ESTIMATION AND INFERENCE IN SIMULTANEOUS EQUATION MODELS

ALASTAIR HALL

Thesis submitted for Ph.D. degree

University of Warwick
Department of Economics

February, 1985

CONTENTS

Acknowledgements

I am greatly indebted to Ken Wallis for his supervision of this work. He provided invaluable guidance at all stages of the research, and without his support the work would not have been completed.

I have benefitted from very useful discussions on various aspects of this thesis with Peter Burridge, Peter Crouch, Jan Magnus, Peter Phillips, George Rowlands and Mark Salmon.

I would also like to thank Carol Jones for her skillful, and patient, typing of the original manuscript.

This research was undertaken during the tenureship of a studentship from the Economic and Social Science Research Council.

SUMMARY

We examine the asymptotic properties of the full information maximum likelihood estimator (MLE) under the assumption of normality in the general nonlinear simultaneous equations model. The initial analysis is for the static model, and then the conditions which allow the generalisation of the results to the dynamic model are explored.

We concentrate on the question of the consistency of the MLE when the normality assumption is erroneous. The conditions for asymptotic normality are also considered, but are given less emphasis because any tests based on the MLE require consistent estimates of its covariance and so also of its mean. It is demonstrated that if it is possible to write down an explicit reduced form, then we can find families of true nonnormal distributions for which the estimator is consistent. However if the reduced form is implicit, then, apart from some special cases, the estimator can only be proved to be consistent if the model is correctly specified. The nature of the reduced form in nonlinear models is rarely considered, and we examine conditions for its uniqueness. It is demonstrated that this entails more stringent conditions on the Jacobian than are usually acknowledged.

Finally we argue that the information matrix test is a natural choice of specification test for the pseudo MLE strategy suggested by Gourieroux, Monfort and Trognon (1984a), which estimates the parameters of the nonlinear regression model by maximising the likelihood from a member of the exponential family. The test statistics are

calculated for the Poisson model example discussed in Gourieroux, Monfort and Trognon (1984b), and their performance contrasted with that of goodness of fit tests. Also tests based on the Edgeworth expansion are compared with tests based on higher derivatives of the standard normal likelihood.

1. INTRODUCTION

1.1 The econometric model and the data generating process.

The question of how to explain the behaviour of economic series is one of fundamental importance. The choice of policy instruments, and the appropriate magnitude by which to adjust them, to achieve a particular goal depends on our understanding of the economy. The central problem is that whilst the outcomes of economic agents actions are observed, it is only possible to hypothesise the decision making process from which these outcomes result. This has naturally led to the use of statistical models to attempt to explain the interrelationship between economic series. It is hoped that by using data to explore the nature of this interrelationship in the past, sufficient information can be acquired to provide useful forecasts of what may happen in the future.

In econometrics it is customary to think of the data as having been generated by a process of the form

$$q(y_t, x_t, \alpha) = u_t, \quad t = 1, \ldots T, \tag{1}$$

where y_t, x_t, u_t are vectors of endogenous, exogenous and error variables respectively, in period t, and α is a vector of unknown parameters. The functional form $q(\cdot)$ is assumed time invariant but is of unknown form. Typically its structure is determined by a mixture of economic theory and prior experience of the variables concerned. Having chosen $q(\cdot)$ the next step is to estimate the unknown parameters. Three main estimation strategies are employed: least squares (LS), instrumental variables (IV)

and maximum likelihood (ML). The latter requires an assumption about the error distribution, and this is usually that it is normal. It is argued that the transformation $q(\cdot)$ of the underlying series represents the mechanism that generated the data and so, on average overtime, the observed values of y_t, x_t satisfy $q(y_t,x_t,\alpha) = 0$. However in any time period $q(y_t,x_t,\alpha)$ may be subject to a random deviation from zero. This deviation is considered equally likely to be positive or negative and decreasingly likely as its absolute value increases. This suggests u_t should be modelled as a bell shape distribution centered on zero. The normal is one such distribution and has the added advantage of making analysis of the model tractable. The properties of LS & IV estimates have been analyzed in the literature, but little is known of the properties of ML in nonlinear models.

In this dissertation we are concerned with the situation in which y takes on values in R^m and $q(\cdot)$ is an unspecified function but subject to certain regularity conditions. Necessarily some nonlinear models, for instance qualitative response models, are not encompassed by our analysis. Within this framework we examine the conditions under which the full information ML estimator is consistent and asymptotically normally distributed. From standard likelihood theory it is known that the MLE is consistent, and both asymptotically normally distributed and the most efficient when the model is correctly specified. In this thesis we concentrate on the degree to which the MLE retains these properties when the true distributon is nonnormal, and so can be considered robust to departures from normality.

The question of the robustness of an estimator is of considerable importance. The eventual power of the model for either forecasting or policy analysis, as well as its accuracy in explaining the data, depends on the use made of our a priori knowledge, which is at best tentative, and specification searches consisting of a succession of diagnostic tests of model adequacy. There is no unique ordering for applying tests, nor any guarantee that different permutations of the sequence lead to the same conclusion. There is, consequently, no guarantee that the original specification was correct nor that the model selection procedures are sufficiently sophisticated to indicate directions in which it might be improved. This is particularly true of the assumed error distribution. The normality specification captures a symmetric, or bell shape, error process in an analytically tractable fashion. As it is not the only choice satisfying this requirement it is important to be aware of any biases in inference caused by its incorrect imposition.

These reservations about test procedures have ramifications for the interpretation of an econometric model. It is important to distinguish between the data generation process (dgp) and approximations to it. If it is possible to find a functional transformation $q(\cdot)$, subject to the conditions in (1), that represents the exact mechanism by which a change in the economic environment effects the behavior of y_t, then this particular representation is the dgp. In the abscence of knowledge about the appropriate choice of $q(y_t, x_t, \alpha)$, the model specification used by practitioners to explain the

interrelationship between series is a synthesis of a priori economic theory and diagnostic tests. It has been noted above that such a procedure lacks the sophistication to infallibly determine the dgp. Therefore the econometric model is best regarded as an approximation to the dgp, whose accuracy depends on the estimation and model selection procedures employed.

This is at the centre of the debate on the Lucas policy critique. Lucas (1976) argued that econometric models could not be used for policy analysis as they were by their very nature self-falsifying. "Given that the structure of an econometric model consists of optimal decision rules of economic agents" (Lucas, 1976, p. 41) any change in a policy variable will alter the economic environment and therefore agents' reaction functions. The structure of the econometric model is consequently, he argued, changing with the policy variable over time. However only the outcomes, and not the decision making processes themselves, are observed. Given the reservations cited above about the genesis of a model specification, the equations are, therefore, better interpreted as approximations to the underlying reaction functions. In this case, as Sims (1982) notes, Lucas' conclusion reduces to the point that

> "Statistical models are likely to be come unreliable when extrapolated to make predictions for conditions for outside the range experienced in the sample" (Sims, 1982, p. 122)

1.2 The linear model as an approximation

The eventual model formulation depends in part on our

original specification. A lot of attention has focused on the use of linear models to explain economic series. These have the advantage of relative computational ease compared to nonlinear models, and so it is important to consider in what situations the choice of a linear model may be suitable. Our arguments suggest that in a large number of cases such models are inappropriate, and so, there is a need to develop the theory of their nonlinear counterparts. For this section we confine attention to scalar y_t and a vector of exogenous variables, but the arguments can be generalised to vector y_t. We consider two justifications for the linear form

$$y_t = x_t'\alpha + u_t, \qquad (2)$$

as an approximation to a nonlinear dgp: the normality of (y_t, x_t') and first order Taylor series expansions.

