Order Number 9513451

Autoregressive pursuit: A new approach to nonstationarity in macroeconomics and finance

Orszag, J. Michael, Ph.D.

The University of Michigan, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

AUTOREGRESSIVE PURSUIT: A NEW APPROACH TO NONSTATIONARITY IN MACROECONOMICS AND FINANCE

by

J. Michael Orszag

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
1994

Doctoral Committee:

Professor Carl P. Simon, Chair
Professor Paul Federbush
Professor E. Phillip Howrey
Professor Saul Hymans

To my parents.

ACKNOWLEDGEMENTS

Carl Simon (Univ. of Michigan), E.P. Howrey (Univ.of Michigan), Saul Hymans (Univ. of Michigan), Stephane Mallat (Courant Institute), , Paul Federbush (Univ. of Michigan), Hong Yang (Princeton University), and Zhifeng Zhang (Stanford University) provided helpful comments or suggestions on various portions of this work. My thesis committee was instrumental in improving the presentation and readability of the thesis.

In addition, I would like to thank Geoff Davis(New York University), Isaac Goldhirsch (Tel Aviv University), Wen Liang Hwang (Univ. of California, Irvine), Jong Hwa Kim (Univ. of Michigan), Peter Orszag, Pasquale Rocco (Michigan), Jonathan Raz (Michigan), Ananias Tomboulides (Princeton University), Uzi Zrahia (Princeton University) and the faculty and fellow students at Michigan for conversations. Funding for research was provided by a Sloan dissertation fellowship. Some of the preliminary ideas were developed when the author was visiting Mallat at the Courant Institute of Mathematical Sciences at New York University. The Yale wavelet packet software of Coifman and Wickerhauser was used for some of the comparisons in the thesis.

Some results from the thesis were presented at the 1993 North American Summer meetings of the Econometric Society.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

THE PROBLEM

This thesis presents a new approach to the analysis of nonstationary macroeconomic and financial time series. By nonstationarity, we mean here specifically situations in which the data generating process is time-dependent. The problem of nonstationarity is thus quite broad and we cannot hope to deal thoroughly with all possible forms of nonstationarity here. Instead, we focus on situations in which the conditional expectation of the time series is *linear* in its lag variables. We have no doubt that the approach here can be extended in various ways to more complex settings and we suggest ways the basic approach can be extended in a chapter on extensions towards the end of the thesis. However, since the problem of estimating nonstationary time series models is a difficult one, we have thought it best to focus clearly on a relatively specific setting and develop the methods through instructive examples, theory and comparisons with the existing literature.

To introduce our specific formulation, we recall that most popular parametric methods of time series analysis in a variety of fields including engineering, medicine, oceanography, geophysics and economics, assume at least in a weak sense time-invariance of statistical relations. Both Box-Jenkins ARMA models and spectral analysis rely on the same basic principle: that univariate data can be modelled as if they were generated by an autoregressive linear process so that:

$$y(t) = \sum_{j=1}^{J} \beta_j y(t-j) + \epsilon(t) \tag{I.1}$$

where $y(t)$ is the time series, β_j is a time invariant set of parameters such that the complex series $1 - \sum_j \beta_j z^{-j}$ has all its zeroes strictly inside the unit circle and $\epsilon(t)$ is a normally and independently distributed noise term.[1] The general model can be viewed as a first-order Taylor series approximation to an arbitrary nonlinear time series model where the function generating the data is not explicitly a function of time. Methods for linear stationary time series models such as Eq. (I.1) and the appropriate multivariate extensions are surveyed in [33] [90] [169] [126] [181] [36]. Appendix B reviews some technical background related to these models.

In this thesis, we are primarily concerned with the problem of estimating Equation (I.1) where the parameters β_j are time-varying. Specifically, we shall develop a general approach for estimating a *time-varying* autoregressive model:

$$y(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon(t) \tag{I.2}$$

in which the functional forms of the true $\beta_j(t)$ are not known. For the most part, we restrict attention to the case in which the operator $1 - \beta_j(t)L^j$ (where L is the lag operator) has a bounded inverse. Since the representation is causal, only positive lags are considered and $\epsilon(t) \sim N(0, \sigma^2)$. The problem appears untractable. If we were to consider all possible $\beta_j(t)$ functions and estimate this equation directly there would seem to be no possible way to achieve consistent estimates. In addition, the redundancy of the $\beta_j(t)$ may result in a very poorly conditioned regressor matrix. We would also expect similar problems with maximum likelihood estimation.

In this thesis, we develop a technique for estimating Eq. (I.2). The framework

[1] For some theoretical results, it is necessary to impose more restrictive assumptions such as absolute summability of the moving average representation.

shares many of the advantages of nonparametric analysis and reduces to the standard stationary time series methodology as a special case. The model we are concerned with, summarized in Eq. (I.2), seems somewhat simple and idealized. However, such a model provides a first approach to the more general problem of nonstationarity which is of potential importance in macroeconomics and finance. An important reason for studying time-varying parameters is that the parameters of the models may depend in some unknown way on some state variables $x(t)$ for the system. A natural extension of Eq. (I.2) is:

$$y(t) = \sum_{j=1}^{J} \beta_j\left(x(t)\right) y(t-j) + \epsilon(t). \tag{I.3}$$

Eq. (I.2) is a special case of Eq. (I.3) where time is the only state variable entering the functions β_j. In addition to the issues of estimation of time-varying relationships between variables, the thesis also takes the further step of showing how to use the estimates in Eq. (I.2) and Eq. (I.3) to define parametric estimates of a *nonstationary* spectrum and relates these ideas to other approaches in the economics and signal processing literatures.

Our analysis makes the twin assumptions of linearity and nonstationarity. Hence, the problem we are considering is only of interest in economics and finance if it is reasonable in practical applications to ignore nonlinearity *and* if nonstationarity is a feature endemic to economic data and of theoretical interest.

To see whether our assumptions are relevant in practice, we will consider some illustrative examples from macroeconomics and finance. We consider first Fig. (I.1) which plots growth rates of GNP versus their lags; from the plot, the relationship between GNP and its lag appears quite noisy. The straight line in Fig. (I.1) represents the statistical relationship between lagged growth rates on the x axis and

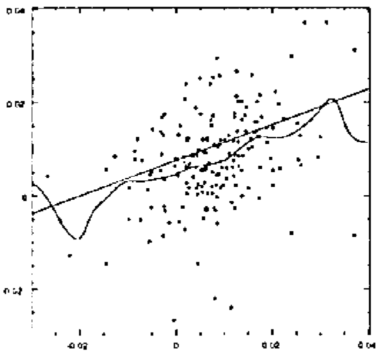

Figure I.1: Plot of gnp growth rates (vertical axis) versus lags (CITIBASE GNPQ quarterly data series for real GNP). Kernel and ordinary autoregressive estimates are superimposed.

contemporaneous growth rates on the y axis.[2] Superimposed in Fig. (I.1) are kernel estimates[3] of the relationship between lagged growth rates and growth rates of GNP; if nonlinearity were important statistically, the kernel estimates would differ significantly from the ordinary autoregressive estimates. While kernel regression estimates may change GNP growth forecasts by several tenths of a percentage point, simulations indicate that kernel estimates for data generated by a parametric model may look approximately the same. The implication of Fig. (I.1) seems to be that our assumption of linearity does not seem to be contradicted strongly by the data.

To examine the relevance of the assumption of nonstationarity with the same dataset, we have constructed in Fig. (I.2) a crude time-varying kernel estimate of the correlation between GNP growth rates and their first lag by considering local

[2] Examination of the autocorrelation function of GNP growth rates (Citibase series GNPQ) indicates that an autoregressive model is more appropriate than a moving average model. With a first order autoregressive model, the data in Fig. (I.1) would be clustered around a the straight line representing autoregressive estimates in Fig. (I.1).

[3] A kernel estimate is a type of local least squares estimate which seeks to capture local features in the data (for an applied introduction to kernel estimation for economists, see [100]). Our estimates were constructed using a Gaussian kernel with $\sigma = 0.0025$ (standard deviation of one quarter of a percentage point).

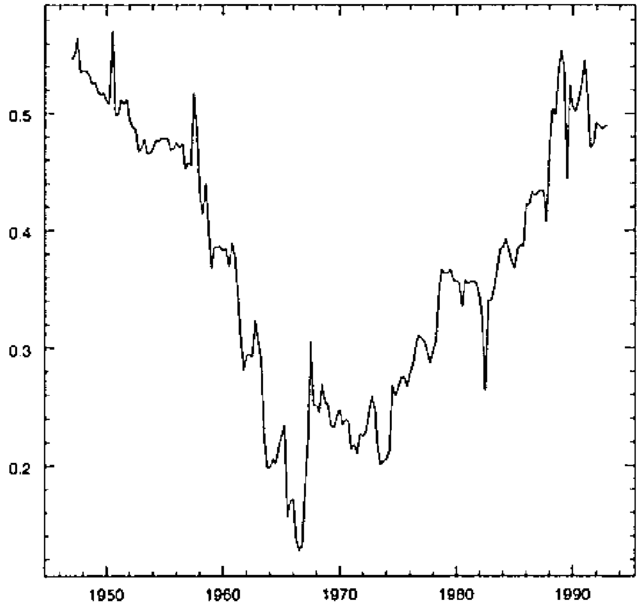

Figure I.2: Plot of estimates of the correlation between GNP growth rates and the lagged growth rate as a function of time.

averages.[4] The estimates of the first lag coefficient in Fig. (I.2) vary from over 0.5 to slightly over 0.1; structural change of this magnitude implies that the effects of shocks vary over time and that forecasts from a model with time-varying parameters would differ somewhat from forecasts from a stationary time series model. The implication of Fig. (I.2) is that there is a presumption that nonstationarity may be of interest in practice.

However, one important question to ask is: are calculations such as those which lead to Fig. (I.2) reliable? Most definitely not; since Fig. (I.2) is constructed using local least squares estimates, the confidence intervals around the parameter estimates

[4] Moving average estimates were used; Gaussian kernels generate roughly similar results but the estimates are smoother. We consider 60 point local averages for both the numerator and denominator of the least squares estimate of the regression of the growth rate of GNP on its lag. These estimates have an unusual statistical property which is reviewed in Appendix E.

can be easily computed and they are wide.[5] Even though there is considerable variation in the estimated correlations, it is entirely possible that such variation is statistical. Indeed, if our crude kernel estimate could reliably capture the time-varying properties of economic time series such as GNP or the stock market, there would be little need for our thesis.

We now turn to an example from finance to buttress our case for our dual assumptions of linearity and nonstationarity with an example from finance. The example is of daily stock market data for the New York Stock Exchange from 1962 to 1992. Fig. (I.3) shows the data on returns[6] for 7675 trading days; a quick look at Fig. (I.3) and comparison with the simulated data in Fig. (I.4) indicates that the data is not Gaussian white noise. Evidence for deviations from white noise comes from a variety of sources including, for instance, thick tails in the data [133] [73].[7] More fundamentally, there is much evidence that means and variances of stock returns vary over time. Many important financial models such as the intertemporal asset pricing model (ICAPM) of Merton [137] are based on the assumption that means and variances are influenced by a finite number of state variables which change over time in a known manner. A wide variety of explanations for deviations from white noise have been proposed in the finance literature and include, for instance, time-varying conditional volatility [8] [70] [30] [194] and chaotic price movements. With financial

[5] In time series of this size, for instance, it is thus always likely to be unclear whether the first order autoregressive coefficient follows for instance a random walk process or is constant over time. For instance, a test we have developed for structural change called the *cumulative waveletgram*, often cannot reject (the results depend on the choice of wavelet function) at the 5% level the null hypothesis that GNP growth rates are governed by a time-invariant first order autoregressive process. Similar results are reported by Hansen [94] using a likelihood ratio test. Nevertheless, Hamilton has shown that models of time-varying parameters are useful in characterizing the peaks and troughs of business cycles [91]. Other macroeconomic time series show significant evidence of structural change.

[6] CRSP (University of Chicago Center for Research on Security Prices) cum dividend returns.

[7] Clark [41] has argued that thick tails can be explained by changing how time is measured in the market. Stock [191] has investigated similar questions with reference to the business cycle.

[8] The importance of nonstationarity in finance is underscored by the fact that time-varying con-

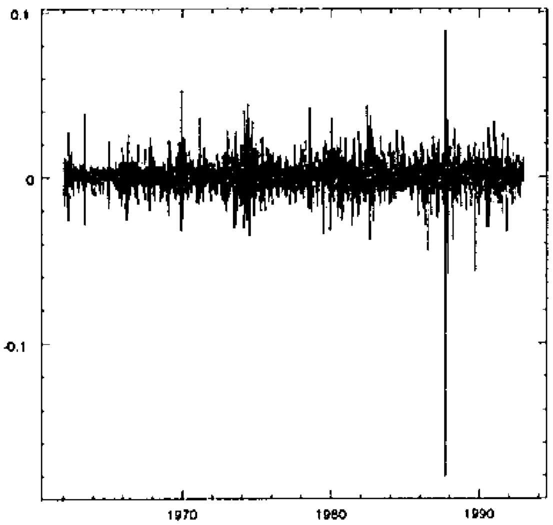

Figure I.3: Daily stock returns (Standard & Poor's value weighted including dividends).

data, therefore, nonstationarity is well-accepted in the literature so we need to focus especially on the case for linearity.

Fig. (I.5) plots the lag return (horizontal axis) against the return.[9] Since the data occurs in a temporal sequence, one possible explanation for the evolution of the data is that there is some low dimensional nonlinear model which generates the data. In some economic time series, a linear model such as a first order autoregressive model can explain much of the variation of the data, but for this asset market data, correlations are weak so that that a good time-invariant low-dimensional model (if it exists) will be nonlinear.[10] However, Fig I.5 reveals no discernible linear relationship

ditional variances have been integrated into standard asset pricing models in finance. For instance, see [136] [186] [208] [184] [5].

[9] Outliers such as the 1987 stock market crash have been removed from the plot.

[10] There has been much work in trying to discover a simple nonlinear model which determines asset returns, but success has to date been limited. A further complication is that the data generating process is unlikely to depend on only the first lag but on many lags of the data; thus, a rather high-dimensional surface may have to be estimated from noisy data. Among the methods developed to deal with the problem of estimating nonlinear relationships in time series models are threshold

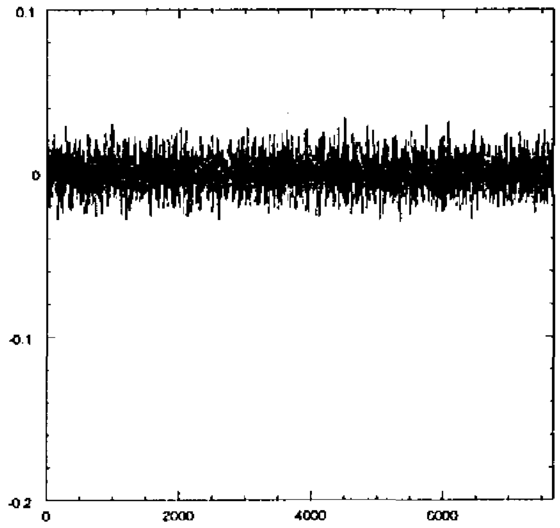

Figure I.4: Independent Gaussian noise drawn from distribution with same variance as daily stock market data.

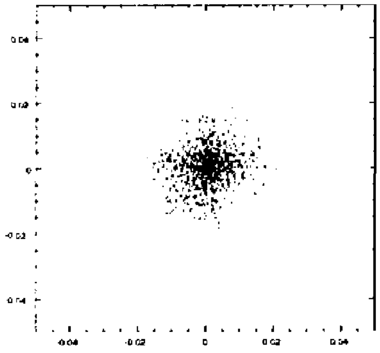

Figure I.5: Plot of daily stock returns (vertical axis) vs. lagged stock returns.

in the stock market data.

We have just reviewed some illustrative examples and have provided some empirical support for our focus of attention on nonstationarity. First, the more developed literature on nonlinear economic time series has tended to find that empirically most economic relationships, at least in the macroeconomy, do not appear to be statistically far from linearity [32].[11] While we believe these findings are premature, one possible explanation for them would be that economic agents conceptualize state variables in a linear regime as this facilitates application of control. There is no inherent problem with instability when state variables follow a linear stochastic process with time-varying coefficients. Second, there is quite a bit of evidence of various types of nonstationarity in economic time series, especially financial time series.[12] Development of nonstationary time series methods hence may be very useful in trying to understand the mechanisms underlying economic dynamics. Third, development of nonstationary time series methods has important theoretical implications for rational expectations economics which assumes stationarity of economic relationships.[13]

autoregressive methods [196], nonparametric estimation models [180], semiparametric estimation methods [24] [179] [79] [147] and bilinear time series models [175] [83] [192]. Some of these different approaches to uncovering nonlinear relationships in time series are reviewed in Tong's recent monograph [197]. Some examples can also be found in the recent Santa Fe monograph on time series prediction [203]. See also [84].

[11] There is even stronger evidence that macroeconomic time series do not exhibit low dimensional chaotic behavior [75].

[12] For instance, Bossaerts and Hillion [31] find that much of the overfitting in finance models such as for exchange rates is due to time variation in coefficients. Barsky and DeLong [13] also find that much of the reported excess variation in stock prices [206] can be explained with time-varying dividend growth forecasts. Benjamin Friedman [76] finds considerable time variation in relationships between money, interest and prices in the U.S. economy. Perron [163] finds that the results of unit root tests change dramatically when structural breaks are included in the analysis.

[13] For explicit arguments along these lines, see [124]. This argument is perhaps somewhat weaker than the others. Given the precedent of the past few decades, it seems unlikely that the macroeconomic theory of twenty years from now will look anything like the macroeconomic theory today. However, it is the case that both New Keynesian and New Classical economists make many of the same assumptions about stationarity. To some degree, as pointed out by Lucas [123], a careful reading of Modigliani and Grunberg [139] or Muth [142] leads to different conclusions about the

While the approach we will develop has certain advantages over both kernel estimates and standard parametric approaches, it will be the case that statistical analysis and inference on small datasets such as those commonly encountered in macroeconomics is subject to doubt. Since we would like the approach we develop to be applicable to both large and small datasets, we are sensitive to the issue of robustness. In some sense, both standard parametric and nonparametric approaches are extreme cases of the approach we advocate. We think in practice researchers should consider some fixed number of possible parametric models and pick the best one. As the number of models becomes large, the approach becomes fully nonparametric. As the number of possible models gets small, we come closer to the parametric case. Nevertheless, even in the case of simple stationary time series models, there are often major robustness problems in terms of properties of forecasts.[14] In the case of small datasets such as occur in macroeconomics, one needs to approach *any* time series analysis with caution.[15]

What do we have to contribute which is new to the literature? ARCH models and cointegrated models assume some form of stationarity or time-invariance; in the case of ARCH models it is the conditional variance which follows a stationary processs and in the case of cointegrated models, it is a linear combination of nonstationary variables which follows a stationary process. One possible model of nonstationarity is the hidden Markov model [173] [172] which was developed for the processing of speech signals[16] and further developed by James Hamilton [91] and applied to macroeconomic

stationarity of structural economic relationships. In the beginning of Ch. IX, we review some of the implications of nonstationarity for economic theory.

[14] For a good exposition of some of these problems, see [146]. Useful examples are in [202].

[15] On the other hand, one reason cited for the use of time series methods in macroeconomics and finance is concern about the robustness and empirical predictive performance [144] of large econometric models with many more parameters. The newer structural econometric models based on Euler equations [97] have even more serious problems empirically [80] [67] [96].

[16] This method was developed at the Communications Research Division of the Institute for De-

data. This model assumes that structural change is abrupt and there is a finite number of regimes. The major models in the literature hence make strong identifying assumptions and are not really tools for exploratory data analysis and comparing different potential models. There are good reasons for this. Nonparametric methods which make as few identifying assumptions as possible are often not practical on the short time series available in macroeconomics. Engineers require nonstationary time series methods in their work on speech analysis and other areas in nonstationary signal processing but they have generally used nonparametric methods based on the Cohen's class of estimators ([28] reviews this literature in detail; Chapter 7 contains also contains a section which is a review of part of this literature).

In the thesis, we thus suggest a framework which represents a compromise between a fully nonparametric approach and a standard parametric approach. We do this by considering a possibly large family of parametric models and using a fast algorithm to search for the best possible parametric model. The method we use is in some sense a combination of the pathbreaking Matching Pursuit algorithm developed independently by Mallat and Zhang [131] and Qian and Chen [171] for the analysis of functions in terms of waveforms with ideas from projection pursuit regression [77] in statistics. The approach is also related to sieve regression [85] [207] and neural network models [141] [102] for time series. We hope that one contribution of the thesis is to show that such sophisticated methods which were developed to estimate *time-invariant nonlinear* relations are also relevant for estimation of *time-varying linear* relations.

fense Analysis in Princeton and at Carnegie-Mellon University by Raj Reddy. This method is reviewed below in our literature review on nonstationary models.

CHAPTER II

FORMULATION OF PROBLEM

We recall from Ch. I that we are interested in estimating the regression equation:

$$y(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \tag{II.1}$$

where $\epsilon_1(t)$ is a serially independent and identically distributed noise term. We do not know the functions $\beta_j(t)$ and wish to estimate them. We now sketch our approach to estimating Eq. (II.1). We assume Eq. (II.1) can be rewritten as:

$$y(t) = \sum_{k=1}^{K} \alpha_k h_k(t) + \epsilon_2(t) \tag{II.2}$$

where:

(1) the α_k are parameters,

(2) the h_k are referred to as model components,

(3) $\epsilon_2(t)$ is a serially independent and identically distributed noise term.

The concept of a model component will be fully defined in Chapter III; suffice it to note here that that the h_k must be appropriately defined functions of lagged y values. The number of model components, K, can be quite large, even larger than sample size T.

We propose the following method to estimate the parameters α_k. In the initial step, we define $y^0 = y$ where y is the data. We run a series of univariate regressions against each of the model components and pick the regression which maximizes some predefined criterion function such as the r^2 of the regression. We call the model component which maximizes our r^2 criterion h^1. We call the residual from this regression y^1. We then run separate univariate regressions of y^1 against each of the model components and pick the regression with the highest r^2. We call the model component in the regression which maximizes our r^2 criterion h^2. We then run a multiple regression of y against h^1 and h^2, yielding:

$$y = C_1^2 h^1 + C_2^2 h^2 + y^2. \tag{II.3}$$

The notation C_j^k refers to the regression estimate at stage k of the model component selected at stage j.

We then run univariate regressions of y^2 against each of the model components and pick the regression with the highest r^2. We call the model component in the regression which maximizes our r^2 criterion h^3. We then run a multiple regression of y against h^1, h^2 and h^3 and call the residual y^3, yielding:

$$y = C_1^3 h^1 + C_2^3 h^2 + C_3^3 h^3 + y^3. \tag{II.4}$$

We continue at stage j of the estimation procedure by running univatiate regressions of y^{j-1} against each of the model components to select h^j. We then run a multiple regression of y against the model components h^k for $k \leq j$ and call the residual from the regression y^j. We continue until either the maximal r^2 becomes zero, we reach some pre-specified limit defined by a fixed number (e.g., $T - 10$) of iterations or by a statistical test or procedure.

In other words, given Eq. (II.2), the procedure concludes with an estimated model:

$$y \;=\; \sum_{i=1}^{N} C_i^N h^i + y^N$$

$$= \sum_{k=1}^{K} \hat{\alpha}_k h_k + y^N \tag{II.5}$$

where N is the final step of the algorithm and K is the number of model components. The regression coefficient $\hat{\alpha}_k$ on a model component h_k is equal to zero if the model component h_k was not selected and the regression coefficient $\hat{\alpha}_k$ is equal to the appropriate C_i^N if the model component h_k was selected.

What makes this estimation procedure attractive is that it is computationally feasible, simple in that is based on ordinary least squares regressions and produces estimates $\sum_k \hat{\alpha}_k h_k(t)$ which converge to $\sum_j \beta_j(t) y(t-j)$ in a sense to be made precise in the theoretical chapter below.

Important problems which need to be addressed include: (1) the choice of model components h_k with which to work (i.e., the "innocuous" step from Eq. (II.1) to Eq. (II.2)); and (2) The procedure spelled out above can actually involve an enormous number of steps; a practical stopping rule is required. These are among the issues we deal with in the thesis.

We now review what is to come in the remainder of the thesis, chapter by chapter:

- In Ch. 3 on *Model Development and Simulation Examples*, we lay out some families of potential nonstationarities in the $\beta_j(t)$, derive the corresponding model components and carry out some some simulation examples.

- In Ch. 4 on *Theoretical Analysis*, we analyze issues of convergence and consistency of point estimates for two broad classes of autoregressive models with time-varying coefficients.

- In Ch. 5 on *Auxiliary Results*, we prove some auxiliary results including: (1) some basic results on use of nonlinear regression confidence intervals for our data, (2) convergence rates.

- In Ch. 6 on *A Stopping Rule*, we provide a new test for randomness in economics which is particularly relevant for time-varying parameter models. We believe it provides a useful 'stopping rule' in the case of nonstationary time series models.

- In Ch. 7 on *Time-Frequcncy Spectral Analysis*, we show how to define a nonstationary spectrum in terms of our estimated models. This provides a useful point from which to review the literature on other methods of nonstationary time series.

- In Ch. 8 on *Economic Examples*, we provide applications of the procedure to macroeconomic and financial time series data.

- In Ch. 9 on *Implications and Extensions*, we review the implications of nonstationarity for economic theory as well as some possible research projects which follow from the results in the thesis.

- In Ch. 10 on *Conclusion*, we summarize our main results.

In addition to the results in the body of the thesis, there are some results of independent interest in an Appendix on "pursuit methods" (Appendix F), which discusses in detail how the method here relates to other pursuit methods in nonparametric regression such as "projection pursuit" [77] and "matching pursuit" [131] [171].

CHAPTER III

MODEL DEVELOPMENT AND SIMULATION EXAMPLES

The purpose of this chapter is to explain and demonstrate the mechanics of our approach to estimation of time series models. This approach focuses on the idea of a "model component", a concept which is discussed in depth in this chapter. We also provide some simulation examples to show how the method works in practice.

From Statistical Models to Model Components

The thesis presents a general approach to estimating a *time-varying* autoregressive model:

$$y(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \tag{III.1}$$

where $\epsilon_1(t)$ is a serially independent and identically distributed disturbance term and in which the functional forms of the true $\beta_j(t)$ are not known. The purpose of this section is to show how to move from Eq. (III.1) to "model components" which we can use in our procedure. This transformation is useful not only because it provides a means of estimating Eq. (III.1) but also because many of the ideas can be generalized to other more complex models such as:

- *State-dependent models* in which the $\beta_j(t)$ depend on state variables $x(t)$ instead of time:

$$y(t) = \sum_{j=1}^{J} \beta_j\,(x(t))\; y(t-j) + \epsilon(t). \tag{III.2}$$

- *Structural time series models* in which the regressor variables are other time series instead of lags of the data. An example of such a model is:

$$y(t) = \sum_{j=1}^{M} \beta_j(t)\, x_j(t) + \epsilon(t) \tag{III.3}$$

  where x_j is another economic variable, perhaps lagged. A simple case is where $y(t)$ is cointegrated with another variable $x_j(t)$ but the cointegrating relationship varies slowly over time.

Our goal is to represent Eq. (III.1) which is nonlinear in the variable 't' with an equivalent model which is linear in transformed regressor variables h_k which we call *model components*:

$$y(t) = \sum_{k=1}^{K} \alpha_k\, h_k(t) + \epsilon_2(t) \tag{III.4}$$

where K is the number of model components and $\epsilon_2(t)$ is a noise term.[1] Eq. (III.4) is a useful representation because the transformed variables $h_k(t)$ enter linearly and the coefficients α_k are time-invariant. Thus, once we have chosen appropriately the transformed regressor variables h_k, we can estimate the coefficients α_k by ordinary regression methods.

If we knew the precise parametric form of the $\beta_j(t)$ in Eq. (III.1), we could use nonlinear regression methods to estimate $\beta_j(t)$. However, we are concerned with situations in which we *do not* know the precise parametric form of the $\beta_j(t)$ and, in these situations, we do not know the form of the likelihood function or the sum

[1] We have used the notation $\epsilon_2(t)$ instead of the $\epsilon_1(t)$ in Eq. (III.1) because we wish to allow for the possibility that there might be approximation error in moving from Eq (III.1) to the representation in Eq. (III.4).

of squares function. Thus, we cannot use the nonlinear optimization procedures commonly used in econometric applications. In such cases, we shall show we can use a representation such as Eq. (III.4) to choose a model.

We now forget for a moment that we might not know the true regression model and pretend that we do know its form. Our goal is to illustrate particular economic problems and the corresponding model components which allow one to transform the nonlinear estimation problem of Eq. (III.1) into the simpler linear framework of Eq. (III.4). After considering some basic problems, we proceed to consider what to do in practical situations where we are unsure about the true form of the model.

Time-Invariant Autoregressive Models

The simplest example of a model of the form Eq. (III.1) is when the $\beta_j(t)$ are constant over time. This is the standard autoregressive model:

$$y(t) = \sum_{j=1}^{J} \beta_j \, y(t-j) + \epsilon_1(t). \tag{III.5}$$

In this case, we can define the model component $h_k(t)$ to equal $y(t-j)$ for $j = k$ and $1 \leq k \leq J$. In this case, the α_k in Eq. (III.4) equal the β_j in Eq. (III.5) and the functions $\beta_j(t)$ in Eq. (III.1). We now proceed to the next simplest (and first nontrivial) example.

Switching Autoregressive Models

The simplest type of time-variation in parameters one might imagine is a situation in which the autoregressive parameters change at some point in time t_0 to new values. This simple model is a commonly used econometric model of structural change.[2]

In this case, the $\beta_j(t)$ in Eq. (III.1) are:

$$\beta_j(t) = \begin{cases} \gamma_j & 0 < t < t_0 \\ \gamma_j' & \text{otherwise.} \end{cases} \tag{III.6}$$

[2] The theory of this model is developed in [7] [34]. For an application, see [163].

Thus, we can represent the $\beta_j(t)$ as:

$$\beta_j(t) = \gamma_j \, 1_{[0,t_0-1]}(t) + \gamma'_j \, 1_{[t_0-1,T]}(t) \tag{III.7}$$

where the indicator function $1_{[a,b]}(t)$ is defined as:

$$1_{[a,b]}(t) = \begin{cases} 1 & a < t \leq b \\ 0 & \text{otherwise} \end{cases} \tag{III.8}$$

In terms of the basic model equation, Eq. (III.1), we have:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \left[\gamma_j \, 1_{[0,t_0-1]}(t)\, y(t-j) + \gamma'_j \, 1_{[t_0-1,T]}(t)\, y(t-j)\right] + \epsilon_1(t). \end{aligned} \tag{III.9}$$

We now show how we can use the last line of Eq. (III.9) to define natural model components for our problem. We define:

$$h_{2j}(t;t_0) = 1_{[0,t_0-1]}(t)\, y(t-j) \tag{III.10}$$

$$h_{2j-1}(t;t_0) = 1_{[t_0-1,T]}(t)\, y(t-j) \tag{III.11}$$

and use Eq. (III.9):

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \left[\gamma_j 1_{[0,t_0-1]}(t)\, y(t-j) + \gamma'_j \, 1_{[t_0-1,T]}(t)\, y(t-j)\right] + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \left[\gamma_j \, h_{2j}(t;t_0) + \gamma'_j \, h_{2j-1}(t;t_0)\right] + \epsilon_1(t) \\ &= \sum_{k=1}^{2J} \alpha_k \, h_k(t;t_0) + \epsilon_1(t) \end{aligned} \tag{III.12}$$

where in the last line of Eq. (III.12):

$$\alpha_k = \begin{cases} \gamma_{\frac{k}{2}} & k \text{ even} \\ \gamma'_{\frac{(k+1)}{2}} & \text{otherwise} \end{cases} \tag{III.13}$$

Thus, by estimating the linear equation:

$$y(t) = \sum_{k=1}^{2J} \alpha_k h_k(t; t_0) + \epsilon_2(t) \tag{III.14}$$

we can recover estimates of $\beta_j(t)$. We also note that Eq. (III.14) is of the form Eq. (III.4). To summarize, in constructing the model components h_k, we have built the nonlinearity in the autoregressive parameter functions $\beta_j(t)$ into the definition of the regressor variables.

Now, suppose that we do not know the true value of t_0. In this case, we construct different model components for each possible value of t_0. The iterative procedure outlined in Ch. II and reviewed in more detail in this chapter defines a mapping between a set of potential model components h_k (which might include all possible values of t_0) and a selected model (which we would hope would contain an implicit estimated value of t_0 as close as possible to the population parameter).

It may also be the case that there may be more than one structural break. In this case, we construct model components of all different lengths so as to allow for multiple structural breaks. These model components have the explicit form:

$$h_k(t) = 1_{[b_k, e_k]}(t)\, y(t - j_k) \tag{III.15}$$

where b_k and e_k represent the beginning and end of a given period of structural stability and j_k is the corresponding lag. Since there may be many possible model components of the form Eq. (III.15), a procedure is required to pick out a model from a set of possible model components. Again, the iterative procedure proposed in the thesis solves this problem.

Models with Gradual Onset of Structural Change

To consider how we might generalize the concept of a model component to handle other variants of structural change, we note that it is reasonable to think of the indicator function $1_{[a,b]}(t)$ as a type of window function. While indicator functions are the simplest possible window functions, they are by no means the most desirable functions if there are complex dynamics in the adjustment process to a new economic regime.

Consider once again the time series model:

$$y(t) = \sum_{j=1}^{J} \beta_j(t)\, y(t-j) + \epsilon_1(t) \tag{III.16}$$

but let the $\beta_j(t)$ adjust only slowly in response to changes in underlying economic fundamentals. With slow adjustment, it seems reasonable to replace the sharp edges of the flat autoregressive parameter functions $\beta_j(t)$ used in the case of abrupt structural change with flat autoregressive parameter functions proportional to window functions with smoothed edges $g_k(t)$:

$$g_k(t) = \begin{cases} e^{\frac{-(t-b_k)^2}{2c_k}} & \text{if } t \leq b_k \\ 1 & \text{if } b_k < t < e_k \\ e^{\frac{-(t-e_k)^2}{2c_k}} & \text{if } t \geq e_k \end{cases} \tag{III.17}$$

An example of a function of this type is shown in Figure (III.1).[3]

Thus, we have:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k\, g_k(t)\, y(t-j_k) + \epsilon_1(t) \end{aligned} \tag{III.18}$$

[3] The parameters b_k and e_k here are set so that $e_k - b_k$ is 3/4th the length for the corresponding flat window function. The parameter c_k in Equation (III.17) is set as $\frac{(e'_k - b'_k)^2}{100}$ where e'_k and b'_k are the end and beginning of the corresponding flat window.

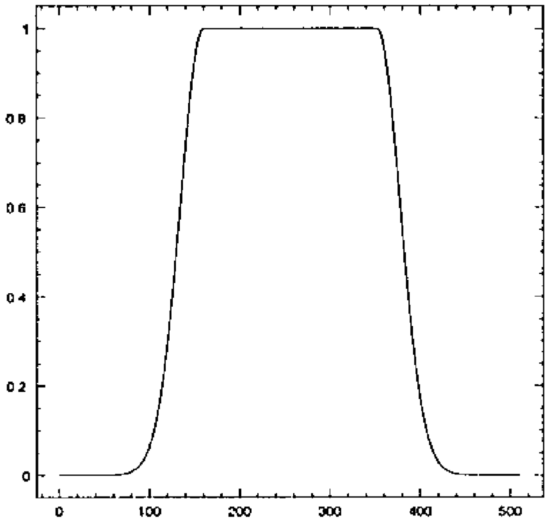

Figure III.1: A flat window function with smooth Gaussian edges.

where j_k is the lag associated with the window function $g_k(t)$. We now use Eq. (III.18) to move to an equation which involves model components. We have:

$$\begin{aligned} y(t) &= \sum_{k=1}^{K} \alpha_k \, g_k(t) \, y(t - j_k) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k \, h_k(t) + \epsilon_1(t) \end{aligned} \tag{III.19}$$

where:

$$h_k(t) = g_k(t) \, y(t - j_k) \tag{III.20}$$

so that the $h_k(t)$ defined by Eq. (III.20) are the natural model components for this type of problem. Therefore, once again we can use model components $h_k(t)$ as an intermediate step in constructing estimates. Since the parameters c_k, b_k and e_k might not be known *a priori*, we may want to construct model components for the different values of these parameters we think are likely to occur. We note that either the e_k or b_k might be past the end of the time series so that the window function $g_k(t)$ can represent slow adjustment to a new regime which is permanent.[4] We may also wish for precision to write $\epsilon_2(t)$ instead of $\epsilon_1(t)$ at the end of the last line of Eq. (III.18) since in practice our specified model components may not capture the true model exactly.

Models with Slow Structural Change

The choice of Eq. (III.17) as a shape for the autoregressive functions $\beta_j(t)$ presumes that adjustment to a new regime occurs relatively rapidly and that there is relative parameter stability once adjustment occurs. It may in fact be the case that structural change occurs as a continuous process; for instance, it may be that technological advances in managing inventories are speeding up the transmission mechanism in an inventory equation so that economic factors which might previously have contributed to long lags begin to produce shorter lags. Whatever the underlying economic cause, there are a wide variety of forms the functions $\beta_j(t)$ might take in an environment with slow structural change.

One possibility is that the $\beta_j(t)$ are proportional to members of a family of Gaussian window functions or, more generally, might each be given by a weighted sum of different window functions $g_k(t)$ corresponding to the effects of different economic factors. For this case, in Eq. (III.1) we have:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k\, g_k(t)\, y(t-j_k) + \epsilon_1(t) \end{aligned} \tag{III.21}$$

where, for instance:

$$g_k(t) = e^{-\frac{(t-c_k)^2}{2\sigma_k^2}} \tag{III.22}$$

[4] For an example, see Fig. (III.19)

In Eq. (III.22), σ_k is a width parameter for window k and c_k is a centering parameter. Using Eq. (III.21), we have:

$$\begin{aligned} y(t) &= \sum_{k=1}^{K} \alpha_k \, g_k(t) \, y(t - j_k) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k \, h_k(t) + \epsilon_1(t) \end{aligned} \tag{III.23}$$

which is in the form of Eq. (III.4) with model components:

$$h_k(t) = g_k(t) \, y(t - j_k) \tag{III.24}$$

where j_k is the lag associated with the model component h_k. If we did not know the values of c_k and σ_k, we could include many different $h_k(t)$ corresponding to different choices of lags and c_k and σ_k.

Another model of structural change may feature a permanent jump in the value of the autoregressive function $\beta_j(t)$. Such a model may be relevant in cases in which there are permanent changes in policy. We have:

$$\beta_j(t) = \begin{cases} \gamma_j & \text{if } t \leq b_j \\ \gamma_j + \gamma_j' \left(\frac{t-b_j}{e_j-b_j}\right) & \text{if } b_j < t < e_j \\ \gamma_j + \gamma_j' & \text{if } t \geq e_j \end{cases} . \tag{III.25}$$

An example of a $\beta_j(t)$ function of this form is shown in Fig. (III.2) where $\gamma_j = 0.2$, $\gamma_j' = 0.6$, $b_j = 300$, and $e_j = 410$.

In terms of our linear time-varying autoregressive model:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} (\gamma_j \, 1_{[0,T]}(t) \, y(t-j) + \gamma_j' \, 1_{[e_j-1,T]}(t) \, y(t-j) + \\ &\quad \gamma_j' \, 1_{[b_j,e_j-1]}(t) \, \frac{(t-b_j)}{(e_j-b_j)} y(t-j)) + \epsilon_1(t) \end{aligned}$$

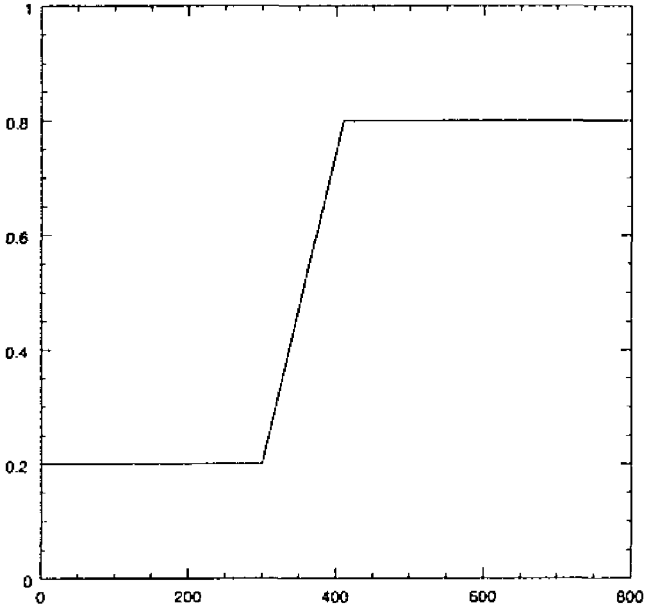

Figure III.2: An example of a $\beta_j(t)$ function with a permanent jump.

$$= \sum_{k=1}^{J} \gamma_k h_k(t) + \sum_{k=J+1}^{2J} \gamma'_{k-J} h_k(t) + \sum_{k=2J+1}^{3J} \gamma'_{k-2J} h_k(t) + \epsilon_1(t). \quad \text{(III.26)}$$

In the last line of Eq. (III.1), the first two sums contain model components which are the model components with flat windows introduced above:

$$h_k(t) = 1_{[b_k,e_k]}(t)\, y(t - j_k) \quad \text{(III.27)}$$

whereas the last sum contains the new model component:

$$h_k(t) = g_k(t)\, y(t - j_k) \quad \text{(III.28)}$$

where:

$$g_k(t) = \frac{t - b_k}{e_k - b_k} 1_{[b_k,e_k]}(t). \quad \text{(III.29)}$$

In this section, we have shown how model components with different window functions capture the effects of slow parameter variation. We have considered two examples: (1) "mean-reverting" structural change in which the autoregressive parameter is described by a smooth Gaussian function; (2) "permanent" structural change with a period of adjustment in which the autoregressive parameter changes linearly.

Models with Periodically Varying Parameters

In addition to models of slow structural change, we may either have periodic structural change due to effects such as seasonality or it may be that the adjustment to new regimes exhibits the oscillatory or "overshooting" behavior predicted by a variety of economic models.[5] In this case, the autoregressive parameter functions $\beta_j(t)$ may exhibit oscillatory behavior. To represent such behavior effectively, we introduce model components with periodic window functions.

The simplest possible model of oscillatory behavior is:

$$\beta_j(t) = \eta_j + \gamma_j \cos(\omega_j t) \tag{III.30}$$

In this case, we have:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} [\eta_j\, y(t-j) + \gamma_j \cos(\omega_j t) y(t-j)] + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \eta_j\, h_j + \sum_{j=J+1}^{2J} \gamma_{j-J}\, h_j + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k h_k + \epsilon_1(t) \end{aligned} \tag{III.31}$$

where:

$$\alpha_k = \begin{cases} \eta_k & \text{if } k \le J \\ \gamma_{k-J} & \text{if } J+1 \le k \le 2J, \end{cases} \tag{III.32}$$

[5] Two examples are: [64] [27].

and the model components $h_k(t)$ are:

$$h_k(t) = y(t - j_k) \tag{III.33}$$

for $k \leq J$ and an associated lag $j_k = k$; and:

$$h_k(t) = \cos(\omega_k t)\, y(t - j_k) \tag{III.34}$$

for $J + 1 \leq k \leq 2J$. A more complicated solution is to introduce model components parameterized (to scale) by η_j, γ_j and ω_j.

In Eq. (III.30), we have assumed that the oscillatory behavior of the autoregressive function $\beta_j(t)$ is time-invariant. However, much of the economic motivation we have given for use of oscillatory models is that such behavior is a response to unusual policy or technological shocks. Thus, it seems reasonable to generalize Eq. (III.30) to:

$$\beta_j(t) = \eta_j + \gamma_j\, 1_{[b_j, e_j]}(t)\, \cos(\omega_j t) \tag{III.35}$$

where b_j and e_j represent the beginning and end of the period of oscillatory behavior. In this case, the autoregressive equation becomes:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \left[\eta_j y(t-j) + \gamma_j 1_{[b_j, e_j]}(t) \cos(\omega_j t) y(t-j) \right] + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \eta_j h_j + \sum_{j=J+1}^{2J} \gamma_{j-J} h_j + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k h_k + \epsilon_1(t) \end{aligned} \tag{III.36}$$

where:

$$\alpha_k = \begin{cases} \eta_k & \text{if } k \leq J \\ \gamma_{k-J} & \text{if } J+1 \leq k \leq 2J \end{cases} \tag{III.37}$$

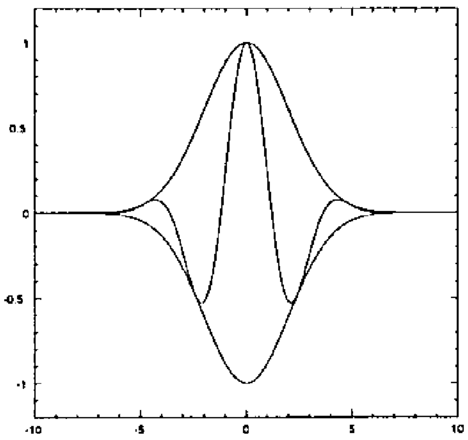

Figure III.3: A low-frequency basis function based on a Gaussian window. The window function lies within an envelope of two Gaussians.

and we use the model components in Eq. (III.33) for $k \leq J$ and the model components:

$$h_k(t) = 1_{[b_k,c_k]}(t)\,\cos(\omega_k t) y(t - j_k) \tag{III.38}$$

to capture the oscillatory component of $\beta_j(t)$ (model components $J + 1 \leq k \leq 2J$).

Since structural change may be occurring slowly, the oscillatory component in Eq. (III.35) may instead take the form:

$$\beta_j(t) = \eta_j + \gamma_j\, g_j(t)\, \cos(\omega_j t) \tag{III.39}$$

where $g_j(t)$ is a window function.

Eq. (III.35) is a special case of Eq. (III.39) when the window function is an indicator function. When the window $g_j(t)$ is a Gaussian window, the oscillatory component of the autoregressive function is proportional to the window function shown in Fig. (III.3).

Using the same steps as in Eq. (III.31) but replacing the indicator function with

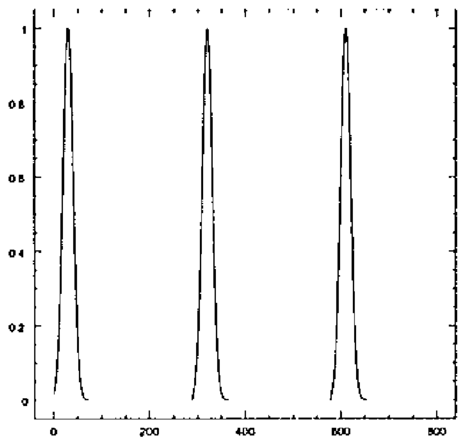

Figure III.4: An example of a periodic $\beta_j(t)$ function.

a smooth window, we have that the appropriate model components are:[6]

$$h_k(t) = g_k(t)\ \cos(\omega_k t)\, y(t - j_k) \tag{III.40}$$

where $g_k(t)$ is a window function such as the Gaussian window superimposed in Fig. (III.3).

As another example, we may know that $\beta_j(t)$ is periodic with some known periodicity K. In this case, we have:

$$\beta_j(t) = \alpha_j\, r_j\left(t - \left[\frac{t-1}{K}\right] K\right) \tag{III.41}$$

where [.] denotes integer part. The function r_j represents behavior within any regime of length K. Thus, for example, if $K = 10$, $\beta_j(1) = \beta_j(11) = \beta_j(21)$ and so on. An example of a $\beta_j(t)$ function of the form Eq. (III.41) is shown in Fig. (III.4). In this case, the $r_j(t)$ function is Gaussian.

In this case, we have:

[6] Based on our experiments, it is not a good idea to include *high-frequency* oscillations in window functions.

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \alpha_j r_j \left(t - \left[\frac{t-1}{K}\right] K\right) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \alpha_j \, h_j(t) + \epsilon_1(t) \end{aligned} \tag{III.42}$$

where:

$$h_j(t) = r_j \left(t - \left[\frac{t-1}{K}\right] K\right) \, y(t-j) \tag{III.43}$$

is the resulting choice for the family of model components for this type of problem. If we do not know the periodicity K or the type of function r_j, we may wish to introduce many model components of this type.

Distributed Lag Models

In this section, we illustrate by example that model components have wider applicability than might be assumed from the examples above. One traditional model used in econometrics is the distributed lag model in which restrictions are placed on the form and shape of the $\beta_j(t)$. Economic explanations given for these models include adaptivity in expectations or delivery lags. Some of the literature on these models is surveyed in ([113], Ch. 9-10) [58]. The basic model is that the β_j are related in the time-invariant sense that:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} r(j,\theta) \, y(t-j) + \epsilon_1(t) \end{aligned} \tag{III.44}$$

where θ is a parameter vector which determines the form of the function r. For example, the function r could be of polynomial form and the parameters θ could be the coefficients on the polynomial expansion:

$$r(j,\theta) = \sum_{i=0}^{P} \theta_i j^i \tag{III.45}$$

where P is the order of the polynomial distributed lag. Other possibilities for the form of $r(j,\theta)$ studied in the literature include spline functions and harmonic functions of lag as well as models which restrict the shape of the lag function to a specific form such as arithmetic or geometric.

Suppose now we think that the distributed lag relationship, whatever form it might take, is varying over time, so that:

$$\beta_j(t) = \alpha\, g(t)\, r(j,\theta). \tag{III.46}$$

where the function $r(j,\theta)$ may be any possible distributed lag function (for example, it may be of the form Eq. (III.45)). In more generality, we might wish to consider situations in which the $\beta_j(t)$ are sums of terms such as occur on the right hand side of Eq. (III.46):

$$\beta_j(t) = \sum_{k=1}^{K} \alpha_k\, g_k(t)\, r_k(j,\theta), \tag{III.47}$$

so that:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j\, y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_k\, g_k(t) r_k(j,\theta)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k\, g_k(t) \sum_{j=1}^{J} r_k(j,\theta)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k\, h_k(t) + \epsilon(t). \end{aligned} \tag{III.48}$$

In this case, Eq. (III.48) suggests that we construct model components of the form:

$$h_k(t) = g_k(t) \sum_{j=1}^{J} r_k(j, \theta) y(t-j) \tag{III.49}$$

where r_k is the form of the distributed lag function associated with model component k and $g_k(t)$ is a window function (such as a flat window).

Since each of the model components in Eq. (III.49) represents behavior within a given regime, a sum of such model components seems to be a reasonable model of an entire time series. The advantage of using such model components is that the functions r_k reduce the degrees of freedom required in estimation.

Summary

In this section, we have reviewed various models of nonstationary time series and their corresponding model components. We have shown that there are natural choices for model components for a variety of models of nonstationary time series. These choices and the special properties of the model components are summarized in Table (III.1).

At the beginning of the section, we urged the reader to forget for a moment that we might not know the true model and imagine that the true model is known. Now, we will discuss what to do when we do not know the form of the model. Two words capture the essential intuition: "be greedy". By this, we mean that the researcher should look at Table (III.1) and choose all the families of model components and parameterizations which he thinks might be relevant to the problem at hand and include them in a set of potential model components. The procedure we propose in the thesis then has the model components 'competing' to be included in the estimated model. The researcher may make competition 'imperfect' by placing less weight on model components which produce estimates with poor statistical properties or which the researcher feels are less plausible theoretically.

The perspective we propose is thus considerably different from that of classical econometrics or statistics in which the researcher gives one model a 'monopoly' over

Economic / Stochastic Model	Model Components
Stationary model $\beta_j(t) = \gamma_j$	$h_k(t) = y(t - j_k)$
Local stationary model $\beta_j(t) = \sum_k \gamma_k 1_{[b_k,e_k]}(t)$	$h_k(t) = 1_{[b_k,e_k]}(t)\, y(t - j_k)$
Fast adjustment to structural change $\beta_j(t) = \sum_k \gamma_k g_k(t)$	$h_k = g_k(t)\, y(t - j_k)$ g_k flat window with Gaussian edges (c.f., Eq.(III.17))
Smooth structural change $\beta_j(t)$: temporary struct. change	$h_k(t) = g_k(t) y(t - j_k)$ $g_k(t)$: Gaussian window
Permanent change $\beta_j(t)$ trending	Use both: (1) $h_k(t) = 1_{[b_k,e_k]}(t)\, y(t - j_k)$ (2) $h_k(t) = \frac{t-b_k}{e_k-b_k} y(t - j_k)$
Oscillatory relationships $\beta_j(t) = \eta_j + \eta'_j \cos(\omega_j t)$	Use both: (1) $h_k(t) = \cos(\omega_k t)\, y(t - j_k)$ (2) $h_k(t) = y(t - j_k)$
Local periodic relationships $\beta_j(t) = \eta_j + \eta'_j\, g_j(t) \cos(\omega_j t)$	Use both: (1) $h_k(t) = g_k(t) \cos(\omega_k t)\, y(t - j_k)$ (2) $h_k(t) = y(t - j_k)$
Periodic relationships $\beta_j(t + \lambda K) = \beta_j(t) \quad \lambda \in \mathbf{Z}$	$h_k(t) = r_k \left(t - \left[\frac{t-1}{K}\right] K\right)\, y(t - j_k)$ r_k window function
Distributed lag relationships $\beta_j(t) = \sum_{k=1}^{K} \alpha_k\, g_k(t)\, r_k(j, \theta)$	$h_k(t) = g_k(t) \sum_{j=1}^{J} r_k(j, \theta) y(t - j)$ r_k distributed lag function g_k window function

Table III.1: Summary Table of Families of Model Components and Corresponding Economic / Stochastic Models.

the data. Both because our procedure lies in between nonparametric and parametric approaches and also because it has some special features, the reader may wonder why we have chosen the particular approach developed in Ch. II and Ch. III. We address these issues in appendices. In Appendix K, we explain that why we have chosen the specific approach outlined in Ch. II instead of other similar possible regression-based approaches which also use model components. In Appendix L, we address the more basic question of why we develop a complex adaptive regression model instead of something more simple.

Examples

We now consider some examples on simulated data to see how our method works and to provide examples of how the method can be used in practice. These examples help provide some crucial "engineering" details about how to use the method proposed in the thesis. Much of our progress in understanding the method came through such experiments. As Huber ([106], pp. 469-470) has pointed out with reference to another method (Projection Pursuit regression), "The situation is analogous to that in numerical spectrum analysis. There the real progress did not come through mathematical statistics in the usual sense, that is, through consistency and asymptotic normality proofs, but through a mathematically much more primitive, qualitative and quantitative understanding."

All the examples in this chapter were computed using a time domain procedure described in Appendix J in which model component details are stored as a series of vectors and the simple regressions are computed by sums over only the relevant time intervals. Appendix J also presents an FFT algorithm which may be used for large datasets and Appendix M contains a formal analysis of computational requirements in order to show formally that implementation of the procedure modern workstations or personal computers not problematic.

First Order Autoregressive Processes

As a simplest possible example, we consider a first order autoregressive process (AR1):

$$y(t) = \beta\, y(t-1) + u(t) \tag{III.50}$$

where:

$$-1 < \beta < 1 \tag{III.51}$$

where $u(t)$ is a normally distributed ($N(0,1)$) noise term, and $y(t)$ is the random sequence we are interested in decomposing. Given $y(0) = 0$ and a sequence $u(t)$ of independent normally distributed random variates we use Equation (III.50) to generate a sequence $\{y(t)\}_{t=1}^{T}$ where T is sample size. We simulate a series with $T = 512$ and $\beta = 0.9$:

$$y(t) = 0.9\, y(t-1) + u(t) \tag{III.52}$$

where $u(t) \sim N(0,1)$. For the realization we consider, the sample variance of $y(t)$ is 6.19 and its sample mean is .1937.

Starting with the simulated data set $\{y(t)\}_{t=1}^{512}$ we use our method to try to recapture Eq. (III.52). To estimate the model, we must select a family of types of window functions g_m which we recall are motivated by the types of nonstationarities we think might potentially occur in the data. We may think that the data may either be stationary or have local stationarity properties. Looking at Table (III.1), the researcher sees that the appropriate model components include flat windows multiplied by lags of the data. If the researcher does not know the precise form of the local stationarity of economic relationships, it is appropriate to include model components with flat windows of different widths and locations in order to pick up such relationships. Since there are many possible widths and locations for such windows, we need to translate the theory into a practical application.

We recall that the appropriate model components have windows of the form:

$$g_{m_i} = 1_{[t_{b_i}, t_{e_i}]}(t) = \begin{cases} 1 & \text{if } b_i \leq t \leq e_i \\ 0 & \text{otherwise} \end{cases} \tag{III.53}$$

where b_i and e_i stand for the beginning and end points of the window respectively. We use L^{s_i} to stand for the lag of the data included in the ith model component. Although it complicates notation, we find it essential to use the extra index m to describe the window for model component i. The reason is that different model components will use the same window (for instance, lag 1 and lag 2 data).

To do this, we consider windows of width of powers of 2 up to the size of the time series. For instance, if the time series has 16 elements, we may wish to consider a window of length 16 and many different windows of length 8, 4, 2 and 1. Since our simulated series has length 512, we can under this scheme consider up to 9 levels of windows since $2^9 = 512$. There are many different windows of length 8, 4, 2 and 1 because we need to consider either all possible locations of these windows or some approximation to all possible locations. There is no *a priori* reason to exclude from analysis intermediate window sizes of lengths such as 5, 9 or 3 but for tractability we must restrict the number of window functions in some way. We also would likely consider a large number of lags for *each window*; for instance, with a time series of 512, it may be reasonable to include a dozen or more lags in the analysis. To illustrate that the procedure does not go astray when we choose an enormous number of model components, we will include fifteen lags. For example, if we consider a given window which is nonzero from points 10 to 266 we will want to multiply this window by the lag one data, the lag two data and so on.[7]

[7] One further technical point. A window with say length 4 at lag 20 does not seem reasonable to

In the construction of model components, there are several adjustable parameters to consider. For instance, we may not wish to include model components for every location at every level of windows; therefore, we may wish to include a model component whose location jumps by a factor; we have found it convenient to set a constant factor and jump by a factor proportional to the window length. We also can include a parameter which considers only every kth lag after some lag m; this is useful in examining financial data.

In general, by restricting the number of possible widths and locations, there is a sense in which we reduce the possibilities of spurious estimates (because there are fewer model components). On the other hand, there is an efficiency cost to the larger approximating expansions for the regression relationship which are required when we have not included the "best" possible model components. The model we have proposed of 'dividing by two' provides a balance between these considerations.

For this example, we have selected 9 levels and 15 lags which results in $22,540$ potential model components $\{h_i\}$. The reason for the large number of model components is that we have to consider many locations of window functions at each level; for instance we will want to include a window which is nonzero from points 10 to 266 as well as for instance a window which is nonzero from points 8 to 264. Ordinarily with Box-Jenkins analysis, it is unusual to consider more than half a dozen lags and by definition there is only one window function so that our procedure here allows for many orders of magnitude of greater flexibility in model selection and estimation.

In the first iteration, we compute that the first chosen model component h^0 is $g_{m^0} L^{s^0} y$ which is the h_i that has lag 1 and whose window begins at the beginning of the sample and ends at the end of the sample. Table (III.2) shows the regression

us because we want some sort of overlap between the length of the window and the lag considered. One implementation we have found to be useful in practice is to consider only lags up to one fourth the length of the window; thus, for a window of size 32 we would not include a lag of greater than 8 in the analysis.

Begin	End	Lag	Regression Coefficient	r^2
1	512	1	0.8968	0.804
1	512	2	0.7969	0.634
1	512	10	0.3248	0.105
230	485	1	0.906	0.427
130	257	1	0.922	0.314
429	492	13	-0.019	0.000
20	275	14	0.2113	0.025

Table III.2: Coefficient estimates and the resulting r^2 for a few possible model components in the first iteration of the procedure as applied to a first order time-invariant autoregressive model. The procedure selects only the model component with the maximal r^2. In this case the true model has the maximal r^2 and is hence selected.

coefficient values for some of the (normalized) nonselected model components on the first iteration as well as the chosen one.

The value of the regression coefficient estimated is 0.8968. We note that from the coefficient estimate, we can determine that a model using only the first selected model component explains 80.4% of the sum of squares (or equivalently, energy or squared norm) of the data.[8]

After subtracting off the projection implied by the selection of the first coefficient, the variance of the residual is 1.01434 which is very close to the variance of the noise driving the autoregressive model. The residual is shown in Figure (III.6). Even if it is unlikely that a reasonable statistical procedure would include more than the first model component in estimates, we continue for one more iteration to see what

[8] In other words, a regression of the data on the first model component would result in a R^2 of 0.804. The theoretical value of R^2 is 0.81. The sample mean was subtracted before analysis.

happens.

