(\mathbb{R}^T)$ completely describes the stochastic process. A sample of size T is denoted as $\mathbf{y}^T \equiv (y_1, \ldots, y_T)'$.

Let σ-finite measure ν^T on $\mathcal{B}(\mathbb{R}^T)$ be given. Assume $P_0^T(B)$ is absolutely continuous with respect to ν^T for all $T = 1, 2, \ldots$, so that there exists a measurable Radon–Nikodým density $g^T(\mathbf{y}^T) = \mathrm{d}P_0^T / \mathrm{d}\nu^T$, unique up to a set of zero measure-ν^T.

Following White [159], we define a probability model $\mathcal{P}$ as a collection of distinct probability measures on the measurable space $(\Omega, \mathcal{F}_T)$. A probability model $\mathcal{P}$ is said to be correctly specified for $\mathbf{Y}^T$ if $\mathcal{P}$ contains P_0^T. Our goal is to evaluate and compare a set of parametric probability models $\{P_\theta^T\}$, where $P_\theta^T(B) \equiv P_\theta[Y^T \in B]$. Suppose there exists a measurable Radon–Nikodým density $f^T(\mathbf{y}^T) = \mathrm{d}P_\theta^T / \mathrm{d}\nu^T$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is a finite-dimensional vector of parameters and is assumed to be identified on $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^k$. See Theorem 2.6 in White [159].

In the context of forecasting, instead of the joint density $g^T(\mathbf{y}^T)$, we consider forecasting the *conditional* density of $\mathbf{Y}^t$, given the information $\mathcal{F}_{t-1}$ generated by $\mathbf{Y}^{t-1}$. Let $\varphi_t(y_t) \equiv \varphi_t(y_t|\mathcal{F}_{t-1}) \equiv g^t(\mathbf{y}^t)/g^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\varphi_1(y_1) \equiv \varphi_1(y_1|\mathcal{F}_0) \equiv g^1(\mathbf{y}^1) = g^1(y_1)$. Thus the goal is to forecast the (true, unknown) conditional density $\varphi_t(y_t)$.

For this, we use an one-step-ahead conditional density forecast model $\psi_t(y_t; \boldsymbol{\theta}) \equiv \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) \equiv f^t(\mathbf{y}^t)/f^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\psi_1(y_1) \equiv \psi_1(y_1|\mathcal{F}_0) \equiv f^1(\mathbf{y}^1) = f^1(y_1)$. If $\psi_t(y_t; \boldsymbol{\theta}_0) = \varphi_t(y_t)$ almost surely for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified, and it is said to be optimal because it dominates all other density forecasts for any loss functions as discussed in the previous section (see [35,67,70]).

In practice, it is rarely the case that we can find an optimal model. As it is very likely that "the true distribution is in fact too complicated to be represented by a simple mathematical function" [133], all the models proposed by different researchers can be possibly misspecified and thereby we regard each model as an approximation to the truth. Our task is then to investigate which density forecast model can approximate the true conditional density most closely. We have to first define a metric to measure the distance of a given model to the truth, and then compare different models in terms of this distance.

The adequacy of a density forecast model can be measured by the conditional Kullback-Leibler [97] Information Criterion (KLIC) divergence measure between two conditional densities,

$$\mathbb{I}_t(\varphi : \psi, \boldsymbol{\theta}) = \mathbb{E}_{\varphi_t}[\ln \varphi_t(y_t) - \ln \psi_t(y_t; \boldsymbol{\theta})] ,$$

where the expectation is with respect to the true conditional density $\varphi_t(\cdot|\mathcal{F}_{t-1})$, $\mathbb{E}_{\varphi_t} \ln \varphi_t(y_t|\mathcal{F}_{t-1}) < \infty$, and $\mathbb{E}_{\varphi_t} \ln \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) < \infty$. Following White [159], we define the distance between a density model and the true density as the minimum of the KLIC

$$\mathbb{I}_t(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*) = \mathbb{E}_{\varphi_t}\left[\ln \varphi_t(y_t) - \ln \psi_t(y_t; \boldsymbol{\theta}_{t-1}^*)\right] ,$$

where $\boldsymbol{\theta}_{t-1}^* = \arg\min \mathbb{I}_t(\varphi : \psi, \boldsymbol{\theta})$ is the pseudo-true value of $\boldsymbol{\theta}$ [133]. We assume that $\boldsymbol{\theta}_{t-1}^*$ is an interior point of $\boldsymbol{\Theta}$. The smaller this distance is, the closer the density forecast $\psi_t(\cdot|\mathcal{F}_{t-1}; \boldsymbol{\theta}_{t-1}^*)$ is to the true density $\varphi_t(\cdot|\mathcal{F}_{t-1})$.