If (y_t, x_t') have a joint normal distribution then $x_t'\alpha = E(y_t|x_t)$. The assumption of normality can be justified quite easily if y_t is an aggregate, by appeal to central limit theorems. However the sample sizes for which these hold will vary from case to case. If y_t is not an aggregate then, from the Edgeworth expansion of its p.d.f., the normality of y_t results from the assumption that all its cumulants higher than the second are zero.

Alternatively it may be argued that if the dgp is $y_t = f(x_t) + v_t$ then if we take a first order Taylor series expansion about the sample means as follows,

$$y_t = f(\bar{x}) + \sum_{i=1}^{k} (x_{it}-\bar{x}_i)\frac{\partial f}{\partial x_{it}}\Big|_{\bar{x}}$$

$$+ \sum_{i,j=1}^{k}\sum (x_{it}-\bar{x}_i)(\bar{x}_{jt}-\bar{x}_j)\ \frac{\partial^2 f}{2x_{it}\partial x_{jt}}\Big|_{\bar{x}} + \ldots + v_t ,$$

then equating higher order terms to a white noise random variable (r.v.) independent of v_t, we have a justification for the linear model. There are two main flaws in this argument. Firstly, as noted by Bowden (1974), the derivatives are state dependent, and therefore not fixed as assumed in the linear model. Secondly, as White (1980) has argued, the Taylor series is only valid as a local approximation whereas we wish to explain behavior throughout the sample space, and use dispersed data to estimate the parameters.

Linear models are also encountered in the time series literature. The Wold decomposition theorem establishes that a stationary series can be split into deterministic and non deterministic components, and that this nondeterministic component has an infinite order moving average representation. The removal of trend and seasonal factors from economic series is usually thought to render them stationary and nondeterministic. A more parsimonious representation of this component is an ARMA model and, by using stationarity to pool information, the appropriate order of the model can be identified by the correlogram and partial autocorrelation function of the series. The model in (2) can be derived as a set of parameter restrictions on a multivariate ARMA model for (y_t, x_t'). The Wold theorem only states that this moving average representation exists and not that it is unique. Recent work by Granger and

Andersen (1978) has demonstrated that identification via the correlogram is only unambiguous within the class of linear models. It can be shown that bilinear models of the form,

$$y_t = \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{i=0}^{q} b_i u_{t-i} + \sum_{k=1}^{r} \sum_{m=1}^{s} c_{km} u_{t-k} y_{t-m}, \qquad (3)$$

with $c_{km} = 0$ for $k > m$, have the same autocovariance structure as an ARMA(p,max{q,s}) model. Higher order correlations will be needed to uniquely identify a model within this class, but the complicated nature of this analysis has tended to result in information criteria being used to discriminate between bilinear models. However Granger and Andersen's results underline that the linear representation, whilst analytically tractable, is not accorded any statistical optimality by the Wold theorem. Rather it is just one model formulation consistent with the sample autocorrelation structure.

The use of linear models may be appropriate in certain cases either because the dgp itself is linear or as an approximation to a nonlinear dgp. Whilst a linear model has the advantage of analytical tractability our review of the theoretical justifications for its use, suggest that it is by no means always a suitable model choice. These are also grounds for expecting traditional model diagnostics to be inadequate indicators of situations in which estimated linear model can be improved on by adopting a nonlinear formulation. The interpretation of specification tests is normally within the context of the linear framework. Tests for incorrect functional form have been developed in the literature but the choice of alternative hypothesis, and its

interpretation if accepted, may be problematical. We do not examine these issues but concentrate on the properties of estimators once a nonlinear formulation is chosen.

1.3 Time varying linear models as an approximation to nonlinear models

Given the data dependence of the derivatives in a Taylor series approximation, the natural extension to the linear approximation is to adopt a time varying linear model. In this case the coefficients on the x_t are regarded as altering overtime with certain properties of their behavior known. An example of this is the state space system, outlined for instance by Harvey (1981), in which parameter estimates are updated after each observation by an updating procedure such as the Kalman filter. This model is suitable for evolutionary processes, but we argue below that its dependence on past observations may make it inapplicable for modelling nonlinear systems. An alternative is to employ switching regression models, which constitute an extreme form of varying parameter model. These have been suggested by Tong and Lim (1980) in the time series literature, and are familiar in econometrics with reference to markets in disequilibrium. Tong and Lim's (1980) threshold autoregression model takes the form

$$y_t = B(J_t)y_t + A(J_t)y_{t-1} + e_t(J_t) + c(J_t),$$

where y_t is a vector of endogenous variables in period t, $A(j)$, $B(j)$ are matrices of fixed coefficients and $e_t(j)$ is strict white noise. The model changes according to the

value of the indicator variable J_t which determines the value of $B(J_t)$, $A(J_t)$, $C(J_t)$ and the distribution of $e_t(J_t)$.

Whilst this formulation is of little practical use in most econometric settings it does highlight the potential weakness of time-dependent parameter models. The problem is that knowledge of an appropriate indicator is required, but this is unlikely to be available due to the unknown nature of the dgp. This approach is, however, more consistent with the idea of different linear approximations to an underlying nonlinear dgp. In any neighbourhood of a particular point, $\bar{y}_t$, the behaviour of y_t can be explained by a linear Taylor series approximation with fixed coefficients. However as y_t, and so the centre of the expansion $\bar{y}_t$, moves through the sample space the coefficients of the linear expansion change. However there is no reason to suppose they evolve by a particular stochastic law. If we regard the appropriate linear approximation as being indexed by some state dependent variable, then in varying parameter models in which the coefficients are presumed to evolve over time by some stochastic process, past observations from other regimes are still affecting the estimates. For instance if we pass the hypothetical switch point, the varying parameter model still bases its coefficient estimates on the previous regime. Harrison and Stevens (1976) have sought to adapt the state space representation to a Bayesian framework. This allows the intervention of subjective information in the updating to weight more heavily the last observation when there is reason to expect previous experience to be misleading. The examples they give for this model are short term sales forecasting, when information about market

climate in the next period may well be available. However we typically do not know when the neglect of the underlying nonlinearities of the system will make our model unreliable.

1.4 Summary

We have argued above that linear models with or without time varying parameters are not necessarily always suitable approximations to the dgp. In this thesis we consider situations in which a more general nonlinear model is deemed appropriate. The majority of our analysis deals with models of the generality of equation (1) and is concerned with the properties of the MLE once a functional form has been chosen, and not with methods of selecting the functional form. The consistency and asymptotic normality of the estimator are, of course, prerequisites for specification searches for a better approximation using conventional test procedures such as the Wald, likelihood ratio or score tests.