At the second stage the percentage of the sum of squares explained by both selected model components is 80.95% which is barely more than the percentage of the sum of squares explained by the first model component alone. The value of the regression coefficient selected is -0.582496 (std. error is 0.15344) and the selected model component has lag 2 and a window which begins at point 456 and ends at point 472. Since this spurious estimate is captured by a model component with a relatively short window, this chosen model component adds little explanatory power. Based on our experience, we do not recommend including such short model components in the analysis, and if included, we recommend that they be weighted according to the criterion that we suggest (which has a theoretical basis) in order to minimize the effect of spurious estimates.[9]

It is helpful to examine next a slightly different autoregressive (AR1) model with $y(t) = -0.3\,y(t-1) + u(t)$. We first generate a realization of $y(t)$ with sample size 512 by using a Gaussian random number generator to generate values of $u(t)$.[10] We consider 11,348 model components which correspond to up to seven lags and five levels. The first iteration leads to a model component of lag one with length of the whole time series. The coefficient estimate is -0.34288 with standard errors of 0.0415152.[11] Further iterations of the algorithm lead to spurious selection of a lag coefficient of -0.418486 at lag 4 between points 0 and 32 with a standard error of 0.139454. In

[9] As discussed above and in Ch. IV, this criterion is to weight by factor inversely related to the average of the fourth power of the window function for the model component. For model components with flat windows, the average of the fourth power of the model component turns out to be the inverse of the fraction of the time series covered by the window function (see Chapter IV, Eq. (IV.17)). This suggests that we should weight short model components less in order to reduce the variance of estimates.

[10] We subtract the sample mean (-0.0207842) before analysis.

[11] The standard errors are the OLS standard errors and therefore are conditional on the choice of model.

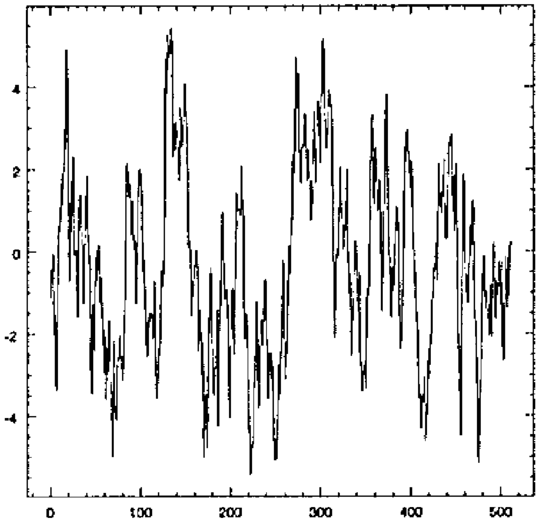

Figure III.5: Data for autoregressive model of sample size 512 with AR1 parameter 0.9: $y(t) = 0.9y(t-1) + u(t)$

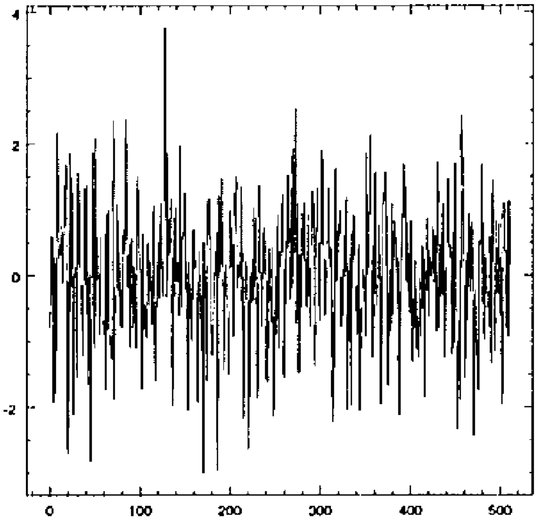

Figure III.6: Residual after subtracting off the first projection from the first order autoregressive model with $\beta = 0.9$. Statistically it is close to white noise.

terms of additional R^2 added at an iteration,[12] the first model component generates 0.118 whereas the second model component generates 0.0152 which is substantially lower. The stopping rule we propose in Ch. VI would stop the algorithm after the first iteration; the coefficient of the second lag in the autoregression is -0.0363728 with a standard error 0.0442195.

We can prove (see Chapter IV) that if flat windows with the length of the time series are included in the analysis, in the large sample limit we (with probability 1) will select only flat windows over the length of the time series if the true model is stationary.[13]

Whereas our method seems in a sense to select the correct model when the underlying model is stationary, conclusions drawn from use of the local weighted least squares (or 'kernel') estimator:

$$\hat{\beta}_1(t) = \frac{\sum_s g(t-s)y(s)y(s-1)}{\sum_s g(t-s)y(s-1)^2} \tag{III.54}$$

(where g is a 'kernel' function) are more problematic because the properties of estimates depend on bandwidth. A Gaussian kernel was used in the analysis and, after experimentation with bandwidth, a Gaussian with $\sigma = 5.0$ was chosen.[14] Estimates

[12] We use R^2 instead of r^2 because on the second iteration we wish to examine the total explanatory power added to the regression.

[13] We make the (weak) assumption that the weights used in computing r^2 do not weight shorter model components more.

[14] The convolutions in the numerator and denominator were computed in the time domain using a kernel which sums to 1 at each point. The sample mean was subtracted from the data before analysis. The data used to generate the kernel estimate was from a different random sample than that used to generate Fig. (III.1). Thus, the ends of the sample are corrupted by small sample bias. If we assume β is constant over time under the null, we can compute approximate confidence intervals for $\hat{\beta}(t)$ using standard error estimates $\frac{\sigma_\epsilon}{\sqrt{T\sum_s g(t-s)y(s-1)^2}}$. This follows since:

$$(\hat{\beta}_1(t) - \beta)^2 = \frac{\frac{1}{T}\sum_{s,s'} g(t-s')g(t-s)\epsilon(s)\epsilon(s')y(s-1)y(s'-1)}{\left[\frac{1}{T}\sum g(t-s)y(s-1)^2\right]^2}. \tag{III.55}$$

and $\epsilon(s)$ is assumed to be white noise and we consider a large sample approximation. The theoretical value of $\sigma_\epsilon = 1$ was used in constructing confidence intervals where we multiply $\sigma_{\hat{\beta}}$ by 1.96. An

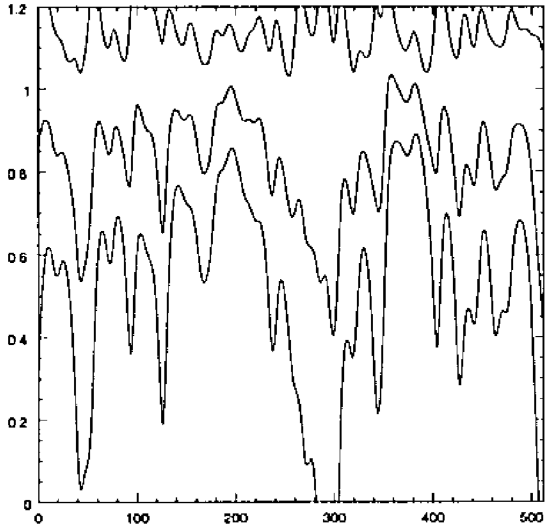

Figure III.7: Kernel estimates for first order autoregressive model.

are shown in Figure (III.7). Fig. (III.8) shows estimates with $\sigma = 20.0$.

To summarize what we have learned from the figures, we can see that one of the issues is that our answers seem to depend on the choice of bandwidth or a smoothing parameter. A similar issue arises in the spectral analysis of nonstationary time series data; we could use 'rolling' estimates of the spectrum which compute estimates for only $L < T$ points at a time but then our estimates depend on L. There therefore seem to be benefits to mixing effective window widths depending on the observed properties of the time series; Ch. VII on time-frequency spectral estimation reviews some ways to do this in the frequency domain whereas the thesis focuses on the time domain.

We point out that there are some other new methods which might be applicable to the analysis of nonstationary economic and financial data. In general, our experience is that more sophisticated methods such as the Matching Pursuit algorithm [131], wavelet packet methods [47] [44] [45] [46] and wavelet transforms [52] [138]

approximate value of T of 2σ was used in constructing standard errors of β.

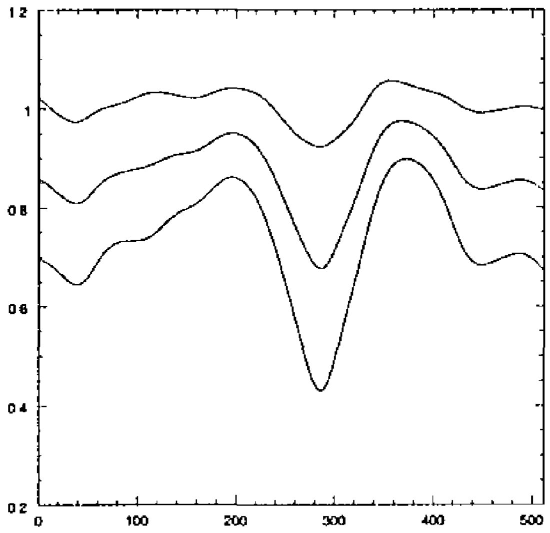

Figure III.8: Kernel estimates for first order autoregressive model with a wider bandwidth.

also achieve uneven performance when the data generating process is a stationary stochastic process. We do not expect the reader to be familiar with these methods which are introduced and reviewed in Chapters VI and VII. The basic idea of these methods is to work with different sorts of representations analogous to the spectrum except that the estimates vary over time.

Switching Autoregressions

We next consider a simple but fundamentally nonstationary process. The simple model we consider is:

$$y(t) = \beta(t)\, y(t-1) + u(t) \tag{III.56}$$

$$\beta(t) = \begin{cases} -\beta_0 & \text{if } T_0 \leq t \leq T_1 \\ \beta_0 & \text{otherwise} \end{cases} \tag{III.57}$$

where $u(t)$ is noise and $-1 < \beta_0 < 1$. T_0 and T_1 denote the beginning and end points of a structural shift in which the sign of the autoregressive parameter $\beta(t)$ changes.

We consider a simple example in which $\beta_0 = 0.9$, $T_0 = 0$, $T_1 = 256$, $T = 512$. The data are shown in Fig. (III.9). Autoregressive estimates for the data are:[15]

$$\begin{aligned} \hat{y}(t) &= \underset{(0.044063)}{0.0788397}\, y(t-1) \end{aligned}$$

$$\begin{aligned} \hat{y}(t) &= \underset{(0.025453)}{0.0145828}\, y(t-1) + \underset{(0.025456)}{0.817745}\, y(t-2) \end{aligned}$$

$$\begin{aligned} \hat{y}(t) &= \underset{(0.0441837)}{-0.00361366}\, y(t-1) + \underset{(0.0254572)}{0.817434}\, y(t-2) + \underset{(0.0442003)}{0.022267} y(t-3) \end{aligned}$$

so that a researcher would probably select an $AR(2)$ model out of the class of autoregressive models. We point out that if our model components are only flat windows with the length of the time series, we would likely select a model with only the second lag coefficient so that there is a sense that, even within the class of stationary time series models, our procedure seems to be a reasonable method of model selection.

Given the form of the model, it is reasonable to use the appropriate model components for locally stationary processes which, as noted in Table (III.1), include flat windows multiplied by lags. Hence, we again use flat windows and choose 5 lags and 5 levels of windows for a total of 8106 potential model components. After 3 iterations, our estimates are:

[15] We have subtracted the sample mean of 0.0533404.

$$\begin{aligned}
\hat{y}(t) \;=\; & -0.003332474\, y(t-2) + 0.948662\, y(t-1) 1_{t\in[255,511]} \\
& (0.0432604) \qquad\qquad (0.0466964) \\
& -0.932788\, y(t-1)\, 1_{t\in[0,256]}. \\
& (0.0473276) \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(III.58)}
\end{aligned}$$

Fig. (III.10) shows our estimates of β_1 as a function of time. We note that in the procedure we have spuriously selected a second lag at the first iteration; this does not happen if β_0 is closer to zero as we analyze in Appendix H. The reason we have incorrectly picked a model component is that the second autocorrelation is constant across the time series; as the size of the data grows, we will continue to pick the incorrect model component if $|\beta_0|$ is larger than a critical value we derive; however, the coefficient on this incorrect model component will get smaller as the sample size increases.

By comparison, we show in Fig. III.11 a simple 'weighted local least squares' estimate of the first lag coefficient. Our method produces more accurate estimates than either the local least squares method or ordinary autoregressive model precisely because it allows the size of the effective 'window' to adapt based on the local properties of the time series.

Smoothly Time-Varying Autoregressions

In many cases, structural change may be slow rather than dramatic so that autoregressive parameters may change slowly with time rather than abruptly. An example of a model with slow structural change is:

$$y(t) = \gamma(t)\, y(t-1) + u(t) \qquad \text{(III.59)}$$

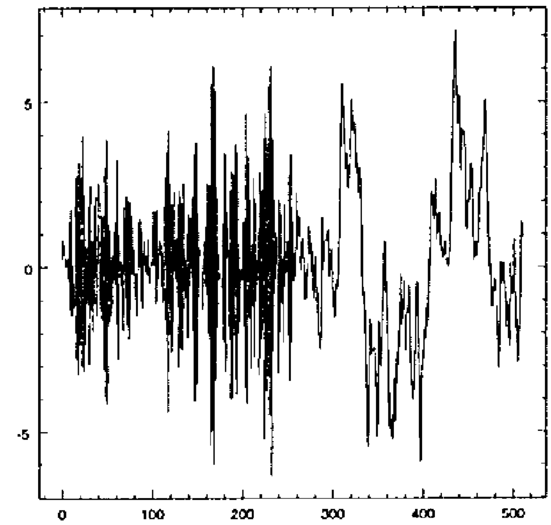

Figure III.9: Data for the switching autoregression with parameter $\beta_0 = -0.9$ and $T_0 = 0$, $T_1 = 256$.

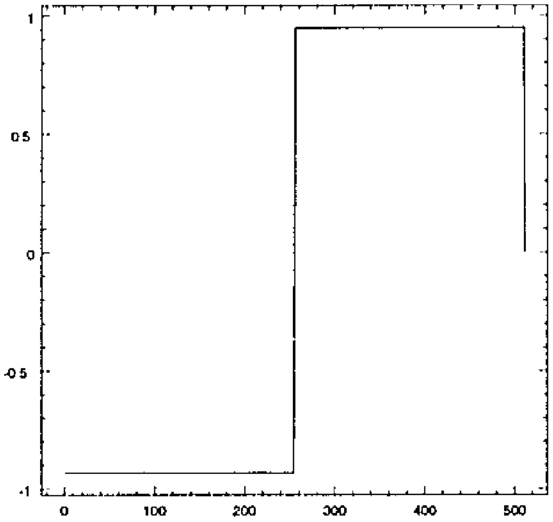

Figure III.10: Data for the switching autoregression with parameter $\beta_0 = -0.9$ and $T_0 = 0$, $T_1 = 256$.

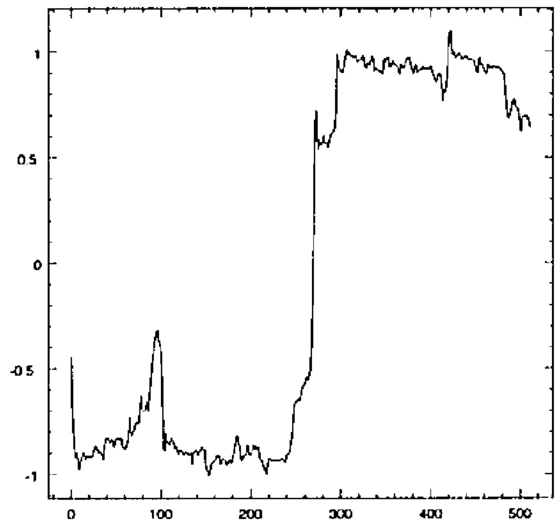

Figure III.11: Rolling least squares estimates for the first lag coefficient with a window of length 30.

$$\gamma(t) = \gamma_0 \, e^{-\frac{(t-t_0)^2}{\lambda_0^2}} \tag{III.60}$$

where $u(t)$ is a noise term. The smoothness of the gaussian function presents non-trivial problems in identification when it is not *a priori* known that the model is of the type (III.59) (III.60). We consider a simulated sample of size $T = 512$ with $\gamma_0 = 1$, $t_0 = 200$, $\lambda_0 = 20$. The simulated data is shown in Figure III.12. The autocorrelation function of the data is shown in Figure III.13. The reason the autocorrelation function shows little dependence is that the parameter variation in $\gamma(t)$ is well-localized.

Since the model is of slow but mean-reverting structural change, Table (III.1) suggests the use of model components with smooth window functions such as Gaussian functions. Here, we first suppose that theoretically we expect local stationarity so, based on these theoretical considerations, we use the wrong set of model components. We perform an analysis with flat windows for 5 levels and 10 lags for a total of 14,681 potential model components. On the first iteration, we pick a model component which

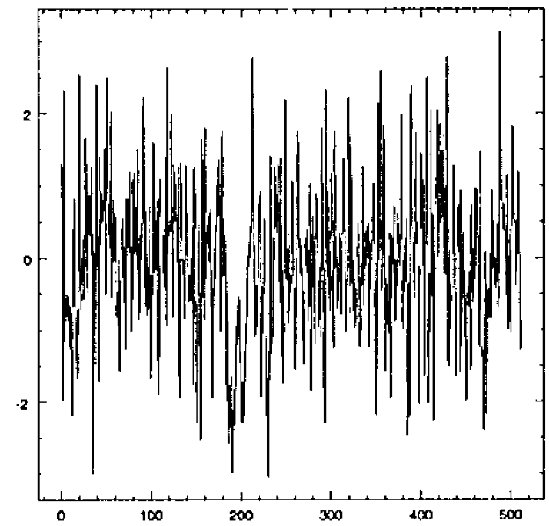

Figure III.12: Simulated data from the smoothly varying autoregressive model with $\lambda_0 = 20$, $\gamma_0 = 1.0$, $t_0 = 200$

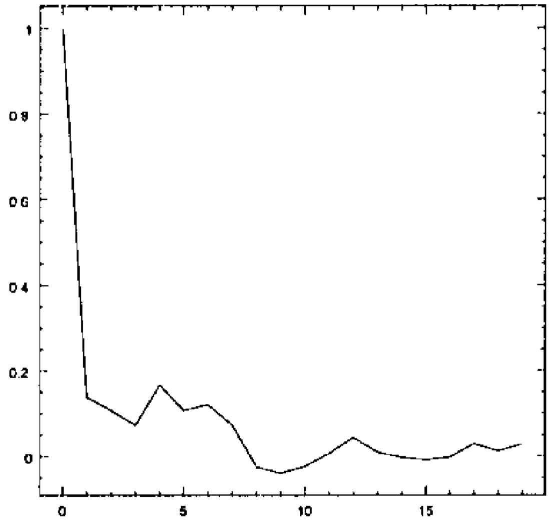

Figure III.13: Autocorrelation function of the smoothly varying autoregressive model. Since only part of the time series has nonzero autocorrelations, the autocorrelation function indicates only weak dependence in the data.

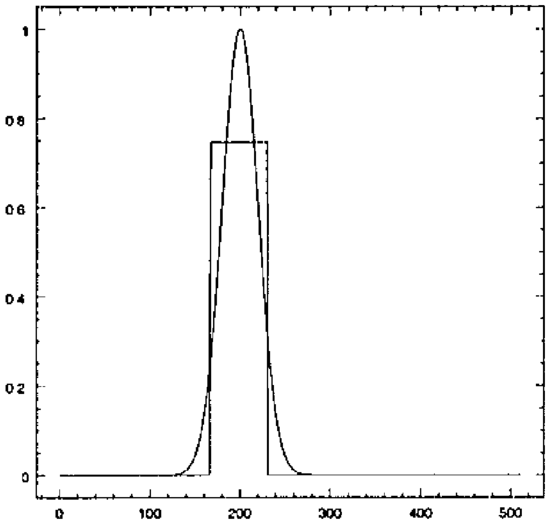

Figure III.14: Graph of the smooth time-varying autoregressive parameter $\gamma(t)$ versus estimates with flat windows.

a flat window from point 167 to 231. Our estimates are compared with the underlying population parameter in Fig. (III.14). The estimate we get has a coefficient of 0.746239 and a standard error of 0.0953892. On the next iteration, we picked a lag 3 component (from 387 to 419 with a coefficient of -0.542851 and a standard error of 0.137286); the percentage of additional variation explained as a result of including a second model component is only 2.6% as opposed to the 10.68% gained by the first model component on the first iteration.

Now, we turn to the case in which we include the correct model family in our analysis. As we reviewed above, this involves use of model components with smooth windows such as a Gaussian:

$$g_m(t) = e^{\frac{-(t-b_m)^2}{2c_m}} \tag{III.61}$$

where the parameter c_m is set as 2^{2k} where k is the level of the window $1 \leq k \leq k_{\max}$ where $k_{\max}$ is the maximum level. In this case, we include 7 lags and 5 levels. We

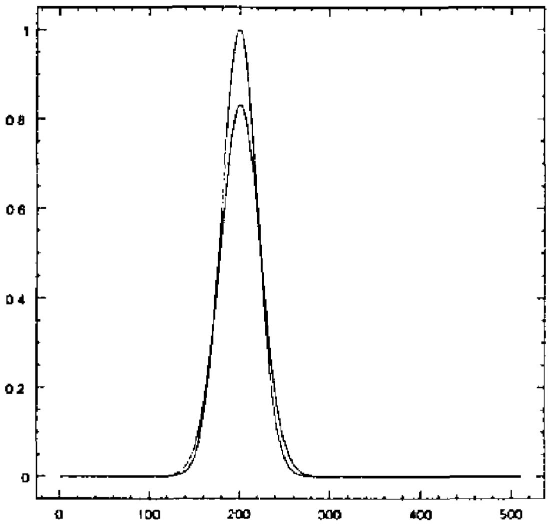

Figure III.15: Graph of the smooth time-varying autoregressive parameter $\gamma(t)$ versus estimates with Gaussian windows.

naturally pick a first order Gaussian window on the first iteration which matches the true parameter quite well; results are shown in Fig. (III.15). Our estimate is centered at the point 201 which is remarkably close to the true value of 200; our point estimate on the model component parameter γ_0 is 0.858678 (standard error of 0.115864) which is below the true value of 1.0. If we had continued the analysis, we would have picked next a lag 4 window which adds only 1.2% of explanatory power.[16]

If one were to estimate a Box-Jenkins autoregressive model (spuriously since the true model is *not* time-invariant), three possible estimated equations would be (standard errors of estimates are in parantheses):

$$\hat{y}(t) = \underset{(0.0440702)}{0.139798}\, y(t-1) \qquad \text{(III.62)}$$

[16] The stopping rule we propose in Ch. VI would have ended the procedure after the first model component was selected.

$$\hat{y}(t) = \underset{(0.0440702)}{0.127281}\, y(t-1) + \underset{(0.0440702)}{0.0894933}\, y(t-2) \tag{III.63}$$

$$\hat{y}(t) = \underset{(0.0441996)}{0.122991}\, y(t-1) + \underset{(0.0443783)}{0.0834274}\, y(t-2) + \underset{(0.44202)}{0.476624}\, y(t-3) \tag{III.64}$$

From the results, a researcher would identify a first or second order autoregressive model with weak correlations whereas the true model has a very strong but well-localized correlation structure. This example points to the danger of using stationary time series models blindly on nonstationary data.

Kernel estimates of the first autoregressive parameter are also difficult to interpret. We define the kernel estimate:

$$\hat{\beta}_1(t) = \frac{\sum_s g(t-s)y(s)y(s-1)}{\sum_s g(t-s)y(s-1)^2} \tag{III.65}$$

We let g be a Gaussian kernel with variance σ^2. The parameter estimates for β_1 for one values of σ is shown in Figure III.16. Estimates can either be too smoothed out or too choppy depending on choice of bandwidth.

Lag Switching Autoregressions

As one additional example, we consider another type of nonstationarity which is natural in economic and financial time series but for which it is inconvenient to use methods such as kernel regression. This situation is one in which lag relationships change over time. An appropriate model to capture such phenomena is:

$$y(t) = \beta_0\, y(t - \gamma(t)) \tag{III.66}$$

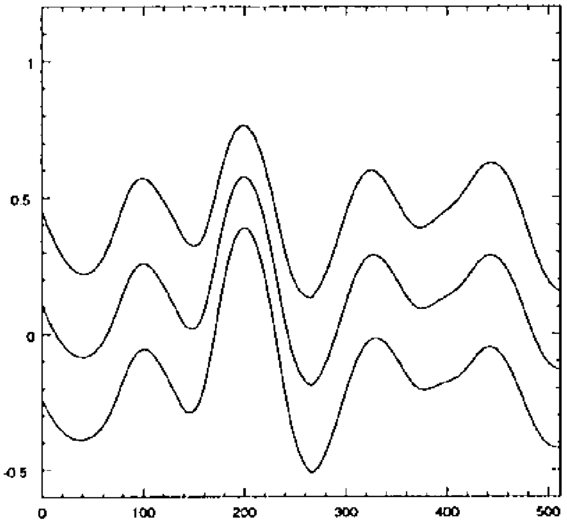

Figure III.16: Kernel regression estimates with a Gaussian kernel and $\sigma = 50.0$.

$$\gamma(t) = \gamma_0 + \gamma_1(t) \quad \gamma_0 \in \mathbb{Z}^+ \tag{III.67}$$

$$\gamma_1(t) = \begin{cases} 0 & t \leq T_0 \\ 1 & T_0 < t \leq T_1 \\ 2 & T_0 \leq t < T \end{cases} \tag{III.68}$$

We simulate data with $\beta_0 = 0.7$, $\gamma_0 = 1$, $T_0 = 300$, $T_1 = 400$, $T = 512$ (see Figures III.17 and III.18).

If we were to estimate ordinary autoregressive models on the simulated data, we would find:

$$\hat{y}(t) \;=\; \underset{(0.0432854)}{0.203731}\, y(t-1) \tag{III.69}$$

$$\hat{y}(t) \;=\; 0.128682\, y(t-1) + 0.37132\, y(t-2) \tag{III.70}$$

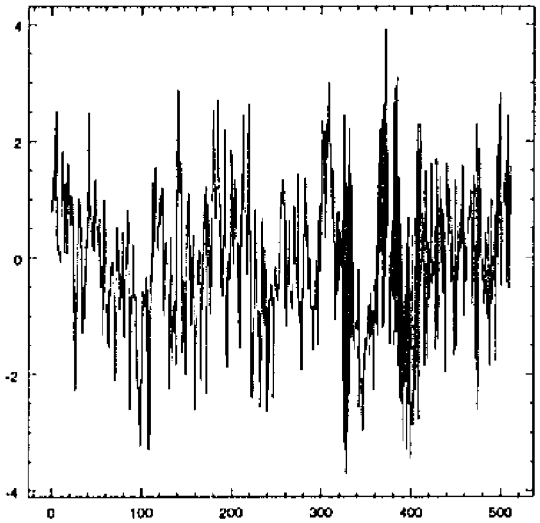

Figure III.17: Simulated data for the lag switching autoregressive process with $\beta_0 = 0.7$, $\gamma_0 = 1$, $T_0 = 300$, $T_1 = 400$, $T = 512$.

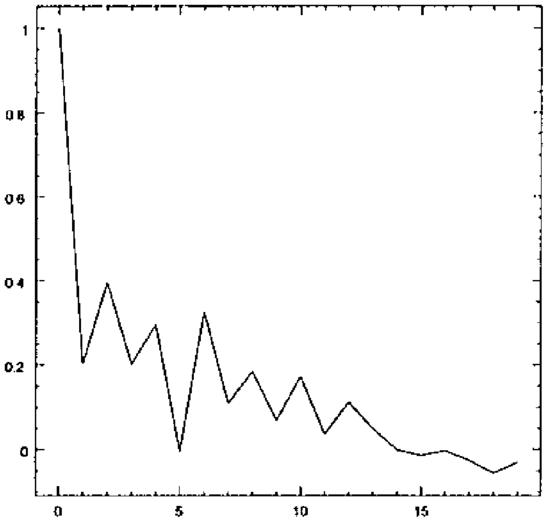

Figure III.18: Autocorrelation function for simulated data for the lag switching autoregressive process with $\beta_0 = 0.5$, $\gamma_0 = 1$, $T_0 = 300$, $T_1 = 400$, $T = 512$.

Lag	Coefficient	Std. Error
1	0.0816962	0.0437613
2	0.308413	0.043828
3	0.0782034	0.0438642
4	0.1427	0.0438529

Table III.3: Estimates for an AR(4) model for the lag switching model.

$$(0.0410522) \qquad (0.041191)$$

$$\begin{aligned}\hat{y}(t) \;&=\; 0.0946797\,y(t-1) + 0.359532\,y(t-2) + 0.0917562\,y(t-3) \quad \text{(III.71)}\\ &\quad\;\;(0.0440274) \qquad\qquad (0.0413369) \qquad\qquad (0.0441152)\end{aligned}$$

Estimates for two higher order models are shown in Table (III.3) and Table (III.4). The researcher who estimated an autoregressive model for this dataset would probably spuriously pick a high order model (the sixth lag is statistically significant with a t statistic of 5.577).

In this case, the model is locally stationary, but due to the shifts in the lag relationships, we might reason that observed structural change would be rapid but not immediate. As summarized in Table (III.1), this suggests model components which have flat windows with smooth Gaussian edges. We consider such model components up to five lags and five levels. Estimates after three iterations are shown in Fig. (III.19).[17] After the third iteration, there is a large drop in explanatory power, from 9.5% explained to 1.7% additional variance explained.

[17] t-statistics for the lag coefficients are: 10.79, 14.261, and 8.129 after three iterations.

Lag	Coefficient	Std. Error
1	0.107812	0.0434284
2	0.322125	0.0431824
3	0.136291	0.045185
4	0.157851	0.043233
5	-0.187147	0.0437018

Table III.4: Estimates for an AR(5) model for the lag switching model.

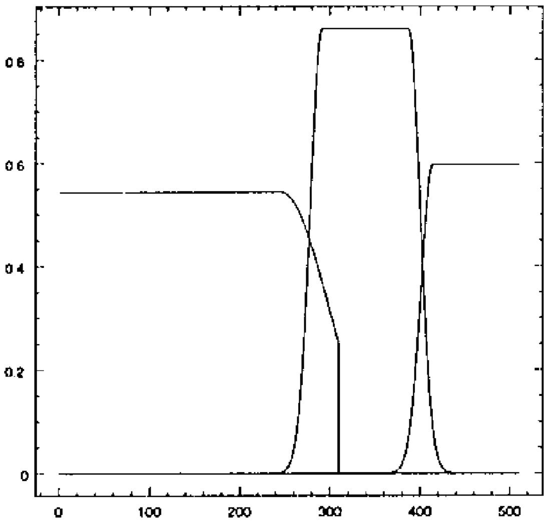

Figure III.19: Estimates of β_1, β_2 and β_3 for the lag switching autoregressive model. The first window starting at $T = 0$ corresponds to lag 1, the second window corresponds to lag 2 and the third window corresponds to lag 3.

Summary

In this chapter, we have reviewed the construction of model components and other issues relating to implementation of our method and have provided "proof of concept" that on certain model problems, our approach leads to some insight into the nature of the data generating process. In the next chapter, we provide some theoretical justifications for the use of the method.

CHAPTER IV

THEORETICAL ANALYSIS

The purpose of this chapter is to examine some of the theoretical properties of our procedure. Although classical statistical methods may seem conceptually incompatible with nonstationary time series models, we shall attempt to make some general statements about our method in terms of classical criteria. In terms of the basic equation:

$$y(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t), \tag{IV.1}$$

we analyze the properties of estimates from two broad classes of time series models:

- *Fraction Models.* Time series models with parameters which are functions of the *fraction* of time rather than time. In this case, the $\beta_j(t) = b_j(\frac{t}{T})$ where T is sample size and b_j is a measurable function of $\frac{t}{T}$. Thus, for example, $\beta_j(t)$ might assume one value for one half of the time series and another value for the other half.

- *Replication Models.* Models with periodic data generating processes. In this case, we can consider possibilities such as: (1) $\beta_j(t + rK) = \beta_j(t)$ for some integers r and K where K is the period of the data generating process; (2) there are multiple *replications* from the same nonstationary data generating process, a case which has been used for analysis of estimators of nonstationary models in fields outside economics.

Both broad classes of nonstationary models have different theoretical uses. Analysis of fraction models provides information on what will occur if we stretch out the data generating process so that we assume we have locally an infinite number of observations. Thus, from analysis of fraction models, we learn what properties of model components and estimation procedures are good as we increase the amount of data.

Analysis of replication models provides information on what will occur if we have very little data locally, but we can make inference by comparing two very similar processes. While replication models might at first glance seem to be only useful for engineering and medical processes, there are economic applications of such models. For instance, we may want to estimate the time series behavior of options prices as a function of time to maturity and it may not be unreasonable to assume that different options contracts exhibit very much the same statistical properties so that we can learn about local properties of the data by averaging across contracts.

We now outline the work to be done. We begin by reviewing some theoretical issues behind both types of models which will be useful to researchers; these theoretical issues include conditions on data generating processes and choices for families of model components.

Our next step is to prove that, under technical assumptions, our procedure converges in the sense that, as we increase the number of model components when there is an infinite amount of data, we approach the underlying model. The convergence proof is a probabilistic interpretation and slight variation on the proofs for the convergence of matching pursuit expansions for nonorthogonal expansions of functions in terms of waveforms due to [131] and for convergence of projection pursuit regression by Jones [112].

We then proceed to examine whether the assumptions behind the convergence proofs apply for the types of models we consider. The primary technical machinery here is the use of mixingale theorems due to McLeish [135] and Andrews [6]. Stochastic integrals are also used to approximate the properties of estimators. We include a review of mixingale theory so as to make this chapter as self-contained as possible. The chapter concludes with a synopsis of the main results.

Characteristics of Models

In this section, we review some of the theoretical issues and assumptions for the data generating processes and the choice of model components. As we shall discuss later, many specific theoretical results hold in a broader setting than described in this section.

Data Generating Process

For the data generating process:

$$y(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t), \tag{IV.2}$$

we assume that $\epsilon_1(t)$ is serially independent and identically distributed. We assume that:

$$E(\epsilon_1^2(t)) = \sigma^2 < \infty \tag{IV.3}$$

and that $\epsilon_1(t)$ has 'more' than four bounded moments.

We also assume that process generating y also has more than four bounded moments:[1]

[1] What really is required is that $y(t-1)\epsilon(t)$ has more than two bounded moments. We have used the Cauchy-Schwartz inequality.

$$\sup_t E(\,y(t)^{4+\delta}\,) < \infty \tag{IV.4}$$

for some $\delta > 0$. We note that this assumption places implicit restrictions on the types of admissible functions $\beta_j(t)$ in Eq. (IV.2).[2]

Fraction Models

We will discuss the choice of model components in the context of both fraction and replication models. For fraction models, we have:

$$\beta_j(t) = b_j\left(\frac{t}{T}\right) \tag{IV.6}$$

so that in terms of Eq. (IV.2):

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} b_j\left(\frac{t}{T}\right) y(t-j) + \epsilon_1(t). \end{aligned} \tag{IV.7}$$

We assume that $s \mapsto b_j(s)$ is a well-defined (measurable) and square integrable function on the interval $[0,1]$.[3]

Given our assumptions, $b_j(s)$ admits an expansion of the form:

$$b_j(s) = \sum_{r=1}^{R} \nu_r^j e_r^j(s) + \epsilon_3^j(s) \tag{IV.8}$$

where: [4]

[2] For instance, this assumption rules out a regression equation with a unit root such as:

$$y(t) = \beta y(t-1) + \epsilon_1(t) \tag{IV.5}$$

with $\beta = 1$.

[3] The notation $s \mapsto b_j(s)$ means b_j is a function of s.

[4] A technical aside: the function b_j is an element of $\mathrm{L}^2[0,1]$ and this footnote explains the sense in which the expansion in Eq. (IV.8) converges as $R \to \infty$. We will discuss both cases where the

(1) ν_r^j are constant coefficients;

(2) $e_r^j(s)$ is a basis function (the $e_r^j(s)$ need not be orthonormal; in fact, we will relate the $e_r^j(s)$ to the window functions used in model components).

(3) $\epsilon_3^j(s)$ is the approximation error from: (a) *truncation error* – using only R terms in the expansion, and (b) norm convergence as $R \to \infty$ need not imply uniform convergence.

Thus, we may consider:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} b_j\left(\frac{t}{T}\right) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J}\sum_{r=1}^{R} \nu_r^j e_r^j\left(\frac{t}{T}\right) y(t-j) + \epsilon_2(t) \\ &= \sum_{k=1}^{K} \alpha_k h_k(t) + \epsilon_2(t) \end{aligned} \tag{IV.11}$$

where $\epsilon_2(t)$ incorporates the approximation error in using the approximating functions e_r^j for the lag functions $\beta_j(t)$.[5] In this case, the α_k are the appropriate ν_r^j and the model components $h_k(t)$ are:

e_r^j are orthonormal and cases where they are not. This added generality is important because the window functions in our analysis (e.g., Gaussian windows or flat windows) are rarely orthonormal.

The sense in which Eq. (IV.8) is true is that, given a complete set $\{e_r\}_{r=1}^{\infty}$ of basis fucntions (which as noted in Eq. (IV.8) depend on j), there exist coefficients $\{\nu_r\}_{r=1}^{\infty}$ such that:

$$\lim_{N\to\infty} ||b_j - \sum_{r=1}^{N} \nu_r e_r||^2 = 0. \tag{IV.9}$$

The result holds as long as for all $b_j \in \mathbf{L}^2[0,1]$, we have:

$$A||b_j||^2 \leq \sum_{r=1}^{\infty} |<b_j, e_r>|^2 \leq B||b_j||^2 \tag{IV.10}$$

for some $A > 0$ and $B < \infty$; in this case, the e_r form a frame of $\mathbf{L}^2$ (c.f. [211] [19]). The e_r need not be linear independent. When the e_r are an orthonormal basis, they satisfy Eq. (IV.10) with $A = 1$ and $B = 1$.

[5] This approximation error comes from the ϵ_3^j in Eq. (IV.8).

$$\begin{aligned} h_k(t) &= e_{r_k}^{j_k}\left(\frac{t}{T}\right) y(t-j_k) \\ &= g_k\left(\frac{t}{T}\right) y(t-j_k) \end{aligned} \tag{IV.12}$$

where j_k is the lag associated with term k in Eq. (IV.12) and r_k indexes the expansion function associated with term k. Eq. (IV.12) suggests that to analyze the fraction model theoretically, we need to make the window functions g_k for model components h_k functions of the fraction of time rather than time itself.

We assume that the associated window functions g_k are normalized so that:

$$\frac{1}{T}\sum_{t=1}^{T}\left|g_k\left(\frac{t}{T}\right)\right|^2 = 1. \tag{IV.13}$$

This normalization assumption is made for consistency with the underlying assumption that the window function is a functic·· of the fraction of time and hence 'grows' as sample size is increased. The window functions g_k technically depend on sample size T; we do not indicate the dependence at this point to simplify notation.

To illustrate the effect of the constraint in Eq. (IV.13), we show in Fig. (IV.1) an example of a normalized flat window where $T = 512$, $b_k = 0.25$, $e_k = 0.50$ in Eq. (IV.14). Thus, a flat window is defined for technical purposes as:

$$g_k\left(\frac{t}{T}\right) = 1_{[b_k,e_k]}\left(\frac{t}{T}\right)\frac{1}{\sqrt{e_k - b_k}} \tag{IV.14}$$

where b_k and e_k are the fractions for the beginning and end of the window g_k.[6] We

[6] We note that for the flat window described by Eq. (IV.14):

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\left|g_k\left(\frac{t}{T}\right)\right|^2 &= \frac{1}{T}\sum_{t=1}^{T} 1_{[b_k,e_k]}\left(\frac{t}{T}\right)\frac{1}{e_k - b_k} \\ &= \frac{1}{T}(e_k - b_k)T\frac{1}{e_k - b_k} = 1. \end{aligned} \tag{IV.15}$$

as required by Eq. (IV.13).

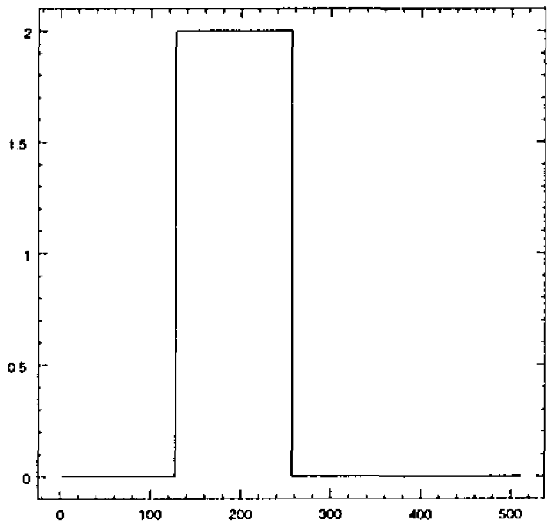

Figure IV.1: A normalized flat window (t is on the horizontal axis).

note that when $b_k = 0$ or the beginning of the time series and $e_k = 1$ or the end of the time series, the window functions are the same as are used in the time-invariant autoregressive Box-Jenkins model.

We also assume that all window functions are bounded and satisfy:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \left| g_k\left(\frac{t}{T}\right)\right|^4 < \infty \tag{IV.16}$$

We can verify that Eq. (IV.16) holds for the choice of constant windows in Eq. (IV.14) as long as the length of the window is a fraction of the length of the time series.[7] In fact, we will show that, in selecting a model, it may be desirable to weight r^2 (simple correlation) by some function of the inverse of:

[7] To see this, we note that (using Eq. (IV.14)):

$$\frac{1}{T}\sum_{t=1}^{T} \left| g_k\left(\frac{t}{T}\right)\right|^4 = \frac{1}{T}\sum_{t=1}^{T} 1_{[b_k, e_k]}\left(\frac{t}{T}\right)\frac{1}{(e_k - b_k)^2} = \frac{1}{e_k - b_k} \tag{IV.17}$$

For this to be finite, $(e_k - b_k)$ must be bounded away from zero.

$$\mu_k^4 = \frac{1}{T}\sum_{t=1}^{T}\left| g_k\left(\frac{t}{T}\right)\right|^4 \tag{IV.18}$$

such as $\frac{1}{\sqrt{\mu_k^4}}$.

To effectively approximate the process generating the data, the window functions in our analysis must lie in the same class as the lag coefficients so that we require that there is at least one way of expressing the true lag functions $\beta_j(t)$ in terms of the window functions we include in our model components.

Replication Models

For replication models, we may wish to consider: (1) situations in which the $\beta_j(t)$ are periodic with period K; or (2) situations in which the data generating process on the interval $[1, K]$ is repeated many times. In the former case:[8]

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t)\, y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \lambda_j r_j \left(t - \left[\frac{t-1}{K}\right] K\right) y(t-j) + \epsilon_1(t) \\ &= \sum_{j=1}^{J} \alpha_j\, h_j(t) + \epsilon_1(t) \end{aligned} \tag{IV.19}$$

so that (as we have already reviewed in Ch. III), the model components we use are:

$$h_j(t) = r_j\left(t - \left[\frac{t-1}{K}\right] K\right)\, y(t-j). \tag{IV.20}$$

In the latter case where the data generating process is repeated many times, we have for each replication that:

$$y(t) \;=\; \sum_{j=1}^{J}\beta_j(t)y(t-j) + \epsilon_1(t)$$

[8] We recall from Ch. III that the notation $[z]$ is used to refer to the integer part of z.

$$
\begin{aligned}
&= \sum_{j=1}^{J} \lambda_j r_j(t) y(t-j) + \epsilon_1(t) \\
&= \sum_{j=1}^{J} \alpha_j h_j(t) + \epsilon_1(t) \qquad \text{(IV.21)}
\end{aligned}
$$

so that we use model components:

$$
h_j(t) = r_j(t)\, y(t-j). \qquad \text{(IV.22)}
$$

In such a situation, there are no regularity conditions on the window functions for the model components $h_j(t)$ other than that they are bounded. We receive a little bit more insight into useful conditions by considering continuous time models such as are commonplace in finance. In such cases, we require that the processes have mean square continuous sample paths with probability one on the interval $[1, K]$. Continuous time models are attractive because they allow for irregularly spaced data, natural stock/flow distinctions and, in finance, follow naturally from the continuous time theory. A review of the use of continuous time models in econometrics can be found in [22].

An example of such a model is a nonstationary continuous time $AR(1)$ process:[9]

$$
\begin{aligned}
dY(t) &= -\lambda_1(t) Y(t) dt + \sigma dW \\
&= -\left[\sum_{j=1}^{\infty} \alpha_j e_j(t)\right] Y(t) dt + \sigma dW \\
&= -\sum_{j=1}^{\infty} \alpha_j h_j(t)\, dt + \sigma dW \qquad \text{(IV.23)}
\end{aligned}
$$

where we have assumed that $\lambda_1(t)$ is measurable and square integrable (on the interval $[0, K]$) in moving from Eq. (IV.23), hence suggesting that in the limit of continuous

[9] There is a theory of such equations (c.f. [115], p. 354-363) and it is known that with mild smoothness conditions on $\alpha_1(t)$, on any finite interval the solution has sample paths with probability 1 which are continuous.

time, we need the measurability of windows with respect to time t in the same way as in the fraction model we required measurability with respect to the fraction of time. Clearly, when a time series is discrete and the interval is finite, there is not an issue of measurability with respect to time, hence the lack of need for formal regularity conditions. Nevertheless, in spirit, the continuous time limit suggests that our model components should not be too 'short'.

Convergence Proof

In the previous section, we have made specific technical assumptions on types of models. In this section, the goal is to prove that our procedure converges in the sense that, if we keep on adding model components, we approach the underlying model. Since the models described in the above section are different, it is useful to define a broader umbrella of models which encompasses the models in the previous section and then explain precisely how the models in the previous section fit in. The use of unifying notation in this section has an added advantage (besides not having to prove the same theorem twice) in that the theorem seems applicable quite generally to estimation of various forms of nonlinear regression procedures including state-dependent models in a time-series context as well as nonstationary time series models outside the scope of this thesis such as those with time-varying cointegrating relationships.

The general model we will use has the following elements:

- **Data.** The dependent variable is denoted by y_t where t indexes the observation. The observed regression variables (which will be lagged y_t in our context) are a vector denoted by x_t. We assume there are J elements in the vector x_t and the j th element is indexed by x_t^j. There is also a vector of auxiliary variables v_t In our case, v_t denotes time or some related variable; in other models, v_t might represent other state variables.

- **Regression Function.** We assume there is a regression function f:

$$f(v_t, x_t) = E(y_t | V_t = v_t, X_t = x_t) \tag{IV.24}$$

where X_t is a vector of explanatory variables indexed by t and V_t is a vector of auxiliary variables and that this function is of the form:

$$f(v_t, x_t) = \sum_{j=1}^{J} c_j(v_t) x_t^j \tag{IV.25}$$

This assumption may seem strong but, if the conditional expectation is not given by Eq. (IV.25), estimates will converge to the best projection of the form Eq. (IV.25) under the regularity conditions described below.

- **Model Components.** We assume we have a family of model components indexed by k:

$$h_k(t) = g_k(v_t) \cdot x_t^j \tag{IV.26}$$

where g_k is a window function. The dot in Eq. (IV.26) allows the model components to consist of a (finite) sum over many different variables (such as was necessary for the distributed lag model in Ch. III).

We now show how our two classes of models fit into this framework. For fraction models, $v_t = \frac{t}{T}$ (e.g., v_t is the fraction). When relationships are locally stationary so that the $\beta_j(t)$ are flat, it is also possible to let the v_t index the particular stationary regime; thus, for instance, we may have for some fractions b_0 and e_0:

$$v_t = \begin{cases} 0 & b_0 \le \frac{t}{T} < e_0 \\ 1 & \text{otherwise} \end{cases} . \tag{IV.27}$$

In the fraction model, the x_t^j variables are the lagged j data generated from y_t. Thus, the $c_j(v_t)$ functions are the autoregressive functions $b_j\left(\frac{t}{T}\right)$, the $g_k(v_t)$ functions are the window functions $g_k\left(\frac{t}{T}\right)$ used in the analysis, and the $h_k(t)$ are the model components.

For replication models, $v_t = t - \left[\frac{t-1}{K}\right] K$ if the data generating process is periodic with period K and $v_t = t$ if we are considering repeated runs from a data generating process with length K. In such cases, the $c_j(v_t)$ are the appropriate autoregressive functions and the $g_k(v_t)$ are the window functions.

Given models of this class, the next step is to define some basic rules for model selection which narrow down the class of procedures we will consider. These technical assumptions include:

Inner Products. The convergence theorem is a Hilbert space convergence theorem which holds more generally (see Appendix A for a review of the relationship between Hilbert spaces and probability theory; Brockwell and Davis ([33], Ch. 2) use Hilbert space methods in time series). The inner product is defined (in its most general form) as:

$$\mathcal{E}(f\,g) = \int f(x,v)\,g(x,v)\,d\mu(x,v) \tag{IV.28}$$

since f and g are functions of x and v. Here, $\mu(dx,dv)$ is the measure for x and v. The symbol $\mathcal{E}$ *is not* in general an expectation operator. For $g = f$, Eq. (IV.28) defines a norm on a Hilbert space $\mathcal{H}$:

$$||f||_{\mathcal{H}}^2 = \mathcal{E}(f^2) = \int |f(x,v)|^2\,d\mu(x,v). \tag{IV.29}$$

We now define a number of specific cases of inner product (IV.28) for the special models we consider. For discrete *replication* models, v_t is time or time modulo a period. In this case, v_t takes values from 1 to K so that:[10]

$$\begin{aligned}
\mathcal{E}(f\,g) &= \frac{1}{K}\sum_{v=1}^{K}\int f(x,v)\,g(x,v)\,d\mu_v(x)dv \\
&= \frac{1}{K}\sum_{v=1}^{K} E_v(f(x,v)g(x,v)). \qquad \text{(IV.31)}
\end{aligned}$$

For the *fraction* model, we define the inner product for two separate cases. In the first case, we assume a finite number R of regimes (because there may be several structural changes), indexed by r, which each cover a fraction $\nu(r)$ of the sample. In this case, Eq. (IV.28) becomes:

$$\begin{aligned}
\mathcal{E}(f\,g) &= \sum_{r=1}^{R}\int f(x_t,v_t)\,g(x_t,v_t)\,\nu(r)\,1_{\{t\in r\}}(r)\,d\mu_r(x) \\
&= \sum_{r=1}^{R}\nu(r)E_r(f(x_t,v_t)g(x_t,v_t)) \qquad \text{(IV.32)}
\end{aligned}$$

where there is a separate (stationary) joint distribution function $\mu_r(dx)$ for each r. We use the terminology E_r to refer to the expectation taken in regime r. With an infinite amount of data, the definition of such a stationary distribution function is usually achieved through a spectral representation (c.f., Appendix A). This representation is useful because we have some theoretical results to be reviewed below for models with a finite number of stationary regimes.

We also define another special case of Eq. (IV.28):

$$\mathcal{E}(f\,g) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\int f(x_t,v_t)g(x_t,v_t)d\mu_t(x)$$

[10] In the continuous case, the sample paths of x are measurable by assumption (as we have pointed out above, this is not restrictive for models in finance). Here, $v_t = t$ or time modulo a period:

$$\mathcal{E}(f\,g) = \frac{1}{K}\int_1^K dv\int f(x,v)\,g(x,v)\,d\mu_v(x). \qquad \text{(IV.30)}$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E(\, f(x_t, v_t)\, g(x_t, v_t)) \tag{IV.33}$$

We can relate this inner product to those of special Hilbert spaces called *Hilbert spaces of almost periodic functions* or continuous functions with finite average squared variation which were developed by mathematicians Bohr and Bochner and others (c.f. [4], Vol. I, Sec. V, pp. 132-138; [178], pp. 254-260) and are related to the spectral representations of stationary stochastic processes. There are a number of technical issues such as the definition of a zero element. These issues are reviewed in Appendix N.

We can also define inner products (in the sense of Eq. (IV.28) which do not involve taking expectations. For instance, we can define for any T (including in the limit $T \to \infty$):

$$\mathcal{E}(f\, g) = \frac{1}{T} \sum_{t=1}^{T} f(x_t, v_t) g(x_t, v_t). \tag{IV.34}$$

Model component selection. Let the estimates at iteration k be y^k (where $y^0 = y$). We define $m^k = y^{k-1} - y^k$ so that m^k represents the terms added to the regression estimates at iteration k. We assume at each iteration k:

$$\mathcal{E}((m^k)^2) \geq \alpha \sup_{\theta \in C} \mathcal{E}((m^k_\theta)^2) \tag{IV.35}$$

for some number $\alpha \in (0, 1]$.[11] The term m^k_θ represents what $y^{k-1} - y^k$ would have been if we had used the model component h_θ instead of the one we chose h^k. The set C may be uncountable; as an example, we show in Ch. V that there are cases in which we can select model components by nonlinear regression. We note that Eq.

[11] We note that $\mathcal{E}((m^k)^2)$ is defined by Eq. (IV.29). The number α may be less than one if we use suboptimal procedures (such as choosing model components which maximize r^2 instead of additional R^2) in selecting the next model component to include in the model.

(IV.35) allows us to select models which are suboptimal; this condition allows us to use statistical weights in model selection instead of always choosing the maximum.

Completeness. We assume that if the regression function $f(x,v) \in \mathcal{H}$, then the span of the model components in the analysis is $\mathcal{H}$. This condition can be difficult to verify in practice since the model components in the analysis depend on the regression function $f(x,v)$. It is thus useful to provide alternative conditions for completeness. In Appendix O, we show that (under regularity conditions) if the window functions $g_k(v)$ for model components containing each explanatory variable x^j span the space containing all the possible $c_j(v)$ (c.f., Eq. (IV.25)), then the resulting set of model components is complete.[12]

Projection Operators. At any stage of the procedure, we need to compute $m^k = y^{k-1} - y^k$. To do this, we decompose:

$$\begin{aligned} y^{k-1} &= P_k y^{k-1} + (I - P_k) y^{k-1} \\ &= m^k + y^k \end{aligned} \tag{IV.37}$$

where P_k is the projection operator at iteration k. For our procedure, this is captured by the predicted value from a linear regression of the residual y^{k-1} on all previously

[12] Specifically, we assume that, for each explanatory variable, there exists a countable subset of window functions $\mathcal{W}$ which are incorporated into model components in the analysis such that:

$$A||c_j||^2 \leq \sum_{g_k \in \mathcal{W}} |< c_j, g_k >|^2 \leq B||c_j||^2 \tag{IV.36}$$

for some $A > 0$ and $B < \infty$ and where the inner product is weighted by the measure $d\mu(v)$. In this case, the g_k form a frame of $\mathbf{L}^2$ (c.f. [211] [19]). In our analysis, the v_t either take on a finite number of values (for the discrete replication model) or take values on a finite interval ($[0,1]$ for the fraction model and $[0,K]$ for a continuous time replication model). In such cases, the coefficient functions c_j lie in separable spaces so that they are spanned by a countable set of window functions. The measure $d\mu(v)$ need not be uniform for the discrete replication model because regimes may be of different lengths.

selected model components h^j for $j < k$ and the newly selected model component h^k. Other definitions of the projection operator P_k are also consistent with convergence.[13] The properties of linear projection operators on Hilbert space and the relationship with regression analysis is reviewed for instance in ([33], Ch. II) ([90], pp. 113-4) ∎

We now are ready to proceed to a proof.

Theorem 1 *Assume: 1. the regression function $f(x, v)$ is measurable with respect to an inner product P generating the values of x and v.*

2.

$$\mathcal{E}(f(x,v)^2) < \infty. \tag{IV.38}$$

3. We pick a set of model components $h^k(x, v)$ which at each iteration k satisfy Eq. (IV.35) so that:

$$\mathcal{E}((m^k)^2) \geq \alpha \sup_{\theta \in C} \mathcal{E}((m_\theta^k)^2) \tag{IV.39}$$

for some fixed $\alpha \in (0,1]$. Here, $m_\theta^k = P_{k,\theta} f^{k-1}$ is the projection on the residual induced by the choice of a model component h_θ at iteration k. This projection may for instance be against the span of all model components chosen before stage k and a new model component h_θ. We define m^k to be m_k^k. The supremum in Eq. (IV.39) is taken with respect to $h_\theta \in C$, their induced projections $P_{k,\theta}$ and the projections $P_{j,j}$, $j < k$, induced by the selected model component h^j at iteration $j < k$.

4. The set of model components is complete in the sense that $f \in \mathcal{H}$ and the span of the set of model components is $\mathcal{H}$.

[13] For instance, convergence still occurs if we define the projection operator to be the linear projection of the residual against the span of the chosen model component alone. Thus, in a practical sense, this added generality may serve to justify a broader class of procedures such as researchers are indeed likely to use in practice; for example, it allows for the standard of subtraction of trends and seasonal terms from data before analysis (but, perhaps constructively, suggests that the residual from may need to be detrended or deseasonalized itself).

5. Each member of the set of potential model components is measurable with respect to P and satisfies:

$$0 < \mathcal{E}(h(x,v)^2) < \infty. \tag{IV.40}$$

6. No model component is correlated (with respect to P) with the error term for the regression function:[14]

$$\mathcal{E}(h_k(x,v)\epsilon) = 0 \tag{IV.41}$$

for all $h_k \in C$. The error term ϵ has nonzero variance and is independently and identically distributed.

Given these assumptions, the procedure converges to estimates:

$$f(x,v) = \sum_k C_k h^k(x,v) \tag{IV.42}$$

in the sense that:

$$\lim_{N\to\infty} \mathcal{E}\left(f(x,v) - \sum_{k=1}^{N} C_k h^k(x,v)\right)^2 = 0. \tag{IV.43}$$

Proof: This proof follows Jones [112] as well as its translation by Mallat and Zhang [131] into the context of their matching pursuit decomposition. If the data are generated by the model:

$$y_t = f(x_t, v_t) + \epsilon_t \tag{IV.44}$$

where ϵ_t is independently and identically distributed and $\mathcal{E}(f(x,v)\epsilon) = 0$, then:

$$\mathcal{E}(y^2) = \mathcal{E}(f^2) + \mathcal{E}(\epsilon^2) = \mathcal{E}(f^2) + \sigma^2 \tag{IV.45}$$

[14] In our case, this implies no contemporaneous correlation.

where σ^2 is the variance of ϵ_t. By multiplying Eq. (IV.44) by any model component h_k, taking "expectations", and applying Assumption (6):

$$\mathcal{E}(h_k y) = \mathcal{E}(h_k f) + \mathcal{E}(h_k \epsilon) = \mathcal{E}(h_k f). \qquad \text{(IV.46)}$$

We can thus restrict our attention to the analysis of f and prove that the residual of the regression function converges to zero.

At any stage j of the procedure, we pick a model component h^j and as a result add explanatory power m^j to the model. By the definition of m^j, we have for any choice of k:

$$f = \sum_{j=1}^{k} m^j + f^k \qquad \text{(IV.47)}$$

where f^k is the residual at stage k. We define $f^0 = f$.

The proof follows the following broad outline:

- **Bound $\mathcal{E}(f^N)^2$ for N large.** This follows since the residual from the procedure is monotonically decreasing.

- **Use triangle inequality.** This step is used to bound $\mathcal{E}(f^{N+K} - f^N)^2$.

- **Bound $\mathcal{E}(f^N - f^M)^2$.** Properties of the procedure are used to provide a bound for any N and M. This step uses Lemma 1.

- **Bound individual terms.** Here we use the bound from the previous step to bound individual terms which arise in the second step from the use of the triangle inequality. This step uses Lemma 2.

- **Cauchy convergence implies result.** We show that f^N is a Cauchy sequence implies result.

Before proceeding to the rest of the proof, we will first prove Lemma 1 and Lemma 2.

Lemma 1 *Let f^k be the residual of the regression function $f(x, v)$ at iteration k and let $f^0 = f$. Define $m_j^k = P_j f^{k-1}$ where P_j is a projection induced by the model component chosen at iteration j and the model components chosen before iteration j. We denote m_j^j by m^j.*

It follows then that for any $k \geq j \geq 1$:

$$|\mathcal{E}(m^j f^{k-1})| \leq \frac{1}{\sqrt{\alpha}} |\mathcal{E}((m^k)^2)|^{\frac{1}{2}} |\mathcal{E}((m^j)^2)|^{\frac{1}{2}} \tag{IV.48}$$

Proof:

This proof is a slight variation on a proof of a similar lemma by Jones [112] as well as a related lemma of Mallat and Zhang [131]. We have by Eq. (IV.39) that:

$$\begin{aligned} |\mathcal{E}(m^j f^{k-1})| = |\mathcal{E}(m^j m_j^k)| &\leq \left(\mathcal{E}((m^j)^2)\right)^{\frac{1}{2}} \left(\mathcal{E}((m_j^k)^2)\right)^{\frac{1}{2}} \\ &\leq \left(\mathcal{E}((m^j)^2)\right)^{\frac{1}{2}} \frac{1}{\sqrt{\alpha}} \left(\mathcal{E}((m^k)^2)\right)^{\frac{1}{2}} \qquad \text{(IV.49)} \end{aligned}$$

where m_j^k is the addition to the regression equation if at iteration k we use the model components used for calculating regressions at iteration j.

The first step follows from computing the expectation of f^{k-1} conditional on the use of the same regressors as are included in m^j and the Cauchy-Schwartz inequality. To see this, we note that we can write:

$$\begin{aligned} \mathcal{E}(m^j f^{k-1}) &= \mathcal{E}\left(m^j (P_j f^{k-1} + (I - P_j) f^{k-1})\right) \\ &= \mathcal{E}(m^j P_j f^{k-1}) = \mathcal{E}(m^j m_j^k) \qquad \text{(IV.50)} \end{aligned}$$

where the last step in Eq. (IV.50) follows from the definition of m_j^k.[15] Using Eq. (IV.50), we complete the first line of Eq. (IV.49) by using the Cauchy-Schwartz inequality to bound $|\mathcal{E}(m^j m_j^k)|$.