However, $\mathbb{I}_t(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*)$ is unknown since $\boldsymbol{\theta}_{t-1}^*$ is not observable. We need to estimate $\boldsymbol{\theta}_{t-1}^*$. If our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we split the data into two parts, one for estimation and the other for out-of-sample validation. At each period t in the out-of-sample period ($t = R + 1, \ldots, T$), we estimate the unknown parameter vector $\boldsymbol{\theta}_{t-1}^*$ and denote the estimate as $\hat{\boldsymbol{\theta}}_{t-1}$. Using $\{\hat{\boldsymbol{\theta}}_{t-1}\}_{t=R+1}^{T}$, we can obtain the out-of-sample estimate of $\mathbb{I}_t(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*)$ by

$$\mathbb{I}_P(\varphi : \psi) \equiv \frac{1}{P} \sum_{t=R+1}^{T} \ln[\varphi_t(y_t)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})]$$

where $P = T - R$ is the size of the out-of-sample period. Note that

$$\mathbb{I}_P(\varphi : \psi) = \frac{1}{P} \sum_{t=R+1}^{T} \ln\left[\varphi_t(y_t)/\psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)\right] + \frac{1}{P} \sum_{t=R+1}^{T} \ln[\psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})] ,$$

where the first term in $\mathbb{I}_P(\varphi : \psi)$ measures model uncertainty (the distance between the optimal density $\varphi_t(y_t)$ and the model $\psi_t(y_t; \boldsymbol{\theta}_{t-1}^*)$) and the second term measures parameter estimation uncertainty due to the distance between $\boldsymbol{\theta}^*_{t-1}$ and $\hat{\boldsymbol{\theta}}_{t-1}$.

Since the KLIC measure takes on a smaller value when a model is closer to the truth, we can regard it as a loss function and use $\mathbb{I}_P(\varphi : \psi)$ to formulate the loss-differential. The out-of-sample average of the loss-differential between model 1 and model 2 is

$$\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2) = \frac{1}{P}\sum_{t=R+1}^{T} \ln\left[\psi_t^2\left(y_t; \hat{\boldsymbol{\theta}}^2_{t-1}\right) / \psi_t^1\left(y_t; \hat{\boldsymbol{\theta}}^1_{t-1}\right)\right] ,$$

which is the ratio of the two predictive log-likelihood functions. With treating model 1 as a benchmark model (for model selection) or as the model under the null hypothesis (for hypothesis testing), $\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2)$ can be considered as a loss function to minimize. To sum up, the KLIC differential can serve as a *loss* function for density forecast evaluation as discussed in Bao, Lee, and Saltoglu [10]. See Corradi and Swanson [31] for the related ideas using different loss functions.

Using the KLIC divergence measure to characterize the extent of misspecification of a forecast model, Bao, Lee, and Saltoglu [10], in an empirical study with the S&P500 and NASDAQ daily return series, find strong evidence for rejecting the Normal-GARCH benchmark model, in favor of the models that can capture skewness in the conditional distribution and asymmetry and long-memory in the conditional variance. Also, Bao and Lee [8] investigate the nonlinear predictability of stock returns when the density forecasts are evaluated/compared instead of the conditional mean point forecasts. The conditional mean models they use for the daily closing S&P500 index returns include the martingale difference model, the linear ARMA models, the STAR and SETAR models, the ANN model, and the polynomial model. Their empirical findings suggest that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models while we find no such evidence for the left tails or the entire distribution.

Conclusions

In this article we have selectively reviewed the state-of-the-art in nonlinear time series models that are useful in forecasting financial variables. Overall financial returns are difficult to forecast, and this may just be a reflection of the efficiency of the markets on processing information. The success of nonlinear time series on producing better forecasts than linear models depends on how persistent the nonlinearities are in the data. We should note that though many of the methodological developments are concerned with the specification of the conditional mean and conditional variance, there is an active area of research investigating other aspects of the conditional density – quantiles, directions, intervals – that seem to be promising from a forecasting point of view.