This work is based on a synthesis of two areas of the literature, and develops new analytical results to answer questions previously unexplored in those areas. Existing work on the properties of estimators in linear and nonlinear models tends to assume the model specification is correct and explores what parts of the specification can be relaxed without losing the desirable properties of the estimator. This is different from the approach taken by White (1982) who examines the properties of the MLE when it is admitted from the outset that the model is misspecified (in this case the estimator is called the quasi MLE (QMLE)). White (1982) derives conditions for the convergence in probability of

this estimator to the value that minimises the Kullback Liebler (1951) information criterion (KLIC). Our work follows the practice of the simulataneous equations model (SEM) literature and considers conditions for the convergence of the QMLE to the true value in nonlinear models.

In chapter 2 we discuss the literature on linear SEM's and the interrelationship between the three stage least squares, full information MLE and full information instrumental variables estimator. The aim is to demonstrate the line of argument by which previous authors have established the consistency and asymptotic normality of the MLE in this situation. This work would appear a logical starting point for deriving analogous results for nonlinear models, and so we need to identify at which stages of these arguments linearity is crucial. We also consider the advantages of estimating equations simultaneously (full information (FI) estimation) as opposed to individually (limited information (LI) estimation). In this thesis we focus purely on full information estimators.

In chapter 3 we survey previous explorations of the properties of these three estimators in nonlinear models. Amemiya (1977) has shown that the instrumental variable interpretation of MLE does not persist to nonlinear models, and so Hausman's (1974) proof of the consistency of the MLE does not generalise from linear to nonlinear models. Phillips (1982) has shown that there must exist classes of the distributions for which ML estimation under normality provides consistent estimates. However very little is known about the size of this class of true distributions and we

argue that the approach taken by Phillips (1982) cannot be extended to provide information on this issue. We also consider the conditions under which an asymptotic theory for nonlinear models can be developed.

Chapter 4 contains an outline of the necessary results from the misspecified model literature. We show that the focus of our work is different from that of White (1982). He derives conditions for the convergence in probability of the QMLE to the KLIC minimising value, whereas we examine the conditions under which this value is in fact the true value. We also explore the difficulty of verifying second order conditions for consistency, and the use of distribution free identification criteria to check these conditions. Attention is focused on the criteria developed by Brown (1983) for nonlinear-in-variables models.

In chapter 5 we consider various alternative analytical approaches to that of Phillips (1982) for deriving conditions for the consistency of the MLE. We establish that there exists a family of weakly stationary true error processes whose conditional distribution varies overtime, for which the MLE under the assumption of independently and identically distributed (i.i.d.) normal errors provides consistent estimators. However the analytical derivation of nonnormal i.i.d. true error distributions, for which ML estimation under normality retains these desirable properties, depends on the nature of the reduced form. If it can be written down explicitly then we can find true distributions for which NLFIML is consistent, although the class is likely to be much narrower than its linear model counterpart, as it depends on the nonlinearities in the

system. We provide some examples of economic interest to illustrate this point.

In chapter 6 we consider the case where the reduced form is implicit. We show that the condition for consistency involves all the moments of the distribution. In this case the analytical results available are that NLFIML is consistent when the model is correctly specified or if the error is from the class of distributions considered by Phillips. However Phillips' proof only establishes the existence of such a class, and as its exact nature varies from case to case, our results suggest that if we require consistent and asymptotically normal estimates, NLFIML should not be used when the reduced form is implicit.

We explore the conditions for a set of structural equations, such as (1), to imply an uniquely defined reduced form. An examination of the work of Gale and Nikaido (1968) shows that these conditions are more strigent than is usually recognised in the econometrics literature. Finally we consider the conditions for the asymptotic normality of NLFIML. White (1983) observes the importance of consistent estimation of the first moment for that of the covariance of the QMLE. Whilst White's analysis contains an algebraic slip, the essence of his comments retains its validity. Without consistent estimates of the covariance, traditional testing procedures based on the parameter estimates break down. In contrast NL3SLS is consistent and asymptotically normal under the same moment conditions as in the linear model, and so would appear the prefered estimator.

Chapter 7 contains a discussion of the conditions under which our conclusions about the properties of NLFIML can be

extended to dynamic models. We examine the types of dynamic processes for which we can apply a version of the strong law of large numbers and so replicate our earlier analysis for static models. Current practice is to employ either martingale or mixing process arguments. McLeish (1975) has shown both types of processes to be mixingales for which the desired law of large numbers can be derived. White and Domowitz (1983) have advocated the use of mixing processes as they have the advantage that functions of them are themselves mixing processes, and so their use involves one basic assumption about y_t. Whereas the martingale arguments entail a series of assumptions about functions of y_t invariably without examining their consequence for the underlying series. However we argue using some results due to Jones (1976) that, contrary to the view apparently expressed by White and Domowitz, the theoretical validation of whether a particular series generated by a model is in fact a mixing process, is likely to prove impossible.

This chapter also contains an extension of a proof by Heijmans and Magnus (1983a) of consistency of the MLE, under weak conditions on the underlying process, in correctly specified models to the case of misspecified models. We show that the MLE converges to the KLIC diminishing value in their framework. Finally, we consider the conditions for asymptotic normality of the QMLE in dynamic models. In particular we focus attention on the choice of scaling factor. White and Domowitz (1983) present a central limit theorem that requires a constant scaling factor multiplied by the increase of the square root of the sample size. They hypothesise that a non constant scaling factor may induce a

nonnormal asymptotic distribution. We argue, using the work of Hall and Heyde (1981), that this need not be the case.

In chapter 8 we argue that the information matrix test suggested by White (1982) is a natural test of model specification when employing the pseudo maximum likelihood estimation strategy, advocated by Gourieroux, Monfort and Trognon (1984a), for the nonlinear regression model. We calculate the appropriate tests for the Poisson model example considered by Gourieroux, Monfort and Trognon (1984b). The resulting tests of distribution are compared with goodness of fit tests. We compare the higher order likelihood derivative tests (suggested by Chesher, 1983) based on the standard normal likelihood with the tests based on Edgeworth expansions (Keifer and Salmon, 1983) and show that they coincide for tests of the third and fourth moments but not for the fifth. Finally it is shown that the decomposition of the information matrix test in the linear model regression model, demonstrated by Hall (1982), can be extended to its nonlinear counterpart.

Chapter 9 contains some conclusions, after which some proofs are presented in the appendix.

2. STATISTICAL PROPERTIES OF ESTIMATORS AND LINEAR MODEL RESULTS

2.1 INTRODUCTION

The properties of and relationship between maximum likelihood (ML), least squares (LS) and instrumental variables (IV) have been explored at length in the literature for the linear model. It is well known that all three can be considered IV estimators, which provides a convenient proof of their consistently and asymptotic normality provided the error process has mean zero. Whereas ML under normality is the most efficient if the specification is correct, a class of IV estimators, including LS, are asymptotically equivalent. In this chapter we outline the basis of these results to illustrate both why linearity delivers such powerful results and why the type of arguments used cannot necessarily be generalised to the nonlinear setting. We also introduce and discuss the criteria for choice of estimators, identification and full or limited information estimation of systems of equations, the basic theoretical issues of which are relevant to all models.