[15] We note that for the case of the simplified algorithm introduced in Appendix F, P_j is a projection onto the space spanned by the model component selected at iteration j. This point is discussed in more detail below.

The last step in Eq. (IV.49) follows since:

$$\mathcal{E}((m_j^k)^2) \leq \sup_{\theta \in \mathcal{C}} \mathcal{E}((m_\theta^k)^2) \leq \frac{1}{\alpha}\mathcal{E}((m^k)^2) \tag{IV.51}$$

The first inequality in Eq. (IV.51) follows since m_j^k (or the residual we do explain) cannot explain more than the best model component we could have chosen; the second inequality follows by dividing Eq. (IV.39) by α. ∎

Lemma 2 *Suppose r_n is a nonnegative sequence of real numbers such that $\sum_{n=0}^{\infty} r_n^2 < \infty$ then* $\liminf_{N \to \infty} r_N \sum_{n=0}^{N} r_n = 0$.

Proof: This proof is from [112]. For any $\epsilon > 0$ we select an N such that:

$$\sum_{n=N}^{\infty} r_n^2 < \frac{\epsilon}{2} \tag{IV.52}$$

Since $r_k \to 0$ we can always find some $k > N$ such that:

$$r_k \sum_{n=0}^{N} r_n < \frac{\epsilon}{2} \tag{IV.53}$$

We set r_i as the minimum term of the sequence r_j from $j = N+1, ...k$. Then:

$$r_i \sum_{n=0}^{i} r_n = r_i \sum_{n=0}^{N} r_n + r_i \sum_{n=N+1}^{i} r_n \leq \frac{\epsilon}{2} + \sum_{n=N+1}^{i} r_n^2 < \epsilon \tag{IV.54}$$

The first term in the last inequality follows from Eq. (IV.53) and $r_i \leq r_k$. The second term in the last inequality follows from the fact that r_i is the minimum term for r_n, $n = N+1...i$, so that:

$$r_i \sum_{N+1}^{i} r_n < \sum_{n=N+1}^{i} r_n^2 < \frac{\epsilon}{2} \tag{IV.55}$$

where the last inequality in Eq. (IV.55) follows from Eq. (IV.52). ∎

We now proceed to the remainder of the convergence proof.

Step 1: Bound $\mathcal{E}((f^N)^2)$. At any stage, we have that the average variance of the residual for the next stage f^N is equal to the average variance of the original regression function f^{N-1} minus the average variance of what we have subtracted off:

$$\mathcal{E}((f^N)^2) = \mathcal{E}\left((f^{N-1} - m^N)^2\right) = \mathcal{E}((f^{N-1})^2) - \mathcal{E}((m^N)^2) \tag{IV.56}$$

since $m^N = P_N f^{N-1}$. Iteration yields for any N:

$$\mathcal{E}((f^N)^2) = \mathcal{E}(f^2) - \sum_{k=1}^{N} \mathcal{E}((m^k)^2) \tag{IV.57}$$

Thus, we note that $\mathcal{E}((f^N)^2)$ is a bounded and monotonically decreasing sequence. It is bounded because $0 \leq \mathcal{E}((f^N)^2) \leq \mathcal{E}(f^2) < \infty$ and it is monotonically decreasing by Eq. (IV.56). Since every bounded, monotone sequence has a limit ([35], Thm. 6, p. 16), $\mathcal{E}((f^N)^2)$ converges as N gets large to some value which we will call λ_*.

For $\epsilon > 0$, we can choose some W (which depends on the choice of ϵ) such for all $Q > W$, $\mathcal{E}((f^Q)^2) \leq \lambda_* + \epsilon^2$ and:

$$|\mathcal{E}((f^Q)^2) - \mathcal{E}((f^R)^2)| \leq \epsilon^2 \tag{IV.58}$$

for all $Q, R > W$.

Although we know that $\mathcal{E}((f^Q)^2)$ converges to some limit, to prove convergence of f^Q as Q gets large, we need to show that $\{f^Q\}$ is a Cauchy sequence.[16]

Step 2: Use the triangle inequality. Our goal is to show Cauchy convergence of $\{f^Q\}$ so that for some large $Q > W$ and all R,

$$(\mathcal{E}(f^Q - f^{Q+R})^2)^{\frac{1}{2}} \tag{IV.59}$$

[16] We need to show that f^Q is a Cauchy sequence because $||f^Q||$ converging to λ_* does not imply that f^Q converges (where f^Q is a sequence of functions). To give an example, let $f^Q = e^{2\pi i \omega Q x}$, for $x \in [0, 1]$ then $||f^Q|| = 1$ for all Q but $f^Q(x)$ does not converge.

is small. Use the triangle inequality to write Eq. (IV.59) as:

$$(\mathcal{E}(f^Q - f^{Q+R})^2)^{\frac{1}{2}} \leq (\mathcal{E}(f^Q - f^Z)^2)^{\frac{1}{2}} + (\mathcal{E}(f^{Q+R} - f^Z)^2)^{\frac{1}{2}}, \tag{IV.60}$$

for some $Z > Q + R$.

Step 3: Bound $\mathcal{E}\left((f^N - f^M)^2\right)$. We consider some $M \geq 0$, $N > M$. Then,

$$\begin{aligned} \mathcal{E}\left((f^N - f^M)^2\right) &= \mathcal{E}\left((f^N - (f^N + \sum_{j=M}^{N-1} m^{j+1}))^2\right) \\ &= \mathcal{E}((f^N)^2) + \mathcal{E}((f^M)^2) - 2\mathcal{E}((f^N)^2) - \\ & \quad 2\sum_{j=M}^{N-1} \mathcal{E}(f^N m^{j+1}) \end{aligned} \tag{IV.61}$$

We now use the Lemma 1 to bound the last term. By the first lemma, each individual term in the series:

$$\sum_{j=M}^{N-1} \mathcal{E}(f^N m^{j+1}) \tag{IV.62}$$

satisfies:

$$|\mathcal{E}(m^{j+1} f^N)| \leq \frac{1}{\sqrt{\alpha}} \mathcal{E}((m^{j+1})^2)^{\frac{1}{2}} \mathcal{E}((m^{N+1})^2)^{\frac{1}{2}} \tag{IV.63}$$

Thus:

$$\mathcal{E}\left((f^N - f^M)^2\right) \leq \mathcal{E}((f^M)^2) - \mathcal{E}((f^N)^2) + \frac{2}{\sqrt{\alpha}} \mathcal{E}\left((m^{N+1})^2\right)^{\frac{1}{2}} \sum_{j=M-1}^{N} \mathcal{E}\left((m^j)^2\right)^{\frac{1}{2}} \tag{IV.64}$$

Step 4: Bound individual terms. We now wish to bound both terms on the right hand side of Eq. (IV.60). To do this, we will use use Eq. (IV.64) and set $M = Q$, $N = Z$, so that we have:

$$\begin{aligned}\mathcal{E}\left((f^Q - f^Z)^2\right) &\leq \mathcal{E}((f^Q)^2) - \mathcal{E}((f^Z)^2) + \frac{2}{\sqrt{\alpha}}\mathcal{E}((m^{Z+1})^2)^{\frac{1}{2}} \sum_{j=Q-1}^{Z} \mathcal{E}((m^j)^2)^{\frac{1}{2}} \\ &\leq \epsilon^2 + \frac{2}{\sqrt{\alpha}}\epsilon^2 \qquad \text{(IV.65)}\end{aligned}$$

The first inequality follows from the fact that for $Z > Q + R$, Eq. (IV.58) holds. The second inequality follows because from the second lemma, we can choose some $Z > Q + R$ such that:

$$(\mathcal{E}((m^Z)^2))^{\frac{1}{2}} \sum_{j=0}^{Z} \mathcal{E}((m^j)^2)^{\frac{1}{2}} \leq \epsilon^2. \qquad \text{(IV.66)}$$

We can use the second lemma because:

$$\sum_j \mathcal{E}((m^j)^2) \leq \mathcal{E}(f^2) < \infty \qquad \text{(IV.67)}$$

and $\mathcal{E}((m^j)^2)$ is non-negative.

Similarly:

$$\mathcal{E}\left((f^{Q+R} - f^Z)^2\right) \leq \epsilon^2 + \frac{2}{\sqrt{\alpha}}\epsilon^2. \qquad \text{(IV.68)}$$

Therefore:

$$(\mathcal{E}(f^Q - f^{Q+R})^2)^{\frac{1}{2}} \leq \sqrt{2}\epsilon\sqrt{1 + \frac{2}{\sqrt{\alpha}}} \qquad \text{(IV.69)}$$

which proves that f^Q is Cauchy.

Step 5: Cauchy convergence implies result.

We note that by Step 4:

$$\mathcal{E}\left((f^Q - f^{Q+R})^2\right) \rightarrow 0 \qquad \text{(IV.70)}$$

which implies by completeness that f^Q converges to an equivalence class of functions or sequences f^* ([193], pp. 98-99, 74-76).

Using Eq. (IV.57), we note that $\sum_{j=1}^{Q} \mathcal{E}((m^j)^2) \leq \mathcal{E}(f^2)$ which implies that:

$$\mathcal{E}((m^Q)^2) \to 0. \tag{IV.71}$$

We define: $m^* = \lim_{N\to\infty} P_N f^{N-1}$. By the regularity condition on model component choice, Eq. (IV.39):

$$0 = \mathcal{E}((m^*)^2) \geq \alpha \sup_{\theta\in\mathcal{C}} \mathcal{E}((m_\theta^*)^2). \tag{IV.72}$$

Therefore, $f^* = (I - P_{\mathcal{C}})f$ where $P_{\mathcal{C}}$ is the projection on the span of all model components in the analysis. If the set of model components is complete, $\mathcal{E}((f^*)^2) = 0$. Thus:

$$\lim_{N\to\infty} \mathcal{E}((f^N)^2) \to 0 \tag{IV.73}$$

which substituting Eq. (IV.47) for f^N in Eq. (IV.73) leads to the result (Eq. (IV.43)). ∎

Justifying the Assumptions

Since application of the convergence theorem relies on an absence of sampling error in any of the computed sample averages, we need to use (weak) laws of large numbers for dependent heterogeneous sequences to prove convergence in probability of sample averages. We shall focus here on issues relating to fraction models since replication models involve averaging over independent (or perhaps weakly correlated) realizations. Since we may have an infinite number or even an uncountable number of model components, we need to prove that the maxima of estimated coefficients for spurious model components converge in probability to zero. One practical result of our analysis is the existence of natural theoretical weights in model selection. We can also say something fairly general about weak convergence of parameter estimates. In fact, in the stationary case we prove that, under fairly general conditions, we never will select a time-dependent window function. We begin with a remark describing a method we will use to prove convergence in probability of averages of the time-dependent variables.

Remark 1 $\mathbf{L}^1$ Mixingale Theory *For proving convergence of sums of variables where the individual terms in the series depend on time in some way, we need some special technical methods. One approach we will take is to use theorems on* $\mathbf{L}^1$ *mixingales (integrable random variables at each* t*; see Appendix A) due to McLeish [135] and Andrews [6].*[17]

For some examples of the use of these theorems, see Hamilton's time series book ([90], pp. 190-2). This remark introduces the idea of a $\mathbf{L}^1$ *mixingale and the relevant results we will need to use.*

Suppose we consider a sequence of zero mean random variables $X(t)$ *which satisfy:*

$$E|E[X(t)|I(t-k)]| \leq c_t \eta_k \tag{IV.74}$$

where $I(t-k)$ *is the information available at time* $t-k$[18] *and* c_t *and* η_k *are bounding sequences. Then Andrews' theorem ([6], p. 459-460) says that if* $\eta_k \to 0$ *as* $k \to \infty$*,* $c_t \geq 0$ *for all* t *and:*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c_t < \infty \tag{IV.75}$$

then we call such an $X(t)$ *an* $\mathbf{L}^1$ *mixingale. It then follows that:*

[17] In Andrews [6], there is also a result on $\mathbf{L}^2$ mixingales which can be used.

[18] $I(t-k)$ is a set of σ-fields or a stochastic basis. We might use as this basis $\epsilon(t-j)$ for $j \geq 0$ or we might use lagged values of X.

$$\frac{1}{T}\sum_{t=1}^{T} X(t) \xrightarrow{\mathrm{P}} 0 \tag{IV.76}$$

as $T \to \infty$.

We shall also use the followinf result for doubly-indexed arrays ([6], p.461). Suppose $X_T(t)$ is zero mean and the information set depends on T and:

$$E|E[X_T(t)|I_T(t-k)]| \leq c_{t,T}\,\eta_k \tag{IV.77}$$

as above. If:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} c_{t,T} < \infty \tag{IV.78}$$

and $\eta_k \to 0$, it follows that:

$$\frac{1}{T}\sum_{t} X_T(t) \xrightarrow{\mathrm{P}} 0. \tag{IV.79}$$

We first will prove a remark about the convergence in probability of estimates for individual model components.

Theorem 2 *Suppose:*

(1) the true model (c.f., Eq. (IV.1) is described by a stationary autoregressive process with an absolutely summable moving average representation and an uncorrelated error term with two bounded moments,

(2) Each model component $h_{p,T}$ has a window function $g_{p,T}$ (the T subscript is used to indicate the dependence of the window function on time) which satisfies:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} g_{p,T}^{2}\left(\frac{t}{T}\right) = 1 \tag{IV.80}$$

then:

(1) All model components $h_{p,T}$ which include a lag r variable and a window function $g_{p,T}(t)$ produce estimates at the first iteration which converge in probability to:

$$\rho(r)\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}g_{p,T}\left(\frac{t}{T}\right) \tag{IV.81}$$

where $\rho(r)$ is the rth autocorrelation coefficient.

(2) For any particular iteration,

$$\frac{1}{T}\sum_{t=1}^{T}h_{p,T}(t)\,\epsilon(t)\xrightarrow{\text{P}}0. \tag{IV.82}$$

(3) At any iteration I define:

$$\hat{\alpha}=\left[\frac{1}{T}H'H\right]^{-1}\left[\frac{1}{T}H'y\right] \tag{IV.83}$$

where $H=(\,h^1 \quad h^2 \quad \cdots \quad h^I\,)$ and $\hat{\alpha}$ is a vector of regression coefficients determined from a regression of the data y on model components h^k, $k\leq I$, then

$$\hat{\alpha}\xrightarrow{\text{P}}\left[\mathcal{E}(H'H)\right]^{-1}\mathcal{E}(H'y) \tag{IV.84}$$

where $\mathcal{E}(h)$ is defined by Eq. (IV.33).

Remark 2 *Result (2) verifies that asymptotically, our sample averages produce the same result as assumed by Assumption (6) of the theorem.*

Proof: For any model component h_p, we compute the probability limit of the denominator in the least squares estimate. We claim that for any individual h_p:

$$\frac{1}{T}\sum_{t=1}^{T}h_{p,T}(t)^2 \quad \xrightarrow{\text{P}} \quad \sigma_y^2 \tag{IV.85}$$

where σ_y^2 is the variance of y.

We now show this result. Since $y(t)$ is stationary, it admits a moving average representation of the form:

$$y(t) = \sum_{j=0}^{\infty} \tau_j \epsilon(t-j) \tag{IV.86}$$

Conditional on information at time $t-k$, the process, $z(t) = y(t)^2 - \sigma_y^2$, satisfies (see [90], pp. 192-3):

$$E|E[z(t)|I_{t-k}]| \;\leq\; \sum_{i,j=k}^{\infty} |\tau_j||\tau_i|M \tag{IV.87}$$

where:

$$M = 2\sigma_\epsilon^2 \tag{IV.88}$$

Since we assume that the window function $g_{p,T}\left(\frac{t}{T}\right)$ is bounded, we set the coefficients in Eq. (IV.77) as:

$$\eta_k = \left(\sum_{i=k-r}^{\infty} |\tau_i|\right)^2 M \tag{IV.89}$$

$$c_{t,T} = \sup_{w\in[1,T]} \left|g_{p,T}\left(\frac{w}{T}\right)\right|^2 . \tag{IV.90}$$

Since the average of c_t is finite and $\eta_k \to 0$ as $k \to \infty$, Eq. (IV.85) follows. We note that absolute summability of the moving average representation is used in Eq. (IV.89) and finite second moments are used in Eq. (IV.88).

We now show:

$$\frac{1}{T}\sum_{t=1}^{T} h_p(t)y(t) \;\xrightarrow{p}\; \gamma(r)\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right) \tag{IV.91}$$

where $\gamma(r)$ is the autocovariance function at lag r. For any lag r, we can use the L^1 mixingale theory to provide the following bounds for the sequence:

$$g_{p,T}\left(\frac{t}{T}\right)\left[E(y(t)y(t-r)) - y(t)y(t-r)\right]. \tag{IV.92}$$

The bounds are:

$$\eta_k = \sum_{i=k}^{\infty} \sum_{j=k-r}^{\infty} |\tau_i||\tau_j| M \tag{IV.93}$$

$$c_{t,T} = \sup_{w \in [1,T]} \left| g_{p,T}\left(\frac{w}{T}\right) \right| \tag{IV.94}$$

where:

$$M = 2\sigma_\epsilon^2 \tag{IV.95}$$

(see [90], pp. 192-3). Eq. (IV.85) and Eq. (IV.91) imply that the least squares estimator on any model component at the first iteration behaves as:

$$\hat{\beta}_r \quad \xrightarrow{\mathrm{P}} \quad \rho(\tau)\, S(g_p) \tag{IV.96}$$

where $S(g_p) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_{p,T}(\frac{t}{T})$.

Since this result applies only for the first iteration, it is useful to prove a result which holds for all iterations and which is necessary for point estimates at any later iterations to be consistent, namely that:

$$\frac{1}{T} \sum_{t=1}^{T} h_p \epsilon_t \xrightarrow{\mathrm{P}} 0 \tag{IV.97}$$

for all window functions h_p.

Using the moving average representation for y

$$\eta_0 = \sum_{j=0}^{\infty} |\tau_j|\, M \tag{IV.98}$$

$$\eta_k = 0 \qquad k \geq 1 \tag{IV.99}$$

$$c_{t,T} = \sup_{w\in[1,T]} \left| g_{p,T}\left(\frac{w}{T}\right)\right| \tag{IV.100}$$

where $M = 2\sigma_\epsilon^2$. Thus, Eq. (IV.97) follows by an application of $\mathbf{L}^1$ mixingale theory.

To show result (3), we use the result in Eq.(IV.91) that for any model component containing any lag:

$$\frac{1}{T}\sum_{t=1}^{T} h^k(t)y(t) \quad \xrightarrow{\mathrm{p}} \quad \frac{1}{T}\sum_{t=1}^{T} E(h^k(t)y(t)). \tag{IV.101}$$

We have already shown that the diagonal terms in the inverse matrix $\mathcal{E}(H'H)$ converge (Eq. (IV.85)). Convergence of the non-diagonal terms is equivalent to convergence of the numerator terms except that there is now a product of two windows g_m and g_n in the mn term but each window is by assumption bounded so convergence occurs with the mixingale coefficient $c_{t,T}$ set as the product of the maxima of each of the windows. Here:

$$\eta_k = \sum_{i=k-r_m}^{\infty} \sum_{j=k-r_n}^{\infty} |\tau_i||\tau_j|M \tag{IV.102}$$

$$c_{t,T} = \sup_{w\in[1,T]} \left| g_{m,T}\left(\frac{w}{T}\right)\right| \sup_{w\in[1,T]} \left| g_{n,T}\left(\frac{w}{T}\right)\right| \tag{IV.103}$$

where r_m is the lag associated with model component h^i and r_n is the lag associated with model component h^j. ■

The previous result established: (1) that when the process is stationary, the least squares coefficient on any given model component converges in probability to an estimate which depends only on the lag and the mean of the window weights, (2) on any iteration, estimates of least squares coefficients on any particular model component are consistent because of the last result (result (2) of the theorem) and the fact that model components are fixed as iterations of the procedure progress.

Although we have shown that any particular model component converges in probability, to assess the consistency of the procedure we need to show that the maximum over all model components converges to zero. This is considerably harder and we shall want to do this for the case in which there are an infinite number of model components. The result requires two theorems. Our strategy is to first deal with the numerator terms in the least squares estimators and then the denominator. If we can prove both terms are consistent in the sense that, even with an infinite number of model components, all is as it should be, then we are done because by Slutsky's Theorem all terms must converge in probability.

Theorem 3 *Suppose:*

(1) the true model is described by a stationary AR process with a moving average representation $\sum_{j\geq 0}\psi_j\epsilon(t-j)$ which satisfies $\sum_{j\geq 0} j\psi_j^2 < \infty$ and $\sum_{j\geq 0}|\psi_j| < \infty$ and is driven by an independent noise process with four bounded moments,

(2) Each model component h_θ has a window function $g_{\theta,T}$ (the T subscript is used to indicate the dependence of the window function on time) which satisfies:

$$\frac{1}{T}\sum_{t=1}^{T} g_{\theta,T}^2\left(\frac{t}{T}\right) = 1, \tag{IV.104}$$

and is bounded for all T.

(3) There are uncountably many model components. Any measurable square integrable window function $s \mapsto g(s)$ where $s = \frac{t}{T}$ on the interval $[0,1]$ is in the set of model components for any lag.

then:

$$\sup_\theta \left[\frac{1}{T}\sum_{t=1}^{T} h_{\theta,T}(t)\,\epsilon(t)\right] \xrightarrow{P} 0. \tag{IV.105}$$

Proof:

We note that by definition:

$$\frac{1}{T}\sum_{t=1}^{T} h_{\theta,T}(t)\epsilon(t) = \frac{1}{T}\sum_{t=1}^{T} g_{\theta,T}\left(\frac{t}{T}\right) y(t-r)\,\epsilon(t) \tag{IV.106}$$

for some $r > 0$. We know that ([95], Thm. 3.1):

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_{\theta,T}\left(\frac{t}{T}\right) y(t-r)\epsilon(t) \Rightarrow \tau\int_0^1 g_\theta(s)dW(s) + U^* \tag{IV.107}$$

where τ is a constant and U^* is a term which will be investigated below.

Since $g_\theta(s)$ is by Assumption (3), square integrable and measurable on $[0,1]$, we can expand it in an orthonormal basis of the square integrable functions on $[0,1]$. Thus:

$$g_\theta(s) = \sum_{k=1}^{\infty} \alpha_k\,\phi_k(s) \tag{IV.108}$$

where $\phi_k(s)$ are orthonormal basis functions so that by Assumption (2) and (3):

$$\sum_{k=1}^{\infty} \alpha_k^2 = 1 \tag{IV.109}$$

Thus:

$$\int_0^1 g_\theta(s)dW(s) = \int_0^1 \sum_k \alpha_k\,\phi_k(s)dW(s). \tag{IV.110}$$

We can verify that in a mean square sense:

$$\int_0^1 \sum_k \alpha_k\,\phi_k(s)\,dW(s) = \sum_k \alpha_k \int_0^1 \phi_k(s)\,dW(s) \tag{IV.111}$$

This follows because for any T,

$$\sum_{k=1}^{T} a_k \int_0^1 \phi_k(s)\,dW(s) + \int_0^1 \epsilon(s)\,dW(s) \tag{IV.112}$$

where $\epsilon(s)$ is the approximation error from using only T coefficients. Since the integrated square error goes to zero (uniform convergence is not required) as T gets large:

$$\int_0^1 \epsilon^2(s)\,ds \to 0 \tag{IV.113}$$

Since each of the $\phi_k(s)$ are orthonormal, each of the stochastic integrals is an independent Gaussian variable of mean zero and variance 1:

$$E[\int_0^1 \phi_k(s)\,dW(s)] = 0 \tag{IV.114}$$

$$\begin{aligned} E[\int_0^1 \phi_k(s)\,dW(s)]^2 &= E[\int_0^1 \int_0^1 \phi_k(s)\phi_k(t)\,dW(s)\,dW(t)] \\ &= \int_0^1 |\phi_k(s)|^2\,ds = 1. \end{aligned} \tag{IV.115}$$

For any sum with T of the α_k, we can maximize the value of the expression (IV.110) subject to Eq. (IV.109) by choosing α_k to be equal to 1 for the maximum over the T independent Gaussian random variables.

However, it can be shown that for a large number T of independent Gaussian random variables, the probability that the maximum $|M_T|$ of the random variables is greater than $\sqrt{2\log T}$ goes to zero almost surely as T gets large. Formally, for any $\lambda > 1$:

$$P(|M_T| > \sqrt{2\lambda\log T}) \leq \gamma T^{1-\lambda} \tag{IV.116}$$

where γ is a constant. This is a known result which is reviewed in Appendix I. In Eq. (IV.107) there is a normalizing factor of $\frac{1}{\sqrt{T}}$ so that the order of the limit is the maximum divided by $\sqrt{T}$ (because the normalizing factor in Eq. (IV.106) is $\frac{1}{T}$). Since $\sqrt{\frac{\log T}{T}} \to 0$ and the number of lags is assumed finite, the probability of having any model component from an infinite set being correlated with the error term for the regression goes to zero as $T \to \infty$.

We now deal with the problem of jumps or U^*. We know that U^* can be defined as follows ([94], pp. 491-2):

$$\lim_{T\to\infty}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(g_{\theta,T}\left(\frac{t}{T}\right)-g_{\theta,T}\left(\frac{t-1}{T}\right)\right)z(t)+\frac{1}{\sqrt{T}}g_{\theta,T}(1)z(T)\right] \tag{IV.117}$$

where:

$$z(t) := \sum_{k=1}^{\infty} E(y(t-r+k)\epsilon(t+k)|I_t) \tag{IV.118}$$

which is zero for all stationary y. Thus U^* is zero and the effect of jumps can be ignored. ∎

Theorem 4 *Suppose:*

(1) the true model is described by a stationary AR process with a moving average representation $\sum_{j\geq 0}\psi_j\epsilon(t-j)$ *which satisfies* $\sum_{j\geq 0}|\psi_j|<\infty$ *and* $\sum_{j\geq 0}j|\psi_j|<\infty$ *and where* $\epsilon(t)$ *is an independent and Gaussian noise term.*

(2) Each model component h_θ *has a window function* $g_{\theta,T}$ *(the* T *subscript is used to indicate the dependence of the window function on time) which satisfies:*

$$\frac{1}{T}\sum_{t=1}^{T} g^2_{\theta,T}\left(\frac{t}{T}\right) = 1 \tag{IV.119}$$

for all T *and the window function is bounded for all* T.

(3) Each model component h_θ *has a window function* $g_{\theta,T}$ *(the* T *subscript is used to indicate the dependence of the window function on time) which satisfies:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left|g_{\theta,T}\left(\frac{t}{T}\right)\right|^4 < \infty \tag{IV.120}$$

and is bounded for all T.

(4) There are uncountably many model components. Any square summable window function $s\mapsto h(s)$ *where* $s=\frac{t}{T}$ *which is a measurable square integrable function on the interval* $[0,1]$ *is in the set of model components for any lag.*

then:

$$\inf_{\theta}\left[\frac{1}{T}\sum_{t=1}^{T} h_{\theta,T}(t)^2\right] \quad \xrightarrow{P} \quad C > 0 \tag{IV.121}$$

Furthermore, this constant C depends on the maximum fourth moment of the window functions in the analysis.

Proof:

We note that by definition:

$$\frac{1}{T}\sum_{t=1}^{T} h_{\theta,T}^2(t) = \frac{1}{T}\sum_{t=1}^{T}\left|g_{\theta,T}\left(\frac{t}{T}\right)\right|^2 y(t-r)^2 \tag{IV.122}$$

We know that ([95], Thm. 3.1, p. 492):

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left|g_{\theta,T}\left(\frac{t}{T}\right)\right|^2 (y(t-r)^2 - \sigma_y^2) \Rightarrow \tau\int_0^1 g_\theta^2(s)dW(s) + U^* \tag{IV.123}$$

for some $\tau < \infty$. Now we can use a similar argument as for the previous theorem except that since $\int g_\theta^4(s)ds$ is only assumed bounded and not fixed, we let the maximum of all the $\int_0^1 g_\theta^4(s)ds$ be λ. It then follows that if we want the sum in Eq. (IV.122) to be positive for all possible model components, we have to make sure that Eq. (IV.123) is not negative enough to make this zero. We can minimize Eq. (IV.122) by choosing $|\alpha_k| = \sqrt{\lambda}$ where the sign of α_k depends on the sign of the largest absolute value of the Gaussian random variables such as are defined by the stochastic integrals:

$$\int_0^1 \phi_k(s)\,dW(s) \quad \overset{d}{\sim} \quad N(0,1) \tag{IV.124}$$

which were used in the previous result.

We now deal with the problem of jumps or U^* where U^* can be defined as in Eq. (IV.117) except that:

$$z(t) := \sum_{k=1}^{\infty} E\left(y(t-r+k)^2 - \sigma^2_{y(t-r+k)} | I_t\right) \tag{IV.125}$$

This is not zero. From the moving average representation for $y(t)$, we can calculate that $z(t)$ is (where $r \geq 1$):

$$z(t) = \sum_{k=1}^{\infty} \sum_{m=(k-r)\wedge 0}^{\infty} \sum_{n=(k-r)\wedge 0}^{\infty} \psi_m \psi_n (\epsilon(t-m-r)\epsilon(t-n-r) - \sigma_\epsilon^2 \delta_{m,n}) \tag{IV.126}$$

which has bounded variance.[19] We define:

$$U = \left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(g_{\theta,T}\left(\frac{t}{T}\right) - g_{\theta,T}\left(\frac{t-1}{T}\right)\right) z(t) + \frac{1}{\sqrt{T}} g_{\theta,T}\ (1)\, z(T)\right] \tag{IV.130}$$

We wish to show that $\frac{U}{\sqrt{T}} \xrightarrow{\text{P}} 0$. We use $\mathbf{L}^1$ mixingale theory. We recall that, as $T \to \infty$, the difference $g_{\theta,T}\left(\frac{t}{T}\right) - g_{\theta,T}(\frac{t-1}{T})$ will become negligible except at points where the window function has jumps. We suppose that the maximum jump in

[19] To calculate this expression we note that:

$$E\left(\sum_{k=r}^{\infty}\sum_{z=r}^{\infty}\sum_{m}^{\infty}\sum_{n}^{\infty}\sum_{p}^{\infty}\sum_{q}^{\infty} \psi_m \psi_n \psi_p \psi_q \epsilon_{t-m+k-r} \epsilon_{t-n+k-r} \epsilon_{t-p+z-r} \epsilon_{t-q+z-r}\right)$$
$$\leq 3 \left(\sum_{m \geq 0} (m+r) \psi_m^2\right)^2 \sigma_\epsilon^4 \tag{IV.127}$$

which is finite by Assumption (1). In Eq. (IV.127) we note that:

$$(k-r)\wedge 0 \leq m, n \leq \infty \qquad (z-r)\wedge 0 \leq p, q \leq \infty \tag{IV.128}$$

We also note that the simpler sum:

$$2\sigma_\epsilon^2 E\left(\sum_{k=r}^{\infty} \sum_{m=(k-r)\wedge 0} \sum_{n=(k-r)\wedge 0}^{\infty} \psi_m \psi_n \epsilon(t-m-r+k)\epsilon(t-n-r+k)\right)$$
$$= 2\sigma_\epsilon^4 \sum_{n \geq 0} (n+r) \psi_n^2 \tag{IV.129}$$

is also finite by Assumption (1). Using these expressions, we can bound the variance of $z(t)$.

$g_{\theta,T}\left(\frac{t}{T}\right)$ is $C_T < \infty$. Thus, we can choose the $\mathbf{L}^1$ mixingale coefficients $c_{t,T}$ to be this maximum. To prove convergence in probability, we need to calculate the mixing coefficients, η_m, for $z(t)$. We note that we can choose:

$$\eta_m = (\sum_q q|\psi_{q+m-r}|)^2 M \tag{IV.131}$$

where $M = 2\sigma_\epsilon^4$. We need to show that $\eta_m \to 0$. It is enough to show that $\sum_{r=m}^{\infty} r|\psi_r| \to 0$. However, this follows from the assumption (1) of absolute summability of $r\psi_r$.

Since the result in Appendix I applies to functions of normally distributed random variables (c.f. [118], p. 21), it follows that as long as the window function has a finite number of jumps as $T \to \infty$, that estimates on spurious model components will converge in probability to zero. ∎

We note that the previous result implies that when computing simple regressions to determine model components, there is a dimensional factor of $\frac{1}{\sqrt{\lambda}}$. This suggests we should weight model components by some function of the fourth moments of their window functions. For instance, for a flat window, the fourth moment is proportional to $\frac{T}{L}$ where L is the length of the window. To eliminate the dimensional factor $\frac{1}{\sqrt{\lambda}}$, we should then weight by say $\sqrt{L}$ in deciding which model to select. We have found exactly this criteria experimentally though there is a bias tradeoff as well so that weighting functions such as L^α where $\alpha < 0.5$ seem reasonable.

We now have shown that as $T \to \infty$, we never have a single spurious estimate at any iteration when the data generating process is stationary. In fact, this holds in a much more general sense because the variance at any point in time is finite and we can bound the error by the maximum in the case where the variance is at its maximum and constant over time.[20] We now are ready to show that we never will select a nonstationary window asymptotically when the true data generating process is stationary.

Theorem 5 *Suppose: (1) The true data generating process is stationary.*

(2) We include flat model components corresponding to the whole window.

(3) We do not weight short windows more than longer windows.

Then:

We never select any windows which are not flat windows of the length of the whole time series.

Proof:

Consider the first iteration. Given a window function $g_{p,T}$ associated with lag r, our estimates all converge in probability to:

$$\rho(r) \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right) \tag{IV.132}$$

where $\rho(r)$ is the r th autocorrelation. Thus, the theoretical value of the additional sum of squares added to the regression is:

$$\rho(r)^2 \, [\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right)]^2 \sigma_y^2 \tag{IV.133}$$

By the Cauchy-Schwartz inequality and the constant sum of squares of window functions (window function average sum of squares is 1):

$$\left|\frac{1}{T} \sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right)\right| \leq 1 \tag{IV.134}$$

[20] We need also to consider jump components carefully.

for all T. The maximum is achieved only by flat window functions with the length of the time series. Therefore, on the first iteration we will always select a flat window function.

On the second iteration, since we have subtracted off a stationary model component, and the correlation with the error term can be ignored, we again have estimates on which we can use Eq. (IV.134) so that we will always select the appropriate lag window rather than a nonstationary window function as long as the weights in model selection do not favor shorter windows. ∎

We now proceed to deal more generally with models which include a finite but arbitrary number of stationary regimes. A large but finite number of regimes can also be regarded as a good approximation to a general nonstationary process. Our results on consistency for stationary processes also go through when there are a finite number of regimes.

Suppose we consider a data generating process on a large interval $[0, L]$. We define the points in the interval as l_i for $i = 2....L$ and let $l_1 = 0$ and $l_{L+1} = L$. We define m_i for $i = 1....L$ as:

$$m_i = \frac{l_{i+1} + l_i}{2} \tag{IV.135}$$

We define a set of model components over this interval. These model components h_j have windows with values $g_j(m_i)$ which are constant over the interval $I_i = [l_i \frac{T}{L}, l_{i+1} \frac{T}{L}]$.[21] We assume that on the infinite interval there is a data generating process:

[21] Technically, we assume that the last interval from $[l_L, l_{L+1}]$ is closed and all the other intervals are open at the right.

$$y(t) = \sum_{j=1}^{\infty} \gamma_j(m_i)\, 1_{t-j \in I_i}(t)\, y(t-j) + \epsilon(t) \tag{IV.136}$$

We thus expand the intervals $[l_i, l_{i+1}]$ as a function of sample size and at the same time add the appropriate lag functions. We can consider an arbitrary but finite number of stationary regimes and apply Theorem 1. Since expectations are constant over the interval $I_i = [l_i \frac{T}{L}, l_{i+1} \frac{T}{L}]$, we can evaluate expectations at any point in the vicinity $t = m_i \frac{T}{L}$ for each regime. Our definition Eq. (IV.136) also gives us a natural triangular array which is useful theoretically. In general, we believe Eq. (IV.136) provides a useful model to examine for approximation results for more general nonstationary time series.

We now prove some basic results about our method with such semi-stationary processes.

Theorem 6 *Suppose:*

(1) the true model on each interval is described by a stationary AR process with moving average coefficients which are absolutely summable.

(2) Each model component h_k has a window function $g_{p,T}$ (the T subscript is used to indicate the dependence of the window function on time) which is bounded, has a finite fourth moment and which satisfies:

$$\frac{1}{T} \sum_{t=1}^{T} g_{p,T}^2 \left(\frac{t}{T} \right) = 1 \tag{IV.137}$$

then

(1) All model components $h_{p,T}$ which include a lag r variable and a window function $g_{p,T}(t)$ produce estimates at the first iteration which converge in probability to:

$$\frac{\sum_{i=1}^{L} \nu(i)\, g_{p,T}(m_i)\, \gamma_i(r)}{\sum_{i=1}^{L} \nu(i)\, g_{p,T}^2(m_i)\, \sigma_{y_i}^2} \tag{IV.138}$$

where $\gamma_i(r)$ is the rth autocovariance coefficient on the ith interval and $\nu(i)$ is the fraction of the time series covered by the ith interval.

(2) We have for any model component $h_{p,T}$:

$$\frac{1}{T}\sum_{t=1}^{T} h_{p,T}(t)\epsilon(t)\xrightarrow{\mathrm{P}}0 \qquad \text{(IV.139)}$$

(3) Least squares estimates α_k on model components h^k, $k \leq I$ converge in probability to the average of expectations:

$$\hat{\alpha}_k\xrightarrow{\mathrm{P}}\left[\mathcal{E}(H'H)\right]^{-1}\mathcal{E}(H'y) \qquad \text{(IV.140)}$$

where $\mathcal{E}$ is defined by Eq. (IV.33).

Proof: For any model component $h_{p,T}$, we can compute the probability limit of the denominator in the least squares estimate. The result is that for any individual $h_{p,T}$:

$$\frac{1}{T}\sum_{t=1}^{T} h_{p,T}^2 \quad \xrightarrow{\mathrm{P}} \quad \sum_{i=1}^{L}\nu(i)|g_p(m_i)|^2\sigma_{y_i}^2 \qquad \text{(IV.141)}$$

where $\sigma_{y_i}^2$ is the variance of y on interval I_i.

Since y is stationary on each interval, we will use L^1 mixingale theory to prove convergence in probability. Since computations for a general autoregressive representation become rather involved, we continue to use a moving average representation. By Eq. (IV.77) we have:

$$\eta_k = \sum_{i=1}^{L}\Big(\sum_{m=k-r}^{\infty}|\tau_{m,i}|\Big)^2 M \qquad \text{(IV.142)}$$

where $\tau_{m,i}$ are the moving average coefficients at lag m for interval i when we take the interval to ∞ and where $M = 2\sigma_\epsilon^2$. We have:

$$c_{t,T} = \sup_{i=1,\ldots L}|g_{p,T}(m_i)|^2 \qquad \text{(IV.143)}$$

Since the average of c_t is finite and $\eta_k \to 0$ as $k \to \infty$, Eq. (IV.141) follows.

We now show that for any model component $h_p(t)$:

$$\frac{1}{T}\sum_{t=1}^{T} h_p(t)\,y(t) \quad \xrightarrow{\mathrm{P}} \quad \sum_{i=1}^{L} \nu(i)\gamma_i(r) g_{p,T}(m_i) \tag{IV.144}$$

where $\gamma_i(r)$ is the autocovariance at lag r in the ith interval.

For any lag r, we can use the $\mathbf{L}^1$ mixingale theory to provide the following bounds for the sequence $g_{p,T}(m_i)\,E\left(y(t)y(t-r)\right) - y(t)y(t-r))$. We have:

$$\eta_k = \sum_{i=1}^{L}\sum_{z=k}^{\infty}\sum_{j=k-r}^{\infty} |\tau_{z,i}||\tau_{j,i}|M \tag{IV.145}$$

$$c_t = \sup_{w\in[1,T]}\left|g_{p,T}\left(\frac{w}{T}\right)\right| \tag{IV.146}$$

where:

$$M = 2\sigma_\epsilon^2 \tag{IV.147}$$

Eq. (IV.141) and Eq. (IV.144) imply that the least squares estimator on any model component at the first iteration behaves as:

$$\frac{\sum_{i=1}^{L} \nu(i)\, g_{p,T}(m_i)\, \gamma_i(r)}{\sum_{i=1}^{L} g_{p,T}^2(m_i)\, \nu(i)\, \sigma_{y_i}^2} \tag{IV.148}$$

where $\nu(i)$ is the fraction of the time series covered by the interval I_i.

Since this result applies only for the first iteration, it is useful to prove a result which holds for all iterations and which is necessary for consistency to occur. We now prove the final result:

$$\frac{1}{T}\sum_{t=1}^{T} h_k(t)\epsilon_t \xrightarrow{\mathrm{P}} 0 \tag{IV.149}$$

for all window functions h_k.

Using the moving average representation for y

$$\eta_0 = \sum_{i=1}^{I} \sum_{j=0}^{\infty} |\tau_{j,i}|\, M \tag{IV.150}$$

$$\eta_k = 0 \qquad k \geq 1 \tag{IV.151}$$

$$c_{t,T} = \sup_{i=1,\ldots L} |g_{p,T}(m_i)| \tag{IV.152}$$

where $M = 2\sigma_\epsilon^2$. Thus, $\eta_k \to 0$ and the result holds.

The result for numerator in (3) has already been shown as have the necessary results for the diagonal terms in the inverse matrix. The L^1 mixingale bounds need to be modified so that for element x, y:

$$c_{t,T} = \sup_{i=1,\ldots L} |g_{x,T}(m_i)| \sup_{i=1,\ldots L} |g_{y,T}(m_i)| \tag{IV.153}$$

and:

$$\eta_k = \sum_{i=1}^{L} \sum_{z=k-r_x}^{\infty} \sum_{j=k-r_y}^{\infty} |\tau_{z,i}||\tau_{j,i}|M \tag{IV.154}$$

where r_x and r_y are the lags associated with the model components h^x and h^y. ∎

The following theorem suggests there are circumstances under which we will never pick flat windows unless they correspond (in terms of the fraction of time) to at least the minimum regions of stationarity.

Theorem 7 *Suppose we include flat model components in the analysis which are constant over all possible pairs of intervals $[l_i\frac{T}{L}, l_j\frac{T}{L}]$ for all $i = 1 \ldots L$, $j > i$, then we never select model components which do not have support on at least the minimal stationary partition: $[l_i\frac{T}{L}, l_j\frac{T}{L}]$ for some i such that $L \geq i \geq 1$ and j such that $L + 1 \geq j \geq 2$.*

Proof: We can always achieve accurate estimates as $T \rightarrow \infty$ as if one window function has only an infinitesemal portion in some interval i the summation terms will have no weight. If it does have a small portion on an interval, estimates will be consistent.

We choose the model according to the maximum value of $|A|$:

$$A = \frac{\sum_{i=1}^{L} g_k(m_i)\nu(i)\gamma_i(r)}{\sqrt{\sum_{i=1}^{L} g_k(m_i)^2\sigma_i^2\nu(i)}} \tag{IV.155}$$

where $\nu(i)$ is the fraction of the time series covered by regime i. We can rewrite A for any partition of the line P_j as:

$$\sum_{i\in P_j} g_k(m_i)\nu(i)\,\sigma_i N_i \frac{\gamma_i(r)}{\sigma_i} \tag{IV.156}$$

where N_i captures the denominator terms in Eq. (IV.155). Thus, by the Cauchy-Schwartz inequality:

$$|A|^2 \leq \sum_{i\in P_j} \nu(i)\left(\frac{\gamma_i(r)}{\sigma_i}\right)^2 \tag{IV.157}$$

which is an average of the r^2 from a regression with flat window functions. Suppose the properties of γ_i and σ_i are constant over the interval P_j then we will never choose a smaller window than the length of P_j because we can reach the supremum (over all possible partitions of lengths less than P_j) if and only if we choose a model component which has flat weights. At further iterations, the minimal stationary partition P_j may change but by the choice of model components, it may never be smaller than I_i. ∎

Therefore, at any iteration, the method will never pick a window smaller than the region of local stationarity in terms of the autocovariances at any particular lag. This provides an indication that the method may be effective in handling locally stationary or semi-stationary processes.

We now review a basic weak convergence result for a general nonstationary process. We note that Theorems 3 and 4 continue to apply in terms of consistency when there are an infinite number of model components because the variance is bounded at any point in time. We shall examine the properties of parameter estimates in the nonstationary $AR(1)$ case to show that the convergence in probability conditions of Theorem 2 are met.

Theorem 8 *Suppose: (1) the true model is described by a nonstationary $AR(1)$ process which has (a) more than two bounded moments, (b) error term with more than two bounded moments, (c) an autoregressive parameter which is always less than 1 in absolute value. The autoregressive parameter depends on the fraction of time and the autocovariance at lag 1 at time t, $\gamma_T(\frac{t}{T})$, is a function of the fraction of the time series.*

(2) We have a set of model components h_k which include window functions g_k which satisfy:

$$\frac{1}{T}\sum_{t=1}^{T}\left|g_{k,T}\left(\frac{t}{T}\right)\right|^2 = 1, \tag{IV.158}$$

are bounded

then:

(1) All model components h_k which include lag one data produce estimates at the first iteration which converge in probability to:

$$\hat{\beta}_k = \frac{\int_0^1 g_k(s)\gamma(s)ds}{\int_0^1 g_k^2(s)\sigma^2(s)ds} \tag{IV.159}$$

(2) All model components h_k satisfy:

$$\frac{1}{T}\sum_{t=1}^{T} h_{k,T}(t)\epsilon(t)\xrightarrow{\text{P}}0. \tag{IV.160}$$

Proof:

For the numerator, we wish to show that:

$$\frac{1}{T}\sum_{t=1}^{T} h_k(t)y(t) \quad \xrightarrow{p} \quad \frac{1}{T}\sum_{t=1}^{T}\gamma\left(\frac{t}{T}\right) g_{k,T}\left(\frac{t}{T}\right) \tag{IV.161}$$

where $\gamma(t)$ is the first autocovariance of y at time t. We note that:

$$y(t)y(t-1) = \beta(t)y(t-1)^2 + \epsilon(t)y(t-1) \tag{IV.162}$$

Define:

$$N = \left[E\left(\epsilon\right)^2\right]^{\frac{1}{2}} \sup_t \left[E(y(t-1))^2\right]^{\frac{1}{2}} \tag{IV.163}$$

which is finite by Assumption (1).

We define:

$$z(t) := g_k\left(\frac{t}{T}\right)(y(t)y(t-1) - E(y(t)y(t-1))) \tag{IV.164}$$

We wish to show that $z(t)$ is a L^1 mixingale. We use the terms within the brackets to define the mixing coefficients η_m. We have:

$$\eta_0 = N + \sup_t |\beta(t)|\, E\left|y(t-1)^2 - E(y(t-1)^2)\right| \tag{IV.165}$$

Noting that:

$$y(t-1)^2 = \beta(t-1)^2 y(t-2)^2 + 2\beta(t-1)\,\epsilon(t-1)\,y(t-2) + \epsilon(t-1)^2 \tag{IV.166}$$

it follows that:

$$\eta_m = \sup_t |\beta(t)\prod_{j=1}^{m-1}\beta(t-j)^2|E\left|y(t-m)^2 - E(y(t-m)^2)\right| \tag{IV.167}$$

which converges to zero as $m \to \infty$ because of the assumption that $\sup_t |\beta(t)| < 1$. We set:

$$c_{t,T} = \sup_{w \in [1,T]} \left| g_{k,T}\left(\frac{w}{T}\right) \right| \tag{IV.168}$$

which by Assumption (2) is finite. Thus, the result for the numerator holds.

For the denominator of the least squares estimate, we wish to show that:

$$\frac{1}{T}\sum_{t=1}^{T} h_{k,T}(t)^2 \quad \xrightarrow{p} \quad \frac{1}{T}\sum_{t=1}^{T} \sigma^2(t) g_{k,T}\left(\frac{t}{T}\right)^2 \tag{IV.169}$$

By the recursion Eq. (IV.166), it follows that:

$$\eta_0 = \sup_t E|y(t-1)^2 - E(y(t-1)^2)| \tag{IV.170}$$

and for $m \geq 1$

$$\eta_m = \sup_t \prod_{j=0}^{m-1} \beta(t-j)^2 E|y(t-m)^2 - E(y(t-m)^2)| \tag{IV.171}$$

so that $\eta_m \to 0$ by Assumption 1. Since we can set:

$$c_{t,T} = \sup_{w \in [1,T]} \left| g_{k,T}\left(\frac{w}{T}\right) \right|^2 , \tag{IV.172}$$

Eq. (IV.169) follows.

We now prove:

$$\frac{1}{T}\sum_{t=1}^{T} h_{k,T}(t)\epsilon(t) \quad \xrightarrow{p} \quad 0 \tag{IV.173}$$

The model component h_k contains a window function g_k which is multiplied by a lag variable $y(t - r_k)$ where r_k is the lag associated with model component h_k. We note that:

$$E\,|y(t-r)\epsilon(t)| \;\leq\; E\,|\epsilon(t)|\;|y(t-r)|\,. \tag{IV.174}$$

Thus, we have L^1 mixingale coefficients:

$$c_{t,T} = \sup_t \left| g_{k,T}\left(\frac{t}{T}\right)\right| \tag{IV.175}$$

$$\eta_0 = N \tag{IV.176}$$

which is finite by Assumption (1). For $m \geq 1$, $\eta_m = 0$. Since c_t has finite mean and $\eta_m \to 0$, the result in Eq. (IV.173) follows. ∎

For the more general nonstationary case, computations can quickly become quite involved. Therefore, it is useful to work with a nonstationary moving average representation:

$$y(t) = \sum_{j=1}^{\infty} \tau_j(t)\, \epsilon(t-j) \tag{IV.177}$$

where ϵ is an i.i.d. disturbance with four bounded moments. Since we have assumed that the first moment of y is bounded:

$$E|y(t)| = E|\sum_{j=1}^{\infty} \tau_j(t)\epsilon(t-j)| \leq M \sum_{j=1}^{\infty} |\tau_j(t)| < \infty \tag{IV.178}$$

where $M = E|\epsilon|$. Thus, we can assume without loss of generality that the moving average representation for y is absolutely summable at each t.

Theorem 9 *Suppose: (1) the true model is described by a nonstationary AR process which has an absolutely summable moving average representation with an error term with more than two bounded moments*

(2) The true model is a fraction model in which the autocovariance at lag r at time t, $\gamma_{r,T}(\frac{t}{T})$, is a function of the fraction of the time series.

(3) We have a set of model components h_p which include window functions g_p which satisfy:

$$\frac{1}{T}\sum_{t=1}^{T}\left|g_{p,T}\left(\frac{t}{T}\right)\right|^{2} = 1, \tag{IV.179}$$

and are bounded.

then:

(1) All model components h_p which include lag r data produce estimates at the first iteration which converge in probability to:

$$\hat{\beta}_p = \frac{\int_0^1 g_p(s)\gamma_r(s)ds}{\int_0^1 g_p^2(s)\sigma^2(s)ds} \tag{IV.180}$$

(2) All model components h_p satisfy:

$$\frac{1}{T}\sum_{t=1}^{T} h_{p,T}(t)\epsilon(t)\xrightarrow{\text{P}}0. \tag{IV.181}$$

(3) At any iteration I define:

$$\hat{\alpha} = \left[\frac{1}{T}H'H\right]^{-1}\left[\frac{1}{T}H'y\right] \tag{IV.182}$$

and $\hat{\alpha}$ is a vector of regression coefficients determined from a regression of the data y on model components h^k, $k \leq I$, then

$$\hat{\alpha}\xrightarrow{\text{P}}\left[\mathcal{E}(H'H)\right]^{-1}\mathcal{E}(H'y). \tag{IV.183}$$

Proof: For any model component h_p, we can compute the probability limit of the denominator in the least squares estimate. The result is that for any individual h_p:

$$\frac{1}{T}\sum_{t=1}^{T} h_{p,T}\left(t\right)^{2} \quad \xrightarrow{\text{P}} \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right)^{2}\sigma_y^2\left(\frac{t}{T}\right) \tag{IV.184}$$

where σ_y^2 is the variance of $y(t-r)$ (r is the lag associated with model component h_p).

We now show this result. Since $y(t)$ admits a moving average representation of the form:

$$y(t) = \sum_{j=0}^{\infty} \tau_j(t)\epsilon(t-j) \tag{IV.185}$$

Conditional on information at time $t-k$, the process, $z(t) = y(t)^2 - \sigma_y^2$, satisfies (see [90], pp. 192-3):

$$E|E[z(t)|I_{t-k}]| \;\; \leq \;\; \sum_{i,j=k}^{\infty} |\tau_j(t)||\tau_i(t)|M \tag{IV.186}$$

where:

$$M = 2\sigma_\epsilon^2 \tag{IV.187}$$

Since we assume that the window function $g_{p,T}\left(\frac{t}{T}\right)$ is bounded, we can set the mixingale coefficients as:

$$\eta_k = \sup_t \left(\sum_{i=k-r}^{\infty} |\tau_i| \right)^2 M \tag{IV.188}$$

$$c_{t,T} = \sup_{w \in [1,T]} \left| g_{p,T}\left(\frac{w}{T}\right)\right|^2 \tag{IV.189}$$

Since the average of c_t is finite and $\eta_k \to 0$ as $k \to \infty$, Eq. (IV.184) follows. We note that absolute summability of the moving average representation is used in Eq. (IV.188) and finite second moments are used in Eq. (IV.187).

We now show:

$$\frac{1}{T}\sum_{t=1}^{T} h_p(t)\,y(t) \;\; \xrightarrow{\text{P}} \;\; \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} g_{p,T}\left(\frac{t}{T}\right)\gamma_{r,T}\left(\frac{t}{T}\right) \tag{IV.190}$$

where γ_r is the autocovariance function at lag r. For any lag r, we can use L^1 mixingale theory to provide the following bounds for the sequence:

$$g_{p,T}\left(\frac{t}{T}\right)\left(E(y(t)y(t-r)) - y(t)y(t-r)\right) \tag{IV.191}$$

$$\eta_k = \sup_t \left[\sum_{i=k}^{\infty} \sum_{j=k-r}^{\infty} |\tau_i(t)||\tau_j(t)| \right] M \tag{IV.192}$$

$$c_{t,T} = \sup_{w\in[1,T]} \left| g_{p,T}\left(\frac{w}{T}\right) \right| \tag{IV.193}$$

where:

$$M = 2\sigma_\epsilon^2 \tag{IV.194}$$

(see [90], pp. 192-3). Eq. (IV.85) and Eq. (IV.91) imply that the least squares estimator on any model component at the first iteration behaves as:

$$\hat{\beta}_p \quad \xrightarrow{p} \quad \frac{\int_0^1 g_p(s)\gamma_r(s)ds}{\int_0^1 g_p^2(s)\sigma^2(s)ds} \tag{IV.195}$$

Since this result applies only for the first iteration, it is useful to prove a result which holds for all iterations and which is necessary for point estimates at any later iterations to be consistent. We now prove the final result:

$$\frac{1}{T}\sum_{t=1}^{T} h_p\epsilon_t \xrightarrow{p} 0 \tag{IV.196}$$

for all window functions h_p.

Using the nonstationary moving average representation for y:

$$\eta_0 = \sup_t \left[\sum_{j=0}^{\infty} |\tau_j(t)| \right] M \tag{IV.197}$$

$$\eta_k = 0 \qquad k \geq 1 \tag{IV.198}$$

$$c_{t,T} = \sup_{w\in[1,T]} \left| g_{p,t}\left(\frac{w}{T}\right) \right| \tag{IV.199}$$

where $M = 2\sigma_\epsilon^2$. Thus, Eq. (IV.196) follows by an application of L^1 mixingale theory.

We now show result (3). Convergence in probability of the numerator terms has already been shown (Eq. (IV.190)). Convergence in probability of the mn element of the inverse matrix involves:

$$c_{t,T} = \sup_{w \in [1,T]} \left| g_{m,T}\left(\frac{w}{T}\right)\right| \sup_{w \in [1,T]} \left| g_{n,T}\left(\frac{w}{T}\right)\right| \tag{IV.200}$$

$$\eta_k = \sup_t \left[\sum_{i=k-r_m}^{\infty} \sum_{j=k-r_n}^{\infty} |\tau_i(t)||\tau_j(t)| \right] M \tag{IV.201}$$

where r_m and r_n are the lags associated with model components h^m and h^n respectively. Thus:

$$\left[\frac{1}{T}H'H\right] \xrightarrow{\mathrm{P}} [\mathcal{E}(H'H)] \tag{IV.202}$$

and Result (3) follows. ∎

Synopsis

In this chapter, we have provided some general convergence and consistency results about the method proposed in the thesis. It is useful to provide a brief summary of some of the main results.

After defining *fraction models* (c.f., Eq. (IV.7) and *replication models* (c.f., Eq. (IV.19) and Eq. (IV.21)) for nonstationary time series, we show how both classes of models can be expressed in terms of model components in the sense that:

$$\begin{aligned} y(t) &= \sum_{j=1}^{J} \beta_j(t) y(t-j) + \epsilon_1(t) \\ &= \sum_{k=1}^{K} \alpha_k h_k(t) + \epsilon_2(t). \end{aligned} \tag{IV.203}$$

In the thesis, we have proposed an estimation procedure which constructs regression estimates from a collection of h_k, $k = 1, .., M$, where M may be much larger than K. From this large set of M "potential model components", we estimate a model of the form:

$$\hat{y}(t) = \sum_{k=1}^{I} \hat{\alpha}_k h^k(t) \tag{IV.204}$$

where we use the notation h^k (instead of h_k) to indicate that the model components have been selected by the procedure. By comparison, if we knew the true $\beta_j(t)$ function, the expected value of y conditional on the values of the lagged dependent variables $\{y(t-j)\}_{j=1}^{J}$ would be:

$$\tilde{y}(t) = \sum_{j=1}^{J} \beta_j(t) y(t-j). \tag{IV.205}$$

In this chapter, we prove the convergence of model component expansions in the sense that, as the number of model components in the analysis I gets large, the right hand side of Eq. (IV.204) converges to the right hand side of Eq. (IV.205). This convergence occurs in the sense that the mean squared error of the representation goes to zero (Theorem 1, Eq. (IV.43)).

Convergence occurs even though in the estimation procedure, we select the model components from a much broader set of M possibilities and do not know *a priori* the precise functional form of the $\beta_j(t)$. For convergence, we must be able to express each of the h_k in Eq. (IV.203) in terms of a linear combination of a subset of potential model components (Theorem 1, Assumption 4). Asymptotically (as I gets large), the procedure selects the appropriate h_k or their linear representations.

The convergence result (Theorem 1) is interesting because at each iteration, we search for a local optimum (i.e., pick the next model component to include) so it is unclear whether we will converge to a representation which is equivalent (in terms of predictions) to what would result from use of a global optimization procedure (minimizing squared error with respect to all possible combinations of model components). The convergence proof shows that in fact we do achieve the same asymptotic mean squared error as would a global optimization procedure (we note that the proof says nothing about the *parsimony* of the representation). The result is robust in the sense that it applies to a broader class of procedures than that outlined in Ch. II, including even simple stepwise regressions such as are discussed in Appendix F.

While the convergence result provides some indications that the proposed procedure is well-behaved as the number of model components included in estimates gets large, it is also necessary to examine whether as sample size T gets large, point estimates of the coefficients of specific model components are consistent. In practice, we work with point estimates (c.f., Eq. (IV.204)) such as:

$$\hat{\alpha} = \left[\frac{1}{T}H'H\right]^{-1}\left[\frac{1}{T}H'y\right] \tag{IV.206}$$

where $H = (\, h^1 \quad h^2 \quad \cdots \quad h^I \,)$ and $\hat{\alpha} = (\, \hat{\alpha}_1 \quad \hat{\alpha}_2 \quad \cdots \quad \hat{\alpha}_I \,)'$ is the vector of regression coefficients at stage I in Eq. (IV.204) (which is determined from a multiple regression of the data y on model components h^k, $k \leq I$).

Since we work with nonstationary data generating processes, it might seem difficult to verify that $\hat{\alpha}_k$ does converge in probability to the true α_k. Under technical assumptions, we verify that (c.f., Theorems 2,6,8,9) in a finite dimensional setting:

$$\hat{\alpha}_k \xrightarrow{p} \alpha_k \tag{IV.207}$$

as long as the error term is uncorrelated with any of the included regressors in the sense that:

$$\frac{1}{T}\sum_{t=1}^{T} h_k(t)\epsilon_2(t) \xrightarrow{p} 0 \tag{IV.208}$$

In Theorems (2,6,8,9), we examine conditions under which Eq. (IV.208) holds.

Another practical statistical issue is that there may be many different model components in the analysis and we select model components based on maximal explanatory power. We show consistency of point estimates resulting from the choice of extremal model components from a possibly uncountable set of potential model components (Theorems 3-4). In addition, we show (Theorem 5) that if the true model is a stationary one and we use model components associated with the fraction model, the method asymptotically selects stationary model components.