For a more extensive coverage to complement this review, the readers may find the following additional references useful. Campbell, Lo, and MacKinlay [22], Chapter 12, provides a brief but excellent summary of nonlinear time series models for the conditional mean and conditional variance as well and various methods such as ANN and nonparametric methods. Similarly, the interested readers may also refer to the books and monographs of Granger and Teräsvirta [72], Franses and van Dijk [55], Fan and Yao [52], Tsay [153], Gao [57], and some book chapters such as Stock [139], Tsay [152], Teräsvirta [145], and White [161].

Future Directions

Methodological developments in nonlinear time series have happened without much guidance from economic theory. Nonlinear models are for most part ad hoc specifications that, from a forecasting point of view, are validated according to some statistical loss function. Though we have surveyed some articles that employ some economic rationale to evaluate the model and/or the forecast – bull/bear cycles, utility function, profit/loss function –, there is still a vacuum on understanding why, how, and when nonlinearities may show up in the data.

From a methodological point of view, future developments will focus on multivariate nonlinear time series models and their associated statistical inference. Nonlinear VAR-type models for the conditional mean and high-dimensional multivariate volatility models are still in their infancy. Dynamic specification testing in a multivariate setting is paramount to the construction of a multivariate forecast and though multivariate predictive densities are inherently difficult to evaluate, they are most important in financial economics.

Another area of future research will deal with the econometrics of a data-rich environment. The advent of large databases begs the introduction of new techniques and methodologies that permits the reduction of the many dimensions of a data set to a parsimonious but highly informative set of variables. In this sense, criteria on how to combine information and how to combine models to produce more accurate forecasts are highly desirable.

Finally, there are some incipient developments on defining new stochastic processes where the random variables that form the process are of a symbolic nature, i. e. intervals, boxplots, histograms, etc. Though the mathematics of these processes are rather complex, future developments in this area will bring exciting results for the area of forecasting.