2.2 Choice of Estimators in Classical Statistics

The majority of econometric theory is based on classical statistics. Probability statements have a frequentist interpretation as the situation envisaged is one in which the researcher can generate unlimited data by repeating the experiment under identical conditions. In econometrics the data are observed passively and so it is necessary to make regularity assumptions, such as

stationarity, before the classical framework can be used. This done, we hypothesise a probability model of the form $q(y,x,\alpha) = u$, with assumptions about $u,y,x,q(\cdot)$, to explain the observed relationships between economic variables. The model is indexed by an unknown parameter vector $\underline{\alpha}$ and the aim of classical statistics is to reduce our uncertainty about α by point and interval estimation using information in the data. The point estimate of α, $\hat{\alpha}$, is a function of random variables and so is itself stochastic. The interval estimate, or hypothesis test, gives an idea of the sampling distribution of $\hat{\alpha}$ and so of the degree to which $\hat{\alpha}$ evaluated at the realised data values is a "true" reflection of α.

We can construct any number of estimators from the data, but as our inference depends on $\hat{\alpha}$ it is desirable to have some method of "screening out" poor estimators. The classical criterion for achieving this is to require $\hat{\alpha}$ to be (i) unbiased: $E(\hat{\alpha}) = \alpha$ and/ or (ii) consistent: $\text{plim}\hat{\alpha} = \alpha$. The estimator chosen is the most efficient (in the sense of having minimum variance), of those satisfying (i) and (ii).

In econometric models an estimator is usually a complicated function of the error random variables making its small sample distribution analytically intractable and so the discussion is limited to large sample properties, namely consistency and asymptotic efficiency. The problem of interval estimation reduces to finding the conditions for consistency and asymptotic normality of $\hat{\alpha}$ under particular circumstances. The argument is that whilst we may know nothing of its small sample behavior, an estimator is dismissed if its performance is not good in large samples.

However any interval estimation using asymptotic results requires the assumption that indeed the sample size is large enough, although this is rarely checked. Asymptotic theory can be regarded as an approximation to the finite sample result. In any particular situation better approximations can be developed from the asymptotic estimates by using Edgeworth expansions to analyze the effects of the largest asymptotically negligible terms in the distribution function of the estimator.

2.3 Identification

The analysis of the properties of estimators presupposes that the parameters can be uniquely determined from the data or, in statistical parlance, that the model is identified. Economic theory has limited our attention to a particular family of probability distributions, termed the model, but what we seek is the structure, the particular distribution, most likely to have generated the data. The problem of lack of identification is essentially one of observational equivalence. This arises when two structures are identical, and so indistinguishable from sample data. A structure is identifiable if, and only if, there are no observationally equivalent structures, in which case the parameters can be uniquely determined from the data.

A well known example of lack of identification is when the common factor restriction occurs in ARMA models. Consider the stationary ARMA(1,1) model:

$$y_t = \phi y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t, \quad |\phi| < 1. \tag{3}$$

By repeated substitution for lagged values of y, (3) can be written as

$$y_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} + \theta \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j-1},$$

$$= \varepsilon_t + (\phi+\theta) \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j-1}.$$

Any structure for which $\phi = -\theta$ is not identifiable as then y_t is white noise. This problem can occur, with the same consequences for identification, in a more general model

$$H(L)y_t = \Phi(L)\varepsilon_t,$$

if $H(L) = \Psi(L)H^*(L)$ and $\Phi(L) = \Psi(L)\Phi^*(L)$. The model cannot be identified due to the common roots shared by both polynomials H(L) and $\Phi(L)$.

The problem of lack of identification is essentially one of insufficient information to enable the parameters to be determined. This can be offset by introducing additional information into the problem, in the form of parameter restrictions. These can take two forms: nonstochastic restrictions on α and/or stochastic restrictions on the p.d.f. of u. For a structure to be model admissible, it must satisfy these restrictions, and it is hoped that sufficient restrictions can be imposed to reduce the number of model admissible structures to one.

Identification is a general statistical problem, but in econometrics it is normally associated with simultaneous equation models. For illustrative purposes we consider the static linear model

$$B'y_t + \Gamma'x_t = u_t, \qquad t = 1,\ldots,T,$$

where y_t is a N x 1 vector of endogenous variables, x_t is a K x 1 vector of exogenous variables and u_t is a N x 1 vector of mean zero disturbances with contemporaneous covariances matrix Σ and $E(u_t u_s') = 0$. The reduced form for y_t is

$$y_t = \Pi x_t + v_t, \qquad t = 1,\ldots,T,$$

where $v_t = B'^{-1}u_t$. Note that we require B to be nonsingular for there to be a unique reduced form associated with the structural equations. We return to the conditions for such a mapping between y and u in a more general setting in chapter 6. The reduced form is necessarily identified and the identification of the structural equations depends on whether given estimates of Π we can uniquely determine (B,Γ). The relationship between structural and reduced form parameters is given by

$$AW = 0 \quad \text{where} \quad A = [B':\Gamma'],\ W' = [\Pi':I_k].$$

As the system stands there is insufficient information to estimate the parameters of the i^{th} equation, α_i. They must satisfy the restrictions $\alpha_i'W = 0$ but as rank(W) = K there are only k linearly independent restrictions on the N+K elements of α_i. However if we know that the coefficients have linear restrictions between them of the form $\alpha_i'\phi = 0$, then this information can be used to achieve identification. The vector α_i must then satisfy $\alpha_i'(W:\phi) = 0$, and so a necessary and sufficient condition for it to be identified

up to a scalar multiple is that rank$(W:\phi)$ = N+K-1. The matrix $[A':W]$ is a nonsingular matrix of dimension N+K and so its columns form a basis for R^{N+K}. We can therefore write

$$\phi = A'\xi + Wn,$$

and as $A\phi = AA'\xi$, because $AW = 0$, rank$(A\phi)$ = rank(ξ). This enables the condition for identification to be restated in a more convenient form. For rank$(W:\phi)$ to be N+K-1, we require there to be N-1 linear independent, both of themselves and W, columns in ϕ. We therefore require rank$(A'\xi)$ = N-1, but this in term implies that rank(ξ) = rank$(A\phi)$ must equal N-1. A necessary and sufficient condition for identification in this model is therefore rank$(A\phi)$ = N-1.

Note we have sought identification up to a scalar multiple because this type of operation on the parameter vector does not alter the content of the equations. An alternative is to fix one parameter to a set value, for instance unity, and require unique identification because this normalisation of the equation means that the multiplication of the remaining coefficients in the equation by a scalar alters the nature of the structural equations.