Together, these results on convergence and consistency provide another 'proof of concept' that the method is well-defined and potentially useful in applications. Perhaps more significantly, we have defined several different senses in which to do statistical analysis for parametric nonstationary time series. We can either consider repeated experiments or assume periodicity or allow window sizes to grow with sample size. This conceptual framework is a useful starting point for simulation and other statistical analysis of the properties of nonstationary time series estimators.

Though the results in this chapter provide a useful theoretical starting point, there are still a number of questions which we will need to address in the thesis:

- We do not know how fast our expansions for parameter estimates converge as a function of iteration; the method may not work well if expansions decay slowly. We address this issue in Ch. V.

- We need to have some idea about how reliable our estimates are. We provide some preliminary results on confidence intervals for model choice in Ch. V.

- As in standard time series analysis, we need a rule to select a model from estimates in finite samples. We propose one solution in Ch. VI.

- How do we use the estimates from the method to produce nonstationary spectral estimates? Indeed, how do we even define a nonstationary spectrum. These questions are dealt with in Ch. VII.

- Asymptotically, do we ever select spurious model components at any iteration? In Appendix H, we provide a counterexample where we asymptotically always select a wrong model component. This does not contradict our theorems, however, as the coefficient on this model component goes to zero when the right model components are added to the model.

However, since the approach is new, the thesis only provides an introductory treatment of many of the important theoretical issues. Thus, in Ch. IX we define what we feel are some important theoretical questions which may be addressed in future research, whether through pure theoretical analysis or computational simulation.

CHAPTER V

SOME AUXILIARY RESULTS

In this chapter, we review some theoretical issues related to our approach. These issues include: (1) some basic results on confidence intervals, (2) rates of convergence.

Confidence Intervals

Since the method produces estimates which are come from a linear regression against the selected model elements, the appropriate confidence intervals conditional on the data and conditional on the selected model elements are the standard confidence intervals associated with least squares regressions. However, experimentally, these confidence intervals often do not seem appropriate in cases in which we include large numbers of model components in the analysis.

In this section, we show that in some cases it is possible to provide approximate confidence intervals for estimates at any particular stage by using nonlinear regression confidence intervals. We prove an equivalence between certain forms of nonlinear regression and analysis with an infinite number of model components. We then show some operational examples of what the 'sum of squares' functions look like. For the approach to be valid, the model components must be smooth functions in that they must satisfy the same differentiability constraints imposed by nonlinear regression analysis. Any use of nonlinear regression analysis to determine confidence intervals is also conditional on the variable selected (e.g., lag one versus lag two) because model components are not a smooth function of the lags.

Theorem 10 *Suppose that the model components in the analysis are of the form:*

$$h_i(t) = g(\theta_i, t)x_k(t) \tag{V.1}$$

for some $\theta_i \in \mathbf{R}^n$ and some fixed regressor variable x_k. We define $\mathcal{M}$ to be the set (perhaps uncountable) of all θ_i, then model selection by maximizing sample r^2 at any iteration is equivalent to solving the problem:

$$\inf_{\beta \in \mathbf{R}\, \theta \in \mathcal{M}} S\left[(y^n(t) - \beta g(\theta_i, t)x_k(t))^2\right] \tag{V.2}$$

where S is a sum over all observations (indexed by t) and y^n is the residual.

Proof:

$$\begin{aligned}
&\inf_{\beta \in \mathbf{R}\, \theta \in \mathcal{M}} S\left[(y^n(t) - \beta g(\theta_i, t)x_k(t))^2\right] \\
&= \inf_{\beta \in \mathbf{R}\, \theta \in \mathcal{M}} \left[S((y^n(t))^2) - 2\beta S(g(\theta, t)y^n(t)x_k(t)) + \beta^2 S(g(\theta, t)^2 x_k(t)^2)\right]
\end{aligned} \tag{V.3}$$

Thus, our least squares estimate $\hat{\beta}$ of β is:

$$\hat{\beta} = \frac{S(g(\theta, t)y^n(t)x_k(t))}{S(g(\theta, t)^2 x_k(t)^2)} \tag{V.4}$$

Using our estimates of $\hat{\beta}$, minimizing the right hand side of Eq. (V.3) is equivalent to solving:

$$\sup_{\theta \in \mathcal{M}} \frac{S\,(g(\theta, t)y^n(t)x_k(t))^2}{S\,(g(\theta, t)^2 x_k(t)^2)} \tag{V.5}$$

which is equivalent to maximizing r^2 over the set of all potential model components indexed by θ. ∎

If we now let θ_i take values in $\mathbf{R}^p$ where p is the number of parameters in θ_i, we can interpret decisions at each stage in terms of a nonlinear regression. This allows us to use theorems from nonlinear regression analysis to compute asymptotic confidence intervals (c.f., [79]). In general, we require smoothness of the regression function τ which implies that the windows g are twice continously differentiable in θ. We also require our parameter estimates to be in the interior of the Euclidean parameter space. As an example, this condition might be violated in our case if we had only Gaussian window functions and the true model were time invariant. We also require for consistency uniform convergence of the (normalized) matrices of first and second derivatives as sample size grows.

It is known that asymptotic confidence intervals are given in terms of the inverse of the matrix:

$$Z(\theta,\beta) = \frac{1}{\sigma^2}\text{plim}\,\frac{1}{T}\begin{pmatrix} \frac{\partial S}{\partial \theta}\frac{\partial S}{\partial \theta}' & \frac{\partial S}{\partial \theta}\frac{\partial S}{\partial \beta} \\ \frac{\partial S}{\partial \theta}\frac{\partial S}{\partial \beta} & \left(\frac{\partial S}{\partial \beta}\right)^2 \end{pmatrix}_{\theta=\theta_0,\beta=\beta_0} \qquad \text{(V.6)}$$

where S is the (specific) sum of squares defined by Eq. (V.2) and we estimate σ^2 by s^2:

$$s^2 = \frac{S(\hat{\beta},\hat{\theta})}{T-p} \qquad \text{(V.7)}$$

and T is sample size and there are $p-1$ parameters in the vector θ. In Eq. (V.6), θ_0 and β are the population values of θ and β; in applications, estimated values are substituted for population values in computing the matrix in Eq. (V.6). Furthermore, the expected bias of estimated coefficients on the model components can be computed and is asymptotically zero.

It is useful to provide some examples on the issues of smoothness and convexity of the sum of squares function in terms of nonlinear regressions. Our examples show that for a typical sample problem we considered in Ch. III, the sum of squares functions is smooth. However, our example also shows that use of nonlinear regression methods alone to select model components is not always advisable, because when the model is misspecified, the sum of squares function is not globally convex.

Consider a window function $g(\theta, t)$ and a linear time-varying autoregressive process:

$$y(t) = \beta g(\theta, t) y(t-1) + \epsilon(t) \tag{V.8}$$

$$\epsilon(t) \sim N(0,1). \tag{V.9}$$

We define the sum of squares function:

$$S_T(\theta) = \inf_{\beta} \left[\frac{1}{T} \sum_{t=2}^{T} (y(t) - \beta g(\theta, t) y(t-1))^2 \right]. \tag{V.10}$$

The issue is when is $S_T(\theta)$ a smoothly differentiable function of θ; the smoothness of $S_T(\theta)$ depends on the smoothness of the window function g used in the analysis. Here, we consider some experimental evidence. As a simplest possible example, we let y be white noise so that the parameters θ are not identifiable since the true β is zero. We set $T = 512$ and consider a Gaussian window function g defined as:

$$g(t; t_0, \sigma_0) = e^{-\frac{(t-t_0)^2}{2\sigma_0^2}}. \tag{V.11}$$

In Fig. V.1, we have fixed $\sigma_0 = 40.0$ and varied t_0 on a grid with grid spacing $\delta t_0 = \frac{1}{40.0}$. From Fig. V.1, it appears that the sum of squares is a smoothly differentiable function of t_0. However, the sum of squares function is not globally convex so that use of a nonlinear optimization program to estimate t_0 would be ill-advised.

In Fig. V.2, we show data from a model where $\beta = 1$ and where $t_0 = 250$, $\sigma_0 = 40.0$. Fig. V.3 shows the true values of the autoregressive parameter $\beta g(t; t_0, \sigma_0)$ as a function of time. In Fig. V.4, we construct the sum of squares function in a

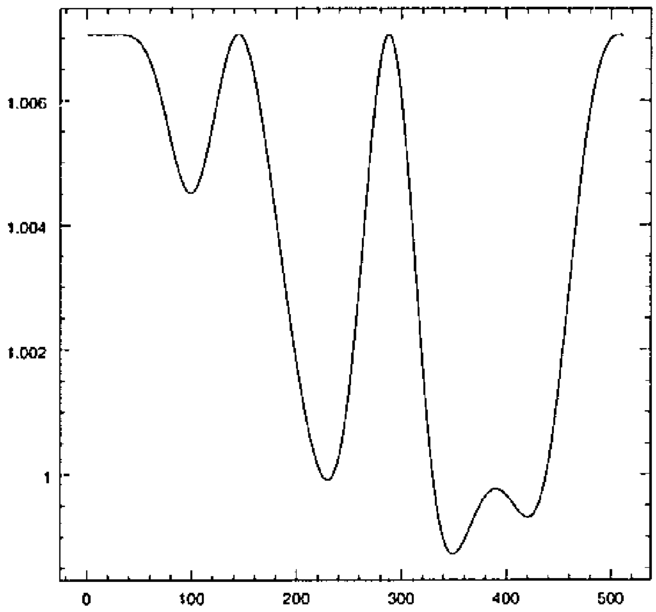

Figure V.1: Sum of squares for Gaussian noise

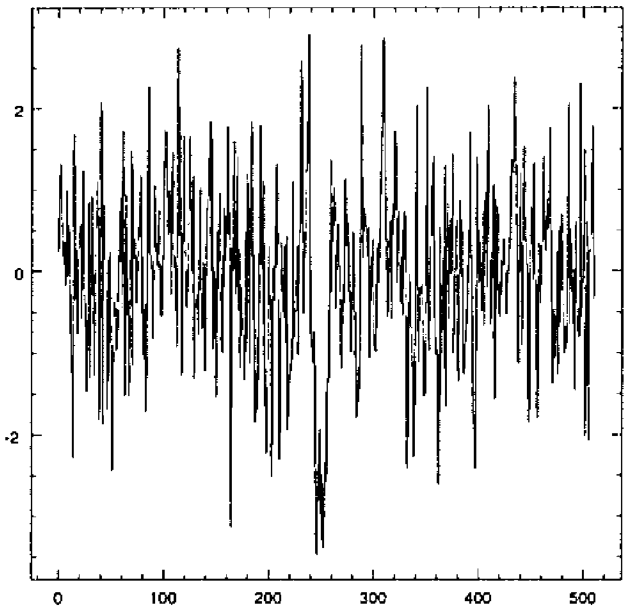

Figure V.2: Data from smoothly varying autoregressive model

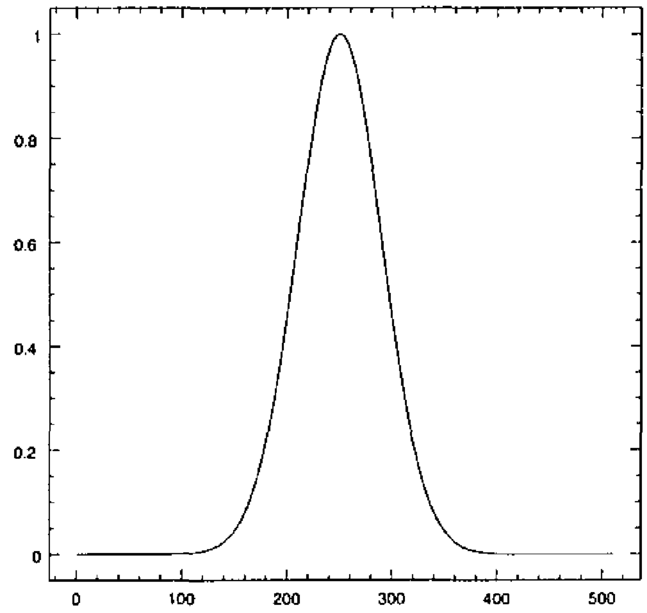

Figure V.3: Autoregressive parameter

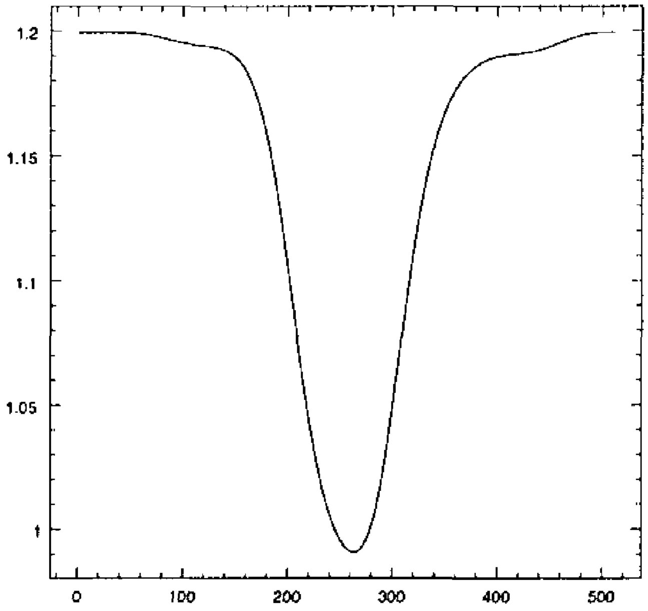

Figure V.4: Sum of squares function with Gaussian window

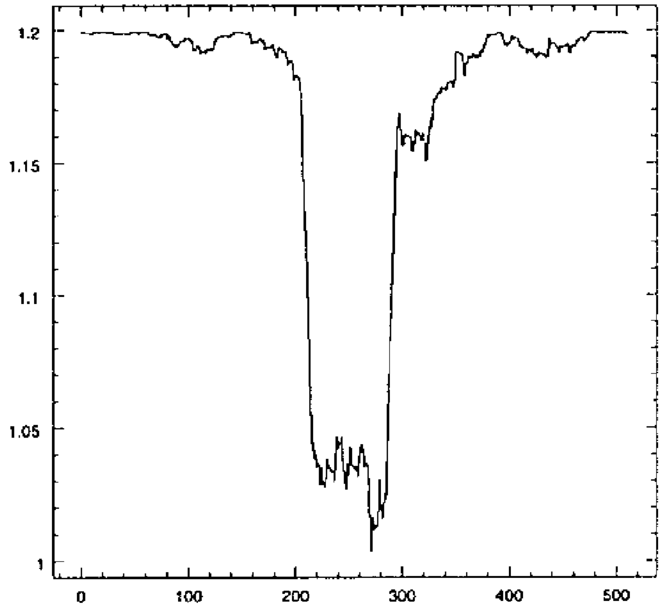

Figure V.5: Sum of squares function with flat window

similar way to Fig. V.1; this sum of squares function appears to have a well-defined minimum and it appears to be smoothly differentiable.

Fig. V.5 shows that the properties of the window function g matter quite a bit for the smoothness of the sum of squares function. When we use a flat window function g to construct the sum of squares function, the sum of squares function in a finite sample is not even C^0. Thus, confidence intervals based on nonlinear regression analysis may be tricky to construct for nonsmooth windows.

Convergence Rates

One factor which might make our method perform poorly in practice is a slow rate of convergence in that it may take many model components to capture the behavior of a given time series. Therefore, it is useful to show that in any finite dimensional space, the residual decays exponentially with the iteration. We will show that this occurs for any regression with a finite amount of data and applies also in the case where simple stepwise regressions are used at each iteration (see Appendix F for discussion).

We define:

$$\lambda^2(y^{n-1}) = \frac{\sup_{h_i \in C} (S(m_i^n)^2)}{(S(y^{n-1})^2)} \tag{V.12}$$

where S means a sum over observations so that λ^2 is the maximal empirical simple correlation r^2 obtainable from a simple regression of y^{n-1} on all model components in the analysis. The set of model components is denoted by C.

We recall that m_i^n is the additional sample variation explained by adding a model component h_i at iteration n.[1] We let:

$$\mu = \inf_{g \in \mathbb{R}^T; g \neq 0} \lambda^2(g). \tag{V.13}$$

Theorem 11 *We let y^0 be data which lies in a finite dimensional space $\mathbb{R}^T$ where T is the number of observations. We define the residual:*

$$y^n = y - \sum_{k=1}^{n} m^k \tag{V.14}$$

where m^k are the contributions to regression estimates from stage k. We define a set of model components C with elements h_i which include window functions which together span the finite dimensional space $\mathbb{R}^T$ then, in Eq. (V.13):

(1) $\mu > 0$.

(2) the residual decays exponentially as the number of iterations increase.

Proof: The proof is a small modification of one in [131]. We suppose that we choose a model component only if for some $0 < \alpha \leq 1$:

[1] Formally, as discussed in Ch. II, $m_i^n = P_i f^{n-1}$ where P_i is a projection of the residual f^{n-1} on the model component chosen at iteration i and all previously selected model components. Thus, m_i^n is the predicted part of a regression of the residual f^{n-1} on each of the previously selected model components.

$$S((m^i)^2) \geq \alpha \sup_{\theta \in C} S((m^i_\theta)^2) \tag{V.15}$$

where $m^i = P_i y^{i-1} = y^{i-1} - y^i$. and m^i_θ is the explained part of the regression if we use model component h_θ at iteration i instead. We note that Eq. (V.15) is similar in appearance to Eq. (IV.35) in Ch. IV but different technically because we use sums instead of expectation values. Then if the data is y^0, we have:

$$S((y^1)^2) = S((y^0)^2) - S((m^1)^2) \tag{V.16}$$

By the definition of μ in Eq. (V.13) we have that (since μ is a minimum):

$$S((y^0)^2)\,\mu \leq \sup_{\theta \in C} S((m^1_\theta)^2) \tag{V.17}$$

By Eq. (V.15):

$$S((m^1)^2) \geq \alpha\mu S((y^0)^2) \tag{V.18}$$

Thus, in Eq. (V.16), we have:

$$S((y^1)^2) \leq S((y^0)^2) - \alpha\mu S((y^0)^2) = (1 - \alpha\mu) S((y^0)^2) \tag{V.19}$$

Similarly, it follows that:

$$S((y^2)^2) \leq S((y^1)^2) - \alpha\mu S((y^1)^2) = (1 - \alpha\mu)^2 S((y^0)^2) \tag{V.20}$$

Therefore:

$$S((y^n)^2) \leq (1 - \alpha\mu)^n S((y^0)^2) \tag{V.21}$$

Thus, the residual decays exponentially in a finite dimensional space. It remains to show that $\mu > 0$ if $S((y^k)^2) > 0$ for any k. Suppose $S((y^k)^2)$ is greater than zero then y^k is a nonzero element of $\mathbf{R}^T$. We show that there must be at least one element h_i in the set of model components such that $|S(y^k h_i)| > 0$. Let the explantory variable included in h_i be denoted as x_i and let the window function of h_i be denoted as g_i then the product $y^k x_i$ lies in the space $\mathbf{R}^T$ which is spanned (by assumption) by the set of window functions g_i. We note that: $|S(y^k x_i g_i)| = |S(y^k h_i)|$ Therefore, it must be the case that there is some h_i and associated x_i, g_i such that: $|S(y^k h_i)| > 0$ if $y^k > 0$. ∎

It is helpful to show that the theorem is operational by considering an example. Since Gaussian noise is the worst possible scenario in terms of explaining the sample variance of a time series with the minimal number of coefficients, it is helpful to consider the properties of a decomposition of Gaussian noise while noting that none of the selected model components are statistically significant as by definition white noise cannot be predicted by its lagged values. This worst case scenario helps us understand the behavior of residual in the case in which we do not properly select model components.

We consider a simple 4096 point decomposition with Gaussian noise:

$$y(t) = \epsilon(t) \tag{V.22}$$

where $\epsilon(t) \sim N(0,1)$. Since we wanted to ensure that there is little possibility of spurious phenomena with Gaussian noise, we chose a large number of potential model components (in this case we use $700,000$) and iterated the algorithm for a long time (in this case $2,000$ times). Since this would be impossible to do computationally using a procedure which computed a linear regression at each iteration, we used a computationally more effective procedure but still convergent procedure which satisfies the conditions of the theorem on exponential decay.

This procedure is the 'simplified approach' described in Appendix F. This procedure has residual which decays at a slower rate than a procedure which uses a linear

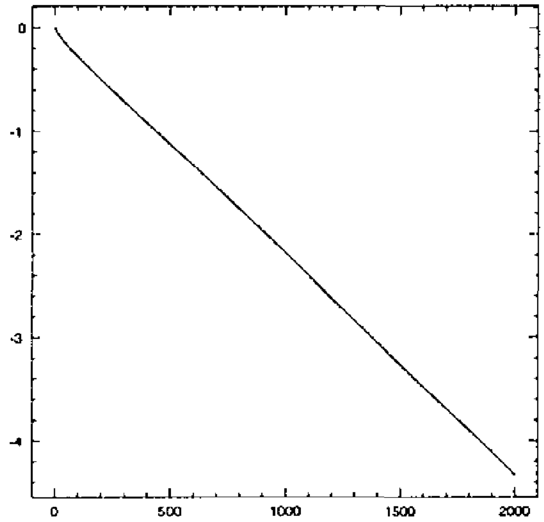

Figure V.6: Decay of residual for 4096 point Gaussian noise with constant filters. Ordinate is logarithm of percentage of energy in residual. Plot is of coefficient number vs. log residual. The rate of decay seems to be exponential.

regression at each stage. The results are illustrated in Figure (V.6) where it appears that the residual is decaying exponentially. The exponential rate of decay on the estimates suggests that the residual is not converging that slowly and that the statistical properties of the estimates are not likely to be all that bad. One would expect and in fact finds exponential decay when a much smaller set of model components is used than the 700, 000 we used in this simulation experiment.

CHAPTER VI

A STOPPING RULE

Given the time series representation:

$$\hat{y}(t) = \sum_{i=0}^{N} \hat{C}_i h^i \tag{VI.2}$$

where h^i are selected model components, the question arises how to choose N optimally so as to achieve the best estimates. One possibility would be to examine the r^2 of the selected coefficient and test whether it is statistically signiciant. Determining the distribution of such a statistic when the residual at iteration n, y^n, is white noise is possible but it is a problem which requires knowledge of all the correlations between model components [118] and the likely extreme-value structure of the distribution may lead to other complications as well as a possible loss of efficiency. Numerical experiments indicated that the problem of the distribution of maxima of r^2 for individual model components also failed to provide any quantitative characterizations in terms of the included model components. While further progress on this problem is expected, it is useful to provide a special stopping rule which also serves as a good test for randomness in economic time series when the alternative involves time-varying parameters.

Tests such as the cumulative periodogram test will be powerful when the alternative hypothesis is a nontrivial covariance stationary process; however, in our method, the alternative is not likely to be a covariance stationary process and instead is likely to involve time-varying coefficients. Therefore, we propose a new test for randomness which simulation experiments and theory suggest good properties in the case where the data is nonstationary. Our test is called a cumulative waveletgram test for randomness, because we use wavelet coefficients instead of Fourier coefficients to construct our test.

Wavelets provide a different type of orthonormal representation than that of Fourier analysis. Our experiments with economic data suggest that in many cases, a given number of wavelet coefficients capture more of the variance of economic time series than do Fourier coefficients; thus, we suggest that wavelets may be a useful tool of broad applicability in the analysis of economic data. Some potential applications include: regression estimators, tests for stationarity, nonstationary spectral estimation, and nonparametric function estimation. Here, of course, we focus on a test for randomness, but this chapter can serve as a short primer on wavelets and their use in time series analysis.

To summarize, our results in this chapter include:

- The cumulative waveletgram has an asymptotic Brownian bridge distribution. In small samples, simulation evidence indicates that the asymptotic confidence intervals need to be substantially corrected.

- There appears to be little loss of power against stationary alternatives from using the cumulative waveletgram instead of the cumulative periodogram.

- There appears to be a large gain in power against nonstationary alternatives such as time-varying variances from using the cumulative waveletgram instead of the cumulative periodogram.

- The cumulative waveletgram does poorly when the data is a sine wave corruped with noise, but we introduce an alternative test called the adaptive waveletgram which can have better performance.

- Empirical examples are provided with economic and financial time series data.

A Primer on Wavelets

Wavelets are basis functions of $\mathbf{L}^2$ (the space of square integrable functions) with two special features: (1) unlike the Fourier basis functions, the wavelet basis functions are local in time (see Fig. (VI.2) for an example and compare with a sine wave), (2) they have frequency dependent bandwidth in that larger 'waves' are used to measure lower frequency movements and shorter waves are used to measure high frequency movements.

There exists a rich and recent mathematical theory of wavelets which is reviewed in monographs by [54] [39] [138]; two seminal theoretical articles on wavelets are: [130] and [52]. There are now a number of surveys of applications and recent theoretical developments; these include: [185] [18] [20] [40].

To describe wavelets, we shall need to use the theory of Fourier transforms. Recall that the frequency domain or Fourier representation of a function is defined by its Fourier transform. The Fourier transform is an operator $T : \mathbf{L}^1 \cap \mathbf{L}^2 \mapsto \mathbf{L}^\infty \cap \mathbf{L}^2$ so that for $f \in \mathbf{L}^1 \cap \mathbf{L}^2$,

$$\hat{f}(\omega) = Tf(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx \tag{VI.3}$$

We thus use the special notation $\hat{f}$ to refer to the Fourier transform of f. It has an inverse: $\hat{f} \mapsto f$ defined by:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x} dx. \tag{VI.4}$$

A wavelet $x \mapsto \psi(x)$ is a function such that:

$$\frac{|\hat{\psi}(\omega)|}{|\omega|^{\frac{1}{2}}} \in \mathbf{L}^2 \tag{VI.5}$$

which requires that (since $\hat{\psi}(0)$ must be zero):

$$\int dx\, \psi(x) = 0. \tag{VI.6}$$

There exist wavelet functions $\psi(x)$ whose 'dilations' indexed by j and translations indexed by k:

$$\{\psi_{jk} = 2^{-\frac{j}{2}}\psi(2^{-j}x - k)\}_{j,k=-\infty}^{\infty} \tag{VI.7}$$

form an orthonormal basis of $\mathbf{L}^2$. We will focus on such wavelets here. Dilation makes the function spread out and hence form a longer wavelet which measures lower frequencies; translation shifts the wavelet function to measure the properties of the time series at some other point in time. In Ch. VII, we will review the idea of *time-frequency* spectral estimation and explain how wavelets fit in a general framework; the analytical framework of Ch. VII will thus provide insight into why wavelets 'work' in certain situations but do not in others. This complex analytical framework is not necessary to understand the mechanics of the wavelet transform or some of its intuition.

To see the intuition of how a wavelet transform works, we consider some function $\phi(x)$ whose translates $\phi_{jk} = 2^{-\frac{j}{2}}\phi(2^{-j}x - k)$ are orthonormal for $k \in \mathbf{Z}$ and any fixed $j \in \mathbf{Z}$. Effectively, ϕ acts as a low pass filter.

Since $\phi(x)$ is a coarse-grained functional representation relative to $\phi(2x)$, we can express $\phi(x)$ in terms of translations of the finer functions $\phi(2x)$:

$$\phi(x) = \sqrt{2}\sum_n \alpha_n \phi(2x - n) \tag{VI.8}$$

for some sequence $\{\alpha_n\}_{n=-\infty}^{\infty} \in \mathrm{l}^2$ (the space of square summable sequences).

It is known[129] that a wavelet function $\psi(x)$ can also be expanded in terms of translations of $\phi(2x)$ for such a ϕ:

$$\psi(x) = \sqrt{2}\sum_{n}\beta_n\phi(2x-n) \tag{VI.9}$$

for some sequence $\{\beta_n\}_{n=-\infty}^{\infty} \in \mathrm{l}^2$. If we take inner products of a function f with Eq. (VI.8) and Eq. (VI.9), we arrive at the algorithm of Mallat [130] for fast computation of orthonormal wavelet transforms.

When we are done, we have a wavelet series representation:

$$y(t) = \sum_{k=-\infty}^{\infty} c_{J,k}\phi_{J,k}(t) + \sum_{j=-\infty}^{J}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t) \tag{VI.10}$$

where:

$$\phi_{J,k}(t) = 2^{-\frac{J}{2}}\phi\left(2^{-J}t-k\right) \tag{VI.11}$$

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t-k\right) \tag{VI.12}$$

and $\psi(t)$ and $\phi(t)$ are the wavelet and 'smoothing' functions described above and in more detail in sources such as [54].

One example of a pair of ϕ and ψ corresponding to a Battle-Lemarie wavelet, is shown in Figures (VI.1) and (VI.2). The function ϕ in Figure (VI.1) is a linear spline. The corresponding ϕ is shown in Figure (VI.2). The explicit formulas for this wavelet in terms of Fourier transforms are ([54], p. 147-8):

$$\hat{\phi}(\omega) = \sqrt{3}(2\pi)^{-\frac{1}{2}}\frac{4\sin^2\left(\frac{\omega}{2}\right)}{\omega^2(1+2\cos^2(\frac{\omega}{2}))^{\frac{1}{2}}} \tag{VI.13}$$

$$\hat{\psi}(\omega) = \hat{\phi}(\frac{\omega}{2})e^{\frac{i\omega}{2}}\sin^2\left(\frac{\omega}{4}\right)\left(\frac{1+2\sin^2(\frac{\omega}{4})}{1+2\cos^2(\frac{\omega}{2})}\right)^{\frac{1}{2}} \tag{VI.14}$$

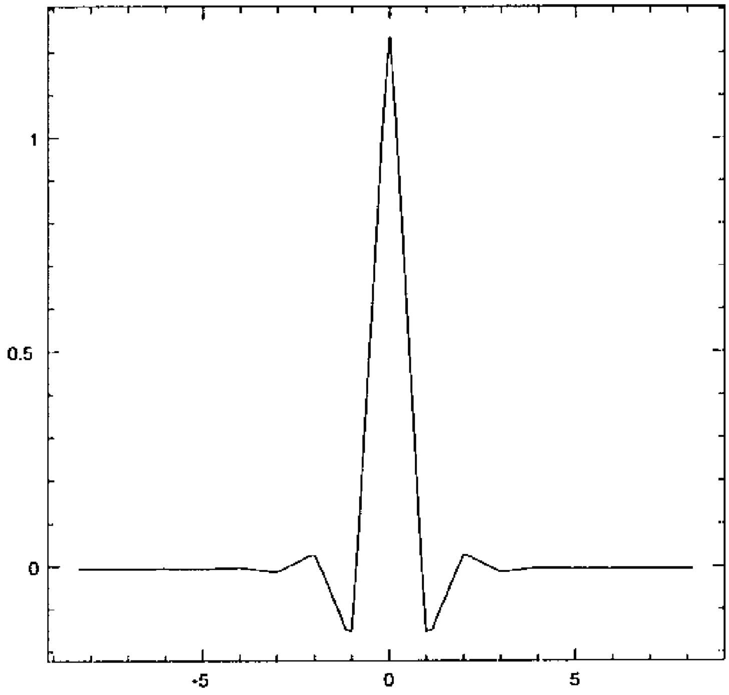

Figure VI.1: The function $\phi(x)$ is a linear spline function which is used to generate the Battle-Lemarie wavelet shown in Figure 3.

Another example of a pair of ϕ and ψ corresponds to the choice of a Haar wavelet [87]:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{VI.15}$$

$$\psi(x) = \begin{cases} -0.5 & \text{if } x \in [0,0.5] \\ 0.5 & \text{if } x \in (0.5,1] \\ 0 & \text{otherwise} \end{cases} \tag{VI.16}$$

As an example, we consider a wavelet decomposition of the simple nonsmooth function shown in Figure (VI.3). The Haar wavelet coefficients are shown in Fig. (VI.4); the Fourier coefficients for the same function are shown in Figure (VI.5). From the figures, it appears that the Haar wavelet representation is more parsimonious than

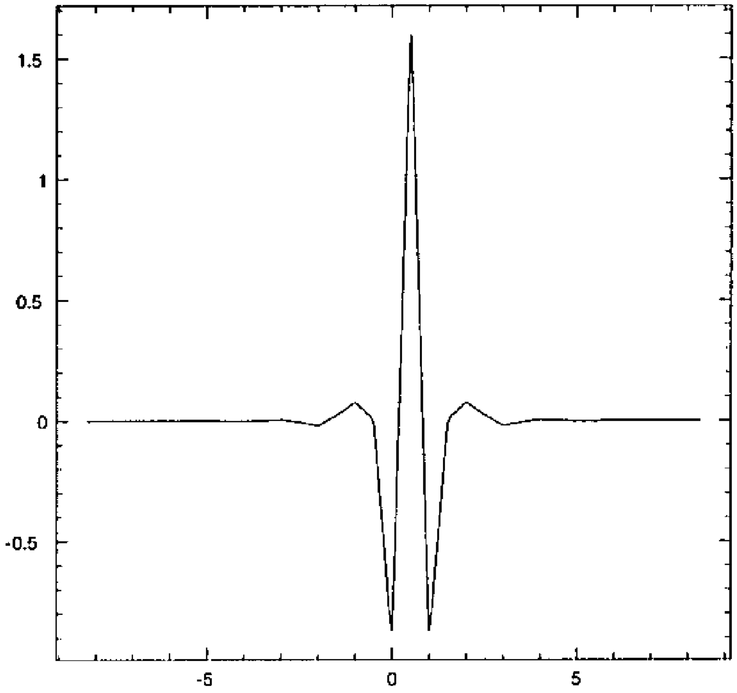

Figure VI.2: The function $\psi(x)$ is a Battle-Lemarie wavelet generated from the linear spline function in Figure 2.

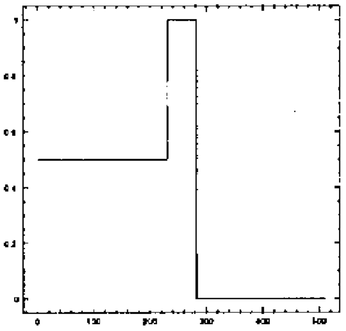

Figure VI.3: Edge function.

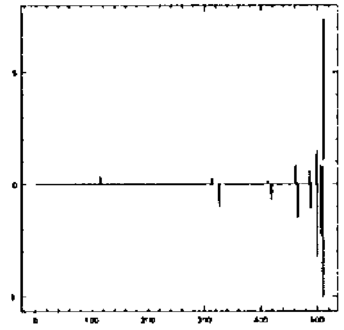

Figure VI.4: Haar wavelet coefficients for edge function.

the Fourier representation because fewer coefficients are required to represent the edge function at a given level of accuracy. The better properties of the Haar representation are due to: (1) locality since the edge function has certain local features not easily expressed in terms of global functions like sine waves, (2) lack of smoothness since the edge function is not differentiable. These special properties of the edge function are of interest in nonstationary time series which almost by definition have sharp local features.

Appendix *C* provides an explicit recipe for calculating wavelet coefficients from a series of data; alternatively, *Numerical Recipes* [166] now contains code to calculate wavelet transforms of discrete sequences.

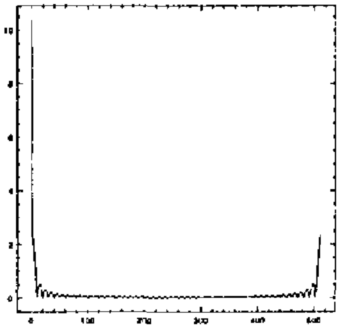

Figure VI.5: Absolute value of Fourier coefficients for edge function.

Computing the Cumulative Waveletgram

To compute the cumulative waveletgram we first compute wavelet coefficients using the fast algorithm developed by Stephane Mallat [130] (again, see Appendix C for a recipe). We define the sample wavelet coefficients of y, $d_{j,k}$, for a time series of length T as:

$$d_{j,k} = \sum_{t=1}^{T} y(t)\psi_{j,k}(t) \tag{VI.17}$$

In appendix C, we show that the sample wavelet coefficients of Gaussian white noise are mean zero and independent random variables.

We define then the cumulative waveletgram (or CWG) as:

$$CWG(t) = \sum_{\text{one path}} |d_{j,k}|^2 \tag{VI.18}$$

where t is an index of the coefficient along a summation path and T is the length of the random series. The summation notation of "one path" is used because sums are taken according to a *single summation path* along which each coefficient enters once and only once. For simplicity two orders of summation may be used. The first sums over time first at each scale and the second sums over scales at each fixed time. The first rule is shown in Figure VI.6. The second summation rule is shown in Figure

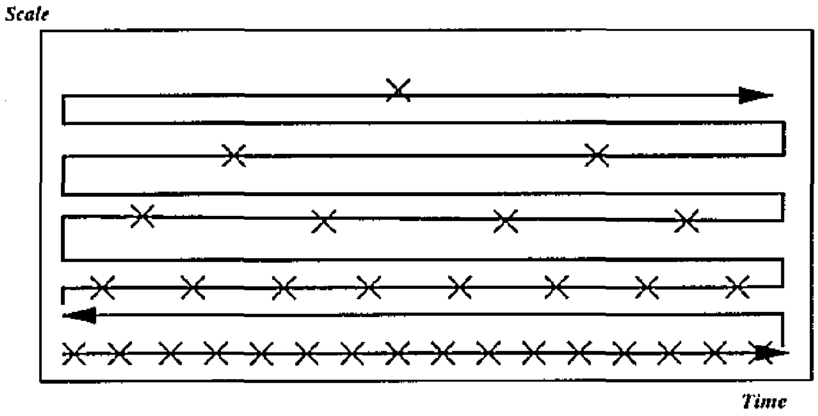

Figure VI.6: First summing rule

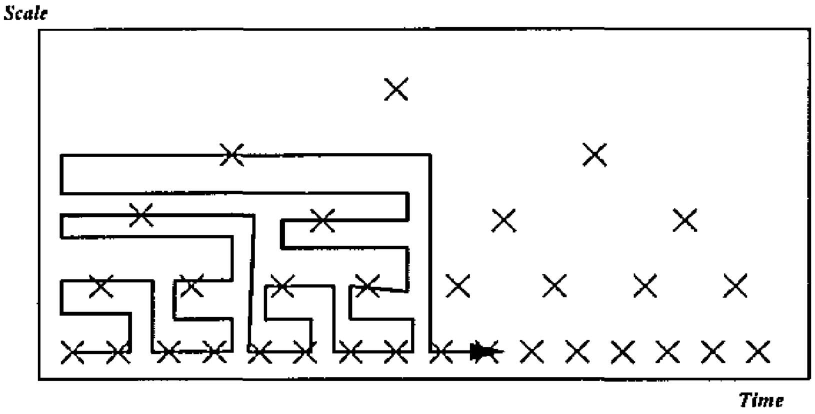

Figure VI.7: Second summing rule

VI.7. The first rule is somewhat closer to Fourier analysis than the second; the first path provides less power against sharp discontinuities than the second path. The second path sums more closely over frequency for each fixed time.

Since the wavelet coefficients are independent, we have:

$$E(CWG(t)) = \sigma^2 t \tag{VI.19}$$

We can also compute:

$$\text{Var}(CWG(t)) = \sigma^4(3t + t(t-1) - t^2) = 2t\sigma^4 \tag{VI.20}$$

To test for randomness we use a test similar to the cumulative periodogram test. We define the cumulative wavelet distribution as:

$$CWD(t) = \frac{CWG(t)}{CWG(T-1)} \tag{VI.21}$$

The cumulative periodogram test is based on the Kolmogorov-Smirnov test but it only has asymptotically the same statistics since the periodgram ordinates only have an asymptotic exponential distribution. For the Kolmogorov-Smirnov test a good approximation for the cumulative periodogram test is to reject the null hypothesis of randomness at level α if the cumulative periodogram, $CPG(t)$, exits from $\frac{t}{T-1} \pm k_\alpha \frac{1}{\sqrt{T-2}}$ where $k_{0.05} = 1.36$ and $k_{0.01} = 1.63$ [190]. These statistics are incorrect for periodogram ordinates in small samples (computationally, $N < 16,000$). We compute the sampling statistics numerically for different orthonormal wavelets. The k_α (which depend weakly on sample size in small samples) are shown in Table VI.1 for 5% intervals. Values of k_α are reported for cases in which one summation is made (by either rule) and in cases in which both summations are made. As can be verified computationally these sampling statistics do not depend on the orthonormal wavelet used; this follows immediately from the equivalence of distributions of the distribution

Sample Size	1 sum	2 sum
64	1.758	1.925
128	1.815	1.981
256	1.845	2.015
512	1.872	2.034
1024	1.884	2.052
2048	1.895	2.064

Table VI.1: 5% simultaneous confidence intervals for the cumulative wavelet distribution test of randomness. Computations are based on 200,000 replications. 1.923 is the asymptotic value for the 1 summation path statistic. The asymptotic value of the 2 summation statistic is an open question.

of the coefficients of white noise decomposed in a real orthonormal basis.

Asymptotic Distribution

The cumulative waveletgram converges in distribution to a Brownian bridge process. Let $W(t)$, $t \in [0,1]$ be Brownian motion. A Brownian bridge $B(t)$ for $t \in [0,1]$ is defined by:

$$B(t) = W(t) - t\,W(1) \tag{VI.22}$$

We let $p(s)$ be a sequence of random variables and define:

$$C_T(t) = \frac{1}{T}\sum_{s=1}^{[Tt]}(p(s) - \sigma^2) \tag{VI.23}$$

where $[Tt]$ is the largest integer less than or equal to Tt. Suppose $p(s) - \sigma^2$ is an independent random variable with four bounded moments and variance τ^2, then:

$$\sqrt{T}C_T(t) \Rightarrow N(0, \tau^2 t) \tag{VI.24}$$

so that $C_T(t)$ converges in distribution by the functional central limit theorem [25] to a Brownian motion on $[0,1]$:

$$\sqrt{T}\, C_T(t) \Rightarrow \tau\, W(t) \tag{VI.25}$$

We now let $p(s)$ be a path of wavelet coefficients. Then the cumulative waveletgram is:

$$CWG(tT) = \frac{Tt}{[Tt]} \left(T\, C_T(t) + \sigma^2\, t\, T\right) \tag{VI.26}$$

and the cumulative wavelet distribution is:

$$CWD_T(t) = \frac{CWG(tT)}{CWG(T)} = \left(\frac{\sqrt{T}\, C_T(t) + \sqrt{T}\sigma^2\, t}{\sqrt{T}\, C_T(1) + \sqrt{T}\, \sigma^2}\right) \frac{Tt}{[Tt]} \tag{VI.27}$$

Subtracting t from both sides, we have:

$$CWD_T(t) - t = \left(\frac{\sqrt{T}\, C_T(t) - t\,\sqrt{T}\, C_T(1)}{\sqrt{T}C_T(1) + \sqrt{T}\, \sigma^2}\right) \frac{Tt}{[Tt]} \tag{VI.28}$$

Multiplying both sides by $\sqrt{T}$ we have:

$$\sqrt{T}\,(CWD_T(t) - t) = \left[\frac{\sqrt{T}C_T(t) - t\,\sqrt{T}\, C_T(1)}{C_T(1) + \sigma^2}\right] \frac{Tt}{[Tt]} \tag{VI.29}$$

The denominator of the term in brackets on the right hand side of Eq. (VI.29) converges to a constant:

$$C_T(1) + \sigma^2 = \frac{1}{\sqrt{T}}\, \sqrt{T}C_T(1) + \sigma^2 \rightarrow \sigma^2 \tag{VI.30}$$

since $\sqrt{T}C_T(1) \Rightarrow \tau W(1)$ by Eq. (VI.25). The numerator of the term in brackets on the right hand side of Eq. (VI.29) converges to a Brownian bridge process proportional to $W(t) - t\, W(1)$:

$$\sqrt{T}\, C_T(t) - t\,\sqrt{T}\, C_T(1) \Rightarrow \tau\, W(t) - t\,\tau\, W(1) \tag{VI.31}$$

by the functional central limit theorem (c.f., Eq. VI.25). Finally, $\frac{Tt}{[Tt]}$ converges to 1. Thus, the cumulative wavelet distribution converges at rate $T^{\frac{1}{2}}$ to a Brownian bridge process $B(t)$:

$$\lambda\sqrt{T}\,(CWD_T(t) - t) \Rightarrow B(t) \tag{VI.32}$$

where $\lambda = \frac{\sigma^2}{\tau}$. For Gaussian noise, we have $\lambda = \frac{\sqrt{2}}{2}$.[1]

Examples

An example of the cumulative wavelet distribution for a Gaussian white noise sample of size 512 is shown in Figure VI.8. 5% confidence intervals are shown and the cumulative wavelet distribution remains within the confidence bounds as indeed it should.

A particularly simple example of the usefulness of the cumulative wavelet distribution test is to consider the sequence such as might arise from a crude structural break or discontinuity in a time series:

$$f(n) = \begin{cases} 1 & n = 256 \\ 0 & 0 \leq n < 256 \quad \text{or} \quad 256 < n < 512 \end{cases} \tag{VI.33}$$

In this case one can calculate that the discrete Fourier transform is: $\hat{f}(\omega_k) \propto (-1)^k \quad k \in \mathbb{Z}$ so that the 'periodogram' of f is always uniform. While the cumulative periodogram test will fail, the cumulative waveletgram test has no difficulty detecting the singularity. Application of the cumulative waveletgram test is shown in Figure (VI.9).

Since the wavelet basis achieves good localization in both time and frequency, it also is capable of performing tests on stationary time series (at some loss of efficiency). For instance, we consider a realization of an an AR1 time series with 512 observations:

[1] The factor λ is the same for all analogous cumulative sum tests in *real* orthonormal bases.

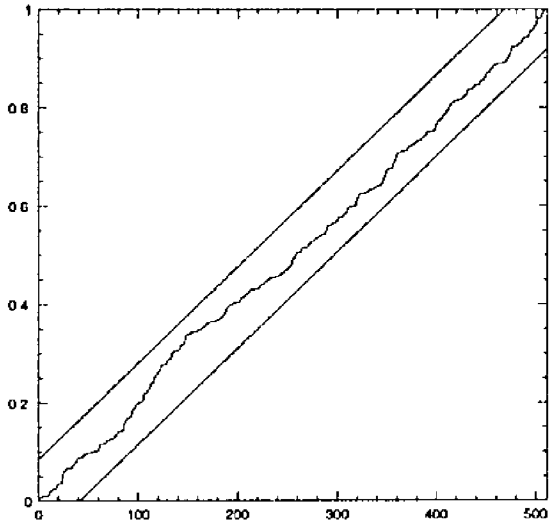

Figure VI.8: The cumulative wavelet distribution and 5% confidence interval bounds for random noise (sample size 512) decomposed in the Daubechies 12 wavelet basis. Summation and statistics are based on first summation path.

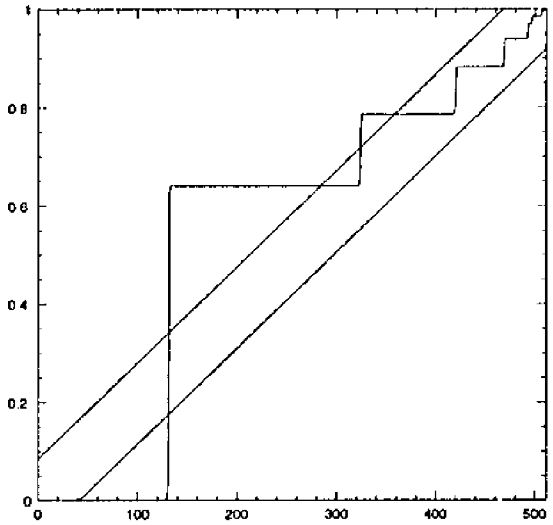

Figure VI.9: The cumulative wavelet distribution and 5% confidence interval bounds for a discrete dirac function decomposed in the Daubechies 12 wavelet basis. Summation and statistics are based on first summation path.

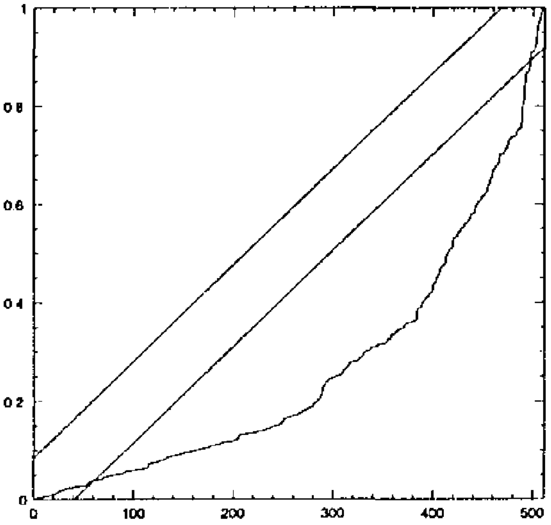

Figure VI.10: The cumulative wavelet distribution and 5% confidence interval bounds for a AR1 time series with AR parameter 0.6 in the Daubechies 12 wavelet basis. The null hypothesis of white noise is clearly rejected. Summation and statistics are based on first summation path.

$$y(t) = 0.6y(t-1) + \epsilon(t) \tag{VI.34}$$

We compute the cumulative waveletgram of a realization of this process and the results are shown in Figure VI.10. The cumulative waveletgram appears to do an excellent job of rejecting the null hypothesis of white noise.

To show the *practical* usefulness of the test, we consider Standard and Poor's log stock returns (from the CRSP daily returns file) which some consider to be random. The cumulative wavelet distribution clearly leads one to reject randomness as shown in Figure VI.11. A cumulative periodogram test also results in a rejection of randomness here.

It is helpful to examine the loss in efficiency from use of the cumulative waveletgram on stationary time series. We generated 200,000 replications of a sample time series of length 512 from the autoregressive model:

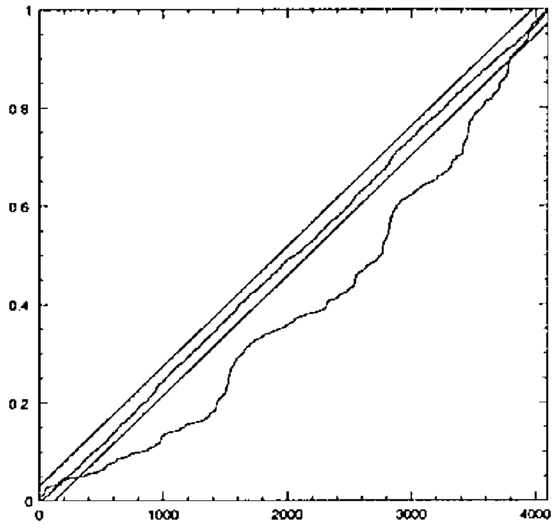

Figure VI.11: Cumulative wavelet distribution of Standard and Poor's returns for 4096 consecutive trading days beginning in 1962. The null hypothesis of a geometric random walk is clearly rejected; the cumulative wavelet distribution of a white noise sample is also shown and it is inside the 5% confidence bounds. Summation and statistics are based on first summation path.

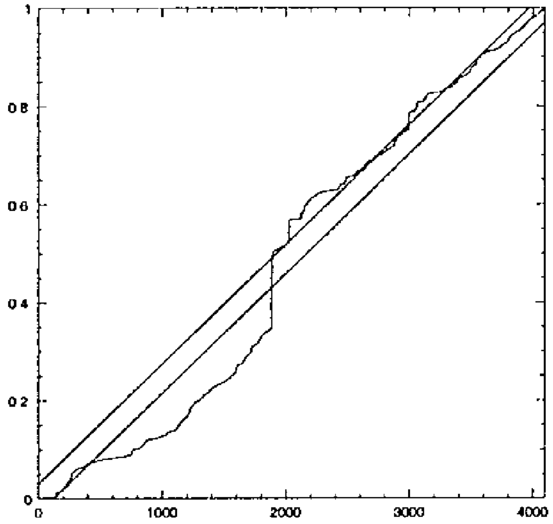

Figure VI.12: Cumulative wavelet distribution of log differences in dollar-DM exchange rates for 4096 consecutive trading days beginning in 1972. The null hypothesis of a geometric random walk is clearly rejected; the cumulative wavelet distribution of a white noise sample is also shown and it is inside the 5% confidence bounds. Summation and statistics are based on first summation path.

$$y(t) = 0.2y(t-1) + u(t) \tag{VI.35}$$

where u_t is white noise with variance 1.0. Adjusted 5 % confidence intervals were used to correct for small sample bias in the cumulative periodogram case ($k_{0.05}$ was set to 1.315 instead of the asymptotic value of 1.36); cumulative periodogram tests rejected 98.93% of the time. For the waveletgram tests we used Daubechies D8 wavelets. The waveletgram test with the first summation path rejected in 96.53% of the simulations and the waveletgram test with two summation paths rejected 94.19 % of the time. When we switched the AR parameter from 0.2 to 0.1 the cumulative periodogram rejected 55.82 % of the time and the waveletgram test with the first summation path rejected the null of randomness in 45.80 % of the simulations and the waveletgram test with two summation paths rejected in 37.32 % of the simulations.

To show the efficacy of the waveletgram test on nonstationary time series we considered a model in which the structure switches midway through the time series; at the midpoint of the time series we changed the autoregressive coefficient from 0.4 to -0.4. We considered in each case 200,000 replications of samples of size 512 and used the Daubechies D8 wavelet for all the cumulative waveletgram tests. All tests were performed at a 5 % significance level. The cumulative periodogram test rejected 61.94 % of the time whereas the cumulative waveletgram test rejected 97.51% of the time with the first summation path and 94.27% of the time when both summation paths were used. When the autoregressive parameter 0.2 was used instead of 0.4, the cumulative periodogram test rejected 10.04% of the time whereas the cumulative waveletgram test rejected 28.19% of the time with the first summation path and 21.59% of the time when both summation paths were used. One expects that the relative performance of the cumulative waveletgram will be somewhat stronger in situations in which there is more than a single structural break or in which the structural change is continuous.

One notes that in all the examples we have presented, inclusion of the second summation path actually results in a deterioration of statistical power. This is because the second summation path is intended to take into account more local features of the data and the first summation path sums across time at each frequency and hence has more global features; the first summation path is therefore the appropriate one when the model is close to stationary as is the case in our examples.

Another possibility is an adaptive waveletgram stopping rule which is based on wavelet packet best bases [47] [45] [44] [46] which are reviewed in Ch. VII. Briefly, wavelet packet bases are orthonormal bases which adapt to the properties of the data; for instance, a wavelet series does not represent some sine waves well, so if the data is a sine wave, the selected wavelet packet basis will be closer to the standard Fourier basis. The adaptive waveletgram test thus has the advantage over the cumulative periodogram tests and the cumulative waveletgram is that it does not depend on any *a priori* assumptions about the nature of the stochastic process; instead the adaptive waveletgram test exploits an orthonormal representation which optimally combines time and frequency elements to minimize a global entropy criterion. This flexibility comes at a cost in that for the adaptive test confidence intervals must be wider than in any test based on a single orthonormal basis. These wider confidence intervals result in suboptimal performance when the type of departures from randomness can be neatly described by a single orthonormal basis or when the time-frequency properties of the time series are not parsimoniously described by waveforms. Confidence intervals differ for the adaptive waveletgram depending on the wavelet used in the analysis. Table VI.2 shows 5% confidence intervals for a Daubechies $D8$ wavelet.

We next consider some specific examples of use of a cumulative waveletgram test and competing tests. The examples we try include a model with time-varying variances and a deterministic periodicity corrupted with noise. All simulations use 200,000 replications. All waveletgram calculations use the Daubechies D8 wavelet.

Sample Size	Position path	Both paths
64	2.14	2.22
128	2.19	2.26
256	2.22	2.28
512	2.25	2.29
1024	2.26	2.32

Table VI.2: 5% simultaneous confidence intervals for the cumulative wavelet packet distribution test of randomness for the D8 wavelet. In the simulations, the mean was not removed prior to analysis; the constants with the mean removed are slightly lower.

In each of the examples the mean was removed from the data before analysis. All waveletgram calculations use summation paths over time and frequency.

Time-Varying Variances

Models of time-varying variances are often used to model asset market behavior in financial economics [30] [70] [29] [145]. We consider a simple model of a time series with a time-varying variance:

$$y(t) = \sigma(t)\epsilon(t) \tag{VI.36}$$

$$\sigma(t) = \lambda_0 + \lambda_1 \left|\frac{(t-t_0)}{T}\right|^{0.5} \tag{VI.37}$$

where $\epsilon(t)$ is an independent Gaussian random variable with variance 1. A sample realization is shown in Figure VI.13 for $\lambda_0 = 0$, $\lambda_1 = 1$, $T = 128$, $t_0 = 64$. In Table VI.3 we show results for fixed values of λ_0, λ_1. We also consider the case in which λ_0 at each point in time is a random variable equal to $|\lambda_0\zeta(t)|$. $\zeta(t)$ is an independently distributed Gaussian random variable. The results for this case are

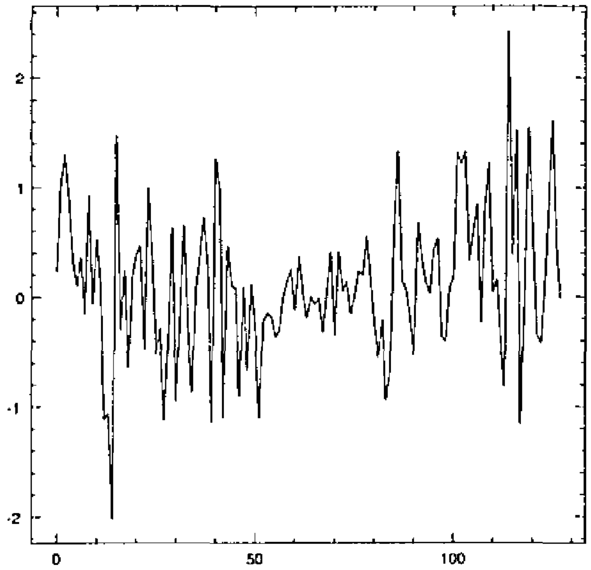

Figure VI.13: Data from a time series with time-varying volatility parameters $\lambda_0 = 0$, $\lambda_1 = 1$, $t_0 = 64$, $T = 128$.

shown in Table VI.4. In all subsequent tables, we use the abbreviations CPG for Cumulative Periodogram Test, CWG for Cumulative Waveletgram Test and AWG for Adaptive Waveletgram Test.

The results show clearly that for this example the cumulative waveletgram test outperforms the cumulative periodogram test, as is to be expected given the non-stationarity in the simulated data. For our example with time-varying volatility, the

λ_0	λ_1	CPG	AWG	CWG
0.0	1.0	0.106	0.262	0.694
0.5	0.5	0.061	0.088	0.273
0.5	1.0	0.071	0.121	0.426

Table VI.3: Power of different randomness tests (level 5%) against the deterministic time-varying volatility model for cumulative periodogram test (CPG), adaptive waveletgram test (APG), and cumulative waveletgram test (CWG).

c_0	λ_1	CPG	AWG	CWG
0.5	0.5	0.058	0.189	0.334
0.5	1.0	0.070	0.165	0.457

Table VI.4: Power of different randomness tests (level 5%) against the stochastic time-varying volatility model.

adaptive waveletgram achieves results intermediate between those of the cumulative periodogram and the cumulative waveletgram. In the case with hidden periodicities it was also basically able to achieve results intermediate between those of the cumulative periodogram test and the cumulative waveletgram test so that from the examples it appears to achieve acceptable performance when applied to two radically different types of time series.

Deterministic Periodicity Corrupted with Noise

A classical application of the cumulative periodogram test is in detection of a hidden periodicity. We first consider a model where the data is generated by a single sinusoidal function corrupted by white noise $\epsilon(t)$:

$$y(t) = \sin\left(2.0\,\omega\,\frac{(t-T/2)}{T}\pi\right) + \sigma\epsilon(t) \tag{VI.38}$$

We consider in particular an example with $T = 128$, $\omega = 20.0$ and $\sigma = 1.0$. The data is shown in Figure VI.14. In Figure VI.15 we show that a cumulative waveletgram test at a 5 % level fails to reject the null hypothesis of randomness. A cumulative periodogram test narrowly rejects randomness (Figure VI.16) as does the adaptive waveletgram test (Figure VI.17). Simulations indicate that a level 5% cumulative periodogram test rejects 92% of the time. Table (VI.5) presents results of some detailed simulations. From the results, it is clear that the cumulative waveletgram test performs particularly poorly with high frequency waveforms; this should not be

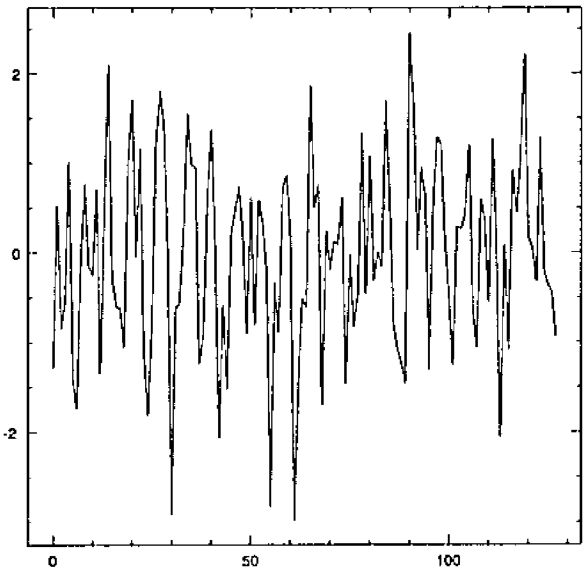

Figure VI.14: High-frequency sine wave corrupted by random noise.

surprising as it is a direct consequence of the time-frequency localization imposed by wavelets (see shaded areas of Figure (VII.1)). From Table (VI.5) it is also clear that in this example the adaptive waveletgram performs significantly worse than either the cumulative waveletgram or the cumulative periodogram for very low frequency sinusoidal deterministic components. In general, however, the adaptive waveletgram achieves a performance level which is intermediate between the cumulative waveletgram and the cumulative periodogram.