Bibliography

1. Ait-Sahalia Y, Hansen LP (2009) Handbook of Financial Econometrics. Elsevier Science, Amsterdam
2. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39(4):885–905
3. Andersen TG, Bollerslev T, Diebold FX, Labys P (2001) The Distribution of Realized Exchange Rate Volatility. J Amer Stat Assoc 96:42–55
4. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71: 579–625
5. Ang A, Bekaert G (2002) Regime Switccheds in Interest Rates. J Bus Econ Stat 20:163–182
6. Bachelier L (1900) Theory of Speculation. In: Cootner P (ed) The Random Character of Stock Market Prices. MIT Press, Cambridge. (1964) reprint
7. Bai J, Ng S (2007) Forecasting Economic Time Series Using Targeted Predictors. Working Paper, New York University and Columbia University
8. Bao Y, Lee TH (2006) Asymmetric Predictive Abilities of Nonlinear Models for Stock Returns: Evidence from Density Forecast Comparison. Adv Econ 20 B:41–62
9. Bao Y, Lee TH, Saltoglu B (2006) Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check. J Forecast 25(2):101–128
10. Bao Y, Lee TH, Saltoglu B (2007) Comparing Density Forecast Models. J Forecast 26(3):203–225
11. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models. J Royal Stat Soc B 64:853–223
12. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH Models: A Survey. J Appl Econ 21:79–109
13. Bera AK, Higgins ML (1993) ARCH Models: Properties, Estimation, and Testing. J Econ Surv 7:305–366
14. Bollerslev T (1986) Generalized Autoregressive Conditional Heteroskedasticity. J Econ 31:307–327
15. Bollerslev T, Chou RY, Kroner KF (1992) ARCH Models in Finance. J Econ 52:5–59
16. Bollerslev T, Engle RF, Nelson DB (1994) ARCH Models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier Science, Amsterdam
17. Bollerslev T, Engle RF, Wooldridge J (1988) A Capital Asset Pricing Model with Time Varying Covariances. J Political Econ 96:116–131
18. Boero G, Marrocu E (2004) The Performance of SETAR Models: A Regime Conditional Evaluation of Point, Interval, and Density Forecasts. Int J Forecast 20:305–320
19. Breiman L (1996) Bagging Predictors. Machine Learning 24:123–140
20. Breiman L (1996) Heuristics of Instability and Stabilization in Model Selection. Ann Stat 24(6):2350–2383
21. Brooks C, Persand G (2003) Volatility Forecasting for Risk Management. J Forecast 22(1):1–22
22. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, New Jersey
23. Campbell JY, Thompson SB (2007) Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? Harvard Institute of Economic Research, Discussion Paper No. 2084
24. Cai Z, Fan J, Yao Q (2000) Functional-coefficient Regression Models for Nonlinear Time Series. J Amer Stat Assoc 95: 941–956
25. Chen X (2006) Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman JJ, Leamer EE (eds) Handbook of Econometrics, vol 6. Elsevier Science, Amsterdam, Chapter 76
26. Chen R, Tsay RS (1993) Functional-coefficient Autoregressive Models. J Amer Stat Assoc 88:298–308
27. Christofferson PF, Diebold FX (2006) Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics. Manag Sci 52:1273–1287
28. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-linear Models. Int J Forecast 20:169–183
29. Clements MP, Smith J (2000) Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. J Forecast 19:255–276
30. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatter Plots. J Amer Stat Assoc 74: 829–836
31. Corradi V, Swanson NR (2006) Predictive Density Evaluation. In: Granger CWJ, Elliot G, Timmerman A (eds) Handbook of Economic Forecasting. Elsevier, Amsterdam, pp 197–284
32. Corsi F (2004) A Simple Long Memory Model of Realized Volatility. Working Paper, University of Lugano
33. Dahl CM, González-Rivera G (2003) Testing for Neglected Nonlinearity in Regression Models based on the Theory of Random Fields. J Econ 114:141–164
34. Dahl CM, González-Rivera G (2003) Identifying Nonlinear Components by Random Fields in the US GNP Growth. Implications for the Shape of the Business Cycle. Stud Nonlinear Dyn Econ 7(1):art2
35. Diebold FX, Gunther TA, Tay AS (1998) Evaluating Density Forecasts with Applications to Financial Risk Management. Int Econ Rev 39:863–883
36. Diebold FX, Li C (2006) Forecasting the Term Structure of Government Bond Yields. J Econom 130:337–364
37. Diebold FX, Mariano R (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–265
38. Diebold FX, Rudebusch GD (1989) Scoring the Leading Indicators. J Bus 62(3):369–391
39. Ding Z, Granger CWJ, Engle RF (1993) A Long Memory Property of Stock Market Returns and a New Model. J Empir Finance 1:83–106
40. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Returns? J Bank Finance 31:279–296
41. Durland JM, McCurdy TH (1994) Duration-Dependent Transi-