This condition relies on the nonstochastic equations $A\phi = 0$ and the stochastic restriction that $E(u_t) = 0$. An alternative motivation for the result is based on the idea of observational equivalence. If the model is identified then the transformed structural equations

$$FB'y_t = F\Gamma'x_t + Fu_t,$$

(F nonsingular) should only be observationally equivalent to the original structure if F = I. This can be checked by examining the first and/or second moments of the transformed system. The first moment approach gives the already derived rank condition. The second moment approach uses the fact that if two structures are observationally equivalent u_t and Fu_t must have the same covariance matrix. However in the unlikely event of our possessing detailed knowledge of the second moment of u_t, this approach yields insufficient restrictions as $E(u_t u_t')$ has only $N(N-1)/2$ distinct off diagonal elements, and so even if we assume $\Sigma = \sigma^2 I$, we only reduce the class of admissible F to be orthogonal matrices. Although identification could then be achieved by assuming the system to be recursive, and so B would be triangular.

Our original derivation is specific to linear systems, makes only a first moment restriction on the errors, and uses no further distributional assumptions. Alternatively we can condition on the distribution of the errors and derive conditions for local identification of the model. Rothenberg (1971) and Bowden (1973) have demonstrated that the parameter vector, α, is identified at α_0 if the information matrix, defined as the expected value of the hessian of the likelihood, is positive definite at that point. Rothenberg (1971) shows that if u_t is distributed normally then the rank condition again results for the linear model. We return to those arguments later in our discussion of the conditions for consistency of an estimator in a general model.

2.4 Information & Estimation

Having considered the identification of a simultaneous equations model, we now examine the methods suggested for its estimation. In practice there are three main approaches: least squares (or minimum distance), instrumental variables and maximum likelihood. Within the normal linear model these three are closely related and before exploring the extent to which this relationship persists in the nonlinear setting, in chapter 3, we first outline the arguments used to establish the properties of these estimators in the linear model.

As in the identification stage, the proposed methods differ in their explicit distributional assumptions. Least squares and instrumental variables are distribution free, in the sense that assumptions are only made about the first two moments of the error process. However the exogeneity of certain variables will be crucial to the construction of these estimators. It has therefore been implicitly assumed that the factorisation of the joint distribution into the conditional and marginal densities has produced a sequential cut on the parameters of this model. Normality is, of course, sufficient for this, but in some cases e.g., the multivariate t, the cut will not occur (see Engle, Hendry and Richard, 1983).

In utilising the extra information about the distribution in ML one would intuitively expect to produce more efficient estimators if the asumption is correct, but at the expense of bias if it is false. This robustness/ efficiency tradeoff is present in the linear model in small samples only, but before considering its origins we must

examine the links between identification, information and estimation, the ideas behind which are relevant to all models.

The efficiency of an estimator clearly depends on the amount of information used. In our discussion of identification we were solely concerned with whether we had sufficient information to be able to determine the unknown parameters uniquely from the data. The distinction then was between just and under identification. For our discussion of estimation we need to distinguish a third situation, namely that of overidentification. This occurs when there is more than enough independent information to identify the parameters. For an estimation procedure to be efficient it will have to take account of all these restrictions, as the use of one set of just identifying restrictions does not guarantee the remaining independent restrictions on an equation will be satisfied. In the linear model the properties of LS estimators are closely related to the degree of identification, as both two and three stage LS (2SLS and 3SLS) are undefined when the system is underidentified, but equal when the system is just identified. The existence of estimators in all models will depend on the number of observations, or rather amount of information, relative to the number of variables. LS and ML break down in the undersized sample case, where there are less observations than exogenous variables, and in the course of this chapter we note the methods used to overcome this problem.

There may similarly be an information loss from estimating each equation in isolation. Such limited

information (LI) techniques ignore the information contained in the rest of the system about a particular equation, and so will never be more efficient than full information (FI) methods which incorporate all restrictions. Against this has to be set the fact that our specification is often tentative, and so some restrictions may be incorrect. The tradeoff to the efficiency of FI may well be a lack of robustness as it allows any erroneous restrictions on one equation to potentially affect the estimation of the whole system. Sims (1980) has argued for the need to match the estimation approach to the manner in which restrictions are placed. If the system is treated equation by equation at the specification stage, which defines the restrictions, it should then be estimated by a LI method. Typically a system with a LI specification but estimated by FI methods will not appear the appropriate formulation when submitted to model diagnostics. The a priori restrictions should therefore be placed by consideration of the entire system. The difficulty of making such restrictions, Sims sees as a further support for his reduced form estimation using vector autoregressions. In this thesis we are concerned purely with the properties of FI estimators.

2.5 LS, IV and ML in the normal linear model

Within the normal linear SEM there is a close relationship between LS, IV and the ML estimators. Hausman (1974) has shown that both 3SLS and FIML can be considered as IV estimators and this approach will prove convenient for examining consistency and normality of the estimators. Hendry (1976) has shown that IV and 3SLS can be considered

as approximations to FIML, and his "estimator generating equation" approach highlights the loss of information, and so (small sample) inefficiencies, of 3SLS and IV. In all of the subsequent analysis systems of equations are assumed to be identified.

Consider the model

$$YB + X\Gamma = U, \tag{4}$$

where Y is a T x N matrix of jointly dependent variables, X is a T x K matrix of predetermined (weakly exogenous) variables, U is a T x N matrix of structural disturbances of the system, T is the number of observations, B is assumed to be nonsingular, $E(X'U) = 0$, and $E(UU') = \Sigma \otimes I_T$. Therefore we are allowing contemporaneous but not intertemporal correlation between disturbances. The equation used in our discussion of identification in SEMS in 2.3 is the transpose of the t^{th} row of (4). If we impose normalisation then the i^{th} equation of the system can be written as

$$y_i = Z_i \delta_i + u_i, \qquad (i = 1,\ldots,N),$$

and the whole system as,

$$y = Z\delta + u, \tag{5}$$

where

$$Z = \begin{bmatrix} Z_1 & & & 0 \\ & Z_2 & & \\ & & \ddots & \\ 0 & & & Z_N \end{bmatrix}, \quad Z_i = [Y_i X_i], \quad \delta_i' = [\beta_i' \gamma_i'],$$

y_i and u_i are the i^{th} columns of Y and U respectively, vecY = y, VecU = u, and β_i, γ_i are the unrestricted coefficients on the endogenous and predetermined variables in the i^{th} equation. Let the reduced form associated with this system be

$$Y = X\pi' + V \tag{6}$$

where $V = UB^{-1}$, $\pi' = \Gamma B^{-1}$

Brundy and Jorgenson (1971) define the instrumental variable estimator of δ as d, the solution to the equations

$$(W'Z)d = W'y, \tag{7}$$

where W is the matrix of instruments satisfying the following conditions:

(i) $\operatorname{plim} \frac{1}{T}W'u = 0$,

(ii) $\operatorname{plim} \frac{1}{T}W'W$ is finite and nonsingular,

(iii) $\operatorname{plim} \frac{1}{T}W'X$ is finite.

Therefore,

$$d = (W'Z)^{-1}W'y,$$

$$d = \delta + (W'Z)^{-1}W'u,$$

and so, given that we can apply the central limit theorem

to $W'u/\sqrt{T}$, we have

$$\sqrt{T}(d-\delta) \overset{a}{\sim} N(0, \text{plim } (\frac{W'Z}{T})^{-1}(\frac{W'W}{T})(\frac{Z'W}{T})^{-1}).$$

The IV estimator is consistent and asymptotically distributed as normal, provided the conditions on W and the first two moments of u are satisfied.