The cumulative waveletgram test has the advantage over the cumulative periodogram test that it does not make the assumption of stationarity of the underlying stochastic process; instead the cumulative waveletgram test uses wavelet basis functions which fix time and frequency elements. Though it is intended for analysis of time series whose properties are time-varying, the cumulative waveletgram test appears to yield acceptable performance against stationary alternatives.

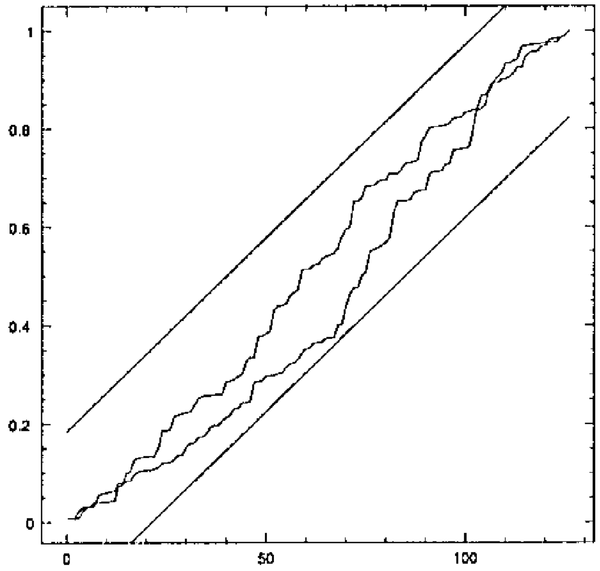

Figure VI.15: Cumulative waveletgram for high-frequency sine wave corrupted by random noise. A Daubechies D8 wavelet was used in computing wavelet coefficients. The cumulative waveletgram test fails to reject the null of randomness at a 5 % level.

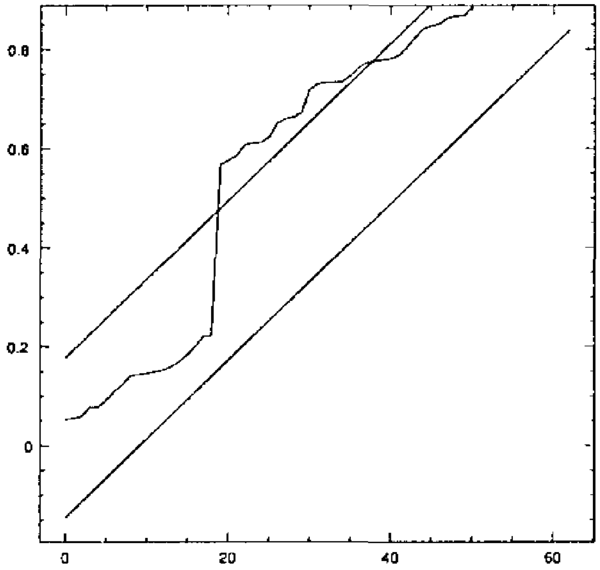

Figure VI.16: Cumulative periodogram for high-frequency sine wave corrupted by random noise. The cumulative periodogram test rejects the null hypothesis of randomness at the 5% level. Simulated 5% confidence intervals were used with a constant equal to 1.27 instead of the asymptotic value of 1.36 due to small sample bias.

σ	ω	CPG	AWG	CWG
1.0	50.0	0.959	0.903	0.562
2.0	50.0	0.203	0.129	0.086
3.0	50.0	0.083	0.065	0.058
1.0	40.0	0.891	0.846	0.295
2.0	40.0	0.136	0.126	0.067
3.0	40.0	0.064	0.064	0.053
1.0	30.0	0.876	0.871	0.047
2.0	30.0	0.123	0.117	0.046
3.0	30.0	0.059	0.061	0.046
1.0	20.0	0.922	0.725	0.440
2.0	20.0	0.164	0.115	0.070
3.0	20.0	0.071	0.064	0.052
1.0	10.0	0.984	0.735	0.903
2.0	10.0	0.248	0.116	0.153
3.0	10.0	0.093	0.064	0.072
1.0	5.0	0.995	0.886	0.978
2.0	5.0	0.294	0.156	0.210
3.0	5.0	0.102	0.072	0.082

Table VI.5: Power of different randomness tests (level 5%) against the corrupted sine wave model.

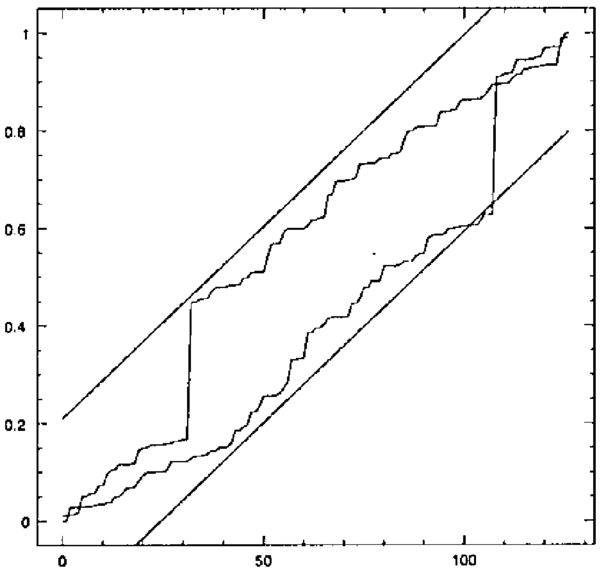

Figure VI.17: Adaptive waveletgram for high-frequency sine wave corrupted by random noise. Like the cumulative periodogram test, this test rejects the null hypothesis of randomness at the 5% level.

Other Stopping Rules

In this chapter, we have reviewed the possibility of a cumulative waveletgram stopping rule and compared performance with cumulative periodogram tests and an adaptive waveletgram test we introduce. One problem with the cumulative waveletgram and adaptive waveletgram tests is that the orthonormal basis obtained is one of $\mathbf{L}^2$ on the real line and not $\mathbf{L}^2$ on a compact interval. It is possible, with some increase in complexity, to use wavelets on a finite interval as developed by [42] and use fast algorithms to calculate the appropriate coefficients.

There is an extensive econometric literature on structural change and model identification. It is useful to review why we have not used these criteria. An alternative approach to choosing a stopping rule would be to refine traditional criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC)

[1] [2] [3]. These criteria balance the total percentage of sample variance explained by a model with the need for parsimony. While such criteria could be measured easily in the autoregressive pursuit procedure, the statistical properties of such measures would have to be studied extensively because we use maximal model components instead of the next lag and also because we expect the statistical distribution to depend on the number of model components; in any case, the statistical interpretation of information criteria is difficult and thus their primary usefulness comes from their practicality.

CHAPTER VII

TIME-FREQUENCY SPECTRAL ESTIMATION

The thesis develops a method for the analysis of nonstationary economic data so it is helpful to provide some special tools with which to visualize the properties of nonstationary economic and financial data. With stationary data, the frequency spectrum provides an excellent and intuitive variance decomposition of the data. With nonstationary data, the properties of the data change over time so we need a new type of variance decomposition analogous to the frequency spectrum

In developing these methods, it is helpful to consider a general review of the literature, both from economics and elsewhere, through the unifying concept of time-frequency spectral estimation; since the goal of classical time series analysis is a variance-covariance decomposition, it seems natural to focus attention on how apparently dissimilar methods work to achieve the same goal. Our broad review is also useful for the comparisons with the literature and for various other developments in the thesis.

Before we do this, it is helpful to state what is the key idea in this chapter: *the time varying spectrum is the Wigner-Ville transform of the estimated covariance matrix; the parametric model estimated by autoregressive pursuit can be used to construct parametric spectral estimates as well as to find the 'eigenvectors' of a time series which define an optimal decomposition.* The goal of this chapter is to clarify the previous sentence and relate it to the literature.

Literature Review

We summarize some of the competing methods which have been developed for the analysis of nonstationary time series. Since the applications of nonstationary time series methods are quite broad, it is not surprising that the methods come from a large variety of fields. Therefore, it is inevitable that some or most of these methods will be unfamiliar; indeed, many of the methods reviewed are based on frontier work and have never been used in economics or finance. The goal of this section is to give our conceptual view of how the method developed in the thesis fits into the broader literature and we will try to provide this view in as self-contained a manner as possible.

It is natural to divide the the field of nonstationary time series methods into two categories: (1) nonparametric and (2) parametric. Our approach lies somewhere in between as it involves selection of a best parametric model from a large family of potential parametric models. Many aspects of the nonparametric approach have developed mainly in the engineering literature and hence will be less familiar to economists than parametric approaches. However, since nonparametric methods are receiving increasing use in microeconometrics, it is reasonable that nonparametric methods for nonstationary time series will receive increasing attention in the economics literature in the future.[1] We begin our review with nonparametric methods as such methods help focus attention on the essence of the problem as well as its difficulty.

[1] Most of the work on nonparametric regression in economics focuses on estimation of nonlinear relationships (c.f., [100]) and this has also been the case in application of such methods to time series (c.f. [86]). Application of nonparametric methods to *nonstationary* time series seems relatively undeveloped at present in the econometrics literature.

Nonparametric Methods

Spectral analysis is a useful variance decomposition for stationary processes; however, for nonstationary processes, the interpretation of the spectrum is unclear. For some nonstationary processes, it is clear how to define the notion of a time-varying spectrum. For instance, if the stationary data generating process changes at some time τ to some other stationary data generating process then it is reasonable to consider definition of a time-varying spectrum as equal to the spectrum of the first process before τ and that of the second process after τ. In most practical situations, change is not instantaneous or the underlying nonstationarity is more fundamental.

For time-varying processes, it is helpful to move away from representation of the data in terms of basis functions which have time-invariant properties (e.g., e^{ikt} where the frequency variable k is constant over time) to representation of the data in terms of basis functions which are *jointly localized* in the time and frequency domain. What "joint localization" means is that the basis function used to represent the data has a relatively compact Fourier space representation yet still has a relatively compact time domain representation. For instance, modulated Gaussian functions:

$$g_{k,b,s}(x) = e^{ikx}e^{-\frac{(x-b)^2}{s}} \tag{VII.1}$$

characterized by parameters k, b, and s are one family of functions which meets these criteria.[2] We note that as $s \rightarrow \infty$ these functions approach Fourier basis functions so that Fourier analysis can be considered as a special case of a more general type of data analysis.

We consider replacing Fourier analysis with expansion with respect to more general basis functions such as the family defined by Eq. (VII.1). To focus the discussion,

[2] This follows since the Fourier transform of a Gaussian is a Gaussian so that the functions decay exponentially in both the time and frequency domains.

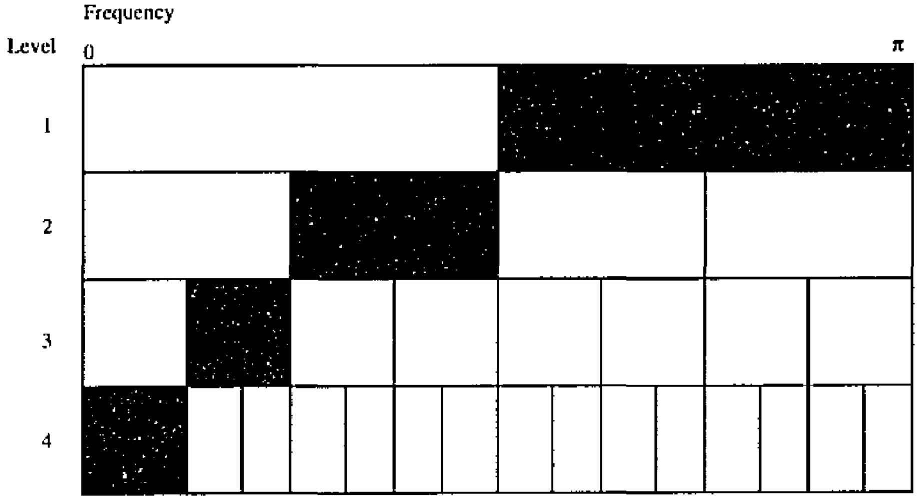

Figure VII.1: The first few levels of a general multilayer decomposition. Boxes represent effective independent frequencies at the different levels. Note that wavelet bases (shaded boxes) have more precise frequency localization at low frequencies than at high frequencies.

we consider data X and define a sample transform U as follows:

$$UX(a,b,\omega) = \int_{-\infty}^{\infty} \phi\left(\frac{t-b}{a}\right) X(t) e^{2\pi i\omega t}\, dt \tag{VII.2}$$

When we fix a we have a frequency based transform; such a transform is often called a "windowed Fourier transform" and is appropriate in applications in which the underlying basis functions (or waveforms) change along identical time scales. When we fix ω/a, we have a scale based method; an example of such a method is the wavelet transform [130] [54] and such a method is appropriate in cases in which high frequency waveforms change much more rapidly than low frequency waveforms.

Eq. (VII.2) defines such a large number of representations for data that we will focus on only the orthonormal representations. The idea is represented in Figure (VII.1) and the next few pages seek to explain at various levels exactly what Figure (VII.1) means. Since the class of possible alternatives to a covariance-stationary system is large, it is helpful to outline in a general sense an overall framework in which to perform time series analysis without the usual covariance-stationarity assumption *and* without at the same time making arbitrary assumptions with just as little theoretical justification. The traditional econometric test for structural change is an F-test [38] performed after splitting the sample into two usually equal portions. Therefore, to begin, we consider a univariate time series with length T which for expository purposes only we will consider to be a power of 2. We now suppose that we are to divide up the time series into $log_2 T$ separate orthonormal decompositions each of which involves dividing the time series into separate time series of length $T/2^j$ where j is an integer ($log_2 T > j \geq 0$) and performing separate spectral decompositions of each partition at each j. We choose an orthonormal decomposition in order to illustrate effective frequency localization; in many cases orthonormality is not desirable or effective. At $j = 0$ there is only one partition and the decomposition is similar or equivalent to ordinary spectral analysis.

Since we will not in general be using Fourier modes, our interpretation of a frequency here will be a loose one. With Fourier basis functions, there is an inherent symmetry in spectral analysis between positive and negative frequencies. Here, we are concerned only with a representation of a time series and have not yet discussed spectral estimation for nonstationary processes; therefore, in the case of Fourier frequencies, we will count both positive and negative frequencies in the total number of frequencies. At $j = 1$, we have two partitions each of which enables us to resolve $T/2$ frequencies, and so on. Figure (VII.1) (ignoring the shading for now) illustrates this idea.

At the highest level of the diagram we have only 2 'frequencies' but we have divided the interval into $T/2$ separate subdivisions. As we progress down the levels of our diagram, we have more frequencies but fewer subdivisions in the time domain. We have put four levels in the diagram for reasons of simplicity; four levels imply a time series of only length 16 so that the corresponding diagram for actual time series such as occur in financial and macroeconomic time series would be extremely complicated. In fact, in practical applications, one may want to include data on a constant intermediate level so that the number of levels would be even more extensive. We only display the frequency representation in $[0, \pi]$ as we assume the frequency representation in $[-\pi, 0]$ is symmetrical.

Let us consider a sample of data $\{f_m\}_{m=1}^{T}$. There are a variety of ways to summarize the data. One way is that, if we feel that the data has time-invariant properties, we can consider the Fourier transform of the data:

$$\hat{f}(k) = \sum_{m=1}^{T} f_m e^{-2\pi i k \frac{(m-1)}{T}} \tag{VII.3}$$

where $k = 0, \ldots T - 1$. With the Fourier transformed data, we might be able to summarize the data effectively in terms of only a few coefficients. However, the properties of the data might be changing over time so we might get more effective information about the data generating the process by dividing up the sample into two segments; one segment is $[1, \frac{T}{2}]$ and the other segment is $[\frac{T}{2} + 1, T]$. On each segment we can take a Fourier transform. On the first segment we have:

$$\hat{f}_1^2(k) = \sum_{m=1}^{\frac{T}{2}} f_m e^{-2\pi i k \frac{(m-1)}{\frac{T}{2}}} \tag{VII.4}$$

for $k = 0, \ldots \frac{T}{2}$ and on the second segment we have:

$$\hat{f}_2^2(k) = \sum_{m=\frac{T}{2}+1}^{T} f_m e^{-2\pi i k \frac{(m-\frac{T}{2}-1)}{\frac{T}{2}}} \tag{VII.5}$$

for $k = 0, \ldots \frac{T}{2}$ and where the notation $\hat{f}_n^p$ refers to the nth segment of a series with p subdivisions. In general, it is simple to consider situations where the series has $p = 2^r$ subdivisions where $0 \leq r \leq \log_2(T)$. The general formula is:

$$\hat{f}_n^{2^r}(k) = \sum_{m=(n-1)\frac{T}{2^r}+1}^{(n+1)\frac{T}{2^r}} f_m e^{-2\pi i k \frac{(m-\frac{T}{2^r}-1)}{\frac{T}{2^r}}} \tag{VII.6}$$

so that n runs from 1 up to 2^r and k from 0 to $\frac{T}{2^r}$.

When the series has two subdivisions, we have only half the number of Fourier coefficients we have than in the case when the series had only one subdivision. More generally, when there are p subdivisions, we have only p^{-1} as many Fourier coefficients at each subdivision. Since when we have more subdivisions, we have fewer 'frequencies', we have more information about 'when' something happens but less information as to what exactly is happening. This tradeoff between the number of subdivisions and the amount of 'frequency' information we can obtain from the data is known as the 'Heisenberg' uncertainty principle; what this says essentially is that it is impossible to come up with a more clever scheme in which the product of the number of independent frequencies p^{-1} and the number of subdivisions p is not a constant.

Time-frequency spectral analysis can be thought of a way of a more sophisticated way of dividing up the time domain and doing separate spectral analyses. For instance, one idea is to allow the frequency estimates to have a bandwidth which may vary across frequencies. Another idea is that, instead of using flat time-domain windows as in rolling spectral analysis, we use smooth time-domain windows. It is possible to construct various smooth frequency domain window functions which pieced together form an orthonormal basis of $\mathbf{L}^2$.

Such constructions are nontrivial; indeed, until several years ago, there were not believed to exist. The next section reviews some of the details of these constructions. It is therefore somewhat technical and can be skipped on a first reading. The material is used implicitly in various places in the thesis, including the choice of stopping rule. The rest of this paragraph reviews the few ideas from the next section which are important for later developments. If a time series has properties which vary over time, it is sensible to use wider windows to measure variations at low frequencies than at high frequencies because more data is required to accurately estimate low frequency movements. The last sentence expresses the idea of a wavelet expansion. The shaded regions in Figure (VII.1) show the effective time-frequency concentration of wavelet-type functions. In some cases, one may want to optimally combine orthonormal basis elements from the different levels to more accurately capture certain properties of a data set such as time-invariant components; this is the idea of wavelet packets. Both wavelets and wavelet packets result in orthonormal expansions which can be used in nonparametric spectral estimation.

Wavelets and Wavelet Packets

Wavelets are basis functions of $\mathbf{L}^2$ with joint time-frequency localization and frequency dependent bandwidth [54] [130] [39] [138] (for applications and recent theoretical developments, see [185]). Wavelet packets decompose wavelets into longer and shorter "waves" to better capture sharp spectral components and other data properties not well represented by the ordinary wavelet transform. A wavelet $x \mapsto \psi(x)$ is a function such that:

$$\frac{|\hat{\psi}(\omega)|}{|\omega|^{\frac{1}{2}}} \in \mathbf{L}^2 \tag{VII.7}$$

which requires that:

$$\int dx\, \psi(x) = 0. \tag{VII.8}$$

There exist wavelet functions $\psi(x)$ whose dilations and translations:

$$\{\psi_{jk} = 2^{-\frac{j}{2}}\psi(2^{-j}x - k)\}_{j,k=-\infty}^{\infty} \tag{VII.9}$$

form an orthonormal basis of $\mathbf{L}^2$. We have focused on such wavelets in Ch. VI which begins with a primer on wavelets.

In terms of Figure (VII.1), it will be helpful to discuss the sort of time-frequency localization wavelets impose. The localization of wavelets is illustrated by the shaded boxes. The type of localization we choose is to use a lot of information to estimate low frequency components and much less data to estimate high frequency components. A method which combines all information in the levels of Figure (VII.1) would essentially use waves with different numbers of oscillations and would include as special cases (either approximately or exactly) the localization properties of wavelets and Fourier analysis. The number of possible decompositions given a fixed wavelet function is enormous (for instance with a sample size of 1024 there are approximately 1.8×10^{308} possibilities) so that the problem is difficult in general. Wavelet packets present one approach to an orthonormal decomposition which attempts to approximately meet these criteria.

We consider the following equations:

$$P_{2r}(x) = \sqrt{2}\sum_{k} h_k P_r(2x - k) \qquad \text{(VII.10)}$$

$$P_{2r+1}(x) = \sqrt{2}\sum_{k} g_k P_r(2x - k) \qquad \text{(VII.11)}$$

where P_0 is some function ϕ which corresponds to a wavelet function ψ. A given function is expanded in terms of each of the P_r as well as its dilations. In terms of Figure (VII.1) any combination of basis functions which span frequency space (wavelets are one possibility, as are basis functions at a single level such as the bottom level) will form an orthonormal basis.

Writing Equations (VII.10) and (VII.11) in terms of Fourier transforms we have:

$$\hat{P}_{2r}(\omega) = m(\omega)\hat{P}_r(\omega) \qquad \text{(VII.12)}$$

$$\hat{P}_{2r+1}(\omega) = n(\omega)\hat{P}_r(\omega) \qquad \text{(VII.13)}$$

where:

$$m(\omega) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2}}\sum g_k e^{ik\omega} \tag{VII.14}$$

$$n(\omega) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2}}\sum h_k e^{ik\omega} \tag{VII.15}$$

There exist filters m and n which lead to orthonormal wavelet bases. Such filters and related literature are reviewed in [130] [52] [200].

Since the set of possible orthonormal basis functions (c.f., Fig (VII.1) has a tree structure, the idea of the Coifman, Wickerhauser and Meyer algorithm is to compare the entropy of function expanded in terms of P_r at scale[3] s with the sum of the entropies of a function expanded in P_{2r} at scale $s+1$ and P_{2r+1} at scale $s+1$, and to select the sequence with the minimum entropy and store its entropy. We begin at $s = 0$ and continue comparisons until we reach $s = s_{max}$ or we select $V_{r,s}$ for each r. As we shall see in some of the comparisons between methods, the entropy minimization method does not perform particularly well on noisy data.[4]

To summarize, the wavelet packet transform is an expansion of a function or a sequence such as a time series f such that:

$$f(x) = \sum_{r,s,k \in B \subset \mathcal{B}} < f, D^s P_r(. - k) > D^s P_r(x-k) \tag{VII.16}$$

where a B is one of a set $(\mathcal{B})$ of 2^N admissible orthonormal bases with components:

$$\{D^{s_i} P_{r_i}(.\, k_i)\}_{i=1}^{N}. \tag{VII.17}$$

[3] Scale s means that the function $P_r(x)$ is dilated to $2^{\frac{s}{2}} P_r(2^s x)$.

[4] Another cost function based on a L^p norm ($p < 2$) for wavelet packets was announced at Stanford University in late Fall, 1993 by statisticians David Donoho and Ian Johnstone. The motivation for the new choice of cost function was to improve performance with noisy data. Donoho and Johnstone [62] [61] have also developed a nonparametric estimator for noisy functions and densities based on shrinkage of wavelet coefficients; this method would work for an arbitrary wavelet packet basis but to our knowledge its statistical properties have not been developed in the case in which disturbances are not independent, the relevant case in time series.

Wavelets correspond to one such possible orthonormal set with $r_i = 1, s_i = i$ for all $1 \leq i < N$ and $r_N = 0, s_N = \log_2 N$. A windowed Fourier transform corresponds to $s_i = \bar{s}$ and $r_i = i$ for $1 \leq i \leq N$. The best basis is the orthonormal basis $B^* \in \mathcal{B}$ which achieves the minimum entropy of any orthonormal basis in the set $\mathcal{B}$.

As estimates of a time-varying spectrum, we find wavelet and wavelet packet estimates are very difficult to interpret in the presence of noise (because white noise has a wavelet packet decomposition which is spread out over all possible basis sets, most of which vary over time); on the other hand, wavelet and wavelet packet decompositions seem to provide more parsimonious representations of many economic time series than Fourier decompositions. Such decompositions can be used in various ways to measure how stationary a time series is; for instance, the 'best' level of Fig. (VII.1) according to a Coifman/Meyer/Wickerhauser entropic criterion is a division of monthly money supply data (CITIBASE FM1D2) into two year segments whereas the best division for U.S. GNP data is the full sample, suggesting stationarity.[5]

In this section, we have provided examples of some members of the family of transforms defined by Eq. (VII.2). We have yet to review technical definitions of nonstationary spectral estimators. The goal of the next section is to address these questions.

Nonstationary Spectral Estimation

With ordinary spectral analysis, the periodogram is estimated by squaring the Fourier transform of the data. This works because the spectrum is the Fourier transform of the autocovariance function which is a convolution of the data with itself and

[5] U.S. GNP data from CITIBASE series GNPQ (log differences) from 1961:4 to 1993:3. Money supply data from 1951:2 to 1993:9. Calculations used a Daubechies $D8$ wavelet. On the issue of division of a time series with time domain algorithms, the development of so-called lapped orthogonal transforms [132](briefly, flat time domain windows with smooth dropoff but which still result in an orthonormal basis when cosine or sine functions are used) present an alternative to wavelet packets if a smooth time domain window function is to be used.

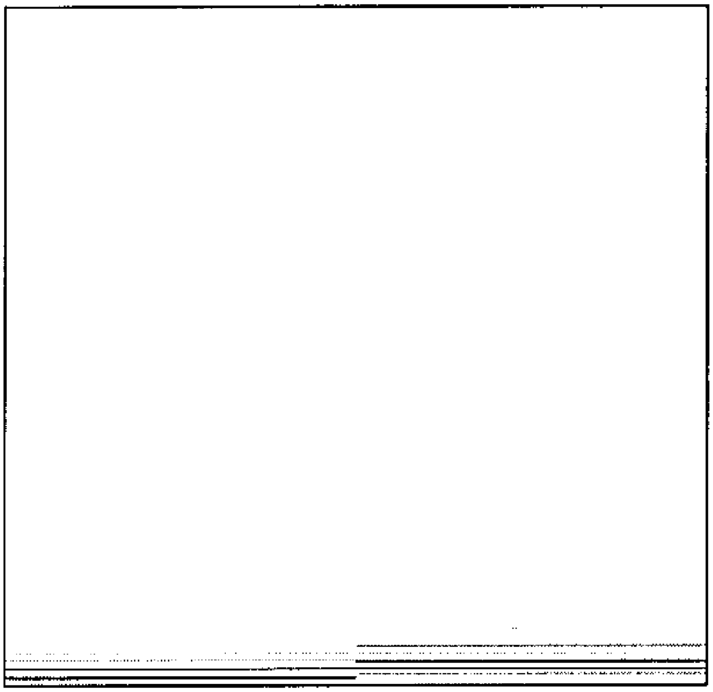

Figure VII.2: Wavelet packet best basis decomposition of an AR(1) process in the time-frequency plane. Frequency is on the vertical axis.

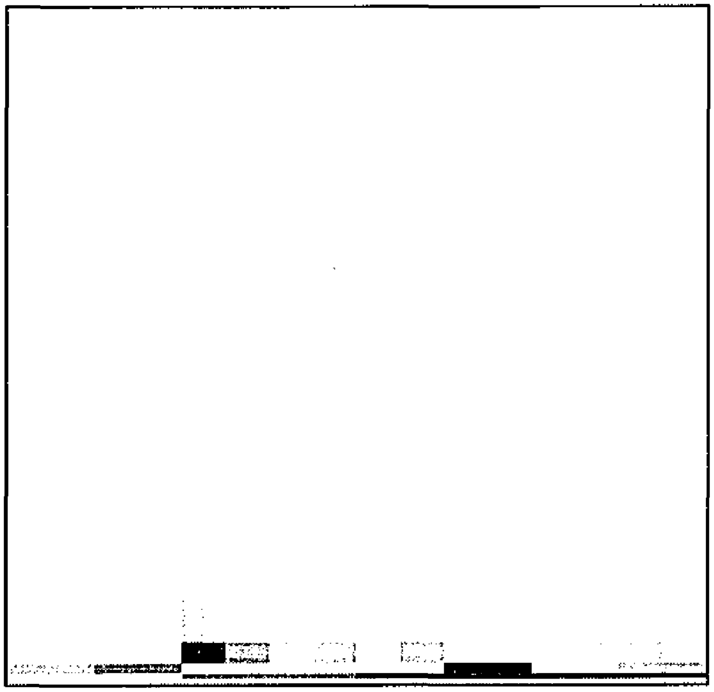

Figure VII.3: Wavelet decomposition of an AR(1) process in the time-frequency plane. Frequency is on the vertical axis.

convolution in the time domain is equivalent to multiplication in the frequency domain. When the time series is not globally stationary, the standard approach which assumes stationarity may lead to misleading conclusions about the spectral properties of the data. A further problem is that there is no accepted definition of a time-varying spectrum in the literature. We now suggest how wavelets and wavelet packets relate to spectral estimates and more traditional methods of nonstationary spectral analysis such as rolling spectral analysis. Furthermore, we explain how such methods as well as traditional methods and the method developed in the thesis fit into our defintion of a time-varying spectrum.

A useful starting point is to consider Cohen's class of time-frequency distributions [43] [71]. The Cohen's class representation for a function f is defined as follows:

$$C_f(t,\omega) = \int \int \Omega(r-t,n) f(r+\frac{n}{2}) f^*(r-\frac{n}{2}) e^{-i2\pi\omega n} dr dn \qquad \text{(VII.18)}$$

Here $C(t,\omega)$ is the time-frequency distribution, $\Omega(r-t,n)$ is the kernel of the time-frequency distribution, and $f(t)$ is the time series. Cohen's class and the time-frequency estimators are for univariate time series. The Cohen's class is usually defined in a slightly different but equivalent way:[6]

$$C_f(t,\omega) = \int\int\int K(\alpha,n)f(r+\frac{n}{2})f^*(r-\frac{n}{2})e^{-i2\pi\omega n+2\pi i\alpha(r-t)}d\alpha dr dn \qquad \text{(VII.19)}$$

There is a relationship between the Cohen's class representation of a function and a nonstationary spectral estimator. To see this relationship, we let the function f be a stochastic process X corresponding to the observed data. We then take expectations:

$$\begin{aligned} C_X(t,\omega) &= E\left(\int\int\Omega(r-t,n)X(r+\frac{n}{2})X^*(r-\frac{n}{2})e^{-i2\pi\omega n}\,dr\,dn\right) \\ &= \int\int\Omega(r-t,n)E\left(X(r+\frac{n}{2})X^*(r-\frac{n}{2})\right)e^{-i2\pi\omega n}\,dr\,dn \qquad \text{(VII.20)} \end{aligned}$$

where for notational simplicity we have used the same notation to refer to the expected and sample values of the Cohen's class time-frequency distribution. We define the local autocovariance kernel for the data X:

$$K_X(t_1,t_2) = E(X(t_1)X^*(t_2)) \qquad \text{(VII.21)}$$

We define:

$$W_X(t,\omega) = \int_{-\infty}^{\infty} e^{-2\pi i\omega m}K_X(t+\frac{m}{2},t-\frac{m}{2})dm \qquad \text{(VII.22)}$$

$W_X(t,\omega)$ is a Fourier transform of a local autocovariance kernel and it is an example of a Cohen's class distribution with $\Omega(\alpha,n)=\delta(\alpha)$.[7] This particular distribution

[6] The original Cohen's class definition was presented in terms of the phase space approach to quantum mechanics. Position and momentum (or capital stock and shadow value in economic terminology) are the variables there instead of time and frequency.

[7] $\delta(\alpha)$ is Dirac's delta function.

is called the Wigner-Ville distribution. The theory of random functions or stochastic processes differs from that of ordinary functions in that a bilinear form such as the Wigner-Ville distribution is necessary to characterize the stochastic process whereas a linear form is sufficient to characterize each realization. This distinction is important to some degree for understanding the way in which we use some methods for the analysis of ordinary functions such as Mallat and Zhang's matching pursuit algorithm and wavelet transforms.

Since the Wigner-Ville distribution has the problem that it can sometimes be negative and that the Wigner-Ville transform of two separate time series is not additive, it is often useful to consider a smoothed Wigner-Ville distribution with $\Omega(\alpha, n) = \delta(\alpha)h(n)$:

$$W(t,\omega) = \int_{-\infty}^{\infty} e^{-2\pi i\omega m} K_X(t+\frac{m}{2}, t-\frac{m}{2})h(m)dm \qquad \text{(VII.23)}$$

For discrete time series, the natural modification of the Cohen's class distribution is [134]:

$$C_f(t,\omega) = 2\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \Omega(n-t, m)f(t+m)f^*(t-m)e^{-4\pi jm\omega} \qquad \text{(VII.24)}$$

which also has a clear interpretation in terms of a time-varying spectral estimate of a locally stationary process. In terms of practical applications, the issue would then seem to be only the choice of an appropriate kernel Ω.

For illustrative purposes, we consider a Wigner-Ville distribution of a complex exponential $f(t) = e^{2\pi\omega' t}$:

$$\begin{aligned} C_f(t,\omega) &= \int_{-\infty}^{\infty} e^{-2\pi i\omega m} f^*(t-\frac{m}{2})f(t+\frac{m}{2})dm \\ &= \int_{-\infty}^{\infty} e^{-2\pi i\omega m} e^{-2\pi i\omega'(t-\frac{m}{2})} e^{2\pi\omega' i(t+\frac{m}{2})}dm \\ &= \int_{-\infty}^{\infty} e^{-2\pi i(\omega-\omega')m}dm = \delta(\omega-\omega') \qquad \text{(VII.25)} \end{aligned}$$

However, if we were to consider a function which is the sum of complex exponentials such as:

$$f(t) = e^{2\pi i\omega' t} + e^{2\pi i\omega'_2 t} \tag{VII.26}$$

the Wigner-Ville distribution will *not* be the sum of the Wigner-Ville distributions of individual elements. [8] In fact, the Wigner-Ville distribution need not be positive. [9]

This and other properties of basic time-frequency distributions are reviewed in [28]. The nonadditivity comes from cross-terms which can be controlled but not eliminated through an appropriate choice of kernel for the Cohen's class [28]. Indeed, the development of Cohen's class estimators was motivated by the need to eliminate or minimize the effects of cross-terms in the Cohen's class distribution. A fundamental

[8] Thus, the Wigner-Ville representation of a time-invariant process usually *will* not reduce to the periodogram. The statements by Priestley in [169] [170] about this matter, while not incorrect, refer to a slightly different situation (in which the correlation function of the data is known) and hence are misleading.

[9] For an example, let us consider the Wigner-Ville distribution of a periodic function:

$$f(t) = \begin{cases} \sin(\omega_0 t) & \text{if } -\frac{T}{2} \le t \le \frac{T}{2} \\ 0 & \text{otherwise} \end{cases} \tag{VII.27}$$

In this case, since the sine function is an odd function, at $t = 0$, $\omega = 0$, the Wigner-Ville distribution is:

$$\begin{aligned} W(0,0) &= \int_{-\frac{T}{2}}^{\frac{T}{2}} \sin(\omega_0 \frac{\tau}{2}) \sin(-\omega_0 \frac{\tau}{2}) d\tau \\ &= -\int_{-\frac{T}{2}}^{\frac{T}{2}} \sin^2\left(\omega_0 \frac{\tau}{2}\right) d\tau \\ &= -2\|f\|^2 < 0. \end{aligned} \tag{VII.28}$$

As can be verified computationally, the Wigner-Ville distribution of a sine wave also has some other peculiarities. Thus, even for representing very simple functions such as a sine wave, the Wigner-Ville distribution has some unusual properties. Windowing is hence needed in practice to achieve better estimates. Another solution is to use the Hilbert transform:

$$Hf = -\frac{1}{\pi t} * f \tag{VII.29}$$

where * represents convolution. Since the Hilbert transform of a sine is a cosine, the analytic signal $g(t) = f(t) + iHf(t)$ avoids any pathologies in the Wigner-Ville representation.

practical problem in using Cohen's class of time-frequency distributions is that the results may depend on the choice of kernel.

One outgrowth of the matching pursuit methodology of Mallat and Zhang [131] is a new approach to time-frequency spectral estimation. Mallat and Zhang focused on non-orthonormal expansions of functions; here, we focus on orthonormal expansions and stochastic data, but the ideas are the same. It seems reasonable, given an orthonormal expansion of a function such as that in a wavelet basis or a wavelet packet basis, to define the nonstationary spectra as:

$$S(t,\omega) = \sum_i C_i^2 C_{e_i}(t,\omega) \tag{VII.30}$$

where C_i are the coefficients of an orthonormal expansion and $C_{e_i}(t,\omega)$ is a Cohen's class distribution of the basis function e_i.[10] One motivation for Eq. (VII.30) is that the total sample variance in the data is the sum over i of the squared coefficients C_i^2 so that the spectrum defined by Eq. (VII.30) is meaningful as a variance decomposition. Furthermore, the decomposition is economically meaningful in that it is in terms of building blocks which capture jointly the time and the frequency variation in the data. The spectral estimator defined by Eq. (VII.30) in addition has the advantage that when the orthonormal basis is the Fourier basis and the Wigner-Ville distribution is used, the spectral estimate is the same as the periodogram. When a windowed Wigner-Ville distribution is used and the basis functions are complex exponentials, the estimate is a weighted periodogram. On the other hand, if wavelet functions or wavelet functions are used, the spectral estimates will likely vary over time.[11]

To see the ideas clearly, we revert to the finite-dimensional setting in which the

[10] For instance, for standard Fourier basis functions and a Wigner-Ville distribution kernel, see Eq. (VII.25).

[11] The disadvantage of an approach based on Eq. (VII.30) is that 'marginals' may not be consistent; by this what is meant is that the integral of the spectral estimate over time may not be equal to the periodogram or the smoothed periodogram of the time series.

time series is a vector f with elements f_i, $i = 1,..T$. Assuming f is mean zero and real, the covariance matrix of f is given by:

$$C_{i,j} = E(f_i f_j). \tag{VII.31}$$

We assume the matrix $C_{i,j}$ is positive-definite. We recall the expansion:

$$f(t) = \sum_i < f, e_i > e_i \tag{VII.32}$$

where e_i are orthonormal basis functions. We recall also that this equation was used to derive a limiting stochastic integral representation of f (such as Eq. (N.7)). The "increment" here is $< f, e_i >$. For orthogonality of increments and a positive "spectral" measure, we require that:

$$E(< f, e_i >< e_j, f >) = \delta_{i,j}\lambda_i \geq 0 \tag{VII.33}$$

We note that:

$$\begin{aligned} E(< f, e_i >< e_j, f >) &= E(\sum_k f_k(e_i)_k \sum_l f_l(e_j)_l) \\ &= \sum_{k,l} C_{k,l}(e_i)_k(e_j)_l \end{aligned} \tag{VII.34}$$

where $(e_i)_k$ denotes the kth element of the vector e_i. Suppose the e_i are the eigenvectors of the covariance matrix of the time series. Then it follows that:

$$\begin{aligned} \sum_{k,l} C_{k,l}(e_i)_k(e_j)_l &= \sum_k \lambda_j(e_i)_k(e_j)_k \\ &= \lambda_j < e_i, e_j >= \lambda_j\delta_{i,j} \end{aligned} \tag{VII.35}$$

as required. Furthermore, if the covariance matrix is positive definite, all eigenvalues are positive. For a stationary time series, the eigenvectors are complex exponentials and thus Fourier analysis (which uses complex exponentials as basis functions) is optimal. However, for nonstationary time series, the eigenvectors are no longer complex exponentials so that different orthonormal basis functions are required to achieve an orthogonal increment process representation of the data.

The infinite-dimensional analog of the finite dimensional theory is to compute the eigenvalues of the (self-adjoint) covariance *operator* and define the spectrum in terms of the eigenfunctions of this operator.

Another nonparametric approach to nonstationary time series is the evolutionary spectrum of Priestley [167] [169] [170]. The idea of the evolutionary spectrum is to follow [15] and consider a generalized stochastic integral representation of the data X. This involves using stochastic integral representation:

$$X(t) = \int \phi_t(\omega) dZ(\omega) \tag{VII.36}$$

where dZ is an orthogonal increment process such that:

$$E(dZ(\omega)dZ(\omega')) = dF_X(\omega)\delta(\omega - \omega') \tag{VII.37}$$

where F_X is the cumulative 'spectral' distribution of the data X in terms of complex exponentials. As stated in Priestley's original paper ([167], p. 205-7) earlier approaches such as that of Cramer [50] assumed only harmonizability so that the dZ need not be an orthogonal process.[12]

Priestley considers functions $\phi_t(\omega)$ of the form: [13]

[12] One extant definition of a time-varying spectrum at the time Priestley developed 'evolutionary spectral' analysis was that of the instantaneous spectrum [158]. While we do not review it in detail here, it has shares similar conceptual problems with evolutionary spectral analysis in that the spectrum is only well-defined in the limit of infinite samples.

[13] Priestley considers a somewhat more general family of "oscillatory functions" with nonlinear phase but justifies Eq. (VII.38) in terms of change of variables in integration ([169], p. 823-4.)

$$\phi_t(\omega) = A(t,\omega)e^{i\omega t}. \tag{VII.38}$$

He defines the evolutionary spectrum to be:

$$g(\omega,t) = dF_X(\omega)|A(t,\omega)|^2 \tag{VII.39}$$

The first main example in Priestley's original paper is ([167], p. 209):

$$x(t) = \int_{-\infty}^{\infty} S(t-u)h_t(u)du \tag{VII.40}$$

$$A(t,\omega) = \int_{-\infty}^{\infty} e^{i\omega u}h_t(u)du \tag{VII.41}$$

$$S(t) = \int_{-\infty}^{\infty} e^{i\omega t}dZ(\omega) \tag{VII.42}$$

so that S is a stationary process with spectrum $d\mu(\omega)$. The evolutionary spectrum $E(t,\omega)$ of x is given by:

$$E(t,\omega) = d\mu(\omega)|A(t,\omega)|^2 \tag{VII.43}$$

which follows since x is a convolution of S with h; for any fixed t, the Fourier transform diagonalizes the convolution operator and one can then use Parseval's Theorem at any fixed t to derive the result. The problem with this example is that the convolution operator is represented by a matrix which is now time-dependent so that it is not diagonalized by a Fourier transform.

Another example in Priestley's paper ([167], p. 210) illustrates the same point. Let y be a stationary process and define $x(t) = c(t)y(t)$ where it is assumed that $c(t)$ is slowly varying. It then follows that the evolutionary spectrum is $E(t,\omega) = |c(t)|^2 S_y(\omega)$ where $S_y(\omega)$ is the spectrum of y. Now, we suppose that $y(t)$ is a process with mainly high-frequency variation; specifically we let:

$$y(t) = -0.5\epsilon(t-1) + \epsilon(t). \tag{VII.44}$$

We define:

$$c(t) = \sum_{j=0}^{\infty} 0.8^j \epsilon(t-j) \tag{VII.45}$$

so that $c(t)$ is slowly varying and time-dependent. However, the spectrum of $x(t)$ is independent of time.

To see the technical problem with Priestley's representation, we replace the expression:

$$f(t) = \int \phi_t(\omega) dZ(\omega) \tag{VII.46}$$

with the finite sample size expression:

$$f(t) = \sum_i < f, e_i > g_i \tag{VII.47}$$

where $g_i = \phi_t(\omega)$ and e_i are complex exponentials. It is clear then that dZ is no longer the inner product of f with complex exponentials so that Eq. (VII.46) is no longer really an orthogonal increment representation for $f(t)$.

For dZ to be an orthogonal increment process for the data (and not for some phantom function), Eq. (VII.38) must be the eigenfunctions of the (self-adjoint) covariance operator for the time series. One alternative definition of an evolutionary spectrum would be to define an orthogonal increment process in terms of inner products of the data f with $\phi_t(\omega)$. The problem with this is that in the most likely cases where $A(t, \omega) = h(t)$, the logic of the approach is weakened by some recent mathematical results which show that orthonormal bases of the form:

$$e_{mn} = g(t-m)e^{2\pi i \omega n t} \tag{VII.48}$$

for $m, n \in \mathbb{Z}$ must have g with somewhat pathological properties either in the time or frequency domain. Thus, there does not exist a well-defined local Fourier expansion in terms of complex exponentials multiplied by a slowly varying function, For instance, g may be an indicator function in which case its Fourier transform oscillates much and decays slowly. The window function g may also be a sinc function: $g(x) = \frac{sin(\pi x)}{\pi x}$. Here, the window function decays slowly and oscillates quite a bit. Another function which has slightly better properties is developed in [109]. The relevant theorem is due to Balian [10] and Low [122]. Some technical corrections to the original proof were found by Coifman and Semmes and are reviewed in [53] [17]. The proof has subsequently been extended to certain nontrivial nonorthonormal bases. The statement of the theorem is:

Theorem 12 *Suppose $e_{m,n}$ $(m, n \in \mathbb{Z})$ where:*

$$e_{mn} = g(t-m)e^{2\pi i\omega nt} \tag{VII.49}$$

form an orthonormal basis of $\mathbf{L}^2$ on the real line. Either the window function is irregular in the time domain:

$$\int x^2 |g(x)|^2 \, dx = \infty \tag{VII.50}$$

or in the frequency domain:

$$\int \omega^2 |\hat{g}(\omega)|^2 \, d\omega = \infty \tag{VII.51}$$

What this theorem says essentially is that there does not exist a well-defined local Fourier basis which uses complex exponentials and a separable local time factor as is implicitly assumed by methods such as evolutionary spectral analysis and complex demodulation [198]; this theorem provided the motivation for investigating other types of local spectral analysis such as the wavelet transform and the wavelet packet transform. Since the theoretical covariance matrix is usually not known, methods such as the wavelet packet decomposition aim to provide a representation of the data in terms of an approximate orthogonal increment process. The point of this section is that since each independent orthonormal basis function has a representation in terms of a time-frequency distribution, an appropriate time-frequency representation of stochastic data is in terms of a weighted sum over these individual time-frequency components. As we have suggested, Mallat and Zhang [131] have made the same point for the time-frequency representation of deterministic functions in terms of *nonorthogonal* waveforms.

Parametric Methods

The disadvantage of Fourier analysis for the analysis of economic time series is that it is nonparametric; economic time series data are frequently limited in length so that parametric methods are often more useful. Estimates of the spectrum can be obtained from the parametric models. If we consider the univariate ARMA model:

$$B(L)X(t) = A(L)u(t) \tag{VII.52}$$

where L is a lag operator. If $u(t)$ is white noise the spectrum of the data X is:

$$f_{XX}(\omega) = \frac{|A(e^{i\omega})|^2}{|B(e^{i\omega})|^2}\frac{\sigma^2}{2\pi} \tag{VII.53}$$

Estimates of the coefficients in the lag polynomials A and B hence lead to parametric spectral estimates. We have already defined a notion of time-dependent spectra in terms of sums of Cohen's class distributions of orthonormal basis functions. We have yet to describe how such a definition carries over to the parametric case when coefficients are time-varying.

One possible approach is to compute a parametric spectral estimate for each point in time [195]; this is, however, inconsistent as there is no such thing as a fully local spectrum. However, it is the case that parametric estimates define implicitly an estimated autocovariance matrix. The spectrum is then the weighted sum of Cohen's class distributions of the eigenvectors; the eigenvalues provide the weights. This approach has two advantages: it provides a natural representation of the data in terms of an orthogonal increment process and it concentrates the maximal spectral energy in the largest component.

Once the form of the autoregressive model is known, we can compute time-dependent spectra from the implied covariance matrix. However, estimation of the autoregressive model or the covariance matrix in the nonstationary case is far from straightforward and we have argued that nonparametric estimation of such a model might be difficult. One intuitive reason for this is that analysis of nonstationary time series present many of the same problems as short time series in that nonstationarity implies that the available data is essentially local and effective sample sizes are much shorter than if the data had time-invariant properties. Therefore, parametric methods which impose structure on the data are often even more necessary for nonstationary time series than for stationary time series.

If we knew exactly how the data evolved over time or if we knew an appropriate form of the autoregressive model, we could let the lag coefficients depend explicitly on t and use ordinary least squares. For instance, suppose we believed that the data generating process was a first order autoregressive process but with time-varying coefficients which evolved according to some function $g(t)$. We then could estimate the parameter β in the equation:

$$y(t) = \beta g(t) y(t-1) + \epsilon(t) \qquad \text{(VII.54)}$$

by ordinary least squares. However, we ordinarily do not know the function $g(t)$. One possibility (which in fact we experiment with in Ch. III) would be to estimate the function $g(t)$ nonparametrically:

$$\hat{\gamma} = \hat{\beta}\hat{g}(t) = \frac{\sum_s w_1(t-s)y(s)y(s-1)}{\sum_s w_2(t-s)y(s-1)^2} \qquad \text{(VII.55)}$$

where w_1 and w_2 are kernels. However, while such an approach can capture major modelling errors, the results are heavily influenced by choice of bandwidth and often it is necessary to impose more structure on the data. A related approach to estimation of time-varying parameters would be to discount past observations in estimating the least squares coefficient at every point in time; such a procedure could be performed recursively.

Another way to impose more structure on the data is to make the coefficients in a model state-dependent. Thus, the function $g(t)$ may be replaced by some function $h(x)$ where x is a vector of state variables. Such an approach to time series modelling is we believe correctly labelled by Priestley [170] as non-linear. Indeed, major classes of non-linear time series models such as the bilinear times series model, the exponential autoregressive model and the threshold time series model are *special cases* of the state-dependent model developed in [168]. These models usually can be estimated by maximum likelihood methods, so that provided the parameters are identified, confidence intervals can be calculated. However, since the models are nonlinear and state-dependent, the interpretation of time-frequency spectra which might be implied by such models is somewhat murky.

One interesting example of a state-dependent model is that of the hidden Markov model [173] [91] [172] in which there is a Markovian transition density between a finite number of states. In each state, there is a conditional distribution of the observable given the state variable. Such a model can be estimated by maximum likelihood methods. Some economic applications have been to business cycle dating [91] and exchange rates [68].

In the state-dependent model, parameters depend in a known way on states so that for instance the first autoregressive coefficient may follow a random walk in time. A state-dependent model may be identified by first estimating (using projection pursuit regression or average derivatives) by nonparametric methods the functional dependence of a local kernel estimate of the first autoregressive coefficients on states x.

The most relevant state dependent models for the work in the thesis are models of the form:

$$y(n) = \beta(n)y(n-1) + \epsilon(n) \qquad \text{(VII.56)}$$

$$\beta(n) = \lambda\beta(n-1) + B(L)\eta(n) \qquad \text{(VII.57)}$$

where $B(L)$ is a lag operator and $\eta(n)$ is noise. It is straightforward to generalize Eq. (VII.56) and Eq. (VII.57) to the case where many lags influence the current value $y(n)$. There is a substantial literature in statistics and econometrics about such random coefficient models especially where the regressor values such as $y(n-1)$ are treated as fixed.[14]. Some important results include:

- Random coefficient models can be rewritten as regression models with heteroskedastic disturbances [155] [48] or treated by state space methods; in general, the two approaches are equivalent [155] (in terms of the implied likelihood function).

- There has been a study of the stability of the univariate autoregressive $AR(1)$ model when the autoregressive parameter itself follows an autoregressive process [205].

- Asymptotic properties of estimates, including efficiency, consistency and asymptotic normality, are studied in [48] [155]. A deficiency is that the theorems on

[14] Examples of economic applications of random coefficient models include [176] [117] [199] [177]

efficiency and consistency treat the regressors $y(n-1)$ as nonstochastic.

- There has been extensive study of the case in which $\beta(n) = \beta_0 + \epsilon$ for some random ϵ [148].

In general, it is known that [21] adaptive estimates of the parameter β tend to follow:

$$\beta(n) = \beta(n-1) + K_n H(\beta(n-1), y(n-1)) \tag{VII.58}$$

where K_n is a "gain" factor, H is an updating rule and y is the data. The updating rules for β may also depend strongly on priors. The updating rule for $\beta(n)$ can be, for instance, examined using dynamical systems theory [21].

The key aspect of the state dependent approach is an hypothesized model for the dependence of parameters on states. This parametric restriction, sometimes coupled with Bayesian priors (c.f., [60]), provides the means to effectively capture many types of 'time-varying' parameters.

The unifying concept in our literature survey is the issue of time-frequency spectral estimation. When assessing state dependent models, the issue is whether they produce time-dependent spectra which are nonstationary. For the purpose of comparisons in the thesis, we will largely treat realizations of stochastically time-varying parameters as deterministic functions. Technically, to handle such models in our framework, we would need to consider characterization of the spectral properties of nonlinear operators.[15]

[15] Extensions of definitions of time-varying spectra to nonlinear frameworks such as ARCH models are interesting topics for future research.

Where Autoregressive Pursuit Fits In

In the thesis, we do not develop further nonparametric methods and we do not work with standard parametric methods such as state-dependent models. Rather, we develop a new approach based on the idea of selecting a best parametric model from a large family of potential parametric models.

Nonparametric methods have the disadvantage of weak performance relative to parametric methods when the true model is close to the hypothesized parametric model. In addition, our results in the thesis indicate that nonparametric time series analysis methods such as the wavelet packet transform seem to achieve uneven results in terms of capturing the time varying properties of stochastic data.

On the other hand, parametric methods lack flexibility. If we wish to understand the time-varying nature of economic activity, parametric methods make strong *a priori* assumptions which may be unreasonable given the data. Our results in the thesis indicate that often time-variation in parameters can lead to serious biases in estimates of autoregressive parameters. Furthermore, it is usual in time series research to try many different potential parametric models; this is, for instance, the idea behind the use of model identification criteria such as the Akaike Information Criterion [1]. Such an approach suffers from pre-test selection bias.

We now summarize the new parametric approach to nonstationary spectral estimation which follows from the thesis:

- Compute estimates of an autoregressive or other time series model using Autoregressive Pursuit.

- Compute the implied covariance matrix for the time series.

- Compute the Wigner-Ville or Cohen's class distribution of the covariance matrix. Alternatively, compute the Wigner-Ville or Cohen's class distribution of the implied eigenvectors (or, for continuous time series, eigenfunctions) of the covariance matrix (or, for continuous time series, covariance operator).

Appendix G proves some results relating to the implementation of nonstationary spectral estimation with Autoregressive Pursuit.

CHAPTER VIII

ECONOMIC EXAMPLES

To show that the univariate autoregressive techniques are practical, we illustrate their usefulness with some financial and macroeconomic data and show how the results differ from more standard approaches. In terms of macroeconomic data, we examine GNP data. In terms of financial data, we examine the behavior of the Standard and Poor's daily returns.

Financial Data

We examine Standard and Poor's daily returns data from July, 1962 up to Dec. 1992 or 7675 trading days. Statistics for the data are in Table (VIII.1). We note very strong excess kurtosis, something which led earlier researchers such as Mandelbrot [133] to suggest that stock returns were drawn from distributions with thick tails.[1]

To examine the local properties of the correlations in the data, we consider kernel estimators of local autocorrelations such as we used in Ch. III for our examples. In Fig. (VIII.2) we show rolling autoregressive estimates of the one day autocorrelation where we compute estimates for 400 days at a time. In Fig. (VIII.3) we show

[1] Most of the kurtosis comes from the 1987 stock market crash. When we truncate the sample on the day before the crash, kurtosis drops to 2.54 and when we eliminate the crash; if we just drop the day of the crash from the sample, kurtosis does not drop so much.

Mean	0.000307058
S.Dev.	0.00882708
Skewness	-1.56616
Kurtosis	42.8138

Table VIII.1: Statistics for the Standard and Poor's 500 index 1962-1992.

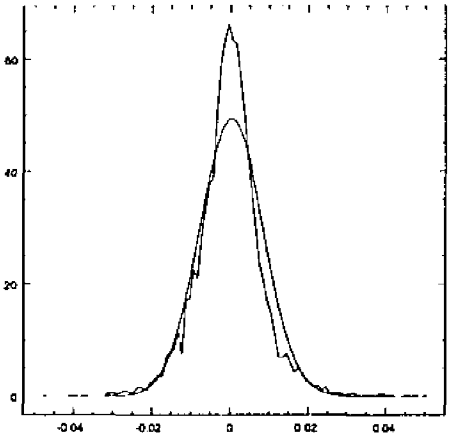

Figure VIII.1: Return density for pre-1987 crash stock market compared with normal distribution with estimated parameters.

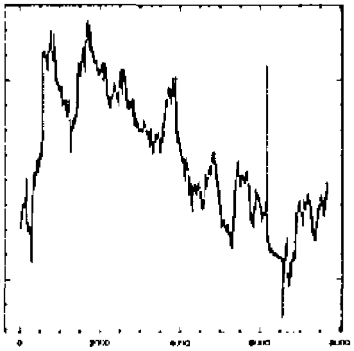

Figure VIII.2: Rolling estimates of autocorrelations for stock returns.

rolling estimates of one day autocorrelations with a Gaussian window with bandwidth $\sigma = 200$.[2] We note that there is apparent random walk behavior in Fig. (VIII.2); the autoregressive parameter appears to have a unit root and first differences of the autoregressive parameter do not seem to have nonzero autocorrelations. In Appendix E, we show that the theoretical autocorrelations for the time difference of autoregressive parameters for stationary $AR(1)$ processes converge to zero as the size of the window over which rolling autoregressions are computed goes to infinity.[3]

First and second order autoregressive model estimates for the data are:

$$\begin{aligned} \hat{y}(t) &= 0.000267914 + 0.127097y(t-1) \\ &\quad (0.0000999) \qquad (0.0113226) \end{aligned} \tag{VIII.1}$$

$$\hat{y}(t) = 0.000278903 + 0.132298y(t-1) - 0.0409129y(t-2)$$

[2] We show 5% confidence intervals which were computed by: (1) using a Gaussian window with bandwidth $\sigma = 200$ on the residual to estimate the variance of the disturbance; (2) assuming an effective bandwidth of $T = 400$. For Fig. (VIII.2) confidence intervals computed by such a method are of similar width but are quite jagged and hence are not shown.

[3] The limiting distribution is *not* a Brownian motion because the lack of autocorrelations only occurs at lags less than the length of the window.

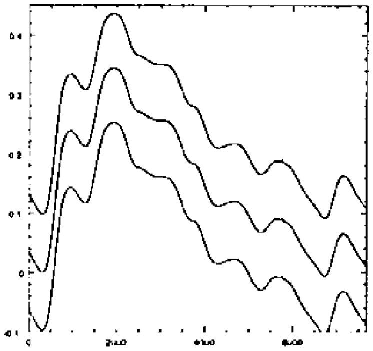

Figure VIII.3: Rolling estimates of autocorrelations for stock returns with Gaussian windows.

$$(0.0000996) \qquad (0.0114056) \qquad (0.0114056) \tag{VIII.2}$$

With the standard approach one would add lag lengths until a stopping criterion such as the Akaike Information Criterion suggested an optimal lag length and then the method would be used for prediction and the like. In the case of the stock market data, after the second lag, most of the lag estimates with small lags do not appear statistically significant. One would then stop the analysis after the second lag. In macroeconomic data, in particular, this is problematic because one might want to include the effects of seasonals and annual effects and small lags might be insignificant.[4]

One simple application of the method proposed here is simply to select the optimal model components to include from the class of all possible Box-Jenkins AR models. We consider a set of potential model components all of which have constant windows which span the whole length of the time series data. We consider 1200 different lags and note the order of the selected model components. The selected model

[4] This point is illustrated well in a paper by A. Hall [88] on monthly data for inventories (in the stock market example, the corresponding point could be made with respect to mean reversion at long horizons). Conventional test criteria often select small lag lengths and thus result in ignoring effects at annual frequencies (lag 12). The advantage of adding a maximal model component at each iteration is that overparameterization is avoided.

Iteration	*Lag Selected*	*Coefficient*	*Std. Error*
1	1	0.127097	0.0113
2	746	-0.049727	0.0118
3	59	0.0407214	0.0113
4	2	-0.0410548	0.0114
5	158	-0.0395326	0.0113
6	32	0.0364563	0.0112846
7	1129	-0.037934	0.01211
8	55	-0.0335363	0.0112783
9	116	-0.0330502	0.0112924
10	26	-0.0320167	0.01127

Table VIII.2: Selected global model components for the Standard and Poor's 500 index 1962-1992.

components are shown in Table (VIII.2).[5] The results show several interesting things. First, the long horizon coefficients are all negative (though weak) and suggestive of mean reversion. Second, the coefficients appear to decay very slowly as more model components are added which suggests some problems with the model. Indeed, ten model components only capture 2.9% of the sample variance of time series which suggests that models with constant coefficients are not very useful with this type of data; by comparison a single model component for a first order autoregressive model with autoregressive parameter 0.9 picks up an average 81% of the sample variance of the data. Still, all the ten model components selected are statistically significant by traditional criteria. Since robustness is an issue here because we have used many model components, we examine a decomposition for an autoregressive process with

[5] Coefficient estimates and standard errors are based on the regression coefficient at the stage at which the model component is added.

v*Iteration*	*Lag Selected*	*Coefficient*	*Std. Err*
1	1	0.201944	0.01118
2	888	0.0433596	0.01117
3	800	0.0388525	0.0117475
4	241	-0.0335004	0.01131
5	926	0.0355094	0.0118
6	6	-0.0321441	0.01115
7	971	0.0341302	0.011876
8	1056	0.03283245	0.0119313
9	775	0.0318119	0.0117014
10	186	-0.0304866	0.0112488

Table VIII.3: Selected global model components for autoregressive control series.

autoregressive parameter 0.2 and exactly the same sample size. The corresponding table to Table (VIII.2) for the autoregressive model is Table (VIII.3). Table (VIII.3) suggests to us that the first model component for the stock market is statistically significant but the others are unlikely to be.