tions in a Markov Model of US GNP Growth. J Bus Econ Stat 12:279–288

42. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least Angle Regression. Ann Stat 32(2):407–499
43. Engel C, Hamilton JD (1990) Long Swings in the Dollar: Are they in the Data and Do Markets Know it? Amer Econ Rev 80(4):689–713
44. Engle RF (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation. Econometrica 50:987–1008
45. Engle RF, Hendry DF, Richard J-F (1983) Exogeneity. Econometrica 51:277–304
46. Engle RF, Kane A, Noh J (1997) Index-Option Pricing with Stochastic Volatility and the Value of Accurate Variance Forecasts. Rev Deriv Res 1:139–157
47. Engle RF, Manganelli S (2004) CaViaR: Conditional autoregressive Value at Risk by regression quantiles. J Bus Econ Stat 22(4):367–381
48. Engle RF, Ng VK, Rothschild M (1990) Asset Pricing with a Factor ARCH Covariance Structure: Empirical Estimates for Treasury Bills. J Econ 45:213–238
49. Engle RF, Russell JR (1998) Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66:1127–1162
50. Fan J, Gijbels I (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London
51. Fan J, Yao Q (1998) Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85: 645–660
52. Fan J, Yao Q (2003) Nonlinear Time Series. Springer, New York
53. Fan J, Yao Q, Cai Z (2003) Adaptive varying-coefficient linear models. J Royal Stat Soc B 65:57–80
54. Fitzenberger B (1997) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. J Econ 82:235–287
55. Franses PH, van Dijk D (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge
56. Gallant AR, Nychka DW (1987) Semi-nonparametric maximum likelihood estimation. Econometrica 55:363–390
57. Gao J (2007) Nonlinear Time Series: Semiparametric and Nonparametric Methods. Chapman and Hall, Boca Raton
58. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get Most out of Returns Data Sampled at Different Frequencies. J Econ 131:59–95
59. Giacomini R, White H (2006) Tests of Conditional Predictive Ability. Econometrica 74:1545–1578
60. Glosten LR, Jaganathan R, Runkle D (1993) On the Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801
61. González-Rivera G (1998) Smooth-Transition GARCH Models. Stud Nonlinear Dyn Econ 3(2):61–78
62. González-Rivera G, Lee TH, Mishra S (2004) Forecasting Volatility: A Reality Check Based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood. Int J Forecast 20(4):629–645
63. González-Rivera G, Lee TH, Mishra S (2008) Jumps in Cross-Sectional Rank and Expected Returns: A Mixture Model. J Appl Econ; forthcoming
64. González-Rivera G, Lee TH, Yoldas E (2007) Optimality of the Riskmetrics VaR Model. Finance Res Lett 4:137–145
65. Gonzalo J, Martíneza O (2006) Large shocks vs. small shocks. (Or does size matter? May be so). J Econ 135:311–347
66. Goyal A, Welch I (2006) A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Working Paper, Emory and Brown, forthcoming in Rev Financ Stud
67. Granger CWJ (1999) Outline of Forecast Theory Using Generalized Cost Functions. Span Econ Rev 1:161–173
68. Granger CWJ (2002) Some Comments on Risk. J Appl Econ 17:447–456
69. Granger CWJ, Newbold P (1986) Forecasting Economic Time Series, 2nd edn. Academic Press, San Diego
70. Granger CWJ, Pesaran MH (2000) A Decision Theoretic Approach to Forecasting Evaluation. In: Chan WS, Li WK, Tong H (eds) Statistics and Finance: An Interface. Imperial College Press, London
71. Granger CWJ, Lee TH (1999) The Effect of Aggregation on Nonlinearity. Econ Rev 18(3):259–269
72. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, New York
73. Guidolin M, Timmermann A (2006) An Econometric Model of Nonlinear Dynamics in the Joint Distribution of Stock and Bond Returns. J Appl Econ 21:1–22
74. Haggan V, Ozaki T (1981) Modeling Nonlinear Vibrations Using an Amplitude-dependent Autoregressive Time Series Model. Biometrika 68:189–196
75. Hamilton JD (1994) Time Series Analysis. Princeton University Press, New Jersey
76. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57:357–384
77. Hamilton JD (1996) Specification Testing in Markov-Switching Time Series Models. J Econ 70:127–157
78. Hamilton JD (2001) A Parametric Approach to Flexible Nonlinear Inference. Econometrica 69:537–573
79. Hamilton JD, Jordà O (2002) A Model of the Federal Funds Target. J Political Econ 110:1135–1167
80. Hansen BE (1996) Inference when a Nuisance Parameter is not Identified under the Null Hypothesis. Econometrica 64: 413–430
81. Hansen PR (2005) A test for superior predictive ability. J Bus Econ Stat 23:365–380
82. Harding D, Pagan A (2002) Dissecting the Cycle: A Methodological Investigation. J Monet Econ 49:365–381
83. Härdle W, Tsybakov A (1997) Local polynomial estimators of the volatility function in nonparametric autoregression. J Econ 81:233–242
84. Harter HL (1977) Nonuniqueness of Least Absolute Values Regression. Commun Stat – Theor Methods A6:829–838
85. Hong Y, Chung J (2003) Are the Directions of Stock Price Changes Predictable? Statistical Theory and Evidence. Working Paper, Department of Economics, Cornell University
86. Hong Y, Lee TH (2003) Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models. Rev Econ Stat 85(4):1048–1062
87. Hong Y, Lee TH (2003b) Diagnostic Checking for Adequacy of Nonlinear Time Series Models. Econ Theor 19(6):1065–1121
88. Hornik K, Stinchcombe M, White H (1989) Multi-Layer Feedforward Networks Are Universal Approximators. Neural Netw 2:359–366
89. Huang H, Tae-Hwy L, Canlin L (2007) Forecasting Output Growth and Inflation: How to Use Information in the Yield

Curve. Working Paper, University of California, Riverside, Department of Economics