Brundy and Jorgenson also prove that for W to yield an asymptotically efficient d, it must be chosen so that the i-j^{th} block of W, W_{ij}, is equal to (W_{ij1}, W_{ij2}), where

a) $$\text{plim } T^{-1}W'_{ij1}X = \sigma^{ij}\pi_j' \text{plim } \frac{1}{T} X'X,$$

b) $$\text{plim } T^{-1}W'_{ij2}X = \sigma^{ij} \text{plim } \frac{1}{T} X_j'X,$$

(where the i-j^{th} elements of Σ and Σ^{-1} are σ_{ij} and σ^{ij} respectively). One possible selection is to put $W_{ij} = [\hat{\sigma}^{ij}X\hat{\pi}_j, \hat{\sigma}^{ij}X_j]$, where $\hat{\pi}_j$, $\hat{\sigma}^{ij}$ are consistent estimators of π_j, σ^{ij}. Of course the 3SLS estimators,

$$\hat{\delta}_{3SLS} = [Z'(S^{-1} \otimes X(X'X)^{-1}X^1)Z]^{-1}[S^{-1} \otimes X(X'X)^{-1}X^1]y,$$

falls into this class. At the first stage the reduced form is estimated by OLS to derive $\hat{\pi}_j$. Each structural equation is then estimated individually by the IV estimator with $W = [X\hat{\pi}_j, X_j]$: this gives the 2SLS (limited information) estimators of δ_i, $i = 1,\ldots,N$. The consistent estimator of Σ, S, is constructed by putting its i-j^{th} element, $\hat{\sigma}^{ij}$, equal to $T^{-1}\hat{u}_i'\hat{u}_j$, where $\hat{u}_i$ is T x 1 vector of residuals resulting when the 2SLS estimators are fitted to the i^{th}

structural equation. Provided the structural equations are just identified 3SLS uses the most efficient* estimator of π_j in the first stage. However it is shown on page 32 that it is not the most efficient estimator in small samples, although all IV of the form above are equally efficient asymptotically.

To derive the ML estimator for this model we assume that U is distributed multivariate normal. The log likelihood for the model in (4) is therefore

$$L(B,\Gamma,\Sigma) = \text{const} + \frac{T}{2}\log\det(\Sigma)^{-1} + T\log\det(|B|)$$

$$- \frac{T}{2}\,\mathrm{tr}\left[\frac{1}{T}\,\Sigma^{-1}(YB + X\Gamma)'(YB + X\Gamma)\right].$$

The first order conditions for optimisation are then

$$\frac{\partial L}{\partial B} = T(B')^{-1} - Y'(YB + X\Gamma)\Sigma^{-1} = 0, \tag{8}$$

$$\frac{\partial L}{\partial \Gamma} = -X'(YB + X\Gamma)\Sigma^{-1} = 0, \tag{9}$$

$$\frac{\partial L}{\partial \Sigma^{-1}} = T\Sigma - (YB + X\Gamma)'(YB + X\Gamma) = 0. \tag{10}$$

To establish the IV interpretation of FIML, Hausman (1974) concentrates the first order conditions with respect to T. From (10),

$$T = \Sigma^{-1}(YB + X\Gamma)'(YB + X\Gamma),$$

and substituting this into (6) gives the equations

$$\begin{bmatrix} -X' \\ (B')^{-1}\Gamma'X' \end{bmatrix}(YB + X\Gamma)\Sigma^{-1} = 0, \tag{11}$$

* Throughout this thesis we refer to the estimator with the minimum (asymptotic) variance as being (asymptotically) most efficient.

in terms of our notation in model (5) in which the coefficients are stacked in vector form, (11) can be rewritten as

$$\begin{bmatrix} \hat{Z}_1 & & & 0 \\ & \hat{Z}_2 & & \\ & & \ddots & \\ 0 & & & \hat{Z}_N \end{bmatrix} (y - Z\delta)(\Sigma^{-1} \otimes I) = 0,$$

which implies the FIML estimator of δ is

$$\hat{\delta} = (W'Z)^{-1}W'y,$$

where $W' = \hat{Z}'(S \times I_T)^{-1}$, $\hat{Z} = \text{diag}(\hat{Z}_1,\ldots,\hat{Z}_N)$, $\hat{Z}_i = [X(\hat{\Gamma}\hat{B}^{-1})_i X_i, X_i]$, and $S = T^{-1}(Y\hat{B} + X\hat{\Gamma})'(Y\hat{B} + X\hat{\Gamma})$.

The equations are nonlinear in $\hat{B}$ and $\hat{\Gamma}$ and so have to be estimated iteratively, giving the estimator after the k^{th} iteration as

$$\hat{\delta}_{k+1} = (W_k'Z)^{-1}W_k'y,$$

the instruments, W_k, being revised at each step by updating $\hat{Z}_i$ and S using the parameter estimates from the last iteration. We have assumed that the second order moments are finite and nonsingular, where appropriate, and so $\hat{\delta}_{k+1}$ may be considered an IV estimator, for every k, as it satisfies all the necessary requirements. The asymptotic normality and consistency of $\hat{\delta}$ follow from the arguments above, and so are guaranteed for a wide class of nonnormal error distributions.

The relationship between the information sets used in

LS and ML has been explored by Hendry (1975) via the estimator generating equations of the system. If we concentrate the log likelihood with respect to Σ^{-1} from (10), and stack the first order conditions on the unrestricted elements of $A = [B,\Gamma]$ in a vector we have

$$[(B'^{-1}:0) - \Sigma^{-1}A'(Z'Z/T)]^{r} = 0, \tag{12}$$

where Z is now [Y:X] and $[D]^{r}$ denotes the operation of stacking the unrestricted elements of the columns of D into a vector.

From (10) we have

$$B'^{-1} = \Sigma^{-1}A'(Z'Z/T)A\ B'^{-1} = T^{-1}\Sigma^{-1}A'Z'(Y-X\Pi'). \tag{13}$$

Taken together (12) and (13) imply

$$q = (\Sigma^{-1}A'(Z'X/T)Q')^{r} = 0,$$

where $Q' = (\Pi':I)$. Therefore ML estimators of A and Σ must satisfy the following equations:

$$\begin{aligned}
&(\Sigma_2^{-1}A_3'(Z'X/T)Q_1')^{r} = 0 && \text{, (i)} \\
&\Sigma_2 = A_2'(Z'Z/T)A_2 && \text{, (ii)} \\
&Q_1' = (\Pi_1':I) && \text{, (iii)} \\
&\Pi_1' = -B_1^{-1}\Gamma_1 && \text{. (iv)}
\end{aligned} \tag{14}$$

These are the estimator generating equations. The subscripts denote the order in which the estimators are obtained in the iterative process. Note the system is linear in A, given Q and Σ.