The model estimated after two model components is:[6]

$$\begin{aligned} \hat{y}(t) &= 0.000281631 + 0.126299\,y(t-1) - 0.0497277\,y(t-746) \\ & \quad (0.0000993) \qquad (0.0113112) \qquad (0.0118563) \end{aligned} \tag{VIII.3}$$

While this model does not fit the data well, it nevertheless illustrates how our methods can be effectively used to select proper models in situations where it would be

[6] We note that the way the regression is calculated, we effectively assume that $y(t) = 0$ for $t < 0$. Thus, we do not drop the first 746 observations.

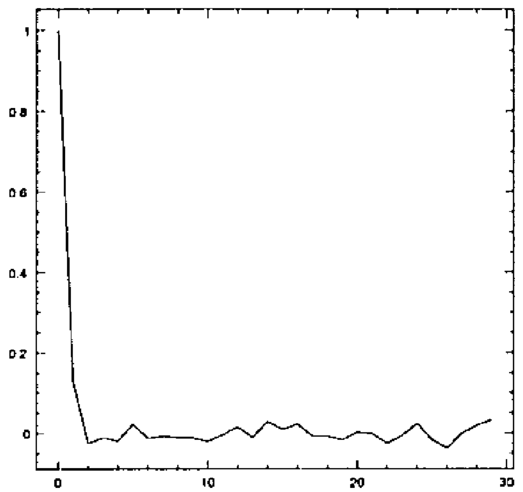

Figure VIII.4: Correlations of Standard and Poor's 500 daily returns (CRSP) from 1962-1992.

impractical to experiment with all possibilities. That global windows are unlikely to be of much use can be seen from the correlation functions of the data shown in Figure (VIII.4). The first autocorrelation is 0.127 and most of the other autocorrelations are close to zero; the autocorrelations before and after the decomposition are thus virtually the same as shown in Figure (VIII.5).

Significant improvement occurs with use of a wider variety of window sizes. Since inclusion of all lags up to 1200 lags would be excessive we subsample after the first fifteen lags and include only every fifteenth additional lag. Consideration of 12 levels of window functions with some subsampling at each layer results in a total of approximately $974,000$ potential model components. In this case the percentage of sample variance explained by the first model component is nearly four times as great as from the largest global component but the dependence is still quite weak. In fact, with 10 included lags ten model components altogether pick up only 14.1% of the sample variance in the data, which is substantially better than the case with only global model components but is still somewhat weak. One striking result is that none

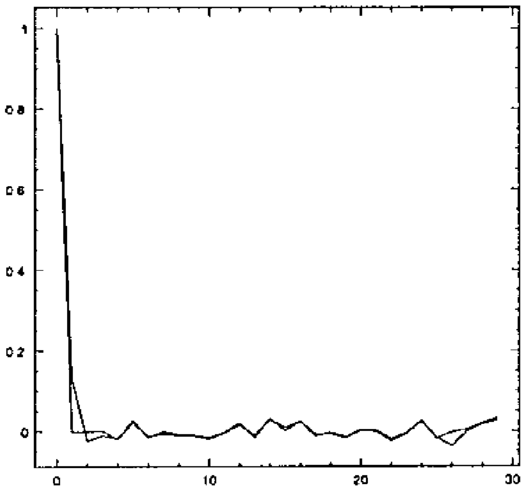

Figure VIII.5: Correlations of Standard and Poor's 500 daily returns (CRSP) from 1962-1992 compared with correlations from residuals after subtracting off first ten projections against global functions.

of the selected model components contains a global window function which suggests the estimates in Equations (VIII.1), (VIII.2) and (VIII.3) are spurious as indeed is suggested by economic theory. The selected components are shown in Table (VIII.4). We note that many of the most strong effects are concentrated around the time of the 1987 stock market and most of the detected relations are on very short time scales.

With flat windows with smooth Gaussian edges, we see that estimates of lag relationships focus on the 1987 stock market crash and otherwise dependence in the data is quite weak but not negligible at some lags. The lag one and lag two coefficients as a function of time are shown in Figures (VIII.8) and (VIII.9) respectively. In addition, even after subtracting off ten projections the data appears not yet to be statistically homogeneous. In Ch. VI, we have developed a stopping rule based on the cumulative sum of orthogonal wavelet coefficients; we call this test the cumulative waveletgram test. We show a cumulative waveletgram test for the data in Figure (VIII.10). This test suggests that there may exist other model components

Iteration	*Lag Selected*	*Begin*	*End*	*Coefficient*	*Estimate*
1	1	6344	6359	0.1602	3.88981
2	1	505	4342	0.1187	0.248673
3	6	6334	6364	0.1144	0.637228
4	1	6342	6357	-0.0770	-3.42743
5	2	6331	6361	-0.0882	-0.430576
6	3	6363	6423	-0.0606	-0.327588
7	5	3039	3069	-0.0528	-0.529741
8	13	6302	6422	-0.0446	-0.152358
9	3	5031	5091	0.0437	0.416004
10	1	6350	6365	-0.0423	-0.178243

Table VIII.4: Selected model components with constant windows for the Standard and Poor's 500 index 1962-1992.

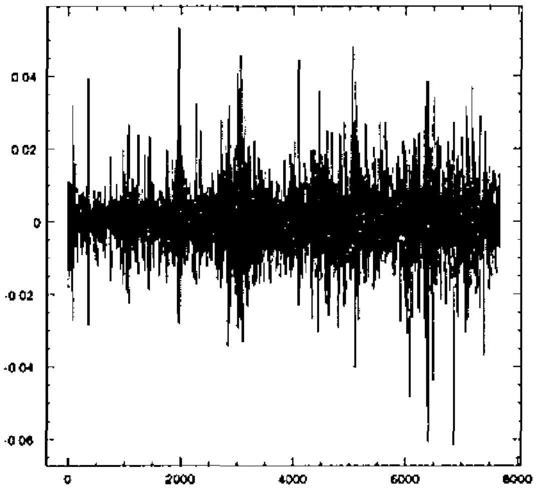

Figure VIII.6: Standard and Poor's 500 daily returns (CRSP) from 1962-1992 after subtracting off first ten projections against model functions with flat windows.

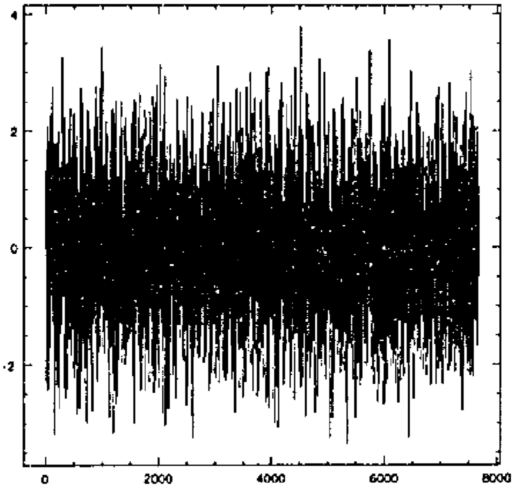

Figure VIII.7: A realization of Gaussian white noise of length 7675. By eye, the residual after subtracting off the first ten projections against model functions appears to have much different stochastic properties.

not included in the analysis which provide a more parsimonious description of the data.

U.S. GNP Data

We next analyze the Citibase *GNPQ* output series based on constant 1987 dollars. The quarterly data runs from 1947 : 1 to 1992 : 4. The raw data are shown in Figure (VIII.11) and the log differeneced data are shown in Figure (VIII.12). In the analysis we shall use the log differenced data after subtracting out the mean.

With constant filters including up to 17 lags and 9 levels of constant filters, the program creates 1437 potential model components. From these model components, a global model component with lag 1 and coefficient picks up about 15% of the sample variance in the data; using a cumulative waveletgram stopping rule, we thus find that GNP is likely to be a stationary process after differencing, though there are some moderate signs of excess kurtosis (1.35). The cumulative waveletgram and 5%

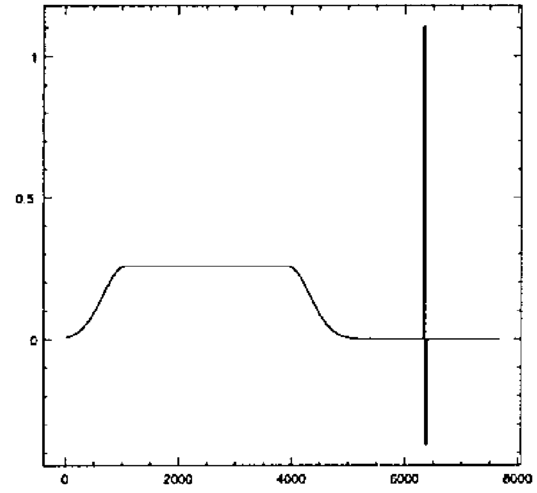

Figure VIII.8: Estimates of β_1 for the CRSP daily data based on model components which include flat windows with smoothed Gaussian edges.

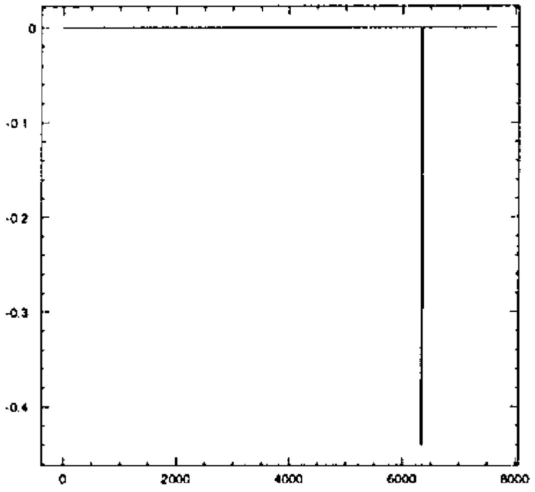

Figure VIII.9: Estimates of β_2 for the CRSP daily data based on use of flat windows with smoothed Gaussian edges.

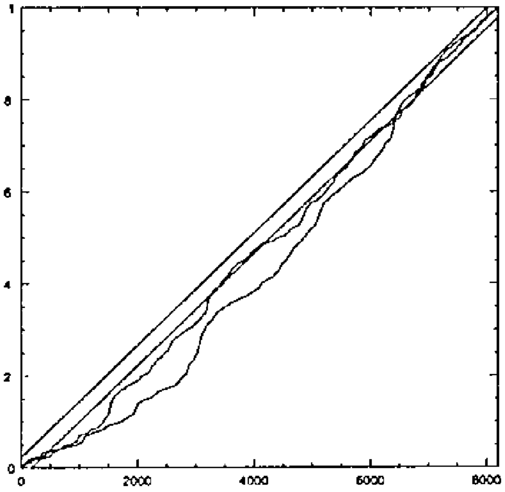

Figure VIII.10: Cumulative waveletgram for residual after subtracting off 10 projections on stock market data and the associated 5% confidence intervals. The test easily rejects randomness at a 5% level, indicating that even after accounting for ten model components, there is still additional information in the data.

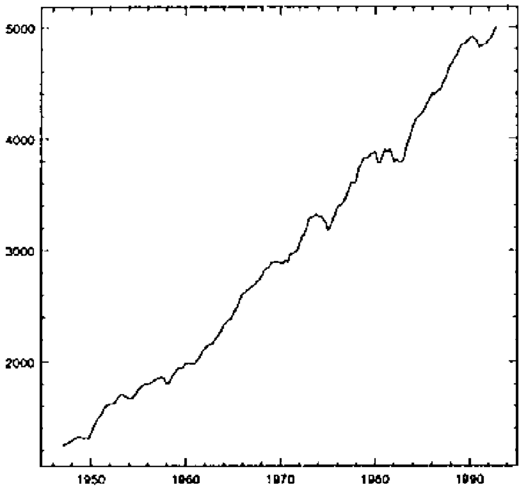

Figure VIII.11: Real GNP series 1947-1992 from CITIBASE *GNPQ* series. Vertical axis is Real GNP in billions of dollars.

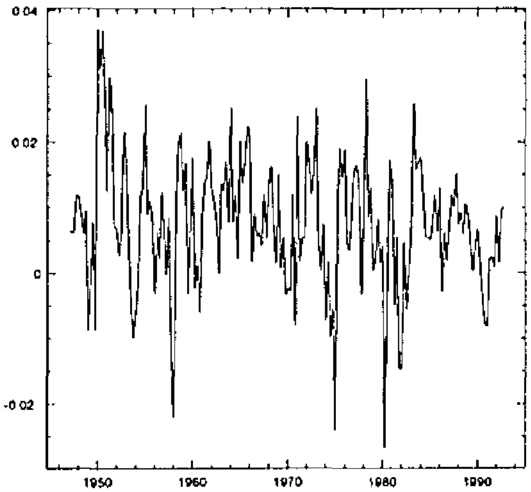

Figure VIII.12: Log differenced GNP data based on CITIBASE *GNPQ* series.

confidence intervals are shown in Figure (VIII.15). In addition, we note that when flat windows with smooth Gaussian edges are introduced, the method continues to pick constant filters, suggesting stationarity.

Somewhat less clear results are obtained through use of the more general adaptive waveletgram test which is discussed in detail in [149]. The adaptive waveletgram utilizes the best orthonormal basis decomposition of Coifman, Meyer and Wickerhauser [47] reviewed in Ch. VII to search for a large variety of deviations from randomness. The adaptive waveletgram test also fails to reject the null hypothesis of randomness at a 5% level but the rejection is somewhat less unambiguous. The adaptive waveletgram is shown in Figure (VIII.16). The probability that white noise would go out of the interval spanned by the position ordered path in the adaptive waveletgram computationally is about 15% so there may be some additional uncaptured nonstationary components in the data.

The second selected model component is a lag two model component with a constant window between 1963 : 4 and 1975 : 1. The coefficient on this model element is

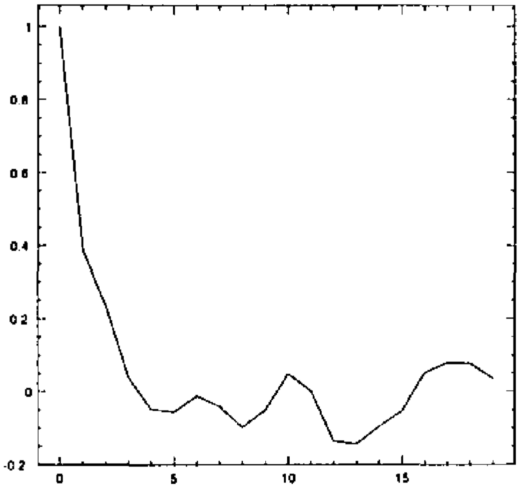

Figure VIII.13: Correlation function of log differenced GNP.

0.469895 whereas the coefficient on the first model element reduces to 0.367467. This suggests an alternative time series model for GNP:[7]

$$\begin{aligned}\hat{y}(t) &= 0.367y(t-1) + 0.4699y(t-2)1_{[1963:4,1975:1]}(t) \\ &\quad\; (0.066) \qquad (0.152884) \qquad \text{(VIII.4)}\end{aligned}$$

which involves a single structural break.

Summary

In this section, we have applied our method to GNP data and stock market data. We find that GNP growth rates appear close to stationary whereas stock market data appears quite nonstationary and requires a large number of model components to describe well. That GNP growth rates appear close to stationary would appear to provide some empirical support for Robert Lucas' idea [124] that all business cycles

[7] The mean growth rate of *GNP* was subtracted off before analysis.

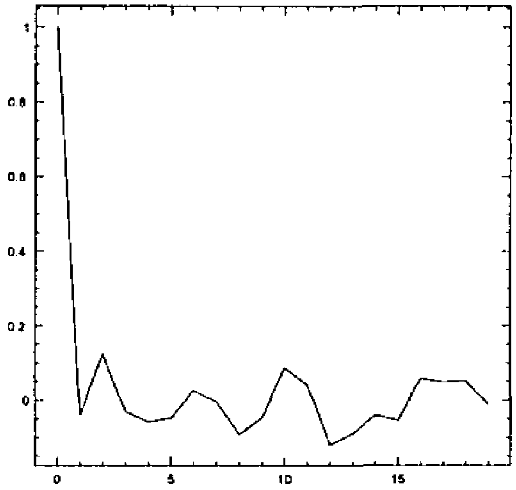

Figure VIII.14: Correlation function of residual from log differenced GNP after subtracting off first projection.

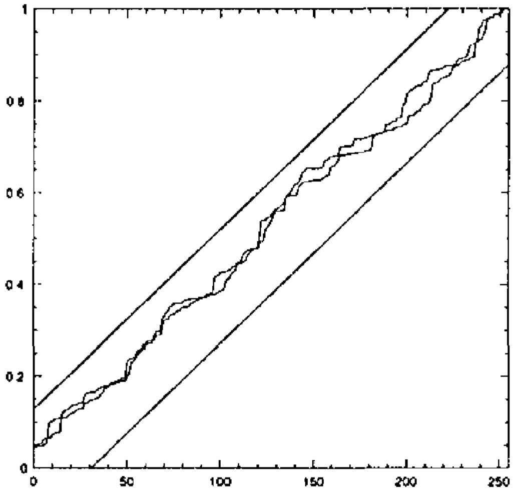

Figure VIII.15: Cumulative waveletgram of residual after subtracting off first projection from log differenced GNP. 5% confidence bounds are shown for two summation paths. A Daubechies $D8$ wavelet was used.

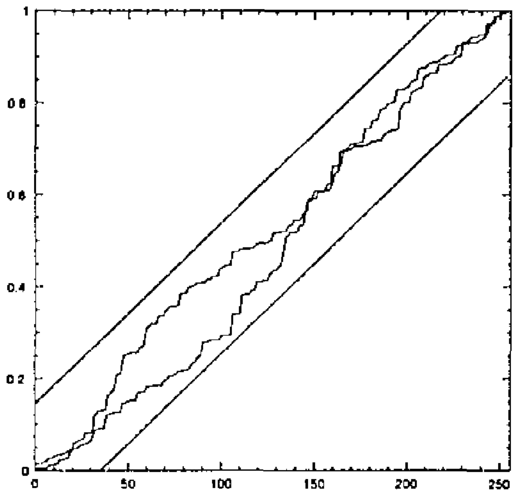

Figure VIII.16: Adaptive waveletgram of residual after subtracting off first projection from log differenced GNP. 5% confidence bounds are shown for two summation paths. A Daubechies $D8$ wavelet was used.

are in some sense the same and that macroeconomic time series are at some level well-described by stationary stochastic processes.

CHAPTER IX

IMPLICATIONS AND EXTENSIONS

Since the thesis introduces a general approach to modeling and measuring nonstationarity in economics, it opens up a number of new research possibilities. The purpose of this chapter is to survey some natural questions raised as a result of the thesis work and some of the issues involved. One point we wish to emphasize is that nonstationarity may have nontrivial implications for *economic theory* and how economic behavior should be modeled. After discussing this point in detail, we turn to econometric issues related to the dissertation.

Nonstationarity and Economic Theory

Following the work of Frisch [78] and Slutsky [188] on impulse propagation mechanisms, most macroeconometric research has proceeded on the grounds that variables follow linear constant coefficient stochastic difference equations. According to Blanchard and Fischer, "The integration of empirical and theoretical work on fluctuations, through the common use of the impulse-propagation mechanism framework and its associated time series implications, is certainly one of the most important achievements of postwar macroeconomics" ([26], p. 28). The impulse propagation mechanism is also a key component of so-called rational expectations econometrics and the related economic theory. Given the wide use of linear time series models in macroeconomics and finance as well as the close integration with economic theory, a natural question to ask is: "does nonstationarity fit into this conceptual framework"?

To answer this question, we recall that one reason we have given for studying time-varying parameters is that the parameters of the models may depend in some unknown way on some state variables $x(t)$ for the system. A natural extension of the model considered in the thesis thus is:

$$y(t) = \sum_{j=1}^{\infty} \beta_j (x(t))\, y(t-j) + \epsilon(t). \tag{IX.1}$$

As we have noted in Ch. I, the method in the thesis is a special case of Eq. (IX.1) where time is the only state variable entering the functions β_j.

At the core of the conceptual framework of macroeconomics and finance is an assumption of rational expectations. Agents are assumed to have a subjective probability distribution for all economic variables and these subjective probability distributions are assumed to be equal to the corresponding objective probability distribution. Therefore, if the true model is defined by Eq. (IX.1), agents might maximize lifetime utility while treating Eq. (IX.1) as exogeneous. Generalized method of moments estimation methods [97] could then be used to test overidentifying restrictions implied by the equilibrium Euler equations of motion.

One reason we may find nonstationarities in economic data is that we have not included enough state variables in the analysis or have otherwise excluded information known to agents in the economy. Agents then would form a prior as to the true form of Eq. (IX.1) and make decisions accordingly. To quote Robert Lucas [124] , "At a purely formal level, we know that a rational agent must formulate a subjective joint probability distribution over all unknown probability distributions which impinge on his present and future market opportunities."

An alternative perspective due to Frank Knight [116] is that agents really do not know perfectly the model of the economy. Parameters thus fluctuate for reasons they partially understand but not fully. This perspective seems more reasonable in that empirically agents in the economy disagree significantly in their predictions about future events. At a minimum, these differences in forecasts lead one to question whether agents use expected values in making decisions and whether agents learn as quickly and as efficiently as is supposed by the theory. Knight drew a distinction between risk and uncertainty and defined uncertainty as cases in which agents have no exact knowledge of the probabilities or the risks that they face.

The reason agents face uncertainty is that the environment is nonstationary and agents have not had time to learn the underlying objective distribution. Knight writes: "It is necessary to stipulate that the fluctuations must be of sufficient extent and irregularity that they do not cancel out or reduce to uniformity or regular periodicity in a time-interval short in comparison with the length of human life" [[116], p. 38]. It follows therefore that progress and change are essential in determining institutional structures:

> "In an unprogressive society knowledge of the future could be perfected to a high degree through actual forecast and control or the effect of certainty secured through the grouping of cases and application of probability reasoning. Under such conditions the problem of management would be indefinitely simplified as activity would follow in the main an established routine and *real* decisions would rarely be required. The actual form of economic control, free contract, and especially private property in material goods, is closely connected with the acute form of the problem of management which arises from the highly "dynamic" character of the society we live in and the extreme degree of uncertainty associated with change. Before the modern industrial era began, as we know, the economic life of Europe was unprogressive and its organization of control was collectivistic..."[[116], p. 370].

Thus, it would seem that the rationale for competitive economic systems is at odds with the basic assumptions of rational expectations economics and the Frisch-Slutsky propagation mechanism.

In regard to Knightian uncertainty, Lucas has dispensed with it by simply saying, "In cases of uncertainty, economic reasoning will be of no value" [124]. Implicit in Lucas' claim is that probabilities commute and that quantification of uncertainty is impossible. One reason that Knight found uncertainty interesting is: "If conditions are subject to unpredictable fluctuations, ignorance of the future will result in the same way and inaccuracies in the competitive adjustment and profits will be the inevitable consequence" [[116], p. 38]. If one wants to establish an approach to macroeconomics based on the theory of competitive general equilibrium, it would be convenient if one could dismiss a major economic factor which leads the theory to break down in practice as being of no economic interest or meaning.

Nonstationarity introduces the modeling question of how to model situations in which information diffuses and learning occurs at a slower rate than the rate at which economic decisions are made. While such situations are opposite to those assumed by rational expectations economics, researchers in macroeconomics and finance are increasingly turning to models of learning and heterogeneous information as a means of understanding empirical phenomena.[1] At stake are a wide range of theoretical results including results on optimal monetary policy, optimal investment policy, optimal capital accumulation in a stochastic Ramsey or Real Business Cycle model, precautionary savings, natural resource management and the Modigliani-Miller Theorem [140] on the invariance of the value of the firm to its capital structure as well as the pre-Ricardian equivalence theorem [12].

[1] For instance, a recent book of Sargent surveys recent work on bounded rationality and learning in economic theory [183]. A growing literature focuses on heterogeneous information and learning in financial markets (c.f., [23])

We now review some new models of invidual and aggregate behavior in an environment of nonstationarity and raise some questions for future research as to how nonstationarity enters these models.

Models of Individual Behavior

We consider the autoregressive model:

$$y(t) = \sum_{j=1}^{\infty} \beta_j(q) y(t-j) + \epsilon(t) \tag{IX.2}$$

where $q \in \mathbf{R}^d$ are a set of parameters which determine the autoregressive coefficients β_j. The standard approach assumes that agents in the economy know the true (population) value of q, q_0. If agents believe the model to be uncertain, they might, for instance, also take into account utility for values of q which they believe might occur.

To illustrate some of the ideas, consider Fig. (IX.1). We assume that the underlying model is characterized by two parameters, β_1 and β_2 (which may for instance be the autoregressive parameters which enter Eq. (IX.2) for particular values of q). Agents might have a subjective belief that the point in the center of Fig. (IX.1) represents the parameter values of the true model. However, because there is a nonstationary environment, agents believe they also may have picked the wrong model and hence wish to behave in a way that insures high utility for all the parameter values in the shaded box in Fig. (IX.1). For instance, agents might choose to maximize the minimum level of utility. They might instead want to insure a certain minimum level of utility when certain model errors occur and treat this minimum level as a constraint.

While one can analyze problems of robustness and parameter uncertainty, it seems that, if agents do not know the model, it is more appropriate for them to be unsure about the overall impulse response than specific parameter values which may have no inherent economic meaning and which, in addition, may have highly nonlinear effects

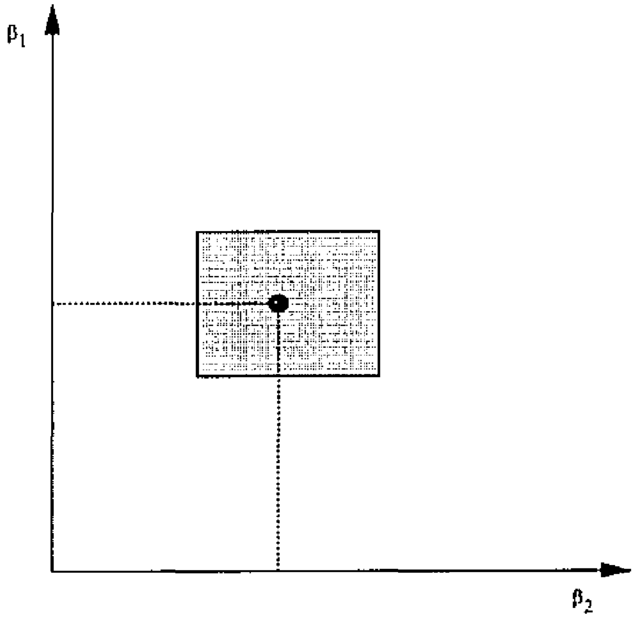

Figure IX.1: Parameter uncertainty.

on the response. An approach to model uncertainty focusing on impulse responses also seems closer conceptually to the ideas behind use of the Frisch-Slutsky impulse propagation mechanism.

In Figure IX we show the impulse response function of a nominal model to a shock (center line). In the standard approach we would assign probability distributions to the deviations of the observed reponse from the true model (the center line) and optimize by considering a weighted average of the objective function under different possible paths. With Knightian Uncertainty we do not know the probability distribution of disturbances so that instead we consider a ball of models which lie within the top line and the bottom line in Figure (IX). We then optimize against the worst possible model in that ball.

When the probability distribution of disturbances is known, it is possible to write down an explicit evolutionary equation for state variables such as:

$$dx = \alpha(x,t)dt + \sigma(x,t)dW \tag{IX.3}$$

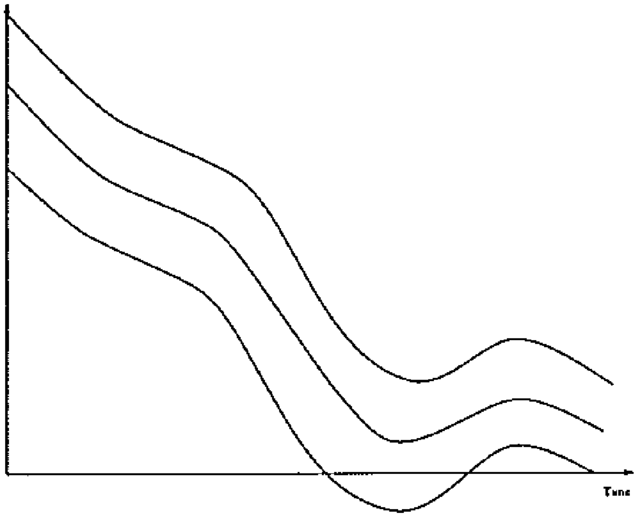

Figure IX.2: A ball around the true model.

In comparison with this parametric model, with Knightian uncertainty all that is known is that:

$$dx = K(x,t)x + Dx + R \tag{IX.4}$$

where $K(x,t)$ includes the effect of control and D and R are disturbances which can lie in balls:

$$||Dx|| < d \tag{IX.5}$$

for multiplicative noise (d is fixed) and:

$$||R|| < d' \tag{IX.6}$$

for additive noise (d' is fixed). That D and R lie in some function space ball may mean they have a dynamic structure different from that associated with the nominal model Eq. (IX.3). Since D is characterized by a function space rather than a finite number of parameters, Eq. (IX.4) involves a nonparametric treatment of uncertainty. To optimize, agents might weight different aspects of additive and multiplicative noise (such as frequency characteristics) and solve a problem of determining the best utility obtainable in the worst case scenario.

Optimization problems of this type fall under the rubric of a new type of control theory called $\mathbf{H}^\infty$ control[2], which because of its emphasis on impulse responses, has natural connections with both linear time series models and the impulse propagation mechanism. Basic references on $\mathbf{H}^\infty$ control and recent results in robust control theory include [63] [212] [204] [11] [65] [16] [201] (see [143], Pt. I, Sec. III for reviews of syntheses of adaptive and robust control); in joint work with Hong Yang [154], we have applied $\mathbf{H}^\infty$ control to the problem of optimal portfolio choice. One important point to note is that minimax control problems of this type result in nonlinear calculations which differ considerably from the linear projections used in the standard approach of setting prices equal to risk-adjusted *expected* values and in theories such as the Capital Asset Pricing Model (CAPM) [187] [120].

Since Eq. (IX.3) and Eq. (IX.4) are so different in nature, there are perhaps benefits from an intermediate semiparametric approach to uncertainty. Finally, we note that more conventional approaches to uncertainty, which consider parameter uncertainties, may be of interest in characterizing the factors which influence how economic agents respond in a nonstationary environment and how predictions differ from the standard theory which makes an implicit stationarity assumption.

Models of Aggregate Behavior

We recall the rational expectations model of Lucas [125] which considers a world in which agents need to separate local and global information. In such a setting, firms observe global variables such as price and aggregate quantities lagged one period but there are also certain local effects. These local effects represent *heterogeneous*

[2] $\mathbf{H}^\infty$ is a Hardy space norm. Hardy spaces represent causal functions like impulse responses.

information, which we have argued is an important aspect of nonstationary economic systems in which agents make decisions at a much faster rate than they learn the details of the model. In this section, we shall discuss how such local effects lead to a "field theoretic" approach to aggregate modeling analogous to that used in modeling multi-body systems in physics.

In this section, we consider a simple model in which firms also observe the quantity decisions of firms similar to them. For instance, in the auto industry, profits will depend on output in a similar industry such as the tire industry and hence there are particularly strong incentives for executives in the auto industry to collect accurate information on outputs in the tire industry and vice versa.

We consider a model where firms are located on a lattice and profits depend on the outputs of nearby firms as well as global variables such as price. We consider a lattice $\mathcal{L}$ and define the set of neighbors of a firm located at lattice site i as $N(i)$. A two dimensional lattice is shown in Fig. IX.3. An important economic question is how the answers depend on the type and dimension of the lattice; in particular, it is necessary for there to be some range of types of lattices which produce the same qualitative answers for there to be useful theoretical predictions obtained from models of this type. Whether this will be the case depends of type of profit function. Clearly, when there are no interactions and all firms are independent, the type of lattice is irrelevant. Since a rough scale invariance appears to be observed in economic activity (c.f. Granger [82] or Barsky and Miron [14]) it may be that the properties of economic equilibrium can be analyzed by looking at the behavior of a relatively small number of economic agents. If this is the case, each firm's profit function will depend only on that of several other firms or neighbors. Aggregate behavior can still be quite correlated and complicated if all firms are linked to different firms.

We might define the profit function for instance as:

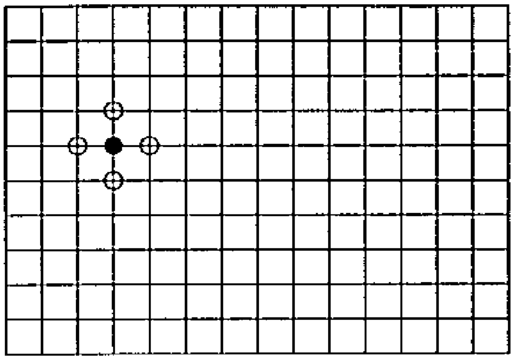

Figure IX.3: A square lattice. The black circle is the location of a firm and the white circles are the location of some of its nearest neighbors.

$$\Pi_i^e = P_i^e \phi_i - a_i \phi_i - \frac{b_i}{2}\phi_i^2 - \frac{\gamma_i}{4!}\phi_i^4 + \sum_{j \in N(i)} \lambda_{i,j}\phi_j\phi_i \qquad \text{(IX.7)}$$

where Π_i^e is expected profits of firm i, P_i^e is firm i's expectation of the market price, ϕ_i is the output of firm i and $\lambda_{i,j}$ are interaction parameters which capture the strength of local relationships between the output of firm j and the profits of firm i. The quartic term is included to approximate technological nonconvexities arising from lump-sum costs.

If the $\lambda_{i,j}$ are all positive and we fix prices or at least price expectations momentarily, a positive shock to a nearby firm's production process will lead to higher output and hence a positive effect on profits and output. In such a setup, the behavior of firms will be positively correlated. In the opposite case where $\lambda_{i,j}$ are negative and price expectations or prices are fixed, we expect the behavior of firms to be negatively correlated. In [153] [152] we have developed a field theoretic approach to statistical macroeconomics and suggested broadly that many lattice and continuum models developed for analysis of interacting particle systems in quantum field theory are relevant for a statistical treatment of economic models with many agents. This models provide potential mechanisms to improve the approximation provided by representative agent models, which replace the value of an idiosyncratic shock with an economy-wide average.

The conceptual problem with representative agent models in economies such as we consider, where fluctuations are large and involve simultaneously many of the factors determining equilibrium, is that the approximation of a representative agent is least satisfactory. Out of equilibrium, agents need only look locally to determine how to adjust prices, but near equilibrium price adjustment relies on a delicate accounting of the preferences and endowments of all agents. Therefore, near equilibrium, behavior is necessarily more correlated and may not be asymptotically independent as the representative agent model assumes.[3] Another facet of "field theoretic" models is that there is a sense, discussed in [152], in which the choice of technology is endogenized.

Models of interacting particle systems in economics have also been developed by Durlauf [66] and Follmer [74]. The case where $\lambda_{i,j}$ are positive are analogous to ferromagnetic field theories and the case where $\lambda_{i,j}$ are negative correspond to anti-ferromagnetic field theories. In economic models, the roughly corresponding terms of *strategic complements* and *strategic substitutes* have been developed in [81] and developed further in terms of their macroeconomic implications by Russel Cooper and Andrew John [49].[4] The difference between the approach in the strategic complements/substitutes literature and in [153] is that such approaches do not generally consider the local nature of interactions and instead focus on a dependence on average behavior.[5]

[3] A statistical model is required to make such a determination. The theoretical distribution function of output derived in [153] is $p(\phi) \propto e^{\frac{\pi(\phi)}{\lambda}}$, where λ is the cost of information and π is aggregate profits.

[4] Examples of macroeconomic models with strategic complementarities and thus Keynesian multiplier features are: [59] [101] [103] and [89]. Models with strategic complements/substitutes are of course ubiquitous in industrial organization.

[5] One reason for the difference in focus is that the empirical evidence for aggregate externalities is weak; for instance, a careful study by Basu and Fernald finds negligible evidence of aggregate production externalities in U.S. industry. In [150], we show that in some sense aggregate produc-

Some technical references on field theory are: [213] [159] [119] [108]. The form of observed spectra in economics [82] suggests a rough scale invariance of economic relations. This has also been confirmed in [14]. In physics, integrals such as we would need to compute to determine theoretical spectra are computed approximately using renormalization group transformations [51] [209], which provide a general framework for describing complex dynamical systems in terms of aggregative equations.[6] Such scaling transformation theory involves considering equilibria on infinitesimally larger markets and iterating until markets become infinitely large. The economic analog of such analysis would seem to be potentially a useful venue for future research.

To recap, we believe nonstationarity may have important implications for how economists construct models. Thus, an improved understanding of the importance (or lack thereof) of nonstationarity may help determine in which directions future theoretical research in macroeconomics will proceed.

---

tion externalities are inconsistent with the existence of equilibrium with free entry whereas this is definitely not the case with local production externalities which seem also to be more economically plausible.

[6] A renormalization group transformation is formally a renormalized sum of random variables. Consider a sequence of random variables $\{X_i\}_{i=0}^{\infty}$. For some $n \in Z^+$ and $\eta > 0$, we define:

$$(R_n X)_i = \frac{1}{n^\eta} \sum_{j=in}^{(i+1)n-1} X_j. \qquad \text{(IX.8)}$$

The transformation R_n is a renormalization group transformation. A fixed point X of the renormalization group transform occurs when $R_n X \overset{d}{\sim} X$ so that the renormalized data has the same distribution as the original random variables. A standard example where a "fixed point" is obtained is for the classical central limit theorem ($\eta = \frac{1}{2}$). Another example is for "stable" laws with scale parameter $\alpha \in (0, 2)$ for which $\eta = \frac{1}{\alpha}$ ([182], Ch. 1).

The basic idea is that, just as the central limit theorem suggests that broad classes of random variables correspond to the same averaged statistical behavior, an economic application of the renormalization group transform may suggest that broad classes of microeconomics correspond to the same macroeconomics.

Econometric Issues

The thesis suggests a number of new approaches to models with time-varying properties in econometrics. Among the interesting econometric issues suggested by the thesis are: effectiveness of different estimators for nonstationary time series, measurement of nonstationarity, definition of a nonstationary spectrum and broad econometric issues related to the use of "greedy"[7] estimation procedures which try to consider many potential models at one time (instead of just one and hence are "greedy"). Furthermore, there are issues relating to how to extend the methods proposed here to other settings such as multivariate models and models with integrated variables. In this section, we review some of these issues.

Alternative Approaches to Linear Time-Varying Models

Commonly-used models of time-variation such as ARCH models, hidden Markov models and random coefficient models rely on an assumption on conditional time-variation so that model characteristics change in response to schanges in underlying state variables or previous values of disturbances. Empirically, it is unclear whether an assumption of no unconditional time-variation is reasonable. Pagan and Schwert [157] [156] have found in split sample tests that unconditional variances of financial data appear to be time-dependent. Loretan and Phillips have found that this result does not change if thick-tails in financial data are accounted for properly in performing split sample tests [121].

It is useful to consider various simple exploratory data analysis methods for linear time-varying models. In Appendix L, we have suggested kernel estimators (see Eq. (L.4)) as one possibility for exploratory data analysis. Another possibility is to construct "spline" estimates by solving variational problems such as:

[7] The word "greedy" is often used to describe estimation procedures like Projection Pursuit [106].

$$\inf_{\beta(t)} \left[\sum_{t=2}^{T} (y(t) - \beta(t)y(t-1))^2 + \lambda T^2 \sum_{t=2}^{T-1} (\beta(t+1) - 2\beta(t) + \beta(t-1))^2 \right] \qquad \text{(IX.9)}$$

The variable λ controls the degree of smoothness of the curve $\beta(t)$ representing the first autocorrelation as a function of time. We note that λ can be interpreted as a Lagrange multiplier on a constraint on the mean curvature of $\beta(t)$. The second difference of $\beta(t)$ in the variational problem is used to approximate the second derivative of $\beta(t)$.

Both kernel and spline estimates are used for nonparametric estimation of nonlinear time-invariant relationships. For a review of spline smoothing procedures in nonparametric regression, see [105] [72] while for computational and technical details see [57]. The properties of spline and kernel estimators have not been closely studied in a nonstationary context. Research issues thus involve development of a comprehensive theory and detailed comparison with other methods such as the adaptive method suggested in the thesis.

State-dependent parameters

Another interesting research topic is application of the methodology in the thesis to state-dependent parameter modeling problems:

$$y(t) = \sum_{j=1}^{J} \beta_j\,(x(t))\; y(t-j) + \epsilon(t). \qquad \text{(IX.10)}$$

where $x(t)$ are state variables which incluence β_j. In this case, the window functions for the model components depend on $x(t)$ instead of t as in the thesis. Research issues involve development of a comprehensive theory and detailed comparison with other methods such as the nonlinear state space models of parameter variation of Young [210], Cooley and Prescott [48], Priestley [168], Hamilton [91] and others.

One particular possibility when there are many state variables $x(t)$ is to consider an *index pursuit model* where we seek to approximate the regression function in Eq. (IX.10) by:

$$y(t) = \sum_{j=1}^{J} \sum_{k=1}^{K} c_{jk} \left(\gamma_k' x(t)\right) y(t-j) + \epsilon(t). \tag{IX.11}$$

where $\gamma_k' \gamma_k = 1$. Here, the c_{jk} are "window functions" which may either be (depending on applications): (1) from a fixed class, (2) parameterized by a finite dimensional set of parameters, (3) arbitrary functions. At each stage of the procedure, we would find a best c_{jk} for each choice of the vector γ_k on the unit sphere (defined by the condition $\gamma_k' \gamma_k = 1$). We then choose the value of γ_k and the function c_{jk} which together best explain the data.

Multivariate Analysis

One interesting research question is how to expand the analysis above to handle the case of multivariate time series. The corresponding problem from the theory of stationary time series is to consider the vector model:

$$y(t) = \sum_{j=0}^{\infty} B_j y(t-j) + \epsilon_1(t). \tag{IX.12}$$

where the dimension of y is $K \times 1$, B_j is $K \times K$ and $\epsilon(t)$ is an independent random variable with a $K \times K$ covariance matrix Σ.

The appropriate version of Equation (IX.12) in terms of a nonstationary model is:

$$y(t) = \sum_{j=0}^{\infty} B_j(t) y(t-j) + \epsilon_1(t). \tag{IX.13}$$

In a manner analogous to the univariate case, we can consider rewriting Eq. (IX.13) as:

$$y(t) = \sum_{k=0}^{\infty} C_k h^k + \epsilon_1(t) \tag{IX.14}$$

where C_k is a coefficient matrix, h^k is a vector-valued model component, and $\epsilon_1(t)$ is a noise term. We can consider adding a vector-valued model component a stepwise fashion by adding at each iteration the model component which adds the most explanatory power. Some interesting research questions on this particular topic include:

- What computational and memory requirements are required for the multivariate case?
- How do the statistical properties of the multivariate case compare with the univariate one?

Time-Varying Spectral Estimation

In the thesis, we have suggested a generalization of the usual concept of a spectrum to handle time variation. Our definition of the spectrum is:

$$S(t, \omega) = \sum_i C_i^2 C_{e_i}(t, \omega) \tag{IX.15}$$

where $C_{e_i}(t, \omega)$ is the Cohen's class distribution of the appropriate eigenvector (eigenfunction) of the covariance matrix (operator) for the time series and C_i^2 is the corresponding eigenvalue.

Some research issues include:

- Definition of cross-spectra in a manner analogous to the way we have defined a time-varying spectrum. What special properties do our cross-spectral estimates have?

- How does the method perform when the time series are stationary but sample size is small? It is known [174] that the exact expressions for eigenvectors of the covariance matrices of some simple autoregressive models result in optimal transforms which are discrete sine and cosine transforms in finite samples rather than the discrete Fourier transform. How does this affect spectral estimates?

- What issues are important in the choice of Cohen's class kernel for smoothing purposes? How do these issues differ, if at all, from the choice of window in smoothing the periodogram?

Moving Average Models

We have developed the idea of estimation of autoregressive models in the thesis. It would be interesting to extend the results here to the more general case of mixed autoregressive-moving average models. Such an extension would seem to introduce tremendous technical problems for moving average model components would change each time a new regressor was added to a model because the residual changes; becuase of the data dependency, it is unclear under which circumstances the proposed method would converge to the correct model. Such research seems quite important as the same issues appear in dealing with *bilinear time series* models [170] [175]. A bilinear $BL(p,q,r,s)$ model is:

$$y(t) = \sum_{j=1}^{p} \alpha_j y(t-j) + \sum_{j=1}^{q} \beta_j \epsilon(t-j) + \sum_{j=1}^{r} \sum_{k=1}^{s} \gamma_{j,k}\, y(t-j)\, \epsilon(t-k). \qquad \text{(IX.16)}$$

Even more general models can be considered which represent higher order approximations to a nonlinear time series model; some models of time-varying conditional variances fit into such a framework [30] [70] [29] [145]. Thus, estimation of such models within the context of methods such as developed in the thesis would seem to be interesting topics for future research.

Measurement of Nonstationarity

An important issue is measurement of how much economic parameters vary over time. Given time-varying spectral or cross-spectral estimates, we suggest a possible measure T:

$$T = -\int w(\omega)\, K(t,\omega)\, \ln K(t,\omega)\, dt\, d\omega \tag{IX.17}$$

where:

$$K(t,\omega) = \frac{|S(t,\omega)|}{\int dt |S(t,\omega)|}. \tag{IX.18}$$

and w is a weight function. The measure T is interesting because it has an entropic interpretation. When there is no information in time variation of spectra, T will be zero and otherwise it will be positive.

Some other research issues include:

- What are the statistical properties of the measure T?
- How much do macroeconomic and financial time series vary over time according to the measure T?
- How does the choice of weight function w affect the results? Is there an optimal choice of weight function w?
- It may be that direct measurement of nonstationarity from the estimated covariance matrix makes more sense than doing so after the intermediate step of constructing time-varying spectral estimates. What are some other possible measures of nonstationarity in a time series based on estimates of the covariance matrix? How do these measures compare with the measure T we have defined above?

- How do other methods of measuring nonstationarity (which may not use the estimation procedure in the thesis) compare? Indeed, what are appropriate methods to use?

Recursive Kalman Filtering

One research topic is to investigate possible uses of recursive Kalman filtering to compute estimates of the linear nonstationary autoregressive model. As data arrives, the first step is to revise estimates based on the new data. The next step is to validate the model to see if any model components can be dropped from the model. The final step is to revise the model to see if additional model components should be added. Such an extension of the basic Kalman filter would be analogous to the use of the "extended" Kalman filter to estimate time-varying parameter and other nonlinear models.

Trends and Cointegration

To analyze time series with trends and unit roots, it has been found useful to use frequency domain methods via Hannan efficient estimators [92] [164] [165]. One merit cited for the frequency domain approach is that avoids the need to specify explicitly the dynamic structure of the errors. Another related advantage is that in many useful circumstance test statistics do not depend on nuisance parameters. In addition, cointegrating relationships are naturally defined at very low frequencies so decomposition in a Fourier basis enables the researcher to separate easily what is important from what is not [93] [69]. Although Hannan estimators are associated with spectral analysis, there is no particular reason why one cannot use other orthonormal basis functions such as wavelets (see Ch. VI for an introduction to wavelets) instead of Fourier basis functions. For instance, if wavelets are approximate eigenvectors of the covariance matrix of the time series then this seems to imply that Hannan-type spectral estimators with wavelet cofficients rather than Fourier coefficients *may be* promising in comparison with estimators for cointegrating regressions such as those developed by Phillips [164] as well as by Phillips and Choi [165].

To illustrate this point, we consider the simplest possible model:

$$y(t) = \alpha y(t-1) + \epsilon(t) \tag{IX.19}$$

where $\epsilon(t)$ has a stationary error structure. The Hannan efficient estimator of α is ([165], p. 266-7):

$$\hat{\alpha} = \frac{\frac{1}{T}\sum_{j=1}^{T} \hat{f}_{y_{t-1}y}(\omega_j)\hat{f}_{uu}(\omega_j)^{-1}}{\frac{1}{T}\sum_{j=1}^{T} \hat{f}_{y_{t-1}y_{t-1}}(\omega_j)\hat{f}_{uu}(\omega_j)^{-1}} \tag{IX.20}$$

where $\hat{f}_{uv}$ is the estimated cross spectrum between u and v. When $\alpha = 1$, the estimator is well-defined and consistent [165].

The reason this estimator is efficient is that the Fourier transform diagonalizes the covariance matrix of a stationary process so that the appropriate weights from the matrix inverse of the covariance matrix in the least squares estimate can be added as in Eq. (IX.20). Consider a scenario in which we compute different transforms of $y_{t-1}y_t$ which might include the Fourier transform as a special case; we may choose to weight estimates by the inverse of some estimated residual from an initial regression. We would then construct regression estimates by recursively subtracting off the best transform components. Such an approach might be an effective way to handle estimation of unit root and cointegration processes while allowing for time-variation of coeffcients.

It also seems worthwhile to investigate application of the methods of the thesis to equations of the form:

$$y(t) = c\left(\frac{t}{T}\right)x(t) + \epsilon(t) \tag{IX.21}$$

where $c(\frac{t}{T})$ is a function to be estimated and $y(t)$ and $x(t)$ are integrated processes.

Other Research Issues

In addition to the range of future research topics discussed above, we also list some other potential topics for future research in time series macroeconomics and finance suggested by this dissertation:

- Comparison of full nonlinear regression with the approach in the thesis.
- Analysis of traditional model selection criteria when the choice of the next model component to include is adaptive rather than fixed as in traditional approaches.
- Analysis of optimal choice of weighting functions and families of windows.
- Optimal stopping rules based on Lagrange multiplier tests.
- Robust estimation of models (in the sense of minimizing absolute deviations or the $\mathbf{L}^1$ norm) and examination of statistical properties.
- Consideration of Bayesian estimation problems by incorporating prior beliefs into choices on weight functions during selection of model components.

In this chapter, we have reviewed a wide variety of topics for future research suggested by this dissertation. While we have focused on theoretical questions here, the motivation for our work was the special properties of empirical macroeconomic and financial time series so that it is our hope that the thesis will raise as many new empirical questions as it does theoretical ones.

CHAPTER X

CONCLUSION

In the thesis, we examine a new approach to the analysis of certain forms of nonstationarity in macroeconomic and financial time series. We address the question of estimation of linear autoregressive models with *time-varying* parameters. Unlike the literature on random coefficient models in econometrics, the approach assumes that the functional forms of the true autoregressive parameters are not known. In our approach, the different possible types of time-varying coefficients are parameterized. The spirit of the idea is in some sense a combination of the ideas of Projection Pursuit Regression in statistics with the idea of Matching Pursuit representation of functions in terms of elementary waveforms which has been developed recently in the pathbreaking work of Mallat and Zhang [131]. Since the method here involves applying ideas from "pursuit" type algorithms to autoregressive processes, we have named the method *Autoregressive Pursuit.* The framework shares many of the advantages of nonparametric analysis and reduces to the standard stationary linear time series methodology as a special case.

The thesis contains a mixture of simulations, theoretical analysis and applications. We develop different estimators and examine their properties on some simulated stationary and nonstationary processes. We examine some theoretical properties of estimates and develop a stopping rule based on the use of wavelet coefficients which we believe to be of independent interest in testing for randomness in nonstationary time series. We also show how the estimation method developed in the thesis can be of use in nonstationary spectral estimation. We develop some applications to macroeconomic and financial data.

While our motivation in developing the method in the thesis was the special properties of economic data, we believe our approach may also have wider appplications in time series analysis. In particular, we think it might be applicable to the numerical estimation of dilation relationships such as occur in the analysis of texture and fractal growth patterns, a problem which has so far proved difficult [9] [8] [107]. It is also may be of some use in the analysis and linear prediction of speech patterns which have been the primary application of hidden Markov models.

In addition to the practical advantages of the approach in the thesis, we also think the approach taken is also interesting from the point of testing hypotheses and models as it allows one to parameterize at once a whole range of possible models and to examine which model or combination of models best represents the data.

APPENDICES

APPENDIX A

Function Spaces

In this appendix, we review the function spaces used in the theoretical parts of the thesis. Let $\mathbf{R}$ and $\mathbf{C}$ denote the set of real and complex numbers and let $\mathbf{Z}$ denote the set of integers. We define $\mathrm{L}^2(\nu)$ to be the Hilbert space of complex functions $f : \mathbf{R} \mapsto \mathbf{C}$ such that f is measurable and:

$$\int_{-\infty}^{\infty} |f(x)|^2 d\nu(x) < \infty \tag{A.1}$$

In the case where ν is a Lebesgue measure, we refer to $\mathrm{L}^2(\nu)$ as L^2. The space $\mathrm{L}^2(\nu)$ is a complex Hilbert space with an inner product $< f, g >$ defined by:

$$< f, g >= \int_{-\infty}^{\infty} f(x) g^*(x) d\nu(x) \tag{A.2}$$

and a norm $||f||$ defined by:

$$\int_{-\infty}^{\infty} |f(x)|^2 d\nu(x) = ||f||^2. \tag{A.3}$$

In the special case where the measure $\nu(dx)$ integrates to one, we use the simplifying notations:

$$E(f^2) = ||f||^2 \tag{A.4}$$

$$E(fg) =< f, g > \tag{A.5}$$

Hilbert spaces are particularly useful in the theory of stationary time series analysis because it is known that any stationary time series X_t can be represented as:

$$X_t = \int_0^{2\pi} e^{i\omega t} dZ(\omega) \tag{A.6}$$

where $dZ(\omega)$ is an orthogonal increment process such that:

$$E(\,dZ(\omega)\,dZ(\omega')\,) = dF_X(\omega)\,\delta(\omega - \omega') \tag{A.7}$$

where F_X is the cumulative spectral distribution of the data X.[1] Therefore, when we have a stationary random variable $X_t \in \mathbf{L}^2(F)$, we can define the $\mathbf{L}^2$ norm as:

$$E(X_t^2) = ||X_t||^2 = \int_0^{2\pi} dF_x(\omega) \tag{A.8}$$

and the inner product of X_t and X_{t-1} as:

$$E(X_t X_{t-1}) =< X_t, X_{t-1} >= \int_0^{2\pi} e^{i\omega} dF_x(\omega) \tag{A.9}$$

A comprehensive review of the use of Hilbert space methods in statistical theory is in [189]. Another reference on Hilbert space methods as used specifically in time series is ([33], Ch. II).

Since we use the theory of $\mathbf{L}^p$ mixingales due to Andrews [6] and McLeish [135], it is useful to define formally $\mathbf{L}^p$ norms as well. A measurable function has an $\mathbf{L}^p(\nu)$ norm of:

$$||f||_{\mathbf{L}^p} = \left(\int_{-\infty}^{\infty} |f(x)|^p d\nu(x)\right)^{1/p} \tag{A.10}$$

for $1 \leq p < \infty$; a measurable function is in $\mathbf{L}^p$ if:

[1] See [162] [169] for detailed discussions of the spectral analysis of stationary time series.

$$\int_{-\infty}^{\infty} |f(x)|^p d\nu(x) < \infty. \tag{A.11}$$

For $p = \infty$ the norm is defined as:

$$||f||_{\mathbf{L}^\infty} = \text{ess sup}\, |f(x)|. \tag{A.12}$$

or the supremum over all measurable sets. When the measure $\nu(dx)$ integrates to one, we can substitute expected values for norms, e.g. $E(f^p) = ||f||^p_{\mathbf{L}^p}$.

Since in various places in the thesis, we make technical comments about the properties of sequences, we review here the proper definitions of norms for sequences. For instance, a sequence: $f : \mathbf{Z} \mapsto \mathbf{C}$ is in $\mathbf{l}^2$ if:

$$\sum_{j=-\infty}^{\infty} |f(j)|^2 < \infty; \tag{A.13}$$

it then has an $\mathbf{l}^2$ norm of:

$$||f||^2_{\mathbf{l}^2} = \sum_{j=-\infty}^{\infty} |f(j)|^2 \tag{A.14}$$

Likewise, a sequence: $f : \mathbf{Z} \mapsto \mathbf{C}$ is in $\mathbf{l}^p$ if

$$\sum_{j=-\infty}^{\infty} |f(j)|^p < \infty; \tag{A.15}$$

it then has an $\mathbf{l}^p$ norm of :

$$||f||^p_{\mathbf{l}^p} = \sum_{j=-\infty}^{\infty} |f(j)|^p \tag{A.16}$$

A sequence which is bounded is in $\mathbf{l}^\infty$.

APPENDIX B

Time Series Background

This appendix briefly reviews some of the background in time series analysis assumed by the thesis. The standard approach to time series analysis relies on a discrete-time linear model with time-invariant coefficients:

$$y(t) = \sum_{j=1}^{\infty} \beta_j y(t-j) + \sum_{k=0}^{\infty} h_k \epsilon(t-k) \tag{B.1}$$

where ϵ is an independently and normally distributed noise term. Letting z denote a backwards shift we can write Equation B.1 as:

$$y(t) = \sum_{j=1}^{\infty} \beta_j z^j y(t) + \sum_{k=0}^{\infty} h_k z^k \epsilon(t) \tag{B.2}$$

We define the polynomials $B(z) = 1 - \sum_{j=1}^{\infty} \beta_j z^j$ and $H(z) = \sum_{j=0}^{\infty} h_j z^j$. We then rewrite Equation B.1 as:

$$B(z)y(t) = H(z)\epsilon(t) \tag{B.3}$$

If $B(z)$ has no zeroes on the unit circle we can rewrite Equation B.3 as:

$$y(t) = \frac{H(z)}{B(z)}\epsilon(t) = R(z)\epsilon(t) \tag{B.4}$$

Eq. (B.3) is referred to as the *moving average* representation of y.

The correlation function of y is defined as:

$$\gamma(k) = E(y(t)y(t-k)) \tag{B.5}$$

The spectrum of the stationary process y is defined as the Fourier transform of its autocovariance function:

$$\hat{y}(\omega) = \sum_{n=-\infty}^{\infty} \gamma(n)e^{-in\omega} \tag{B.6}$$

The spectrum is a convenient tool for analyzing the frequencies of fluctuations which are contributing to the variation in the time series. The spectrum is:

$$\hat{y}(\omega) = |R(e^{i\omega})|^2 \frac{\sigma^2}{2\pi} \tag{B.7}$$

An estimate of the spectrum is the smoothed square of the Fourier coefficients of the data. Smoothing is necessary to achieve consistency of estimates. Some useful references on the material in this appendix are: [33] [126] [169] [181] [36].

APPENDIX C

Computation of Discrete Wavelet Transforms

To compute wavelet transforms on a set of discretely sampled data $\{y(t_i)\}_{i=1}^{T}$ where t_i are located on a uniform grid, we set $c_{0,j} = y(t_j)$ for $j = 1 \ldots T$ and implement the following recursions which follow from Mallat's Fast Wavelet Transform algorithm [130]:

$$c_{j,k} = \sum_n h^*(n-2k)c_{j-1,n} \tag{C.1}$$

$$d_{j,k} = \sum_n g^*(n-2k)c_{j-1,n}. \tag{C.2}$$

Filter coefficients for h and g in Eq. (C.1) and Eq. (C.2) are in Appendix D. We will outline what conditions these filter coefficients must satisfy below. In Eq. (C.1) and Eq. (C.2), we note that the wavelet coefficients are defined as (see Appendix A for the definitions of inner products):

$$d_{j,k} =< y, \psi_{j,k} > \tag{C.3}$$

and the approximation or smoothing coefficients of y, $c_{j,k}$ are:

$$c_{j,k} =< y, \phi_{j,k} > . \tag{C.4}$$

where:

$$\phi_{j,k}(t) = 2^{-\frac{j}{2}}\phi\left(2^{-J}t - k\right) \tag{C.5}$$

$$\psi_{j,k}(t) = 2^{-\frac{j}{2}}\psi\left(2^{-j}t - k\right) \tag{C.6}$$

for a wavelet function $\psi(t)$ and a corresponding 'smoothing' function $\phi(t)$.