90. Huang YL, Kuan CM (2007) Re-examining Long-Run PPP under an Innovation Regime Switching Framework. Academia Sinica, Taipei
91. Inoue A, Kilian L (2008) How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US CPI Inflation, forthcoming. J Amer Stat Assoc 103(482):511–522
92. Jackwerth JC, Rubinstein M (1996) Recovering probability distributions from option prices. J Finance 51:1611–1631
93. Judd KL (1998) Numerical Methods in Economics. MIT Press, Cambridge
94. Kanas A (2003) Non-linear Forecasts of Stock Returns. J Forecast 22:299–315
95. Koenker R, Bassett Jr G (1978) Regression Quantiles. Econometrica 46(1):33–50
96. Kuan CM, Huang YL, Tsayn RS (2005) An unobserved component model with switching permanent and transitory innovations. J Bus Econ Stat 23:443–454
97. Kullback L, Leibler RA (1951) On Information and Sufficiency. Ann Math Stat 22:79–86
98. Lee TH, Ullah A (2001) Nonparametric Bootstrap Tests for Neglected Nonlinearity in Time Series Regression Models. J Nonparametric Stat 13:425–451
99. Lee TH, White H, Granger CWJ (1993) Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests. J Econ 56:269–290
100. Lee TH, Yang Y (2006) Bagging Binary and Quantile Predictors for Time Series. J Econ 135:465–497
101. Lettau M, Ludvigson S (2001) Consumption, Aggregate Wealth, and Expected Stock Returns. J Finance 56:815–849
102. Lewellen J (2004) Predicting Returns with Financial Ratios. J Financial Econ 74:209–235
103. Lintner J (1965) Security Prices, Risk and Maximal Gains from Diversification. J Finance 20:587–615
104. Linton O, Whang YJ (2007) A Quantilogram Approach to Evaluating Directional Predictability. J Econom 141:250-282
105. Lo AW, MacKinlay AC (1999) A Non-Random Walk Down Wall Street. Princeton University Press, Princeton
106. Long X, Su L, Ullah A (2007) Estimation and Forecasting of Dynamic Conditional Covariance: A Semiparametric Multivariate Model. Working Paper, Department of Economics, UC Riverside
107. Lopez JA (2001) Evaluating the Predictive Accuracy of Volatility Models. J Forecast 20:87–109
108. Ludvigson S, Ng S (2007) The Empirical Risk Return Relation: A Factor Analysis Approach. J Financ Econ 83:171–222
109. Lundbergh S, Teräsvirta T (2002) Forecasting with smooth transition autoregressive models. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 21
110. Luukkonen R, Saikkonen P, Teräsvirta T (1988) Testing Linearity in Univariate Time Series Models. Scand J Stat 15:161–175
111. Maheu JM, McCurdy TH (2000) Identifying Bull and Bear Markets in Stock Returns. J Bus Econ Stat 18:100–112
112. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3(3):205–228
113. Markowitz H (1959) Portfolio Selection: Efficient Diversification of Investments. John Wiley, New York
114. Marsh IW (2000) High-frequency Markov Switching Models in the Foreign Exchange Market. J Forecast 19:123–134
115. McAleer M, Medeiros MC (2007) A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. J Econ; forthcoming
116. Money AH, Affleck-Graves JF, Hart ML, Barr GDI (1982) The Linear Regression Model and the Choice of p. Commun Stat – Simul Comput 11(1):89–109
117. Nyguist H (1983) The Optimal L_p-norm Estimation in Linear Regression Models. Commun Stat – Theor Methods 12: 2511–2524
118. Nadaraya ÉA (1964) On Estimating Regression. Theor Probab Appl 9:141–142
119. Nelson CR, Siegel AF (1987) Parsimonious Modeling of Yield Curves. J Bus 60:473–489
120. Nelson DB (1991) Conditional Heteroscedasticity in Asset Returns: A New Approach. Econometrica 59(2):347–370
121. Noh J, Engle RF, Kane A (1994) Forecasting Volatility and Option Prices of the S&P 500 Index. J Deriv 17–30
122. Pagan AR, Ullah A (1999) Nonparametric Econometrics. Cambridge University Press, Cambridge
123. Patton AJ, Timmermann A (2007) Testing Forecast Optimality Under Unknown Loss. J Amer Stat Assoc 102(480): 1172–1184
124. Perez-Quiros G, Timmermann A (2001) Business Cycle Asymmetries in Stock Returns: Evidence form Higher Order Moments and Conditional Densities. J Econ 103:259–306
125. Poon S, Granger CWJ (2003) Forecasting volatility in financial markets. J Econ Lit 41:478–539
126. Powell JL (1986) Censored Regression Quantiles. J Econ 32:143–155
127. Priestley MB (1980) State-dependent Models: A General Approach to Nonlinear Time Series Analysis. J Time Ser Anal 1:47–71
128. Raj B, Ullah A (1981) Econometrics: A Varying Coefficients Approach. Croom Helm, London
129. Riskmetrics (1995) Technical Manual, 3rd edn. New York
130. Romano JP, Wolf M (2005) Stepwise multiple testing as formalized data snooping. Econometrica 73:1237–1282
131. Ross S (1976) The Arbitrage Theory of Capital Asset Pricing. J Econ Theor 13:341–360
132. Ruppert D, Wand MP (1994) Multivariate Locally Weighted Least Squares Regression. Ann Stat 22:1346–1370
133. Sawa T (1978) Information Criteria for Discriminating among Alternative Regression Models. Econometrica 46:1273–1291
134. Schwert GW (1990) Stock Volatility and the Crash of '87. Rev Financ Stud 3(1):77–102
135. Sentana E (1995) Quadratic ARCH models. Rev Econ Stud 62(4):639–661
136. Sichel DE (1994) Inventories and the Three Phases of the Business Cycle. J Bus Econ Stat 12:269–277
137. Sharpe W (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. J Finance 19:425–442
138. Stinchcombe M, White H (1998) Consistent Specification Testing with Nuisanse Parameters Present only under the Alternative. Econ Theor 14:295–325
139. Stock JH (2001) Forecasting Economic Time Series. In: Baltagi BP (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, Chapter 27
140. Stock JH, Watson MW (2002) Forecasting Using Principal Components from a Large Number of Predictors. J Amer Stat Assoc 97:1167–1179
141. Stock JH, Watson MW (2006) Forecasting Using Many Predic-