FIML is the most efficient estimator because it is based on all the above equations. 3SLS, however, ignores (iv) in the construction of its instruments, which come from the unrestricted estimation of the reduced form. If there were overidentifying restrictions on A then this would impose restrictions on Π and so 3SLS implicitly assumes each equation to be just identified. The 3SLS procedure can be iterated as well by either revising Σ_2 from the 3SLS residuals or revising Σ_2 and Q by using (iv) and the 3SLS estimates. Only the second method uses the complete information set, and so gives FIML on convergence.

Solving the equations in (14) is computationally burdensome, and so we may seek algorithms that ease this burden but give FIML estimates, or algorithms that only approximate FIML. IV estimators of the class described by Brundy and Jorgenson fall into this second group. The most efficient of these is 3SLS and although it ignores information, it is asymptotically equivalent to FIML. All IV estimators of this class have the same asymptotic distribution as FIML, provided there are no restrictions on Σ. These in turn are asymptotically equivalent to 2SLS and LIML when each equation is just identified. This underlines the point made earlier about the inefficiencies involved in ignoring overidentifying restrictions and using LI techniques.

In the linear model the efficiency/robustness tradeoff involved in the explicit use of the normality assumption in

estimation is only a small sample phenomenon. The FIML estimator is (asymptotically) robust for a wide class of nonnormal distributions as its consistency depends only on the errors having mean zero. FIML is also asymptotically distributed the same as a class of IV estimators, which can therefore be used to simplify the numerical computations required to produce an asymptotically optimal estimator.

In the above discussion of small sample behavior, we have assumed away the problem of the "undersized" sample by not considering constraints on T. For the first step of 3SLS we require at least as many observations as exogenous variables e.g. $T > K$. Sargan (1978) has shown that if $T < N+K$ then the log likelihood will be infinite and so have no maximum. This follows from an examination of the log likelihood concentrated with respect to Σ,

$$L = T \log|\det B| - \tfrac{1}{2}T \log\det(AZ'ZA').$$

If it is possible to find A_0 satisfying the a priori restrictions such that,

(i) $\det B_0 \neq 0$,

(ii) $\lambda' A_0 = \alpha'$, where $X\alpha = 0$,

then the first term of L is finite and the second infinite. Sargan shows that in the undersized sample case, we can find such an A_0 with probability one. In situations when ML and LS breakdown it may be possible to construct an IV estimator that would be asymptotically efficient. All that was needed

at the first stage was a consistent estimator of the reduced form coefficients. This can be derived by consistent, but inefficient estimation of the structural coefficients as

$$\text{plim } \hat{\Pi}' = - \text{plim } \hat{\Gamma} \text{ plim } \hat{B}^{-1} = \Gamma B^{-1} = \Pi' .$$

The matrices $\hat{B}$ and $\hat{\Gamma}$ can be derived by LIIV estimation of the structural equations using $W_j = [D_j, X_j]$ where D_j is a set of dummy variables associated with division of the sample into m_{j-1} subsets, where m_j is the number of endogenous variables in the j^{th} equation of (4). Each column of D_j has elements equal to unity for the corresponding subset and zero elsewhere. From equation (7) it can be seen that the condition for the block diagonal matrix $W'Z$ to be invertible is that $W_i'Z_i$ be nonsingular for all i. A necessary condition for this is that $T > m_i + K_i - 1$, and so the estimator can be constructed if $T < K$. However when the sample is not undersized, this method produces an inefficient estimator in small samples.

In this chapter we have seen that there is a close relationship between the estimators familiar in the linear model literature. Both 3SLS and FIML can be regarded as FIIV and so are consistent and asymptotically normally distributed. Further if our criterion for choice of estimator within this class is asymptotic efficiency there is nothing to choose between LS and ML estimators. In the next chapter we examine the extent to which the persistance, or lack of it, of these relationships and properties has been explored in the literature on nonlinear models.

3. ASYMPTOTIC THEORY AND EXISTING LITERATURE ON NLSEMS

3.1 Asymptotic theory in nonlinear models

We now consider the extent to which the close interrelationship between LS, IV and ML estimators in linear models can be generalised to nonlinear models. In particular we focus attention on the conditions for consistency and asymptotic normality of these estimators and the degree to which the lines of argument used to establish these properties in the linear model can be extended to this more general framework. However this presupposes that we can construct an asymptotic theory for nonlinear models. It will be seen below that the approach taken in the literature is to make analogous assumptions to those made in the linear model. To develop an asymptotic distribution theory, which must rest on the convergence of functions of the stochastic variables, restrictions will inevitably need to be placed on the class of model considered. In the linear model this is achieved by assuming that the cross product matrices converge to a finite limit and that the Central Limit theorem can be applied to $T^{-\frac{1}{2}} \sum_{t=1}^{T} w_t u_t$ for various w_t. By placing the appropriate restrictions on w we ensure that for any linear combination of u in a particular function of w, say L(u), the weight attached to a value of L(u) decreases to zero as $|u| \to \infty$, faster than $L(u) \to \infty$. If $\phi(u)$ is the p.d.f. of u then, algebraically,

$$L(u)\phi(u) \to 0 \text{ as } |u| \to \infty.$$

This controls the effect of outliers when evaluating the limit in probability of L(u), so that convergence occurs.

These assumptions translate easily into order of probability restrictions on the variables. In the nonlinear model we are concerned with the convergence of nonlinear functions of u, h(u), and the effect of outliers must be similarly restricted. The class of function must satisfy

$$L(u)\phi(u) \to 0 \text{ as } |u| \to \infty,$$

and so the choice crucially depends on the specified p.d.f. of u. These conditions are equivalent to requiring that the series of exogenous variables and errors be a Cesaro sum generator, in the terminology of Burguette, Gallant and Souza (1983). This is a series satisfying the following conditions from Gallant and Holly (1980). "Let v_t, t = 1,2,..., be a sequence of independent and identically distributed s-dimensional random variables defined on a complete probability space (Ω,A,P^*) with common distribution ν. Let ν be absolutely continuous with respect to some product measure on R^S and let b be a nonnegative function with $\int b d\nu < \infty$. Then there exists E with $P^*(E) = 0$ such that if $w \notin E$

$$\lim_{T\to\infty} (1/T) \sum_{t=1}^{T} f[V_t(w)] = \int f(v)d\nu(v),$$

for every continuous function with $|f(v)| \leq b(v)$." In the content of econometric models we have $v_t = (u_t, x_t)$ and, letting $y(u_t, x_t, \alpha)$ be the reduced form for y_t, this theorem gives us

$$T^{-1} \sum_{t=1}^{T} f(y_t, x_t, \alpha) \text{ and } T^{-1} \sum_{t=1}^{T} \int_u f(y(u, x_t, \alpha), x_t, \alpha)dP(u)$$

converge uniformly to

$$\int_x \int_u f(y(u,x,\alpha),x,\alpha)dP(u)d\mu(x),$$

where μ is the probability of measure of x. This of course depends on the existence of the bounding function with finite expectation. Again the arguments depend on the p.d.f. of u. The implications for the underlying variables are less clear in the nonlinear case, and invariably not explored. For the present we follow convention in making the usual assumptions. The question of whether they are necessarily too restrictive to make the results of no practical use in econometric models is considered in chapter 7.