For any $J \geq 0$, we recall that a discrete wavelet expansion for $y \in \mathbf{L}^2$ is of the form:

$$y(t) = \sum_{k=-\infty}^{\infty} c_{J,k}\phi_{J,k}(t) + \sum_{j=-\infty}^{J}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(t) \tag{C.7}$$

To see what conditions the filters h and g must satisfy, we can write [130]:

$$\phi(x) = \sqrt{2}\sum_n h_n\phi(2x - n) \tag{C.8}$$

$$\psi(x) = \sqrt{2}\sum_n g_n\phi(2x - n). \tag{C.9}$$

We take the Fourier transform:

$$\begin{aligned} \hat{\phi}(\omega) &= \hat{m}(\omega/2)\hat{\phi}(\omega/2) \\ &= \prod_{j=1}^{\infty}\hat{m}\left(\frac{\omega}{2^j}\right) \end{aligned} \tag{C.10}$$

where:

$$\hat{m}(\omega) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2}}\sum_n h_n e^{-i\omega n} \tag{C.11}$$

We note that:

$$\begin{aligned} \hat{\psi}(\omega) &= \hat{n}(\omega/2)\hat{\phi}(\omega/2) \\ &= \prod_{j=1}^{\infty}\hat{n}\left(\frac{\omega}{2^j}\right) \end{aligned} \tag{C.12}$$

where:

$$\hat{n}(\omega) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} \sum_{n} g_n e^{-i\omega n} \tag{C.13}$$

Let us define the matrices H and G to correspond to the coefficients in Eq. (C.1) and Eq. (C.2). We then have that:

$$c_{j+1} = Hc_j \tag{C.14}$$

$$d_{j+1} = Gc_j \tag{C.15}$$

To invert the transform or reconstruct the data from the coefficients:

$$c_{j-1} = H^* c_j + G^* d_j \tag{C.16}$$

Certain restrictions are placed on the choice of filter coefficients. Orthonormality requires:

$$HG^* = 0$$

$$GH^* = 0$$

Reconstruction requires in addition:

$$H^* H + G^* G = I \tag{C.17}$$

In terms of Fourier transforms we need:

$$|m(\omega)|^2 + |m(\omega + \pi)|^2 = 1 \tag{C.18}$$

$$|n(\omega)|^2 + |n(\omega + \pi)|^2 = 1 \tag{C.19}$$

$$|m(\omega)|^2 + |n(\omega + \pi)|^2 = 0 \tag{C.20}$$

$$|n(\omega)|^2 + |m(\omega + \pi)|^2 = 0 \tag{C.21}$$

with: $m(0) = 1, m(\pi) = 0, n(0) = 0, n(\pi) = 1$. These conditions imply [130] [200]:

$$\sum_k h_k = \sqrt{2} \tag{C.22}$$

$$\sum_k h_{k-2m} h^*_{k-2n} = \delta(m-n) \tag{C.23}$$

$$\sum_k g_{k-2m} g^*_{k-2n} = \delta(m-n) \tag{C.24}$$

$$\sum_k h_{k-2m} g^*_{k-2n} = 0.0 \tag{C.25}$$

where the δ function in this Appendix is a Kronecker delta:

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{C.26}$$

Appendix D lists some of the filter coefficients which satisfy the above conditions and which are necessary to efficiently compute wavelet transforms as well as wavelet packet transforms.

Let us consider some independent Gaussian noise $y(t)$ and derive some statistical properties of its wavelet coefficients based on the formulas we use to compute the wavelet transform.

First, we note that:

$$E(c_{j,k}) = E\left(\sum_n h(n-2k)c_{j-1,n}\right) = 0. \tag{C.27}$$

for $j > 0$ since $E(c_{0,n}) = E(y(n)) = 0$. This implies that:

$$E(d_{j,k}) = E\left(\sum_n g(n-2k)c_{j-1,n}\right) = 0. \tag{C.28}$$

so that discrete wavelet coefficients are mean zero. To simplify the presentation (without loss of generality), we have assumed that the wavelets are real.

We note:

$$\begin{aligned} E(d_{2,k'}d_{1,k}) &= E\left(\sum_r \sum_l \sum_n h(r-2k')g(n-2k)c_{0,n}h(l-2k')c_{0,l}\right) \\ &= \sigma^2 \sum_r \sum_n h(r-2k')\,g(n-2k)\,h(n-2k') \\ &= 0 \end{aligned} \tag{C.29}$$

since:

$$\sum_n g(n-2k)\,h(n-2k') = \sum_n g(n)\,h(n-p) = 0 \tag{C.30}$$

by the filter requirements and:

$$E(c_{0,n}c_{0,l}) = E(y(n)y(l)) = \sigma^2\delta(n-l) \tag{C.31}$$

since $y(n)$ is white noise. Similarly:

$$\begin{aligned} E(d_{3,k'}d_{1,k}) &= E\left(\sum_s \sum_r \sum_l \sum_n h(s-2k')g(r-2k')g(n-2k)c_{0,n}h(l-2k')c_{0,l}\right) \\ &= \sigma^2 \sum_s \sum_r \sum_n h(s-2k')g(r-2k')g(n-2k)h(n-2k') \\ &= 0 \end{aligned} \tag{C.32}$$

and in general $E(d_{j,k'}d_{j',k}) = 0$ for $j \neq j'$. We also note that:

$$\begin{aligned} E(d_{1,k'}d_{1,k}) &= E\left(\sum_l \sum_n g(l-2k')g(n-2k)c_{0,n}c_{0,l}\right) \\ &= \sigma^2 \sum_n g(n-2k)\, g(n-2k') \\ &= \sigma^2 \delta(k-k') \qquad \text{(C.33)} \end{aligned}$$

since:

$$\sum_n g(n-2k)g(n-2k') = \delta(k-k') \qquad \text{(C.34)}$$

again by the requirements which filters must satisfy. Similarly, $E(d_{j,k}d_{j,k'}) = 0$ for $k \neq k'$. Thus, since the wavelet transform is linear, the computed wavelet coefficients of white noise are independent white noise. This must be the case as white noise decomposed in *any* orthonormal basis is white noise.

APPENDIX D

Wavelet Filter Coefficients

This appendix contains some of the wavelet filter coefficients used in computations. These particular filter coefficients are derived by Daubechies [52].

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	7.07106781186547570e-01	7.07106781186547570e-01
1	7.07106781186547570e-01	-7.07106781186547570e-01

Table D.1: Haar wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	4.82962913144534160e-01	-1.29409522551260370e-01
1	8.36516303737807940e-01	-2.24143868042013390e-01
2	2.24143868042013390e-01	8.36516303737807940e-01
3	-1.29409522551260370e-01	-4.82962913144534160e-01

Table D.2: Daubechies D4 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	3.32670552950082630e-01	3.52262918857095330e-02
1	8.06891509311092550e-01	8.54412738820266580e-02
2	4.59877502118491540e-01	-1.35011020010254580e-01
3	-1.35011020010254580e-01	-4.59877502118491540e-01
4	-8.54412738820266580e-02	8.06891509311092550e-01
5	3.52262918857095330e-02	-3.32670552950082630e-01

Table D.3: Daubechies D6 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	2.30377813309000010e-01	-1.05974017850000000e-02
1	7.14846570553000050e-01	-3.28830116670000010e-02
2	6.30880767930000030e-01	3.08413818359999990e-02
3	-2.79837694169999990e-02	1.87034811718999990e-01
4	-1.87034811718999990e-01	-2.79837694169999990e-02
5	3.08413818359999990e-02	-6.30880767930000030e-01
6	3.28830116670000010e-02	7.14846570553000050e-01
7	-1.05974017850000000e-02	-2.30377813309000010e-01

Table D.4: Daubechies D8 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	1.60102397974000000e-01	3.33572528500000010e-03
1	6.03829269797000020e-01	1.25807519990000000e-02
2	7.24308528437999980e-01	-6.24149021300000020e-03
3	1.38428145901000000e-01	-7.75714938400000050e-02
4	-2.42294887066000000e-01	-3.22448695850000020e-02
5	-3.22448695850000020e-02	2.42294887066000000e-01
6	7.75714938400000050e-02	1.38428145901000000e-01
7	-6.24149021300000020e-03	-7.24308528437999980e-01
8	-1.25807519990000000e-02	6.03829269797000020e-01
9	3.33572528500000010e-03	-1.60102397974000000e-01

Table D.5: Daubechies D10 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	1.11540743350000000e-01	-1.07730108500000000e-03
1	4.94623890397999980e-01	-4.77725751100000020e-03
2	7.51133908021000000e-01	5.53842200999999980e-04
3	3.15250351709000010e-01	3.15820393180000010e-02
4	-2.26264693965000010e-01	2.75228655299999990e-02
5	-1.29766867567000010e-01	-9.75016055869999950e-02
6	9.75016055869999950e-02	-1.29766867567000010e-01
7	2.75228655299999990e-02	2.26264693965000010e-01
8	-3.15820393180000010e-02	3.15250351709000010e-01
9	5.53842200999999980e-04	-7.51133908021000000e-01
10	4.77725751100000020e-03	4.94623890397999980e-01
11	-1.07730108500000000e-03	-1.11540743350000000e-01

Table D.6: Daubechies D12 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	7.78520540849999970e-02	3.53713800000000020e-04
1	3.96539319482000000e-01	1.80164070400000000e-03
2	7.29132090845999950e-01	4.29577973000000010e-04
3	4.69782287405000000e-01	-1.25509985560000000e-02
4	-1.43906003928999990e-01	-1.65745416310000000e-02
5	-2.24036184993999990e-01	3.80299369350000010e-02
6	7.13092192669999990e-02	8.06126091510000060e-02
7	8.06126091510000060e-02	-7.13092192669999990e-02
8	-3.80299369350000010e-02	-2.24036184993999990e-01
9	-1.65745416310000000e-02	1.43906003928999990e-01
10	1.25509985560000000e-02	4.69782287405000000e-01
11	4.29577973000000010e-04	-7.29132090845999950e-01
12	-1.80164070400000000e-03	3.96539319482000000e-01
13	3.53713800000000020e-04	-7.78520540849999970e-02

Table D.7: Daubechies D14 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	5.44158422430000010e-02	-1.17476784000000000e-04
1	3.12871590914000020e-01	-6.75449405999999950e-04
2	6.75630736296999990e-01	-3.91740372999999990e-04
3	5.85354683654000010e-01	4.87035299299999960e-03
4	-1.58291052559999990e-02	8.74609404700000050e-03
5	-2.84015542961999990e-01	-1.39810279170000000e-02
6	4.72484573999999990e-04	-4.40882539310000000e-02
7	1.28747426619999990e-01	1.73693010020000010e-02
8	-1.73693010020000010e-02	1.28747426619999990e-01
9	-4.40882539310000000e-02	-4.72484573999999990e-04
10	1.39810279170000000e-02	-2.84015542961999990e-01
11	8.74609404700000050e-03	1.58291052559999990e-02
12	-4.87035299299999960e-03	5.85354683654000010e-01
13	-3.91740372999999990e-04	-6.75630736296999990e-01
14	6.75449405999999950e-04	3.12871590914000020e-01
15	-1.17476784000000000e-04	-5.44158422430000010e-02

Table D.8: Daubechies D16 wavelet filter coefficients.

Coeff. Number	Smoothing Filter $h(n)$	Wavelet filter $g(n)$
0	2.667005790100000010e-02	-1.326420300000000010e-05
1	1.881768000780000000e-01	-9.358867000000000050e-05
2	5.272011889319999960e-01	-1.164668550000000000e-04
3	6.884590394540000000e-01	6.858566950000000030e-04
4	2.811723436610000020e-01	1.992405295000000020e-03
5	-2.498464243269999990e-01	-1.395351747000000000e-03
6	-1.959462743769999990e-01	-1.073317548300000000e-02
7	1.273693403360000000e-01	-3.606553567000000010e-03
8	9.305736460400000060e-02	3.321267405899999970e-02
9	-7.139414716599999970e-02	2.945753682199999990e-02
10	-2.945753682199999990e-02	-7.139414716599999970e-02
11	3.321267405899999970e-02	-9.305736460400000060e-02
12	3.606553567000000010e-03	1.273693403360000000e-01
13	-1.073317548300000000e-02	1.959462743769999990e-01
14	1.395351747000000000e-03	-2.498464243269999990e-01
15	1.992405295000000020e-03	-2.811723436610000020e-01
16	-6.858566950000000030e-04	6.884590394540000000e-01
17	-1.164668550000000000e-04	-5.272011889319999960e-01
18	9.358867000000000050e-05	1.881768000780000000e-01
19	-1.326420300000000010e-05	-2.667005790100000010e-02

Table D.9: Daubechies D20 wavelet filter coefficients.

APPENDIX E

Statistical Properties of Rolling Autoregressive Estimates

In this appendix, we review an interesting asymptotic property of estimated coefficients in the case when we use a uniform set of overlapping flat windows.

Theorem 13 *Consider a stationary autoregressive process:*

$$y(t) = \beta y(t-1) + \epsilon(t) \tag{E.1}$$

with:

$$|\beta| < 1. \tag{E.2}$$

$$\epsilon \sim N(0, \sigma^2) \tag{E.3}$$

and consider a local estimator:

$$\hat{\beta}_L\left(t + \frac{L}{2}\right) = \frac{\frac{1}{L}\sum_{r=t}^{t+L} y(r)y(r-1)}{\frac{1}{L}\sum_{r=t}^{t+L} y(r-1)^2} \tag{E.4}$$

then as $L \to \infty$, $\hat{\beta}_L(t) - \hat{\beta}_L(t-1)$ is not autocorrelated at lags less than L.

Proof: As a first step, we note that as $L \to \infty$, the denominator of the least squares estimator converges to the unconditional variance of y:

$$\frac{1}{L}\sum_{r=t}^{t+L} y(r-1)^2 \to \tau^2 \tag{E.5}$$

Furthermore, we note that:

$$\sum_{r=t+1}^{t+L+1} \epsilon(r)y(r-1) - \sum_{r=t}^{t+L} \epsilon(r)y(r-1) = \epsilon(t+L+1)y(t+L) - \epsilon(t)y(t-1) \tag{E.6}$$

From Eq. (E.6), we have for large L that:

$$L\tau^2\left(\hat{\beta}_L\left(t+\frac{L}{2}+1\right) - \hat{\beta}_L\left(t+\frac{L}{2}\right)\right) \approx \epsilon(t+L+1)y(t+L) - \epsilon(t)y(t-1) \tag{E.7}$$

The original version of this appendix worked exclusively with Eq. (E.7). However, Eq. (E.7) is not a valid as a limiting expression because the right hand side depends on L and a slightly more complicated technical argument is required.[1]

We first demonstrate the result for $\beta = 0$. Define $s \in [0,1]$ by: $s = \frac{r}{T}$. Define also $k = \frac{L}{T}$. Noting that by definition:

$$\frac{1}{T}\sum_{r=t}^{t+L+1} \epsilon(r)\epsilon(r-1) = \frac{1}{T}\sum_{r=t}^{t+[kT]+1} \epsilon(r)\epsilon(r-1) \tag{E.8}$$

we then define:

$$C_T(s) = \frac{1}{T}\sum_{r=[sT]}^{[sT]+([kT]+1)} \epsilon(r)\epsilon(r-1) \tag{E.9}$$

which is approximately:

$$C_T(s) = \frac{1}{T}\sum_{r=[sT]}^{[(s+k)T]} \epsilon(r)\epsilon(r-1) \tag{E.10}$$

[1] I am grateful to Prof. E.P. Howrey for pointing out this problem and for suggesting some possible ways out.

Since the random variable: $\epsilon(r)\epsilon(r-1)$ has only local dependence, the functional central limit theorem [25] implies:

$$\sqrt{T}C_T(s) \Rightarrow \omega \int_s^{s+k} dW(t) \tag{E.11}$$

where ω is σ^2 (not σ) since ϵ is Gaussian and where $W(s)$ is a Wiener process.

Thus:

$$\sqrt{L}\hat{\beta}\left(\left(s+\frac{k}{2}\right)T\right) \Rightarrow \sqrt{k}\int_s^{s+k} dW(t) \tag{E.12}$$

An infinitesemal change in s results in change in the least squares estimator which is proportional to the Brownian motion increment:

$$dW(s+k) - dW(s) \tag{E.13}$$

This increment will only show correlations with another increment:

$$dW(r+k) - dW(r) \tag{E.14}$$

in some very special cases. First, if $r = s$, then the increments will be positively correlated. Likewise, if $r = s + k$, then the increments will be negatively correlated. Also, if $r = s - k$ the increments will be negatively correlated. However, when $r = s+k$ and $r = s - k$ we have violated the assumptions of the theorem because the window is of length $L = [kT]$ and our claim was only that autocorrelations at *nonzero* lengths *less* than L were zero.

We now proceed to show the result on a more rudimentary level when $\beta = 0$. The purpose of doing this is to show how the result works in practice with large but finite datasets. In addition, the Brownian motion asymptotics only apply for autocorrelations at large lags whereas the calculations below apply for large L but autocorrelations at small lags. We have from Eq. (E.7):

$$L\tau^2\left(\hat{\beta}_L(t+\frac{L}{2}+1)-\hat{\beta}_L\left(t+\frac{L}{2}\right)\right)\approx\epsilon(t+L+1)\epsilon(t+L)-\epsilon(t)\epsilon(t-1) \quad \text{(E.15)}$$

It immediately follows that the expected change in $\hat{\beta}$ is zero. To compute the autocovariances of changes in $\hat{\beta}$, we need to compute the expectation value:

$$\begin{aligned} Z &= E((\epsilon(t+L+1)\epsilon(t+L)-\epsilon(t)\epsilon(t-1))\times \\ &\quad (\epsilon(t+L+m+1)\epsilon(t+m+L)-\epsilon(t+m)\epsilon(t+m-1))) \end{aligned} \quad \text{(E.16)}$$

We note:

$$Z = Z_1 + Z_2 + Z_3 + Z_4 \quad \text{(E.17)}$$

where:

$$Z_1 = E\left(\epsilon(t+L+1)\epsilon(t+L)\epsilon(t+L+m+1)\epsilon(t+m+L)\right) \quad \text{(E.18)}$$

$$Z_2 = -E\left(\epsilon(t+L+1)\epsilon(t+L)\epsilon(t+m)\epsilon(t+m-1)\right) \quad \text{(E.19)}$$

$$Z_3 = -E\left(\epsilon(t)\epsilon(t-1)\epsilon(t+L+m+1)\epsilon(t+m+L)\right) \quad \text{(E.20)}$$

$$Z_4 = E\left(\epsilon(t)\epsilon(t-1)\epsilon(t+m)\epsilon(t+m-1)\right) \quad \text{(E.21)}$$

We first consider Z_1. For the expectation of a product of independent Gaussian variables to be nonzero, it is known that there must be an even number of random variables and all random variables must be in pairs whose indices match. The first ϵ cannot match with the second and can match only with the third when $m = 0$ which produces a variance term rather than a covariance. The first ϵ can match with the fourth ϵ when $m = 1$, but in this case the second and third ϵ do not have matching indices. Thus, Z_1 is zero.

We now consider Z_2. The index of the first term cannot match with the second but can match with the third when $L + 1 = m$. The first term can likewise match with the fourth when $m = L + 2$. When $m = L + 1$ the second and fourth terms match so that $Z_2 = \sigma^4$. However, in this case, $|m| \geq L$ which violates the conditions of our theorem. When $m = L + 2$, the second and third terms do not match so that $Z_2 = 0$. Thus, $Z_2 = 0$.

We now consider Z_3. The index of the first term cannot match the second. It can match with the third when $m = -L - 1$, In this case, the fourth and second terms match. The index of the first term can match the fourth when $m = -L$, but then the indices of the second and third term do not match. Thus, $Z_3 = 0$ unless $m = -L-1$ which contradicts the assumption of the theorem (that the autocorrelation lag $|m| < L$).

We next consider Z_4. The indices of the first term do not match with the second or third terms for $m \neq 0$, but they do match the fourth term when $m = 1$; however, in this case the indices of the third and the second terms do not match. Hence, $Z_4 = 0$.

Therefore, if $|m| < L$, the autocorrelation of increments of $\hat{\beta}_L(t)$ are asymptotically zero so that $\hat{\beta}_L(t)$ behaves asymptotically over short intervals as a Brownian motion.

We now extend the results to models with $\beta \neq 0$. The assumption $|\beta| < 1$ is crucial. We first describe how to extend how the Brownian motion formalism used for the $\beta = 0$ case to the case where $\beta \neq 0$. We define:

$$C_T(s) = \frac{1}{T} \sum_{r=[sT]}^{[(s+k)T]} y(r)y(r-1) \tag{E.22}$$

where k is the proporiton of the data spanned by the window. Eq. (E.22) can be rewritten as:

$$C_T(s) = \frac{1}{T} \sum_{r=[sT]}^{[(s+k)T]} [\beta y(r-1)y(r-1) + \epsilon(t)y(t-1)] \tag{E.23}$$

We note that:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{r=[sT]}^{[(s+k)T]} y(r-1)y(r-1) \to \tau^2 \tag{E.24}$$

We also note that $\epsilon(t)y(t-1)$ has only local dependence because:

$$y(t-1) = \sum_j \beta^j \epsilon(t-j-1) \tag{E.25}$$

implies that correlations with distant ϵ die off at an exponential rate. Therefore:

$$\sqrt{T}(\frac{C_T(s)}{\tau} - \beta) \Rightarrow \frac{\lambda}{\tau} \int_s^{s+k} dW(t) \tag{E.26}$$

where λ^2 is the long-run variance of $y(t-1)\epsilon(t)$. We then have:

$$\sqrt{L}(\hat{\beta}(s + \frac{k}{2}) - \beta) \Rightarrow \frac{\lambda}{\tau^2}\sqrt{k} \int_s^{s+k} dW(t) \tag{E.27}$$

which is of exactly the same form as before for the case $\beta = 0$ so that we can use exactly the same asymptotic arguments for the lack of autocorrelations in changes in the local autoregressive estimate.

The exact calculations on the numerator are somewhat more involved than for the $\beta = 0$ case but we go through them anyway in order to show how the result works in detail for large L. To compute the correlation among the changes in estimates of $\hat{\beta}(t)$ we use Eq. (E.6) to obtain the expression:

$$\begin{aligned} I = E(\sum_{j,k\geq 0} \beta^j \beta^k (\epsilon(t+L+1)\epsilon(t+L-j) - \epsilon(t)\epsilon(t-1-j)) \times \\ (\epsilon(t+L+1+m)\epsilon(t+L+m-k) - \epsilon(t+m)\epsilon(t+m-1-k))) \end{aligned} \tag{E.28}$$

We compute the expectation value of Eq. (E.28) term by term. We have:

$$I = I_1 + I_2 + I_3 + I_4 \tag{E.29}$$

where:

$$\begin{aligned} I_1 &= E(\sum_{j,k\geq 0} \beta^{j+k}\epsilon(t+L+1) \times \\ &\quad \epsilon(t+L-j)\epsilon(t+L+1+m)\epsilon(t+L+m-k)) \end{aligned} \tag{E.30}$$

$$I_2 = -E\left(\sum_{j,k\geq 0} \beta^{j+k}\epsilon(t+L+1)\epsilon(t+m)\epsilon(t+L-j)\epsilon(t+m-1-k)\right) \tag{E.31}$$

$$I_3 = -E\left(\epsilon(t)\sum_{j,k\geq 0} \epsilon(t-1-j)\epsilon(t+L+1+m)\epsilon(t+L+m-k)\right) \tag{E.32}$$

$$I_4 = E\left(\epsilon(t)\sum_{j,k\geq 0} \epsilon(t-1-j)\epsilon(t+m)\epsilon(t+m-1-k)\right). \tag{E.33}$$

We first consider I_1. The term $\epsilon(t+L+1)$ cannot have indices which match $\epsilon(t+L-j)$ since $j > 0$. it cannot have indices matching the third term but it can have indices matching the fourth term if $m - k = 1$. Thus, if $m > 0$, the only possibility is when $k = m - 1$. In this case, the second and the third terms must match but this can only occur if $j = -m - 1$ which can only occur if $m < 0$. Thus, if $m > 0$, I_1 is zero. It is also zero when $m < 0$ because $\epsilon(t+L+1)$ cannot pair up with any other terms.

We next consider I_2. The first two variables match when $m = L + 1$. However, in this case $|m| \geq L$ which contradicts the assumptions of the theorem. In the case where $m = L + 1$, all $j = k - 1$, which results in $-\mu^2\sigma^4$ for some constant $\mu(\beta)$. Similarly, the first variable can never match the third variable. The first matches the fourth when $m = k + L + 1$ in which case the second term cannot match indices with the third.

We next consider I_3. The first term can never have identical indices with the second. If $m = -L - 1$, the first term can have an identical index with the third in which case, we have terms contibuting wherever $j = k$ which leads to $-\mu^2\sigma^4$ for some $\mu(\beta)$. The first term can match with the fourth when $L + m - k = 0$ so that $k = m + l$ which also violates the conditions of our theorem. In this case, the third term cannot match with the second. Thus, when the conditions of the theorem are satisfied, $|m| < L$, $I_3 = 0$.

Finally, we consider I_4. The first term can never have identical indices with the second. It also cannot have identical indices with the third since $m \neq 0$ by assumption. The first term can have identical indices with the fourth term when $k = m - 1$. However, in this case, the second and third terms cannot have identical indices. Therefore, $I_4 = 0$.

Thus, we have shown that over short intervals, estimates of autoregressive models using rolling estimation procedures result in random walk behavior. We have also shown that there is a strong negative correlation among increments spaced by window length. This negative correlation causes the autoregressive estimate to maintain a time invariant variance. ∎

That this theorem works operationally can be shown in a figure (c.f. Fig. E.1) where we have constructed autoregressive estimates using a large window $L = 1000$ on a random sample of size 8000. The autocorrelations of the differences are all statistically zero as expected and a regression of the estimated β on their first lag results in an estimated coefficient of 0.999187 with ordinary standard errors of 0.000285181.

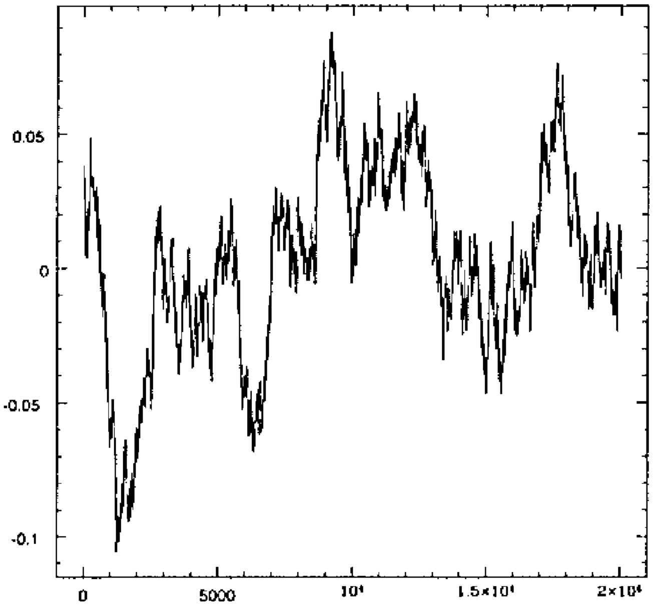

Figure E.1: Autoregressive estimates.

APPENDIX F

Pursuit Methods

In this appendix, we relate the approach to time series analysis in the thesis to *Projection Pursuit* and *Matching Pursuit.* We also show how to use this relationship to define some faster computational procedures which might be useful in practice with large datasets. Readers interested only in the alternative computational procedures should skip to the section labeled "Recipe for a Simplified Algorithm".

Projection Pursuit Regression

Suppose we want to determine the statistical relationship between a set of variables $X \in \mathbf{R}^d$ and a response variable $Y \in \mathbf{R}$. Projection Pursuit [106] is a method of estimating the conditional expectation:

$$f(x) = E(Y|X = x) \tag{F.1}$$

where the dimension d is large. Projection Pursuit estimates $f(x)$ by an expansion in terms of one-dimensional "ridge" functions g_i:

$$\hat{f}(x) = \sum_{i=0}^{\infty} g_i(\beta_i' x) \tag{F.2}$$

where $\beta_i'\beta_i = 1$. Projection Pursuit regression is implemented iteratively. The residual at any stage I is given by:

$$r^I = Y - \sum_{i=0}^{I} g_i(\beta_i' x). \tag{F.3}$$

At any stage I, we fit the best function g_I to the residuals nonparametrically for a given choice of β_I and then choose the β_I which produces the estimated function g_I with the smallest residual sum of squares. This approach is conceptually attractive for exploratory analysis of high-dimensional data and thus frequently used in practice.[1] Nevertheless, as a nonparametric procedure, it is computationally intensive and requires large sample sizes to use effectively. In addition, while the procedure is over twenty years old, little is known about its asymptotic statistical properties.[2]

Matching Pursuit Analysis

The Matching Pursuit analysis of Mallat and Zhang [131] provides a way of expressing functions in terms of elementary waveforms such as that shown in Figure (III.3). If we let a waveform of type i be e_i, Mallat and Zhang developed a way of expanding a function $f(x)$ in terms of a sum of waveforms:

$$f(x) = \sum_{i=0}^{\infty} C_i e_i(x) \qquad \text{(F.4)}$$

where the two functions e_i and e_j are not orthogonal to each other.[3] Consideration of nonorthonormal vectors allowed Mallat and Zhang to develop a simple way of representing functions in terms of nonorthonormal functions which has attractive properties such as translation invariance. The major weakness of orthonormal wavelet

[1] For an example of how the method could be used in time series analysis, consider Fig. I.1. If we wanted to estimate (or just look for) a nonlinear multivariate relationship between $y(t-2)$, $y(t-1)$ and $y(t)$, we would need to estimate a two-dimensional regression surface. Since data are limited, we could not use kernel methods; however, since Projection Pursuit restricts attention to one-dimensional projections, we could estimate a surface using Projection Pursuit methods.

[2] Much more is known about the statistical properties of the first stage estimates or g_0. Index models of the type g_0 are used in econometrics for 'average derivative' estimation of regression relationships (c.f., [99] [100]); perhaps the most impressive empirical application of such methods has been the verification of the 'Law of Demand' by Hildenbrand and others over the past decade [104] [98].

[3] Both the function $f(x)$ and the waveforms are assumed to lie in a Hilbert space.

representations is their lack of translation invariance.[4] Translation invariance of representations is considered very important in applications such as image processing.

Mallat and Zhang [131] propose an iterative procedure to determine the expansion Eq. (F.4). Mallat and Zhang start with a "dictionary" composed of a large set of potential functions e_i to include in a function expansion. At any stage I, Mallat and Zhang propose adding to the function expansion for $f(x)$ the waveform which is most correlated with the residual. In other words, we let the residual at stage n be y^n, then we select the waveform to include at stage n, e_n, as the waveform which solves:

$$\sup_{e_i \in S} | < y^n, e_i > | \tag{F.5}$$

where S is the "dictionary" set and each waveform e_i satisfies $||e_i|| = 1$. We update using the formula:

$$y^{n+1}(x) = y^n(x) - < y^n, e_n > e_n(x) \tag{F.6}$$

and we start by setting $y^0 = y$. Thus, for any $I > 0$:

$$y = \sum_{n=0}^{I-1} C_n e_n + y^I \tag{F.7}$$

where:

$$C_n = < y^n, e_n > . \tag{F.8}$$

For the e_i, Mallat and Zhang use Gabor functions which have attractive features in terms of time-frequency spectral estimation.[5]

[4] To address this problem, Mallat and Zhong (not Zhang) originally considered representing functions in terms of the maxima of redundant nonorthonormal wavelet decompositions [128] [127]. While this representation was translation invariant and found use in applications, it was proven (by counterexample) not to converge.

[5] Among all functions, Gabor functions achieve the minimum product of a measure of "variance" in the time and frequency domains. Specifically, let $||g|| = 1$ and define:

Gabor functions are Gaussian functions which are modulated (multiplied by complex exponentials):

$$e_i(x) = L_i\, e^{i\omega_i x}\, e^{-\frac{(x-c_i)^2}{2\sigma_i^2}} \tag{F.11}$$

where L_i is a normalization factor.[6] Despite their attractive features, Gabor functions are not orthonormal. Nevertheless, Mallat and Zhang were able to show that their expansion had many of the same properties as orthonormal series expansions (such as Fourier series expansions).

To show how this might happen, we return to the update formula Eq. (F.6) from which we wish to show that:

$$||y^{n+1}||^2 = ||y^n||^2 - |C_n|^2. \tag{F.12}$$

This follows because:

$$\begin{aligned} ||y^{n+1}||^2 &= < y^{n+1}, y^{n+1} > = < y^n - C_n e_n, y^n - C_n e_n > \\ &= ||y^n||^2 + |C_n|^2 ||e_n||^2 - C_n < e_n, y^n > - C_n^* < y^n, e_n > \\ &= ||y^n||^2 - |C_n|^2 \end{aligned} \tag{F.13}$$

Using Eq. (F.12) recursively, we have:

$$\mu_g = \int x|g(x)|^2 dx \tag{F.9}$$

$$\Delta_g^2 = \int (x - \mu_g)^2 |g(x)|^2 dx \tag{F.10}$$

then if $\hat{g}$ is the Fourier transform of g, the choice of g as a Gabor function achieves the minimum bound of $\Delta_g \Delta_{\hat{g}}$. For a proof, see ([39], pp. 56-60). The product that the Gabor function minimizes is the area of what is called the "Heisenberg box" which is a means of identifying what area of the time-frequency spectrum each waveform occupies. Such boxes look like the shaded boxes Fig. (VII.2) and Fig. (VII.3) in Ch. VII. For a detailed theoretical discussion, see [53].

[6] We recall from above that all e_i are normalized so that $||e_i|| = 1$.

$$||y||^2 = \sum_{n=0}^{I-1} |C_n|^2 + ||y^I||^2. \tag{F.14}$$

Using a theorem from Projection Pursuit regression (due to Jones [112]), Mallat and Zhang [131] are able to prove that $||y^I||^2 \rightarrow 0$ as $I \rightarrow \infty$. Thus, their procedure converges in the $\mathbf{L}^2$ sense that:

$$||y||^2 = \sum_{n=0}^{\infty} C_n^2. \tag{F.15}$$

This is exactly the sense in which orthonormal expansions such as Fourier series converge.

The reason Mallat and Zhang are able to adopt a theorem from Projection Pursuit Regression is that Matching Pursuit is a special case of Projection Pursuit in which the variable X is one dimensional and the functions $g_i(\beta_i'x)$ are assumed to be proportional to functions $e_i(x)$ in a fixed 'dictionary'. Thus, while Projection Pursuit regression allows any choice of the functions g_i, Matching Pursuit analysis restricts the functions g_i to be in a certain class as defined by a fixed 'dictionary' of functions.

How Autoregressive Pursuit Fits In

We now relate Matching Pursuit and Projection Pursuit to the method proposed in the thesis. In the thesis, we replace the basis functions $e_i(x)$ of Matching Pursuit analysis (in Eq. F.4) with model components which include a window function multiplied by an explanatory variable. Our set of data-dependent model components is hence analogous to the "dictionary" which Mallat and Zhang use in constructing their nonorthogonal series expansion. Our procedure in selecting model components to include in the analysis is also analogous to the procedure proposed by Mallat and Zhang.

Our procedure is similar to Projection Pursuit regression because we are interested in approximating a conditional expectation which depends on a set of explanatory variables. The difference is that we assume that the conditional expectation assumes a particular form in that it is linear in one set of explanatory variables but that the slope of this linear relationship depends on other explanatory variables (in the thesis, we focus on variables related to time).

Thus, there is a sense in which our analysis is more 'parametric' than either Projection Pursuit or Matching Pursuit because we restrict the regression function to be of a specific form. Another important difference between our approach and both Matching Pursuit and Projection Pursuit is that, in the thesis, we consider orthogonal projections whereas both Matching Pursuit and Projection Pursuit work with nonorthogonal projections. The main reason for focus on nonorthogonal projections in these settings is computational; the method of 'back-projection' is sometimes used in Projection Pursuit to orthogonalize and there have also been experiments with an orthogonal version of Matching Pursuit [161] [56].

Since the main reason for using nonorthogonal projections in these other methods is computational, it is useful to provide some details of alternative methods for computing estimates for Autoregressive Pursuit and how the nonorthogonal projections relate to our orthogonal projections. This helps relate nonparametric methods for high-dimensional data (such as can be used to estimate index models in econometrics) to ordinary regression analysis such as we use in the thesis.

Recipe for a Simplified Algorithm

We start with a set of K potential model components or regressors. We multiply each regressor by a number such that each regressor h_i satisfies:

$$\sum_{t=1}^{T} |h_i(t)|^2 = 1 \tag{F.16}$$

where t is time and T is sample size. The researcher chooses the set of model components. One possibility we suggest in the main text is a set of constant windows multiplied by a particular lag of the data. For instance, we can define a family of windows with width $\frac{T}{2^k}$ for $k = 0, ...m$ where $m \leq \log_2(T)$. For each k we might consider windows which start at all positions $t \in [1, \lambda_k T]$ for $\lambda_k < 1$[7] or we may want to consider a more restrictive set such as windows which start at $t = r\frac{T}{2^k} + 1$ for some $0 \leq r \leq r^* < 2^k$.[8] Once we have selected the windows with which to multiply the lagged data, we then choose normalizing constants so that Eq. (F.16) is satisfied.

Once the K model components are selected we define $y^0 = y$ where y is the data. We then run a univariate regression against each of the model components and pick the regression with the the highest r^2. We then call y^1 the residual from that regression. We then run univariate regressions of y^1 against each of the K model components and pick the regression with the highest r^2. We call the residual y^2 and then regress y^2 against the K model components. We call y^3 the residual of the regression of y^2 against the model component which produces the highest r^2. We then run a univariate regression of y^3 against each of the K model components and so on. We continue the procedure iteratively. To summarize, at any stage n, we make y^{n+1} the residual from the regression with the highest r^2. It follows from Theorem 1 (see Ch. IV) that this procedure converges to the orthogonal projection against the span of all included model components. It follows from the results in Ch. V that in any sample the rate of convergence is exponential in the number of iterations.

If we wish to implement a "general to specific" type testing strategy, we can estimate a large model and then eliminate one model component at a time and test for statistical significance. Alternatively, we can implement a procedure which tests whether to stop at each iteration of the estimation procedure. In Ch. VI, we define a new test for nonstationary processes which can be used in either case. One advantage

[7] The restriction $\lambda_k < 1$ is imposed to insure that all normalized windows are bounded as is required by the theory. Without this restriction, window functions could have arbitrarily small widths which would imply that the (mean-square) normalized window functions could be arbitrarily large. In another setting (in which generalized method of moments estimates are applied to a test for one-time structural change), Andrews [7] has recommended that $\lambda_k = 0.85$.

[8] Again, $r^* < 2^k$ so that normalized window functions are bounded.

of the approach here is that model components are selected step by step according to their explanatory power rather than some arbitrary criterion such as what lag they represent. After we have selected a model, we run a single multiple linear regression to estimate the coefficients.

Comments on the Simpfied Algorithm

The simplified algorithm is an effective computational method but it breaks down when there are large correlations among potential model components. We can eliminate many of the deficiencies of the simplified algorithm and retain many of its advantages by using a few fast subiterations. By subiteration, we mean a procedure which is performed as part of the main procedure at each iteration. The idea of subiterations is at any given iteration to construct a new set of model components which includes only the previously selected model components and then in a sense to perform a separate decomposition according to the simplified algorithm; we can run a few subiterations during the procedure if we wish to keep the computed projection at any stage 'close' to the multiple regression which would be implied by doing least squares estimation.

We can present a recipe for implementing such an approach. We again start with a set of K potential model components or regressors. We multiply each regressor by a number such that each regressor h_i satisfies:

$$\sum_{t=1}^{T} |h_i(t)|^2 = 1 \tag{F.17}$$

where t is time and T is sample size. The researcher chooses the set of model components (which may for instance be the family of model components with constant window functions referred to in the section above).

Once the K model components are selected we define $y^0 = y$ where y is the data. We then run a univariate regression against each other model components and pick the regression with the highest r^2 or the highest r^2 weighted by some user-selected weights.[9] We call the model component which maximizes our r^2 criterion h_1. We call the residual from this regression y^1. We then run a univariate regression of y^1 against each of the model components and pick the regression with the highest r^2 (or weight thereof). We call the model component in the regression which maximizes our r^2 criterion h_2. We call the residual $y^{2,0}$.

We now begin subiterations. We run univariate regressions of $y^{2,0}$ against h_1 and h_2 and select the regression with the highest r^2 (h_1 since the regression against h_2 has an r^2 of zero) and call the residual $y^{2,1}$. We run a univariate regression of $y^{2,1}$ against h_1 and h_2 and select the regression with the highest r^2 (h_2 since the regression against h_1 has a r^2 of zero) and call the residual $y^{2,2}$. We continue until $|y^{2,m+1} - y^{2,m}| < \epsilon$ and call the result $y^{2,*}$. We could also consider a fixed number m of subiterations and call the result $y^{2,*}$.

Once we have computed $y^{2,*}$, we continue the procedure by computing a third regular iteration. That is, we run a univariate regression of $y^{2,*}$ against each of the model components and pick the regression with the highest r^2 (or weight thereof). We call the model component in the regression which maximizes our r^2 criterion h_3. We call the residual $y^{3,0}$.

We now begin subiterations for the third iteration. We run univariate regressions of $y^{3,0}$ against h_1, h_2, h_3 (all previously selected model components) and select the regression with the highest r^2 and call the residual $y^{3,1}$. We run separate univariate regressions of $y^{3,1}$ against each of the previously selected model components. We select the regression with the highest r^2 and call the residual $y^{3,2}$. We continue until $|y^{3,m+1} - y^{3,m}| < \epsilon$ or we have reached a prespecified m and call the result $y^{3,*}$.

Thus, to summarize, at any iteration n, we start with data $y^{n,*}$, select a new model component h_{n+1} and then compute subiterations in which the only possible

[9] If we want to reduce the variance of estimates at the expense of some bias, we might want to put more weight on model components which involve more of the sample.

regressor variables are h_j for $j = 1, ...n + 1$. We continue subiterations indexed by m until: $|y^{n+1,m+1} - y^{n+1,m}| < \epsilon$. We call the result $y^{n+1,*}$. We then continue with an ordinary iteration to pick h_{n+2}.... We continue iterations (and the corresponding subiterations) as necessary.

The purpose of subiterations is to keep estimates close to the multiple regression (orthogonal projection) against all included model components. With a few subiterations at every iteration we can improve the properties of the simplified algorithm while retaining its computional flexibility. If we consider arbitrarily many subiterations for each iteration, it follows from Theorem 1 (Ch. 4) that the procedure converges to an orthogonal projection against all previously selected model components which is the procedure proposed in the thesis.[10] Thus, subiterations provide the conceptual link between Projection Pursuit and Matching Pursuit which are based on *nonorthogonal projections* and our main procedure which is based on *orthogonal projections.*

To understand how subiterations work in practice, let us consider an example. We consider what happens when we restrict analysis (for illustrative purposes) of a second order autoregressive model to the two correct model components. While we assume that there are only two model components in the analysis, the results would not change as long as we picked the correct model components in iteration 1 and 2. We consider the model:

$$y(t) = 0.16\,y(t-1) + 0.64\,y(t-2) + u(t) \tag{F.18}$$

where $u(t)$ is white noise. We use the simplified algorithm to decompose a 512 sample realization of the time series. On the first iteration the algorithm chooses the second lag (with a window over the whole sample) with an estimate of 0.74 and on the second

[10] Let $\mathcal{H}_k$ be the space spanned by the model components included up to iteration k. Since $\mathcal{H}_k$ is a Hilbert space, Theorem 1 proves convergence to a projection against all included model components. If the span of the model components is $\mathcal{H}_k$, subiterations thus converge to a projection against the space $\mathcal{H}_k$.

iteration it chooses the first lag with an estimate of 0.137. We now show how the subiterations work to correct the error made in the main iterations due to correlations among the selected model components. When the correction terms summized in Table F.1 are added in, we have estimates of 0.6384 on the second lag and 0.188354 on the first lag which results in a substantial improvement in the quality of the estimate on the second lag.

Coeff.	Parameter Est.	Iter.	Subiter.	Begin	End	Lag
25.404963	0.736441	1	0	0	512	2
4.738204	0.137347	2	0	0	512	1
-2.465702	-0.071476	2	1	0	512	2
1.283135	0.037195	2	2	0	512	1
-0.667735	-0.019356	2	3	0	512	2
0.347485	0.010073	2	4	0	512	1
-0.180829	-0.005242	2	5	0	512	2
0.094102	0.002728	2	6	0	512	1
-0.048970	-0.001420	2	7	0	512	2
0.025483	0.000739	2	8	0	512	1
-0.013262	-0.000384	2	9	0	512	2
0.006901	0.000200	2	10	0	512	1
-0.003591	-0.000104	2	11	0	512	2
0.001869	0.000054	2	12	0	512	1
-0.000973	-0.000028	2	13	0	512	2
0.000506	0.000015	2	14	0	512	1
-0.000263	-0.000008	2	15	0	512	2
0.000137	0.000004	2	16	0	512	1
-0.000071	-0.000002	2	17	0	512	2

Table F.1: Two iterations of the general form of the algorithm on an AR2 model. The algorithm picks up the correct lag lengths and model components and subiterations correct the original parameter estimates for bias.

APPENDIX G

Auxiliary Results on Time-Frequency Spectral Estimation with Autoregressive Pursuit

We consider a time series y described by the equation:

$$T^{-1}y = \epsilon \tag{G.1}$$

where ϵ is a vector of uncorrelated random variables and T^{-1} is a matrix.

In the thesis, we have developed a way to estimate the (operator) inverse of T. We let $T^{-1} = S$. Then $C^{-1} = S^*S$ and, up to a scale factor, C is the covariance matrix for the time series. To find the eigenvectors of C, we need only to find the eigenvectors of C^{-1}

Theorem 14 *Let C be a matrix. The eigenvectors C are the same as those of C^{-1}*

Proof: Consider the eigenvalue equation:

$$C\psi_j = \lambda_j\psi_j \tag{G.2}$$

Multiply both sides by C^{-1} and divide by λ_j:

$$\frac{1}{\lambda_j}\psi_j = C^{-1}\psi_j \tag{G.3}$$

Thus, the eigenvectors (eigenfunctions) of C are the same as those for C^{-1} and the eigenvalues for C are the inverse of those for C^{-1} ∎

We consider the time-varying stochastic process:

$$y(t) = g(t)y(t-1) + \epsilon(t) \tag{G.4}$$

Here, the operator S is $1 - g(t)L$. The matrix representation of S is:

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -g(t_1) & 1 & 0 & 0 & \dots & 0 \\ 0 & -g(t_2) & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \vdots & -g(t_{n-1}) & 1 \end{pmatrix} \tag{G.5}$$

This implies C^{-1} is of the form:

$$C^{-1} = \begin{pmatrix} 1+g(t_1)^2 & -g(t_1) & 0 & \dots & 0 \\ -g(t_1) & 1+g(t_2)^2 & -g(t_2) & \dots & 0 \\ 0 & -g(t_2) & 1+g(t_3)^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & -g(t_{n-1}) \\ 0 & 0 & 0 & -g(t_{n-1}) & 1 \end{pmatrix} \tag{G.6}$$

When the $g(t_i)$ are time-invariant, the limiting eigenfunctions of the covariance matrix are sine and cosine waves. It is useful to examine what small-sample effects are present for the analysis of some of the short time series available in economics. In Fig. G.1 we show the 10th largest eigenvector of a covariance matrix for a time series of length 64 where the first order coefficient is: $g(t) = 0.4$; this and the other eigenfunctions appear close to the theoretical time-invariant limiting eigenfunctions. In Fig. G.2 we show the $13th$ largest eigenvector for the covariance matrix of a

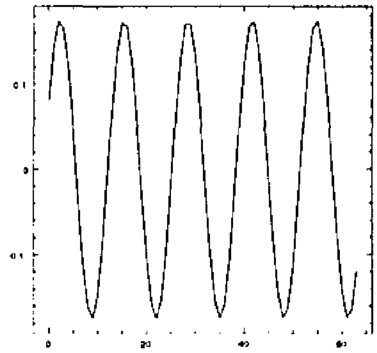

Figure G.1: An eigenfunction of a short stationary autoregressive process.

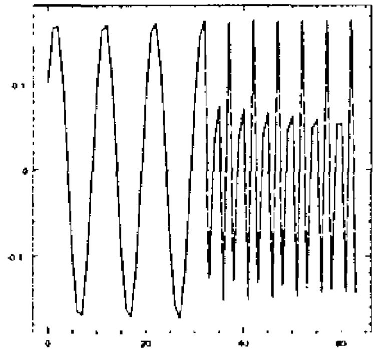

Figure G.2: An eigenfunction of a short nonstationary autoregressive process.

switching autoregressive process of length 64 where the autoregressive parameter switches from 0.4 to −0.4 midway through the time series; there appears to be a clear delineation of the break point so that the eigenvector analysis would appear to be meaningful. In Fig. G.3, we show the second largest eigenvector of a smoothly varying autoregressive process (such as we introduced in Ch. III); this as well as the other eigenvectors look like the wavelet functions used in Ch. II and Ch. VII.

Nonstationary Spectral Estimation

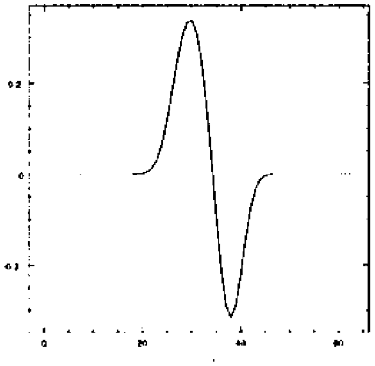

Figure G.3: An eigenfunction of a smoothly varying autoregressive process.

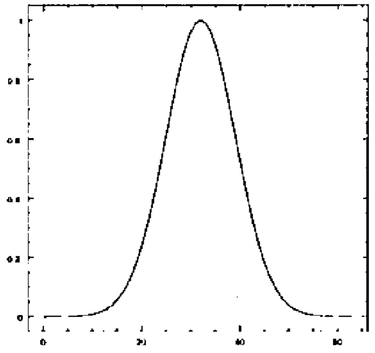

Figure G.4: The first order lag coefficient for the smoothly varying autoregressive process.

Earlier, we defined a nonstationary spectrum in terms of a sum of Cohen's class distributions of eigenvectors or eigenfunctions of the covariance matrix:

$$S(t,\omega) = \sum_i C_i^2 C_{e_i}(t,\omega) \tag{G.7}$$

where e_i are the eigenvectors (eigenfunctions) of the covariance matrix(operator) and C_{e_i} means the Cohen's class distribution of e_i. It is useful to examine some properties of this spectral estimator.

Theorem 15 *When the Cohen's class kernel is a Wigner-Ville distribution, the spectral estimates from the nonstationary spectral estimator:*

$$S(t,\omega) = \sum_i C_i^2 C_{e_i}(t,\omega) \tag{G.8}$$

are the same as those of the Wigner-Ville distribution of the local autocovariance kernel for a discrete time series (c.f. Eq. VII.24):

$$W(t,\omega) = \sum_{s\in\mathbb{Z}} K\left([t+\frac{s}{2}],[t-\frac{s}{2}]\right) e^{-2\pi i\omega s} \tag{G.9}$$

where:

$$K(t+s,t-s) = E\left(X(t+s)X^*(t-s)\right) \tag{G.10}$$

is the local covariance kernel of the random function X and the brackets denote integer part.

Proof:

We note that if e_k are the eigenvectors of the covariance kernel K of the data, we have:

$$K(s,t) = \sum_k \lambda_k e_k(s) e_k^*(t) \tag{G.11}$$

which is linear in the kernels:

$$E_k(s,t) = e_k(s)e_k^*(t). \tag{G.12}$$

The Wigner-Ville distribution of the eigenvector e_k of the covariance kernel K is:

$$W_{e_k}(t,\omega) = \sum_s E_k\left([t+\frac{s}{2}],[t-\frac{s}{2}]\right) e^{-2\pi i\omega s} \tag{G.13}$$

But, by Eq. (G.11) and Eq. (G.12), we have:

$$K\left([t+\frac{s}{2}],[t-\frac{s}{2}]\right) = \sum_k \lambda_k E_k\left([t+\frac{s}{2}],[t-\frac{s}{2}]\right) \tag{G.14}$$

Thus, the Wigner-Ville distribution for the covariance kernel is:

$$\begin{aligned} W(t,\omega) &= \sum_s \sum_k \lambda_k E_k\left([t+\frac{s}{2}],[t-\frac{s}{2}]\right) e^{-2\pi i\omega s} \\ &= \sum_k \lambda_k W_{e_k}(t,\omega) \end{aligned} \tag{G.15}$$

Therefore, in the simple case where the Cohen's class distribution used to decompose the eigenvectors is the Wigner-Ville distribution, our spectral estimator is equivalent to the direct Wigner-Ville distribution of the covariance kernel. ■

In terms of the covariance matrix K in Eq. (G.16), it is useful to explain what terms enter the Wigner-Ville distribution. At time $t = 1$, for instance, the term corresponding to $s = 0$ is $K(1,1)$ and the term corresponding to $s = 1$ is $K(1,0)$. Likewise, the term corresponding to $s = -1$ is $K(0,1)$, the term for $s = 2$ is $K(2,0)$ and for $s = -2$ is $K(0,2)$.

$$K = \begin{pmatrix} K(0,0) & K(0,1) & K(0,2) & \dots & K(0,T-1) \\ K(1,0) & K(1,1) & K(1,2) & \dots & K(1,T-1) \\ K(2,0) & K(2,1) & K(2,2) & \dots & K(2,T-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & K(T-2,T-1) \\ K(T-1,0) & K(T-1,1) & K(T-1,2) & \vdots & K(T-1,T-1) \end{pmatrix} \quad \text{(G.16)}$$

Thus, at each t, the sequence we Fourier transform is symmetric about 0 by symmetry of the matrix Eq. (G.16); thus, our local spectral estimates are *real.* However, these local spectral estimates may not be positive.

Remark 3 *Our definition of a local Wigner spectrum for a discrete series is different from the standard definition in the literature (c.f., Eq. VII.24) because that definition does not make sense at a minimum for $MA(1)$ processes. The definition we use is analogous to that developed previously in [37] (c.f., [110]).*

We can now compare our estimates to those of Tjostheim [195].Given a model:

$$y(t) = \sum_{j\geq 0} c_j(t)\epsilon(t-j) \quad \text{(G.17)}$$

$$E(\epsilon(t-j)\epsilon(t-k)) = \delta_{j-k}, \quad \text{(G.18)}$$

Tjostheim's spectral estimate is:

$$S(t,\omega) = \frac{1}{2\pi}|\sum_{j=0}^{\infty} c_j(t)e^{-i\omega j}|^2 \quad \text{(G.19)}$$

Tjostheim's estimator takes the Fourier transform of a column of the covariance matrix and imposes that negative autocovariances $\gamma(-k) = \gamma(k)$. Thus, at $t = 1$, Tjostheim would use elements $K(0,1)$ for $s = 1$, $K(1,1)$ for $s = 0$, $K(2,1)$ for $s = 1$

and so on. However, for any fixed column of a matrix, Tjostheim's symmetry assumption (for instance, that $K(0,1) = K(2,1)$) can only be imposed if the time series is covariance stationary, which is exactly what Tjostheim wanted to avoid assuming.

APPENDIX H

Inconsistency of Model Choice

In this appendix, we provide a simple example in which the procedure we have chosen selects the wrong model component asymptotically but in which the coefficient estimate on the wrong model component converges to zero in probability.

We consider a time series defined by the switching autoregressive model introduced in Ch. III:

$$y(t) = \beta(t)\,y(t-1) + u(t) \tag{H.1}$$

$$\beta(t) = \begin{cases} \beta_0 & t \leq 0 \\ -\beta_0 & \text{otherwise} \end{cases} . \tag{H.2}$$

As usual, we assume that:

$$u(t) \sim N(0, \sigma^2) \tag{H.3}$$

and that we have observations on the process on the interval $[-L+1, L]$ so that the total sample size for y is $T = 2L$.

We consider only three potential model components:

$$h_1(t) = \begin{cases} \frac{y(t-1)}{\sqrt{\sum_{t=-L+1}^{0} y_{t-1}^2}} & t \leq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{H.4}$$

$$h_2(t) = \begin{cases} \frac{y(t-1)}{\sqrt{\sum_{t=1}^{L} y(t-1)^2}} & t > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{H.5}$$

$$h_3(t) = \frac{y(t-2)}{\sqrt{\sum_{t=-L+1}^{L} y(t-2)^2}} \tag{H.6}$$

Each model component thus satisfies $||h_i|| = 1$. Our first result is that as $L \to \infty$, there are cases in which the model component h_3 is chosen instead of the correct model components h_1 and h_2.

Theorem 16 *If $|\beta_0| > \frac{\sqrt{2}}{2}$, the probability of choosing h_3 on the first iteration converges to* 1. *In the converse case the probability of choosing h_3 converges to zero. Given that h_1 or h_2 is selected on the first iteration, h_3 will be selected with probability zero asymptotically on the second iteration.*

Proof: As $L \to \infty$,

$$\frac{1}{\sqrt{2L}}\sqrt{\sum_{t=-L+1}^{L} y(t-2)^2} \to \frac{\sigma}{\sqrt{1-\beta_0^2}} \tag{H.7}$$

by the continuous mapping theorem and the fact that β_0^2 is equal for $t > 0$ and $t \leq 0$. Similarly:

$$\frac{1}{\sqrt{L}}\sqrt{\sum_{t=-L+1}^{0} y(t-1)^2} \to \frac{\sigma}{\sqrt{1-\beta_0^2}} \tag{H.8}$$

$$\frac{1}{\sqrt{L}}\sqrt{\sum_{t=1}^{L} y(t-1)^2} \to \frac{\sigma}{\sqrt{1-\beta_0^2}} \tag{H.9}$$

We define:

$$\sigma_\beta = \frac{\sigma}{\sqrt{1-\beta_0^2}} \tag{H.10}$$

We next compute the $< y, h_i >$ which we define as:

$$< y, h_i >= \frac{1}{\sqrt{L}} \sum_t y(t) h_i(t) \tag{H.11}$$

We have:

$$< y, h_1 >= \frac{\frac{1}{L} \sum_{t=-L+1}^{0} y(t) y(t-1)}{\frac{1}{\sqrt{L}} \sqrt{\sum_{t=-L+1}^{0} y(t-1)^2}} \tag{H.12}$$

We note:

$$\frac{1}{L} \sum_{t=-L+1}^{0} y(t) y(t-1) \rightarrow \beta_0 \sigma_\beta^2 \tag{H.13}$$

Thus:

$$C_1 =< y, h_1 >\rightarrow \beta_0 \sigma_\beta \tag{H.14}$$

Similarly:

$$C_2 =< y, h_2 >\rightarrow -\beta_0 \sigma_\beta \tag{H.15}$$

Finally, by the same logic:

$$C_3 =< y, h_3 >\rightarrow \sqrt{2} \beta_0^2 \sigma_\beta \tag{H.16}$$

where β_0^2 appears because it is the second order autocovariance of y. We next compute the asymptotic variances of the C_i. We note that because of our normalization factor in the h_i:

$$L \, \mathrm{Var} C_1 \rightarrow \sigma^2 \tag{H.17}$$

$$L \, \mathrm{Var} C_2 \rightarrow \sigma^2 \tag{H.18}$$

$$L\,\mathrm{Var}C_3 \to \sigma^2 \tag{H.19}$$

Thus, each of the C_i are asymptotically normal with the same asymptotic variances.

Given that the variances of the C_i converge to zero, we pick C_3 if $\sqrt{2}\beta_0^2 > |\beta_0|$ or if $\sqrt{2}\beta_0 > 1$. Thus, model choice is inconsistent at the first stage of the algorithm asymptotically if $|\beta_0| > \frac{\sqrt{2}}{2}$.

Otherwise, we pick either h_1 or h_2 first. Suppose we pick h_1 or h_2 first, then we must prove that the other model component will be selected next. After subtracting off the projection against the selected model component, the residual on one half of the time series is white noise. Suppose we have selected h_1. Then we have that:

$$C_3 =< y, h_3 >\to \sqrt{2}\beta_0^2\sigma_3 \tag{H.20}$$

where we have not yet defined σ_3. Noting that:

$$\frac{1}{2L}\sum_{t=-L+1}^{0} y(t-2)^2 + \sum_{t=1}^{L} y(t-2)^2 \to \frac{\sigma^2}{2} + \frac{\sigma^2}{2(1-\beta_0^2)} \tag{H.21}$$

we have that:

$$\sigma_3^2 = \frac{\sigma^2}{2} + \frac{\sigma^2}{2(1-\beta_0^2)} \tag{H.22}$$

We also have that:

$$C_1 =< y, h_1 >\to 0 \tag{H.23}$$

since for the first half of the time series $\beta_0 = 0$. Likewise:

$$C_2 =< y, h_2 >\to -\beta_0\sigma_\beta \tag{H.24}$$

We next compute the asymptotic variances for the C_i. We have:

$$L\,\mathrm{Var}C_1 \to \sigma^2 \tag{H.25}$$

since the first half of the time series is noise,

$$L\,\mathrm{Var}C_2 \to \sigma^2 \tag{H.26}$$

since the model is correctly selected, and:

$$L\,\mathrm{Var}C_3 \to \sigma^2. \tag{H.27}$$

Thus, the variances of the C_i do not change and they converge to zero as sample size gets large. Therefore, given we selected C_1 on the first iteration, we select C_2 as long as:

$$C_2^2 > C_3^2 \tag{H.28}$$

or if $\lambda > 0$ where:

$$\lambda = 0.5\beta^2 \frac{(-2 + 2\beta^2 - \beta^4)}{\beta^2 - 1} \tag{H.29}$$

which requires that: $-2 + 2\beta^2 - \beta^4 < 0$. The equation has complex roots and when $|\beta| < 1$, $2\beta^2 - \beta^4 < 2$ so that h_2 is always selected. ∎

We next show that if we incorrectly select the model component h_3, the regression coefficient on h_3 converges to 0 after we select h_1 and h_2 and iterate an infinite number of times.