tors. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam

142. Stone CJ (1977) Consistent Nonparametric Regression. Ann Stat 5:595–645
143. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
144. Teräsvirta T (1994) Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models. J Amer Stat Assoc 89:208–218
145. Teräsvirta T (2006) Forecasting economic variables with nonlinear models. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, pp 413–457
146. Teräsvirta T, Anderson H (1992) Characterizing Nonlinearities in Business Cycles using Smooth Transition Autoregressive Models. J Appl Econ 7:119–139
147. Teräsvirta T, Lin CF, Granger CWJ (1993) Power of the Neural Network Linearity Test. J Time Ser Analysis 14:209–220
148. Tong H (1983) Threshold Models in Nonlinear Time Series Analysis. Springer, New York
149. Tong H (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford
150. Trippi R, Turban E (1992) Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance. McGraw-Hill, New York
151. Tsay RS (1998) Testing and Modeling Multivariate Threshold Models. J Amer Stat Assoc 93:1188–1202
152. Tsay RS (2002) Nonlinear Models and Forecasting. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 20
153. Tsay RS (2005) Analysis of Financial Time Series, 2nd edn. Wiley, New York
154. Varian HR (1975) A Bayesian Approach to Real Estate Assessment. In: Fienberg SE, Zellner A (eds) Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage. North Holland, Amsterdam, pp 195–208
155. Watson GS (1964) Smooth Regression Analysis. Sankhya Series A 26:359–372
156. West KD (1996) Asymptotic inference about predictive ability. Econometrica 64:1067–1084
157. West KD, Edison HJ, Cho D (1993) A Utility Based Comparison of Some Models of Exchange Rate Volatility. J Int Econ 35: 23–45
158. White H (1989) An Additional Hidden Unit Tests for Neglected Nonlinearity in Multilayer Feedforward Networks. In: Proceedings of the International Joint Conference on Neural Networks, Washington, DC. IEEE Press, New York, II, pp 451–455
159. White H (1994) Estimation, Inference, and Specification Analysis. Cambridge University Press, Cambridge
160. White H (2000) A Reality Check for Data Snooping. Econometrica 68(5):1097–1126
161. White H (2006) Approximate Nonlinear Forecasting Methods. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, chapter 9
162. Zellner A (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. J Amer Stat Assoc 81:446–451
163. Ziegelmann FA (2002) Nonparametric estimation of volatility functions: the local exponential estimator. Econ Theor 18:985–991