Theorem 17 *The estimated second lag relationship converges to zero provided h_1 and h_2 are included in the analysis.*

Proof: We consider the equation:

$$y(t) = \beta_1 1_{t\leq 0} y(t-1) + \beta_2 1_{t>0} y(t-1) + \beta_3 y(t-2) + \epsilon(t) \tag{H.30}$$

The least squares coefficient for β_2 must converge to zero because $y(t-2)$ has no additional explanatory power for y, but it is helpful to review this point analytically.

For our case, the least squares estimates are given by the vector $W^{-1}z'y$ where in the large L limit we have:

$$W = 0.5 \begin{pmatrix} 1 & 0 & \beta_0 \\ 0 & 1 & -\beta_0 \\ \beta_0 & -\beta_0 & 2 \end{pmatrix} \tag{H.31}$$

and:

$$z = \frac{1}{T} \begin{pmatrix} 1_{t\leq 0} y(t-1) \\ 1_{t>0} y(t-1) \\ y(t-2) \end{pmatrix}. \tag{H.32}$$

Using partitioned matrix formulas the inverse of W has a bottom right element of:

$$W^{-1}_{33} = 2(2 - \begin{pmatrix} \beta_0 \\ -\beta_0 \end{pmatrix}' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ -\beta_0 \end{pmatrix})^{-1} = 2(2 - 2\beta_0^2)^{-1} \tag{H.33}$$

Similarly:

$$W^{-1}_{31} = -2\frac{\beta_0}{2 - 2\beta_0^2} \tag{H.34}$$

$$W^{-1}_{32} = 2\frac{\beta_0}{2 - 2\beta_0^2} \tag{H.35}$$

The explicit formula for the least squares estimate of β_2 is

$$\hat{\beta}_2 \rightarrow 0.5\beta_0 W^{-1}_{31} - 0.5\beta_0 W^{-1}_{32} + \beta_0^2 W^{-1}_{33} \tag{H.36}$$

Plugging in our expressions we have:

$$\hat{\beta}_2 \to -\frac{\beta_0^2}{2-2\beta_0^2} + \frac{\beta_0^2}{2-2\beta_0^2} = 0. \tag{H.37}$$

which is what was we wished to show. ∎

APPENDIX I

A Result on Maxima of Normal Variables

Theorem 18 *Let $\{\epsilon_i\}_{i=1}^{N}$ be N normally distributed random variables with a maximum variance of* 1 *and define $M_N = \max[\epsilon_1, \epsilon_2, ...\epsilon_N]$ then:*

$$P\left(|M_N| > \sqrt{2\lambda \log N}\right) \le \gamma N^{1-\lambda} \tag{I.1}$$

for any $\lambda > 1$ and $\gamma = \frac{1}{\sqrt{\pi \log 2}}$.

Proof: This proof comes from ([114], p.218-219):

For $u > 1$:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} \frac{\left(1 - e^{-u}\right) e^{\frac{-u^2}{2}}}{u} \le P(\epsilon \ge u) \le \frac{1}{\sqrt{2\pi}} \frac{e^{\frac{-u^2}{2}}}{u} \tag{I.2}$$

since:

$$\begin{aligned} P(\epsilon \ge u) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \int_0^\infty e^{-uy} e^{\frac{-y^2}{2}} dy \\ &\ge \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \int_0^1 e^{-uy} e^{-\frac{1}{2}} dy \\ &= \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}} e^{\frac{-u^2}{2}} \frac{1 - e^{-u}}{u} \end{aligned} \tag{I.3}$$

and:

$$P(\epsilon \ge u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \int_0^\infty e^{-uy} e^{\frac{-y^2}{2}} dy$$

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \int_0^\infty e^{-uy} dy \\
&= \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \frac{1}{u} \qquad \text{(I.4)}
\end{aligned}
$$

When the ϵ_i are normally distributed random variables, not necessarily independent, we have for any $\lambda > 1$:

$$P(|\epsilon_i| > \sqrt{2\lambda \log N}) \leq \frac{1}{\sqrt{\pi \log 2}} N^{-\lambda} \qquad \text{(I.5)}$$

Thus:

$$P(|M_N| > \sqrt{2\lambda \log N}) \leq \frac{1}{\sqrt{\pi \log 2}} N^{1-\lambda} \qquad \text{(I.6)}$$

for any $\lambda > 1$. Therefore, as $N \to \infty$,

$$P(\lim_{N \to \infty} \frac{|M_N|}{\sqrt{2\lambda \log N}} > 1) = 0. \qquad \text{(I.7)}$$

for any $\lambda > 1$

∎

We need only this result in the thesis, but much more can be said when the random variables have a known correlation structure. For a survey of results on extrema see [118]. For a survey of results on extrema of the normal distribution, see ([160], ch. 8).

APPENDIX J

Note on Computation of Estimates

In this Appendix, we discuss two useful approaches to computation of estimates. The first approach works in the time domain and relies on a choice of normalization factor to reduce computational complexity.

Since at any stage of the analysis, we need to compute many simple regressions, it is useful to have a computational trick with which to compute these regressions and choose the maximal r^2 quickly. This trick can be explained as follows. The simple regression equations we wish to estimate at each iteration n are:

$$y^n(t) = \beta h_i(t) + \epsilon_4(t) \tag{J.1}$$

where $\epsilon_4(t)$ is a residual term which need not be independently and identically distributed. At each stage of the procedure, we need to compute the r^2 for a regression of the type Eq. (J.1) for all model components in the set of model components we are considering.

For Eq. (J.1):

$$\hat{\beta} = \frac{\frac{1}{T}\sum_{t=1}^{T} h_i(t) y^n(t)}{\frac{1}{T}\sum_{t=1}^{T} |h_i(t)|^2} \tag{J.2}$$

Using Eq. (J.1), the explained sum of squares is:

$$\sum_{t=1}^{T} \hat{\beta}^2 h_i^2 = \frac{\left[\sum_{t=1}^{T} h_i(t) y^n(t)\right]^2}{\sum_{t=1}^{T} h_i^2(t)} \tag{J.3}$$

Our goal to avoid computing the denominator at every iteration so we define a new model component: $z_i(t)$ to be:

$$z_i(t) = \frac{h_i(t)}{\sqrt{\sum_{t=1}^{T} h_i^2(t)}} \tag{J.4}$$

Plugging in Eq. (J.4) to Eq. (J.3), we have that the explained sum of squares is:

$$\sum_{t=1}^{T} \hat{\beta}^2 h_i^2 = \left[\sum_{t=1}^{T} z_i(t) y^n(t)\right]^2 \tag{J.5}$$

so that if we wish to find the model component which maximizes r^2, we work with model components normalized as in Eq. (J.4). This way we can choose the best model components quickly by computing the sum on the right hand side of Eq. (J.5) one model component at a time.

Moving back to the original regression equation Eq. (J.1) and using the definition of z_i we have:

$$\begin{aligned} y^n(t) &= \beta z_i(t) \sqrt{\sum_{t=1}^{T} h_i(t)^2} + \epsilon_4(t) \\ &= \alpha z_i(t) \end{aligned} \tag{J.6}$$

where $\alpha = \beta\sqrt{\sum_{t=1}^{T} h_i(t)^2}$. Thus, we can recover estimates for β from estimates of α by dividing by the normalizing factor: $N_i = \sqrt{\sum_{t=1}^{T} h_i(t)^2}$.

This particular method of computing regression coefficients and r^2 is especially valuable because it enables us to use fast convolution algorithms such as the Fast Fourier Transform (FFT). Let us explain. At any stage n, we want to compute the sum on the right hand side of Eq. (J.5). Since our model components h_i are composed of a window function times a given lag l_i, $y(t - l_i)$, we can construct for each lag the function $w_n^l(t) = y^n(t)y(t - l_i)$. Since we ordinarily consider window functions at many possible locations, we can compute the regression coefficient for a given type of window function centered at any given location by:

(1) computing the Fourier transform of the function $w_n^l(t)$ and multiplying with the Fourier transform of the appropriate window function family then inverse Fourier transforming the product.

(2) at each location dividing by the appropriate normalizing factor which also can be computed at the beginning of the analysis through use of the Fast Fourier transform.

Appendix M contains estimates of the amount of computation and storage required for this particular approach; this shows formally that it is indeed possible to implement our approach on personal computers and workstations.

APPENDIX K

Notes on Similar Procedures

The model selection procedure we describe in Ch. II is iterative. As a result, we may not be selecting the most parsimonious representation. Why? Given the choice of a set of potential model components, we must devise a practical way of determining estimates of the univariate time series model. The standard time series approach of adding one lag at a time or examining all possible lags will not do when we need to examine an enormous number of possibilities; choosing N regressors (where N is a number less than sample size T) from P possibilities (where P might be K, the number of model components) is ordinarily not a feasible computational procedure for P or N large. For instance, if P is a constant *multiple* of N (greater than 1), the number of regressions we must compute grows exponentially in N.[1]

Even though we may choose regression equations which are not optimal in terms of parsimony, it is a general theoretical result that the procedure still produces estimates which are equivalent approximations of the regression relationship. In other words, the procedure converges (as the number of model components in the analysis increases) to a projection against the same space as would an exhaustive search of

[1] Use Stirling's approximation for the factorial function in computing the number of combinations. In computer science, an NP-complete problem is a problem for which there exists no algorithm to compute a solution in polynomial time [111]. We have reason to believe that an approach based on selecting the best possible regression is computationally intractable in the sense of being NP-complete. For Matching Pursuit analysis of functions (see Appendix F), Davis [55] has proven formally that under some technical assumptions the procedure is NP-complete.

all possible models. Convergence to a projection does not imply consistent estimates unless further conditions are met which are discussed in Ch. IV.

We now explain some potential alternative procedures similar to the one developed in the thesis and discuss why we have chosen the particular approach outlined in Chapter II. Recall that we have chosen to continue at stage j of the estimation procedure by running simple regressions of the residual y^{j-1} on each of the remaining model components to select the best model component h^j. We then run a multiple regression of y against the model components h^k for $k \leq j$ and call the residual from the regression y^j.

There are a variety of different procedures which also produce statistically meaningful results which we could have advocated. For instance, we could have, for each model component not yet included in the model, run separate multiple regressions of y^{j-1} against the potential new model component and h^k for $k \leq j$. In this case, we would choose to include in the model, the model component which maximizes the R^2 from the multiple regression (or some weight thereof). The reason we did not implement this approach is that it requires significantly more computations; a direct implementation requires approximately K times (where K is the number of model components) more computations than our approach; since K is large, this approach does not yet seem practical for exploratory data analysis applications.

Since we must run linear regressions at each stage of our approach and each linear regression takes $O(N^3)$ operations where N is the number of regressors, the total order of computations from the regressions alone is $O(N^4)$.[2] Thus, our procedure may be relatively expensive computationally for models with large numbers of regressors.

[2] We might consider using block matrix inversion algorithms to reduce the order of computations, but if used in an iterated fashion, these algorithms are numerically ill-conditioned. We note that we could compute regressions in a theoretically equivalent manner by adding up coefficients from regressions of the residual at each stage on all previously selected model components; this is equivalent to Gram-Schmidt orthogonalization of the regressors and may result in numerical instabilities so that it is not recommended.

Fortunately, there is also another procedure which is much faster and is applicable to cases where correlations between selected model components are weak as seems to be the case with large financial datasets where there is not much temporal dependence in the data.[3] This method simply computes the residual at any stage by a simple regression against the selected model component instead of using the multiple regression. This procedure is introduced and reviewed in detail in Appendix F. Since there seem to be some statistical disadvantages to this procedure, we show in Appendix F that there is a fast way to improve this procedure. This faster procedure is well-defined in a technical sense in that it converges to a projection against the span of all potential model components.[4]

Therefore, our approach is a compromise which is designed so as to be implementable on modern workstations and personal computers and so as to be useable as an exploratory data analysis tool. While we have selected one particular approach to focus on, the researcher should be aware of alternative approaches in case of unusual datasets or computational facilities.

[3] This observation is based on numerical experiments, some of which are reported in an earlier paper [151].

[4] Although we show convergence of another method in Ch. IV, our remarks in the proof show how to modify it to handle the special case in Appendix F.

APPENDIX L

Note on Kernel Regression vs. Adaptive Regression

In this appendix, we address the basic question of why we want to use a complex adaptive regression model such as we have developed in the thesis instead of something more simple such as a local weighted moving average estimate.

A simple approach to estimating time-varying models is to simply run a weighted or rolling local least squares model to estimate parameters. Consider for simplicity a *fractional* autoregressive model:[1]

$$y(t) = b_1\left(\frac{t}{T}\right) y(t-1) + \epsilon(t) \tag{L.1}$$

and $b_1(s)$ $(s = \frac{t}{T})$ is a measurable function on the interval $[0,1]$).

We then consider the problem of minimizing the locally weighted sum of squares:

$$S\left(b_1\left(\frac{k}{T}\right)\right) = \frac{1}{2Th}\sum_{t=1}^{T}\sum_{k=1}^{T} K\left(\frac{k-t}{Th}\right)\left(y(t) - b_1\left(\frac{k}{T}\right)y(t-1)\right)^2. \tag{L.2}$$

where the function $K(s) \in C^0$ satisfies:

$$\int_{-1}^{1} K(s)\,ds = 1, \tag{L.3}$$

is positive, and has $K(1) = K(-1) = 0$.

[1] The model is named a fractional model because the autoregressive parameter depends on the fraction of time rather than time.

The least squares estimate for $b_1(s)$ is:

$$\hat{b}_1(s) = \frac{\frac{1}{Th}\sum_{t=1}^{T} K(\frac{s-\frac{t}{T}}{h})y(t)y(t-1)}{\frac{1}{Th}\sum_{t=1}^{T} K(\frac{s-\frac{t}{T}}{h})y(t-1)^2} \tag{L.4}$$

Given that we can use an estimator of a local autoregressive coefficient by a simple local least squares method which averages over more and more points local to a given point $s \in (0,1)$ as the size of the sample increases, why do we need to consider something else?

The main issue is conceptual. When we run a local least squares regression, we implicitly put a window of a given size on the data. Since the properties of the data or components of a model may be constant over an interval much larger or smaller than that of the chosen window width, there is a sense in which we lose statistical precision if the true model has components which evolve over time intervals different than the chosen window width.

The idea of our method is to use a parametric procedure which considers linear combinations of potential model components. As a result, our method lies somewhere between the use of local least squares estimates and ordinary least squares autoregressive estimates. Local least squares are inaccurate if the true model is a linear time-invariant autoregressive process whereas autoregressive estimates are inaccurate if the true model is evolving over time. Since locally we do not have a large amount of data, our intuition is to try to retain a parametric focus as much as possible because of the superior statistical properties of parametric models.

APPENDIX M

Computational Requirements

It is helpful to review issues of computational requirements for a special case of window functions. We consider the family of window functions defined by Gaussians and their derivatives.

$$g_{d,b,s}(t) = \frac{\partial^d}{\partial t^d} e^{-\frac{(t-b)^2}{2s^2}} \tag{M.1}$$

We assume periodicity. We let d be an integer which runs from 0 to D. We assume that a time series has length N and b assumes values at the N points of the time series. We include M lags in the analysis and consider values of s at:

$$s^2 = C2^{j+\frac{k}{V+1}} \tag{M.2}$$

where k runs from 0 to V and j runs from 1 to Q. We expect Q to increase slowly with N, perhaps with its base 2 logarithm.

For such a model, the univariate storage requirements are $O(N[(V+1)\log_2 N+K])$ where K is a constant. Therefore, the storage requirements grow slowly with N and do not present a problem for implementation on modern workstations or personal computers.

If we use the FFT for computations, computational requirements are at each iteration:

$$O(M(V+1)(D+1)N(\log_2 N)Q) \tag{M.3}$$

This shows that computational requirements are quite reasonable so that implementation for a single iteration on modern workstations or personal computers would not seem to be problematic. The computational bottleneck comes from the least squares regressions which if we use L regressors require computations of order L^4 to do all the iterations. As long as the number of regressors L is small, computational requirements are reasonable. However, for large models, the method presented in Appendix F requires only of order N computations at each iteration and also converges.

APPENDIX N

Finite Average Variance Spaces

In this appendix, we provide both technical and intuitive background behind the use of the inner product:

$$\begin{aligned}\mathcal{E}\left(f\,g\right) &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\int f(x_t,v_t)g(x_t,v_t)d\mu_t(x) \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E\left(f(x_t,v_t)\,g(x_t,v_t)\right) \qquad \text{(N.1)}\end{aligned}$$

The technical issue is that there is an infinite normalization factor in Eq. (N.1) because $f(x_t, v_t)$ and $g(x_t, v_t)$ may have infinite 'energy' or total squared variation on the real line.

Let us consider the example of any stationary stochastic process over the interval $[-\infty, \infty]$ or $[0, \infty]$. Since the process has infinite total variance with respect to t, the samples from a random process are not measurable sequences or functions. For instance, if we take the Fourier transform of a discrete data series of length $2T$, we have:

$$x(t) = \sum_{j=0}^{\infty} c_j e_j(t) \qquad \text{(N.2)}$$

where:

$$e_j(t) = e^{2\pi i\omega_j t} \qquad \text{(N.3)}$$

$$\omega_j = \frac{k}{2T+1} \tag{N.4}$$

$$c_j = \frac{1}{2T+1} \sum_{t=-T}^{T} x(t) e_k^*(t) \tag{N.5}$$

where the data x is periodic with period $2T$. Since the series is not square summable, if we were to run any infinite regression with Fourier waveforms as regression variables, Eq. (N.2) would not hold in the sense that the sum of squares of both sides are equal.

The reason is that if $x(t) \neq 0$ for any t, it does not follow that:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} |x(t)|^2 = 0 \tag{N.6}$$

which is required for $x(t)$ to be in a normed space in the classical sense. Thus, measurability with respect to time on the infinite interval is a problem for any analysis.

Since processes with finite mean variance are of intrinsic scientific interest, it is not suprising that there has been intensive work in trying to apply the geometric insights of Hilbert space methods to such processes.

One way around the problem was provided by the theory of stationary stochastic processes which defines a representation of the data in terms of a stochastic integral:

$$x(t) = \int_0^{2\pi} e^{i\omega t} dZ(\omega) \tag{N.7}$$

where $dZ(\omega)$ is an orthogonal increment process which satisfies:

$$E\left(dZ(\omega)\, dZ(\omega')\right) = dF_X(\omega)\delta(\omega - \omega') \tag{N.8}$$

where dF_X is the spectral distribution function of the data. Each $x(t)$ is measurable with respect to the distribution function $dF_X(\omega)$ and we can define a Hilbert space norm by:

$$E\,(x(t))^2 = \int_0^{2\pi} dF_X(\omega) = \sigma^2 \tag{N.9}$$

We can clearly use this representation when there are an arbitrarily large but finite number of regimes which begin at different fractions of time. In this case, the expectation would be defined as the (weighted) average over different regimes.

Another solution to this problem of measurability is to define a mean value norm:

$$\mathcal{E}\,(f(t)g(t)) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)\,g(t)\,dt \tag{N.10}$$

Such a norm is equivalent to that used for special Hilbert spaces called *Hilbert spaces of almost periodic functions* ([4], Ch. V, p. 132-138).[1] We wish to deal with situations with only a countable number of points so that the corresponding definition would be:[2]

$$\mathcal{E}\,(f(t)\,g(t)) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(t)\,g(t) \tag{N.12}$$

In addition to Eq. (N.12) (which is equivalent to Eq. (IV.34), we also consider the alternative definition which is relevant to our problem:

$$\mathcal{E}\,(f(x,v)g(x,v)) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E(f(x_t,v_t)g(x_t,v_t)) \tag{N.13}$$

[1] An almost periodic function is a function such that for every $\epsilon > 0$, it is possible to find some $l > 0$ (which depends on ϵ such that every interval of the x axis of length l contains at least one number τ such that:

$$|f(x+\tau) - f(x)| < \epsilon \tag{N.11}$$

for $x \in [-\infty, \infty]$ ([178], p. 254). We use only the norm and not the definition (e.g., Eq. (N.11) of an almost periodic function.

[2] We note that it follows from a theorem of Bohr ([4], Ch. V, p. 132-138) that since there are countably many frequencies in the Fourier expansions for almost periodic functions $f(t)$ and $g(t)$ that the Fourier series expansion for $f(t)$ converges in the M norm.

This satisfies the triangle inequality and the Cauchy-Schwartz inequality if f and g have finite variance.[3]

In the theory of almost periodic functions, the element zero may be defined as any function λ such that ([4], p. 132):

$$\mathcal{E}\left(\lambda^2\right) = 0 \tag{N.16}$$

We shall define the element zero in the same manner. Thus, convergence in the sense of Theorem 1 implies convergence to an equivalence class of functions or sequences with zero variance.[4] Apart from the normalization factor, this does not seem any different from the usual mode of strong convergence in which convergence is also to an equivalence class of functions (which can differ in a nonmeasurable manner).

Let us consider an example of a fractional cointegrated model for which we can define a probability measure in terms of a fraction of the time series:

[3] The Cauchy-Schwartz inequality is a little bit tricky. For each t:

$$E(f_t g_t) \leq (E(f_t^2))^{\frac{1}{2}}(E(g_t^2))^{\frac{1}{2}} \tag{N.14}$$

the right hand side of which is finite by assumption. Now:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} E(f_t g_t) &\leq \frac{1}{T}\sum_{t=1}^{T}(E(f_t^2))^{\frac{1}{2}}(E(g_t)^2))^{\frac{1}{2}} \\ &\leq \left[\frac{1}{T}\sum_{t=1}^{T}(E(f_t^2))\right]^{\frac{1}{2}}\left[\frac{1}{T}\sum_{t=1}^{T}(E(g_t^2))\right]^{\frac{1}{2}} \end{aligned} \tag{N.15}$$

[4] Define f^N to be a Cauchy sequence and we wish to show that it is equivalent to $f^N + u^N$ where u^N is the element λ (so that it has zero variance), then by the triangle inequality:

$$\left[\mathcal{E}(f^N + u^N - f^M - u^M)^2\right]^{\frac{1}{2}} \leq \mathcal{E}\left((f^N - f^M)^2\right)^{\frac{1}{2}} + \left(\mathcal{E}\left((u^N - u^M)^2\right)\right)^{\frac{1}{2}} \tag{N.17}$$

so that if f^N converges and u^N satisfies $\mathcal{E}\left((u^N - u^M)^2\right) = 0$, $f^N + u^N$ also converges (c.f., [193], pp. 98-9). Suppose u^N, u^M are zero variance then by the Cauchy-Schwartz inequality $\mathcal{E}(u^N u^M) = 0$ and $\mathcal{E}\left((u^N - u^M)^2\right) = 0$. If f^N converges, it must converge to f^* where f^* is an equivalence class of terms which differ by terms of zero variance ([178], pp. 331).

$$y(t) = c\left(\frac{t}{T}\right) x(t) + \epsilon(t) \tag{N.18}$$

where $\epsilon(t)$ is independently distributed noise and $x(t)$ is a variable which is integrated. Therefore, in Eq. (N.18), $y(t)$ and $x(t)$ are cointegrated with a cointegrating relationship which varies slowly over time.

We define:[5]

$$X_T(s) = \frac{1}{\sqrt{T}} x([sT]). \tag{N.19}$$

It is known that $X_T(s) \Rightarrow X(s)$ where $X(s)$ is a Wiener process. Thus, the regression function Eq. (N.18) has the following property:

$$c(s)\frac{1}{\sqrt{T}}x([sT]) + \frac{1}{\sqrt{T}}\epsilon([sT]) \Rightarrow c(s)X(s). \tag{N.20}$$

Thus, we can look at $c(s)X(s)$ as our regression function which is measurable with respect to a Wiener measure on the interval $[0,1]$. We can construct a set of model components for 'cointegrating pursuit' with large windows multiplied by the data and our series expansion for $c(s)$ will converge in a L^2 sense. We can say something more general in that all we need is that $c(s)X(s)$ is measurable on $[0,1]$; therefore, we might consider problems where $X(s)$ follows another process such as a Brownian bridge (let $X(s)$ be an option price) and we might be trying to detect a change in the relationship with the process $Y(s)$.

In the case of our example (where $X(s)$ is a Wiener process), Theorem 1 says that we can construct estimates such that:[6]

$$\int (\hat{c}(s) - c(s))^2\, s\, ds = 0. \tag{N.21}$$

[5] The notation $[W]$ refers to the largest integer less than or equal to W.

[6] This assumes that $\hat{c}$ does not depend on the particular realization of $X(s)$.

For comparison, in the case where we have a stationary time series with a single lag variable (with the fractional autoregressive function $c(s)$) with slowly varying variance $\sigma^2(s)$ on the interval $[0, 1]$ (representing the fraction of the time series), we have heuristically that:

$$\int(\hat{c}(s) - c(s))^2 d\sigma^2(s) = 0 \tag{N.22}$$

Since $\sigma^2(s) \propto s$ for the integrated process, there does not seem to be very much *operational* distinction between situations in which we take a limit of $T \to \infty$ and situations in which the sample paths are smooth enough that we can integrate with respect to the fraction of time instead.

APPENDIX O

Frames and Completeness

In this appendix, we prove that if the window functions g_k for model components for a particular lag span the same space as the coefficient functions for that variable, then under certain conditions, the set of resulting model components spans the same space as the regression function.

In Ch. IV, we defined the auxiliary variable v which in our time series context is defined in several different senses. For replication models, v represents time or time modulu a periodicity. For fraction models, v either represents the current regime or the fraction of time ($\frac{t}{T}$). We define an inner product $< f, g >$ on the space $\mathcal{Q}$ spanned by the coefficient functions $c_j(v)$; the inner product is weighted by the measure for v which we will denote by $\mu(v)$.

Theorem 19 *Suppose:*

(1) for each variable x_j:

$$A_j||c_j||^2 \leq \sum_{k \in \mathcal{W}} | < c_j, g_k > |^2 \leq B_j||c_j||^2, \tag{O.1}$$

for some subset $\mathcal{W}_j$ of window functions for model components associated with the variable x_j where $||g_k|| = 1$.

(2) The variance of x_j is positive and finite and there exists a constant $C < \infty$ such that:

$$\sup_t E\left(x_j(t)^2\right) < C \inf_t E\left(x_j(t)^2\right) \tag{O.2}$$

(3) The model components h_k in the analysis are defined as: $h_k(t) = g_k(v_t)x_j(t)$.

(4) At least one of the included variables x_j is correlated with the regression function f:

$$\sup_j \left[\frac{|\sum_v \mu(v) E_v f x_j|}{\left(\mathcal{E}(x_j^2)\right)^{\frac{1}{2}} \left(\mathcal{E}(f^2)\right)^{\frac{1}{2}}} \right] \geq \lambda > 0 \tag{O.3}$$

(5) $fx_j \in \mathcal{Q}$ for all j,

then the set of model components in the analysis satisfy:

$$A\mathcal{E}(f^2) \leq \sum_{k \in \mathcal{W}} \left| \mathcal{E}\left(f \frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}} \right) \right|^2 \leq B\mathcal{E}(f^2) \tag{O.4}$$

where $\mathcal{W} = \cup_j \mathcal{W}_j$ so that the generated model components form a frame and hence a spanning set.

Proof:

We note that:

$$\left| \mathcal{E}\left(f \frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}} \right) \right|^2 = \frac{|\mathcal{E}(fh_k)|^2}{\mathcal{E}(h_k^2)} \tag{O.5}$$

By the definition of h_k:

$$\mathcal{E}(h_k^2) = \mathcal{E}(g_k^2 x_j^2) \tag{O.6}$$

Using the definition of g_k and $||g_k|| = 1$ and the assumption on the variances of the explanatory variables:

$$0 < \inf_t E(x_j(t)^2) \leq \mathcal{E}(g_k^2 x_j^2) \leq \sup_t E(x_j(t)^2) < \infty \tag{O.7}$$

Thus:

$$\frac{1}{\sup_t E(x_j(t)^2)}|\mathcal{E}(fh_k)|^2 \leq \left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \leq \frac{|\mathcal{E}(fh_k)|^2}{\inf_t E_t(x_j(t)^2)} \tag{O.8}$$

Now:

$$\mathcal{E}(fh_k) = \mathcal{E}(fx_jg_k) = \mathcal{E}(z_jg_k) \tag{O.9}$$

Furthermore, since g_k is nonstochastic:

$$\mathcal{E}(z_jg_k) = < E_v(z_j), g_k > \tag{O.10}$$

Now, $E_v(z_j) \in \mathcal{Q}$ by assumption. In a time series context, $z_j = (y(t) - \epsilon(t))\,y(t - k_j)$ so that taking expectations $E_v z_j = \gamma_{k_j}(v)$ or the autocovariance at lag k_j of y. Since $c_j(v) \in \mathcal{Q}$, g_k spanning $\mathcal{Q}$ in the sense of Eq. (O.1) is equivalent to:

$$A_j||E_v(z_j)||^2 \leq \sum_{k\in\mathcal{W}_j} |< E_v(z_j), g_k >|^2 \leq B_j||E_v(z_j)||^2 \tag{O.11}$$

Using Eq. (O.8), it follows that:

$$\sum_{j=1}^{J}\frac{A_j}{\sup_t E_t(x_j^2)}||E_v(z_j)||^2 \leq \sum_{k\in\mathcal{W}}\left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \tag{O.12}$$

and:

$$\sum_{k\in\mathcal{W}}\left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \leq \sum_{j=1}^{J}\frac{B_j}{\inf_t E(x_j(t)^2)}||E_v(z_j)||^2. \tag{O.13}$$

Now:

$$\begin{aligned}
||E_v(z_j)||^2 &= \sum \mu(v)(E_v(fx_j))^2 \\
&\leq \sum_v \mu(v)E_v(f^2)E_v\left(x_j^2\right) \\
&\leq \sup_t E_t(x_j^2)\,\mathcal{E}(f^2)
\end{aligned} \tag{O.14}$$

Thus, by Eq. (O.13):

$$\sum_{k\in\mathcal{W}} \left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \leq \sum_{j=1}^{J} B_j \frac{\sup_t E(x_j(t)^2)}{\inf_t E(x_j(t)^2)}\, \mathcal{E}(f^2) = B\,\mathcal{E}(f^2) \qquad \text{(O.15)}$$

For a lower bound, we note that:

$$||E_v(z_j)||^2 \geq |\sum_v \mu(v) E_v(z_j)|^2 \qquad \text{(O.16)}$$

By assumption, the maximal correlation between f and x_j is greater than λ:

$$\sup_j \left[\frac{|\sum_v \mu(v) E_v z_j|}{\left(\mathcal{E}(x_j^2)\right)^{\frac{1}{2}} (\mathcal{E}(f^2))^{\frac{1}{2}}}\right] \geq \lambda \qquad \text{(O.17)}$$

Hence:

$$\sup_j ||E_v(z_j)||^2 \geq \lambda^2\, \mathcal{E}(f^2) \inf_t E(x_j(t)^2). \qquad \text{(O.18)}$$

We denote by j^* the j which satisfies the bound in Eq. (O.18) then:

$$A_{j^*}\, \lambda^2 \frac{\inf_t E(x_{j^*}(t)^2)}{\sup_t E(x_{j^*}(t)^2)}\, \mathcal{E}(f^2) \leq \sum_{k\in\mathcal{W}} \left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \qquad \text{(O.19)}$$

Hence:

$$A\,\mathcal{E}(f^2) \leq \sum_{k\in\mathcal{W}} \left|\mathcal{E}\left(f\frac{h_k}{(\mathcal{E}(h_k^2))^{\frac{1}{2}}}\right)\right|^2 \qquad \text{(O.20)}$$

which together with Eq. (O.15) proves the result. ∎

It is useful to give some intuition for why the ratio of minimum to maximum variances of the explanatory variables x_t are relevant. The intuition comes from a simple linear regression. Suppose that the model is:

$$y(t) = \alpha g_0\left(\frac{t}{T}\right) y(t-1) + \epsilon(t) \qquad \text{(O.21)}$$

then:

$$\hat{\alpha} = \frac{\frac{1}{T}\sum_{t=1}^{T} y(t)y(t-1)g_0\left(\frac{t}{T}\right)}{\frac{1}{T}\sum_{t=1}^{T} y(t-1)^2 g_0\left(\frac{t}{T}\right)^2} \tag{O.22}$$

so that the theoretical value of α (the probability limit of $\hat{\alpha}$) is:

$$\alpha = \frac{\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(y(t)y(t-1))g_0\left(\frac{t}{T}\right)}{\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(y(t-1)^2)g_0\left(\frac{t}{T}\right)^2}. \tag{O.23}$$

In the fraction model, $E(y(t)y(t-1))$ is a function of the fraction of time so that:

$$E\left(y(t)\,y(t-1)\right) = \gamma_T\left(\frac{t}{T}\right). \tag{O.24}$$

Likewise:

$$E\left(y(t-1)^2\right) = \sigma_T^2\left(\frac{t}{T}\right). \tag{O.25}$$

We denote the limiting γ_T and σ_T^2 by γ and σ^2.

Using Eqs. (O.23) (O.24), (O.25):

$$\alpha_0 = \frac{\int \gamma(s)g_0(s)ds}{\int \sigma^2(s)g_0^2(s)ds}. \tag{O.26}$$

We shall now try to show how Eq. (O.26) can be part of an abstract *orthonormal* expansion for an arbitrary $c_1(s)$ in:

$$\begin{aligned} y(t) &= c_1\left(\frac{t}{T}\right) y(t-1) + \epsilon_1(t) \\ &= \sum_{j=0}^{\infty} \alpha_j g_j\left(\frac{t}{T}\right) + \epsilon_2(t). \end{aligned} \tag{O.27}$$

For the expansion to be orthonormal, we must be able to come up with a set of window functions (with norm 1 with respect to a measure $\sigma^2(s)ds$)) such that:

$$\alpha_j =< c_1, g_j > . \tag{O.28}$$

We *define* the theoretical value of $c_1(s)$ to be $\frac{\gamma(s)}{\sigma^2(s)}$. Then we note that (using Eq. (O.23):

$$\begin{aligned} \alpha_0 &= \frac{\int \frac{\gamma(s)}{\sigma^2(s)} g_0(s)\sigma^2(s)ds}{\int \frac{\sigma^2(s)}{\sigma^2(s)} g_0^2(s)\sigma^2(s)ds} \\ &= \int c_1(s)g_0(s)\sigma^2(s)ds =< c_1, g_0 > . \end{aligned} \tag{O.29}$$

where the last step follows from the definition of c_1 and the normalization of the window function. Hence, if the g_j form an orthonormal basis of $\mathbf{L}^2$ weighted by $\sigma^2(s)$:

$$\lim_{N\to\infty} ||c_1(s) - \sum_{j=0}^{N} \alpha_j g_j(s)||^2 \to 0 \tag{O.30}$$

in the weighted $\mathbf{L}^2$ norm. When $\sigma^2(s)$ is time-invariant, the ratio of $\sup_t Ex_j(t)^2$ to $\inf_t Ex_j(t)^2$ is 1 so that g_j forming an orthonormal basis of $\mathbf{L}^2[0,1]$ is equivalent to g_j forming an orthonormal basis of the weighted space; otherwise, approximations differ in a sense which depends on how nonstationary is the data generating process for $x_j(t)$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[2] H. Akaike. A Bayesian analysis of the Minimum AIC Procedure of autoregressive model fitting. *Annals of the Institute of Statistical Mathematics*, 30A:9–14, 1978.

[3] H. Akaike. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, 66:237–242, 1979.

[4] N.I. Akhiezer and I.M. Glazman. *Theory of Linear Operators in Hilbert Space, Vol. I.* Frederick Ungar, 1961.

[5] K.I. Amin and V.N. Ng. Option valuation with systematic stochastic volatility. *Journal of Finance*, 48:881–910, 1993.

[6] Donald W.K. Andrews. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4:458–467, 1988.

[7] Donald W.K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61:4:821–856, 1993.

[8] A. Arneado, F. Argoul, E. Bacry, J.F. Muzy, and M. Tabard. Golden mean arithmetic in the fractal branching of diffusion-limited aggregates. *Physical Review Letters*, 68:3456–3459, 1992.

[9] E. Bacry, J. Muzy, and A. Arneado. Singularity spectrum of fractal signals from wavelet analysis: exact results. *Journal of Statistical Physics*, 70:635–674, 1993.

[10] R. Balian. Un principe d'incertidue fort en theorie du signal ou en mecanique quantique. *C.R. Acad. Sci. Paris*, 292, Serie 2, 1981.

[11] B. Ross Barmish. *New Tools for Robustness of Linear Systems.* Macmillan, 1994.

[12] Robert Barro. Are government bonds net worth? *Journal of Political Economy*, 82:6:1095–1117, 1974.

[13] Robert Barsky and Bradford DeLong. Why does the stock market fluctuate? *Quarterly Journal of Economics*, 108:291–311, 1993.

[14] Robert Barsky and Jeffrey Miron. The seasonal cycle and the business cycle. *Journal of Political Economy*, 97: 3:503–534, 1989.

[15] M.S. Bartlett. *An Introduction to Stochastic Processes with Special Reference to Methods and Applications.* Cambridge, 1955.

[16] Tamer Basar and Pierre Bernhard. $\mathbf{H}^{\infty}$-*Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach.* Birkhauser, 1991.

[17] G. Battle. Heisenberg proof of the Balian-Low theorem. *Letters in Mathematical Physics*, 15:175–177, 1988.

[18] G. Belykin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, L. Raphael, and M.B. Ruskai. *Wavelets and their Applications.* Jones and Bartlett, 1992.

[19] John Benedetto. Frame decompositions, sampling and uncertainty principle inequalities. In John Benedetto and Michael Frazier, editors, *Wavelets: Mathematics and Applications.* CRC Press, 1994.

[20] John Benedetto and Michael Frazier. *Wavelets: Mathematics and Applications.* CRC Press, 1994.

[21] A. Beneviste, Michel Metivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, 1990.

[22] A.R. Bergstrom. *Continuous Time Econometric Modelling.* Oxford, 1990.

[23] Sudipto Bhattacharya and George Constantinides, editors. *Financial Markets and Incomplete Information.* Rowland Littlefield, 1989.

[24] Peter J. Bickel, Chris A. Klaasen, Yaacov Ritov, and Jon Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins, 1993.

[25] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, 1968.

[26] Olivier Blanchard and Stanley Fischer. *Lectures on Macroeconomics.* MIT, 1989.

[27] Alan Blinder and Stanley Fischer. Inventories, rational expectations and the business cycle. *Journal of Monetary Economics*, 8:3:277–304, 1981.

[28] Boualem Boashash. *Time-Frequency Signal Analysis: Methods and Applications.* Wiley, 1992.

[29] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[30] Tim Bollerslev, Ray Chou, and Kenneth Kroner. ARCH modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics*, 52:5–59, 1992.

[31] P. Bossaerts and P. Hillion. Selecting models to forecast financial returns: A new criteria. California Institute of Technology, 1993.

[32] W.A. Brock, D. Hsieh, and B. LeBaron. *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence.* MIT Press, 1991.

[33] Peter Brockwell and Richard Davis. *Time Series: Theory and Methods.* Springer-Verlag, 1987.

[34] R. Brown, J. Durbin, and J. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Ser. B*, pages 149–163, 1975.

[35] R. Creighton Buck. *Advanced Calculus.* McGraw Hill, 1956.

[36] Peter E. Caines. *Linear Stochastic Systems.* Wiley, 1988.

[37] D.S.K. Chan. A non-aliased discrete-time wigner distribution for time-frequency signal analysis. *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1333–1336, 1982.

[38] Gregory C. Chow. Tests for equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605, 1960.

[39] Charles K. Chui. *An Introduction to Wavelets.* Academic Press, 1992.

[40] Charles K. Chui. *Wavelets: A Tutorial in Theory and Applications.* Springer Verlag, 1992.

[41] P.K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41:135–55, 1973.

[42] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1:54–81, 1993.

[43] Leon Cohen. Generalized phase-space distribution functions. *Journal of Mathematical Physics*, 7:781–786, 1966.

[44] R. Coifman, Y. Meyer, S. Quake, and V. Wickerhauser. Signal processing and compression with wavelet pakcets. Numerical Algorithms Research Group, Yale Univ., 1990.

[45] R. Coifman, Y. Meyer, and V. Wickerhauser. Size properties of wavelet packets. In M. Ruskai et al., editor, *Wavelets and Their Applications*, pages 453–470. Jones and Bartlett, 1992.

[46] R. Coifman, Y. Meyer, and V. Wickerhauser. Wavelet analysis and signal processing. In M. Ruskai et al., editor, *Wavelets and Their Applications*, pages 153–178. Jones and Bartlett, 1992.

[47] R. Coifman and V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38:2, 1992.

[48] T.F. Cooley and E.C. Prescott. Efficient estimation in the presence of stochastic parameter variation. *Econometrica*, 44:167–184, 1976.

[49] Russell Cooper and Andrew John. Coordinating coordination failures in keynesian models. *Quarterly Journal of Economics*, 103:441–63, 1988.

[50] H. Cramer. On some clases of non-stationary processes. *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 2:57–78, 1960.

[51] R.J. Creswick, H.A. Farach, and C.P. Poole. *Introduction to Renormalization Group Methods in Physics.* Wiley, 1992.

[52] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics*, 41:909–996, 1988.

[53] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Information Theory*, 36:961–1005, 1990.

[54] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM CBMS-NSF Conference Series in Applied Mathematics, 1992.

[55] Geoff Davis. Adaptive non-linear approximations. Ph.D. thesis, New York University, June, 1994.

[56] Geoffrey Davis, Stephane Mallat, and Zhifeng Zhang. Adaptive time-frequency approximations with matching pursuits. Courant Institute of Mathematical Sciences, May, 1994.

[57] Carl de Boor. *A Practical Guide to Splines.* Springer-Verlag, 1978.

[58] Phoebius Dhyrmes. *Distributed Lags: Problems of Estimation and Formulation.* North Holland, second edition, 1981.

[59] Peter A. Diamond. Aggregate demand management in search equilibrium. *Journal of Political Economy*, 90:881–904, 1982.

[60] Thomas Doan, Robert Litterman, and Christopher Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100, 1984.

[61] David L. Donoho, Iain M. Johnstone, Gerard Kerkyachharian, and Dominique Picard. Density estimation by wavelet thresholding. Department of Statistics, Stanford University, April, 1993.

[62] David L. Donoho and Ian M. Johnstone. Minimax estimation via wavelet shrinkage. Department of Statistics, Stanford University, 1992.

[63] Peter Dorato, editor. *Robust Control.* IEEE Press, 1987.

[64] Rudiger Dornbusch. Expectations and exchange rate dynamics. *Journal of Political Economy,* 84:1161–76, 1976.

[65] John C. Doyle, Bruce A. Francis, and Allen R. Tannenbaum. *Feedback Control Thcory.* Macmillan, 1992.

[66] Steven Durlauf. Path dependence in aggregate output. Stanford Univ., January 1992.

[67] M.S. Eichenbaum, L.P. Hansen, and K.J. Singleton. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Technical Report 1981, NBER, 1986.

[68] Charles Engel and James D. Hamilton. Long swings in the dollar: are they in the data and do markets know it? *American Economic Review,* 80:689–713, 1990.

[69] R.F. Engle. Band spectrum regression. *International Economic Review,* 15:1–11, 1974.

[70] Robert F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica,* 50:987–1008, 1982.

[71] B. Escudie and J. Grea. Sur une formulation generale de la representation en temps et en frequence dans l'analyse des signaux d'energie finie. *C.R. Acad, Sci. Paris (Ser. A),* 283:1049–1051, 1976.

[72] R. Eubank. *Spline smoothing and nonparametric regression.* Marcel Dekker, 1988.

[73] E.F. Fama. Mandelbrot and the stable Paretian hypothesis. *Journal of Business,* 36:420–429, 1963.

[74] Hans Follmer. Random economies with many interacting agents. *Journal of Mathematical Economics,* 1:51–62, 1974.

[75] Murray Frank and Thomas Stengos. Some evidence concerning macroeconomic chaos. *Journal of Monetary Economics,* 22:423–438, 1988.

[76] Benjamin Friedman. Money, credit and interest rates in the business cycle. In Robert J. Gordon, editor, *The American Business Cycle: Continuity and Change,* pages 395–458. Chicago, 1986.

[77] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computation,* C-23:881–889, 1974.

[78] R. Frisch. Propagation problems and impulse problems in dynamic problems. In *Economic Essays in Honor of Gustav Cassel.* London, 1933.

[79] A.R. Gallant. Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390, 1987.

[80] Peter Garber and Robert King. Deep structural excavation? a critique of Euler equation methods. NBER technical paper # 31. 1983.

[81] John Geanakoplos, Paul Klemperer, and Jeremy Bulow. Multimarket oligopoly: Strategic substitutes and complements. *Journal of Political Economy*, 93:488–511, 1985.

[82] C.W.J. Granger. The typical spectral shape of an economic variable. *Econometrica*, 34:150–61, 1966.

[83] C.W.J. Granger and A. Anderson. *An Introduction to Bilinear Time Series Models.* Gottingen: Vanderhoeck and Ruprecht, 1978.

[84] C.W.J. Granger and T. Terasvirta. *Modelling Nonlinear Economic Relationships.* Oxford, 1993.

[85] U. Grenander. *Abstract Inference.* Wiley, 1981.

[86] Lazlo Gyorfi, Wolfgang Hardle, Pascal Sarda, and Philippe Vieu. *Nonparametric Curve Estimation from Time Series.* Springer Verlag, 1989.

[87] A. Haar. Zur theorie der orthogalen funktionen-systeme. *Mathematics Annals*, 69:331–371, 1910.

[88] Alastair Hall. Testing for a unit root with pretest data based model selection. Department of Economics, North Carolina State University.

[89] J. Haltiwinger and M. Waldman. Rational expectations and the limits of rationality: an analysis of heterogeneity. *American Economic Review*, 75:326–340, 1985.

[90] James D. Hamilton. *Time Series Analysis.* Princeton, 1994.

[91] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–84, 1989.

[92] E.J. Hannan. Regression for time series. In Murray Rosenblatt, editor, *Time Series Analysis.* Wiley, 1963.

[93] E.J. Hannan and P.M. Robinson. Lagged regression with unknown lags. *Journal of the Royal Statistical Society, Ser. B*, 42:146–60, 1973.

[94] B.E. Hansen. The likelihood ratio test under non-standard conditions: testing the Markov trend model of GNP. *Journal of Applied Econometrics*, 7:S61–S82, 1992.

[95] Bruce E. Hansen. Convergence to stochastic integrals for dependent heterogeneous processes. *Econometric Theory*, 8:489–500, 1992.

[96] Lars P. Hansen and Ravi Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99:2:225–262, 1991.

[97] L.P. Hansen and K.J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50:1269–1286, 1982.

[98] W. Hardle, W. Hildenbrand, and M. Jerison. Empirical evidence on the law of demand. *Econometrica*, 59:1525–1549, 1991.

[99] W. Hardle and T. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84:986–995, 1989.

[100] Wolfgang Hardle. *Applied Nonparametric Regression.* Econometric Society Monograph, Cambridge University Press, 1990.

[101] O. Hart. A model of imperfect competition with Keynesian features. *Quarterly Journal of Economics*, 47:109–138, 1982.

[102] Simon Haykin. *Neural Networks: A Comprehensive Foundation.* IEEE Press, 1994.

[103] W. Heller. Coordination failures under complete markets with applications to effective demand. In W. Heller et al., editor, *Equilibrium Analysis: Essays in Honor of Kenneth J. Arrow.* Cambridge, 1986.

[104] Werner Hildenbrand. *Market Demand.* Princeton, 1994.

[105] P.J. Huber. Robust smoothing. In E. Launer and G. Wilkinson, editors, *Robustness in Statistics.* Academic Press, 1979.

[106] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–475, 1985.

[107] Wen-Liang Hwang and Stephane Mallat. Characterization of self-similar multifractals with wavelet maxima. Technical Report, Department of Computer Science, New York University, July 1993.

[108] Claude Itzykson and Jean-Michel Drouffe. *Statistical Field Theory.* Cambridge, 1989.

[109] H.E. Jensen, T. Hoholdt, and J. Justesen. Double series representation of bounded signals. *IEE Transactions on Information Theory*, 34:613–624, 1988.

[110] Jechang Jeong and William J. Williams. Time-varying filtering and signal analysis. In Boashash, editor, *Time-Frequency Signal Analysis: Methods and Applications*, pages 389–405. Wiley, 1992.

[111] David S. Johnson. A catalog of complexity classes. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, pages 67–161. Elsevier, 1990.

[112] Lee K. Jones. On the conjecture of Huber concerning the convergence of projection pursuit regression. *The Annals of Statistics*, 15:2:880–882, 1987.

[113] George G. Judge et al. *The Theory and Practice of Econometrics.* Wiley, second edition, 1985.

[114] Jean-Pierre Kahane. *Some Random Series of Functions.* Cambridge, 1985.

[115] Ioannis Karatzas and Steven Shreve. *Brownian Motion and Stochastic Calculus.* Springer-Verlag, 1988.

[116] Frank Knight. *Risk, Uncertainty and Profit.* Houghton-Mifflin, 1921.

[117] P.S. Laumas and Y.P. Mehra. The stability of the demand for money function: the evidence from quarterly data. *Review of Economics and Statistics*, 58:464–468, 1976.

[118] M.R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, 1983.

[119] Michel LeBellac. *Quantum and Statistical Field Theory.* Oxford, 1991. translated by G. Barton, orig. pub. in French, 1988.

[120] J. Lintner. The evaluation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47:13–37, 1965.

[121] Mico Loretan and P.C.B. Phillips. Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets. *Journal of Empirical Finance*, 1:2:211–248, 1994.

[122] F. Low. Complete sets of wave packets. In *A Passion for Physics – Essays in Honor of Gregory Chew*, pages 17–22. World Scientific, 1985.

[123] Robert Lucas. Econometric policy evaluation: A critique. In K. Brunner and A. Meltzer, editors, *The Phillips Curve and the Labor Market.* Carnegie-Rochester Series on Public Policy, 1976.

[124] Robert Lucas. Understanding business cycles. In Brunner and Metzler, editors, *Stabilization of the Domestic and International Economy*, pages 7–29. Carnegie-Rochester Series on Public Policy, 1977.

[125] Robert E. Lucas. Expectations and the neutrality of money. *Journal of Economic Theory*, 4:103–124, 1972.

[126] Helmut Lutkepohl. *Introduction to Multiple Time Series Analysis.* Springer-Verlag, 1991.

[127] S.G. Mallat. Zero crossings of a wavelet transform. *IEEE Transactions on Information Theory*, 37: 4:1019–1033, 1991.

[128] S.G. Mallat and Sifen Zhong. Wavelet transform maxima and multiscale edges. In M. Ruskai et al., editor, *Wavelets and Their Applications*, pages 439–451. Jones and Bartlett, 1992.

[129] Stephane Mallat. Multiresolution approximations and wavelet orthonormal bases of $\mathbf{L}^2$. *Transactions of the American Mathematical Society*, 315:69–88, 1989.

[130] Stephane Mallat. A theory of multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[131] Stephane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[132] Henrique S. Malvar. *Signal Processing with Lapped Transforms.* Artech House, 1992.

[133] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36:394–419, 1963.

[134] Wolfgang Martin and Patrick Flandrin. Wigner-Ville spectral analysis of nonstationary processes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33(6):1461–1470, 1985.

[135] D.L. McLeish. On the invariance principle for nonstationary mixingales. *Annals of Probability*, 5:616–621, 1977.

[136] A. Melino and S.M. Turnbull. Pricing options with stochastic volatility. *Journal of Econometrics*, 45:239–265, 1990.

[137] Robert Merton. An intertemporal capital asset pricing model. *Econometrica*, 41:867–888, 1973.

[138] Yves Meyer. *Wavelets and Operators.* Cambridge, 1992.

[139] Franco Modigliani and Emile Grunberg. The predictability of social events. *Journal of Political Economy*, 62:465–78, 1954.

[140] Franco Modigliani and Merton Miller. The cost of capital, corporate finance, and the theory of investment. *American Economic Review*, 48, 1958.

[141] Michael C. Mozer. Neural net architectures for temporal sequence processing. In Andreas S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Proceedings of the Santa Fe Institute (vol. XV)*, pages 243–264. Addison-Wesley, 1994.

[142] John F. Muth. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55:299–306, 1961.

[143] Kumpati Narendra, Romeo Ortega, and Peter Dorato, editors. *Advances in Adaptive Control.* IEEE Press, 1991.

[144] Charles R. Nelson. The prediction performance of the F.R.B.-M.I.T.-PENN model of the U.S. economy. *American Economic Review*, 62:902–917, 1972.

[145] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1990.

[146] P. Newbold, C. Agiakloglou, and J. Miller. Long-term inference based on short-term forecasting models. In T. Subba Rao, editor, *Developments in Time Series Analysis*, pages 9–25. Chapman and Hall, 1993.

[147] Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135, 1990.

[148] D.F. Nicholls and B.G. Quinn. *Random Coefficient Autoregressive Models: An Introduction.* Springer-Verlag Lecture Notes in Statistics, No. 11, 1982.

[149] J. Michael Orszag. The adaptive waveletgram. University of Michigan, Department of Economics. February 1993.

[150] J. Michael Orszag. Economic feasibility of thick market effects. University of Michigan, Department of Economics, November 1992.

[151] J. Michael Orszag. Estimation of linear nonstationary autoregressive processes. Presented at North American Econometric Society Meetings, June 1993.

[152] J. Michael Orszag. A field theoretic approach to statistical macroeconomics. University of Michigan, Department of Economics, August 1993.

[153] J. Michael Orszag. On a field theoretic approach to statistical macroeconomics. University of Michigan, Department of Economics, November 1993.

[154] J. Michael Orszag and Hong Yang. Portfolio choice with Knightian uncertainty. Department of Economics, University of Michigan. Presented at the 1993 Risk Theory Conference, Wharton School, University of Pennsylvania, March 1993.

[155] A.R. Pagan. Some identification and estimation results for regression models with stochastically varying coefficients. *Journal of Econometrics*, 13:341–363, 1980.

[156] A.R. Pagan. Alternative models for conditional stock volatility. *Journal of Econometrics*, 45:267–290, 1990.

[157] A.R. Pagan. Testing for covariance stationarity in stock market data. *Economics Letters*, 33:165–170, 1990.

[158] C.H. Page. Instantaneous power spectra. *Journal of Applied Physics*, 23:103–106, 1952.

[159] Giorgio Parisi. *Statistical Field Theory.* Addison-Wesley, 1988.

[160] Jagdish Patel and Campbell Read. *Handbook of the Normal Distribution.* Marcel Dekker, 1982.

[161] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Proc. 27th Annual Assilomar Conf. on Signals, Systems and Computers*, 1993.

[162] Donald Percival and Andrew Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge, 1993.

[163] Pierre Perron. The great crash, the oil price shock and the unit root hypothesis. *Econometrica*, 57:1361–1401, 1989.

[164] Peter Phillips. Spectral regression for cointegrated time series. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 413–436. Cambridge, 1991.

[165] Peter C.B. Phillips. Testing for a unit root by frequency domain regression. *Journal of Econometrics*, 59:263–286, 1993.

[166] William Press and et al. *Numerical Recipes in C.* Cambridge, 1992.

[167] M.B. Priestley. Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society (Series B)*, 27:204–237, 1965.

[168] M.B. Priestley. State dependent models: A general approach to non-linear time series analysis. *Journal of Time Series Analysis*, 1:47–71, 1980.

[169] M.B. Priestley. *Spectral Analysis and Time Series.* Academic Press, 1981.

[170] M.B. Priestley. *Non-linear and Non-stationary Time Series.* Academic Press, 1988.

[171] S. Qian and D. Chen. Signal representation via adaptive normalized gaussian functions. *IEEE Transactions on Signal Processing*, 36:1, 1994.

[172] Lawrence Rabiner and Bing hwang Juang. *Fundamentals of Speech Rcognition.* Prentice Hall, 1993.

[173] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3:4–16, 1986.

[174] K.R. Rao and P. Yip. *Discrete Cosine Transform: Algoirthms, Advantages, Applications.* Academic Press, 1990.

[175] Subba Rao and M.M. Gabr. *An Introduction to Bispectral Analysis and Bilinear Time Series Models.* Springer Verlag, 1984.

[176] G.C. Rausser and P.S. Laumas. The stability of the demand for money in canada. *Journal of Monetary Economics*, 3:367–380, 1976.

[177] Helmut Reisen and Helene Yeches. Time-varying estimates on the openness of the capital account in Korea and Taiwan. *Journal of Development Economics*, 41:285–305, 1993.

[178] Frigyes Riesz and Bela Sz.-Nagy. *Functional Analysis.* Frederick Ungar, 1955.

[179] P. Robinson. Semiparametric econometrics: A survey. *Journal of Applied Econometrics*, 3:35–52, 1988.

[180] P.M. Robinson. The estimation of a nonlinear moving average model. *Stochastic Processes and their Applications*, 5:81–90, 1977.

[181] Murray Rosenblatt. *Stationary Sequences and Random Fields.* Birkhauser, 1985.

[182] Gennady Samorodnitsky and Murad Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance.* Chapman and Hall, 1994.

[183] Thomas J. Sargent. *Bounded Rationality in Macroeconomics.* Oxford University Press, 1993.

[184] Stephen Satchel and Allen Timmerman. Option pricing with systematic consumption risk. University of London, Birkbeck College, Discussion Paper in Financial Economics, Oct. 1993.

[185] Larry Schumaker and Glenn Webb. *Recent Advances in Wavelet Analysis.* Academic Press, 1994.

[186] L.O. Scott. Option pricing when the variance changes randomly: Theory, estimation and an application. *Journal of Financial and Quantitative Analysis*, 22:417–438, 1987.

[187] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19:425–442, 1964.

[188] E. Slutsky. The summation of random causes as the source of cyclic processes. *Econometrica*, 5:105–146, 1937.

[189] Christopher Small and D.L. McLeish. *Hilbert space methods in probability and statistical inference.* Wiley, 1994.

[190] N.V. Smirnov. Estimate of deviation between empirical distribution functions in two independent samples (in Russian). *Bulletin of Moscow University*, 2:3–16, 1939.

[191] J.H. Stock. Measuring business cycle time. *Journal of Political Economy*, 95:1240–61, 1987.

[192] T. Subba Rao. On the theory of linear time series models. *Journal of the Royal Statistical Society, Ser. B.*, 43:244–55, 1981.

[193] Angus E. Taylor. *Introduction to Functional Analysis.* Wiley, 1958.

[194] S.J. Taylor. *Modelling financial time series.* Wiley, 1986.

[195] D. Tjostheim. Spectral generating operators for non-stationary processes. *Advances in Applied Probability*, 9:831–846, 1976.

[196] H. Tong. *Threshold Models in Non-linear Time Series Analysis.* Springer-Verlag, 1983.

[197] H. Tong. *Nonlinear Time Series Analysis: A Dynamical Systems Approach.* Oxford, 1990.

[198] J.W. Tukey. Discussion emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics*, 3:191–219, 1961.

[199] S.J. Turnovsky. Stochastic stability of short-run market equilibrium under variations in supply. *Quarterly Journal of Economics*, 82:666–681, 1968.

[200] P.P. Vadiyanathan. *Multirate Systems and Filter Banks.* Prentice Hall, 1993.

[201] M. Vidyasagar. *Control Systems Synthesis: A Factorization Approach.* MIT, 1985.

[202] Mark W. Watson. Univariate detrending methods with stochastic trends. *Journal of Monetary Economics*, 18:49–75, 1986.

[203] Andreas S. Weigend and N. A. Gershenfeld. *Time Series Prediction: Proceedings of the Santa Fe Institute (vol. XV).* Addison-Wesley, 1994.

[204] Alexander Weinmann. *Uncertain Models and Robust Control.* Springer-Verlag, 1992.

[205] A.A. Weiss. The stability of the AR(1) process with an AR(1) coefficient. *Journal of Time Series Analysis*, 6:181–186, 1985.

[206] Kenneth West. Bubbles, fads and stock price volatility tests. *Journal of Finance*, 43:639–656, 1988.

[207] Halbert White and Jeffrey M. Wooldridge. Some results on sieve estimation with dependent observations. In Barnett et. al., editor, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 459–493. Cambridge, 1991.

[208] J.B. Wiggins. Option values under stochastic volatility: Theory and empirical estimates. *Journal of Financial Economics*, 19:351–372, 1987.

[209] Kenneth Wilson and J. Kogut. Renormalization group theory and the ϵ-expansion. *Physical Review*, 12C:75, 1974.

[210] P. Young. Time variable and state dependent modeling of nonstationary and nonlinear time series. In T. Subba Rao, editor, *Developments in Time Series Analysis*, pages 374–413. Chapman and Hall, 1993.

[211] R. Young. *An Introduction to Nonharmonic Fourier Series.* Academic Press, 1980.

[212] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms and approximate inverses. *IEEE Transactions on Automatic Control*, AC-26:301–20, 1981.

[213] J. Zinn-Justin. *Quantum Field Theory and Critical Phenomena.* Oxford, 2nd edition, 1993.