3.2 Nonlinear 3SLS

The properties of nonlinear three stage least squares (NL3SLS) have been considered by Jorgenson and Laffont (1974) and Amemiya (1977). Jorgenson and Laffont's original treatment is for a model of the form

$$y_t = f(z_t,B) + u_t,$$

where z_t is a vector of endogenous, exogenous and lagged dependent variables. In this chapter we limit attention to static models, and in chapter 7 consider the assumptions about the variables necessary to generalise the results to dynamic models. This issue is not discussed by Jorgenson and Laffont but the "appropriate" assumptions are made. For the present we are concerned with static models and so leave

the discussion of this problem to chapter 7. Amemiya (1977) extended their work by considering NL3SLS for the model

$$f(y_t, x_t, \alpha) = u_t \qquad t = 1,2,..,T,$$

where y_t, x_t, u_t, and α are vectors of dimension (mx1), (kx1), (mx1) and (px1) respectively. The error process is assumed to satisfy (i) $E(u_t) = 0$ (ii) $E(u_t u_t') = \Omega$ and (iii) $E(u_t u_t') = 0$, $t \neq s$. The whole system can then be stacked in the following way

$$F(y,x,\alpha) = U,$$

where U is of dimension m x T.

The NL3SLS estimates of α are obtained by minimising

$$J(\alpha) = F(\alpha)'AF(\alpha),$$

for some matrix A. Jorgenson and Laffont (1974) consider $A = [\hat{\Omega}^{-1} \otimes X(X'X)^{-1}X']$ where $\hat{\Omega}$ is a consistent estimator of Ω. This is not the most efficient choice of A, and following Amemiya (1977), we consider

$$A = (\hat{\Omega} \otimes I)^{-1}E(\partial F/\partial\alpha')\left[\frac{E\partial F}{\partial\alpha}(\hat{\Omega} \otimes I)^{-1}\frac{E\partial F}{\partial\alpha'}\right]^{-1}\frac{E\partial F}{\partial\alpha}(\hat{\Omega} \otimes I)^{-1},$$

in the discussion of asymptotic efficiency. To establish the consistency and asymptotic normlity of this estimator Jorgenson and Laffont (1974) make the following assumptions:

a) u_t are i.i.d.,

the discussion of this problem to chapter 7. Amemiya (1977) extended their work by considering NL3SLS for the model

$$f(y_t, x_t, \alpha) = u_t \qquad t = 1,2,\ldots,T,$$

where y_t, x_t, u_t, and α are vectors of dimension (mx1), (kx1), (mx1) and (px1) respectively. The error process is assumed to satisfy (i) $E(u_t) = 0$ (ii) $E(u_t u_t') = \Omega$ and (iii) $E(u_t u_t') = 0$, $t \neq s$. The whole system can then be stacked in the following way

$$F(y,x,\alpha) = U,$$

where U is of dimension m x T.

The NL3SLS estimates of α are obtained by minimising

$$J(\alpha) = F(\alpha)'AF(\alpha),$$

for some matrix A. Jorgenson and Laffont (1974) consider $A = [\hat{\Omega}^{-1} \otimes X(X'X)^{-1}X']$ where $\hat{\Omega}$ is a consistent estimator of Ω. This is not the most efficient choice of A, and following Amemiya (1977), we consider

$$A = (\hat{\Omega} \otimes I)^{-1} E(\partial F/\partial \alpha') \left[\frac{E\partial F}{\partial \alpha} (\hat{\Omega} \otimes I)^{-1} \frac{E\partial F}{\partial \alpha'} \right]^{-1} \frac{E\partial F}{\partial \alpha} (\hat{\Omega} \otimes I)^{-1},$$

in the discussion of asymptotic efficiency. To establish the consistency and asymptotic normlity of this estimator Jorgenson and Laffont (1974) make the following assumptions:

a) u_t are i.i.d.,

b) $\lim_{T\to\infty} \frac{1}{T}X'X = M$, a finite nonsingular matrix,

c) $\operatorname{plim} \frac{1}{T}X'\frac{\partial f_i}{\partial \alpha'} = H_i$ uniformly in α.

Then $$\operatorname{plim} \frac{1}{T}X'\frac{\partial f}{\partial \alpha'} = \begin{bmatrix} H_1 \\ \vdots \\ H_p \end{bmatrix} = H \text{ of rank } p,$$

where $$\frac{\partial f_i'}{\partial \alpha} = \begin{bmatrix} \frac{\partial f_{11}}{\partial \alpha_1} & \cdots & \frac{\partial f_{1T}}{\partial \alpha_1} \\ \vdots & & \vdots \\ \frac{\partial f_{p1}}{\partial \alpha_p} & \cdots & \frac{\partial f_{pT}}{\partial \alpha_p} \end{bmatrix}.$$

These assumptions are the nonlinear counterparts of those made in the linear model, and the analysis used to derive the results is also essentially the same.

The mean value theorem applied to $F(\alpha)$ about a point α_0 gives

$$F(\hat{\alpha}) - F(\alpha_0) = \left[\begin{matrix} \partial f_1/\partial \alpha' \\ \vdots \\ \partial f_p/\partial \alpha' \end{matrix}\right]_{\alpha=\alpha^*} (\hat{\alpha}-\alpha_0), \qquad (15)$$

where α^* lies between α_0 and $\hat{\alpha}$. Premultiplying both sides of (15) by $T^{-1/2}(\hat{\Omega} \otimes X'X)^{-1/2}(I \otimes X') = S^{1/2}$, the LHS becomes

$$S^{1/2}F(\hat{\alpha}) - S^{1/2}u.$$

Now plim $S^{1/2}\,U = 0$ and, from the definition of $\hat{\alpha}$, $F(\hat{\alpha}) \leq F(\alpha_0)$ and so $U'SU \geq F(\hat{\alpha})'SF(\hat{\alpha})$. The probability limit of $S^{1/2}\,F(\hat{\alpha})$ is also zero, and as the plim of the right hand side is a finite matrix multiplied by $\text{plim}(\hat{\alpha}-\alpha_0)$, we have shown consistency.

To establish asymptotic normality, it is also necessary to assume

d) $$\text{plim}\ \frac{1}{T}\ X'\frac{\partial^2 f_i}{\partial \alpha_j\, \partial \alpha'} = G^i_j \quad \text{uniformly in } \alpha, \ \begin{matrix} i = 1,..,m, \\ j = 1,..,p \end{matrix}$$

e) we can apply the central limit theorem to $X'u_i/\sqrt{T}$, where u_i is the i^{th} row of U so that

$$\begin{bmatrix} X'u_1/\sqrt{T} \\ \vdots \\ X'u_m/\sqrt{T} \end{bmatrix} \xrightarrow{d} N(0, \Omega \otimes M).$$

Under the above conditions

$$\sqrt{T}(\hat{\alpha}-\alpha_0) \xrightarrow{d} N(0,(H'(\Omega \otimes M)^{-1}H)). \tag{16}$$

This result is derived by considering a mean value expansion of $\partial J/\partial \alpha\big|_{\hat{\alpha}}$ around α_0, and then showing that $\sqrt{T}(\hat{\alpha}-\alpha_0)$, whose distribution is still dependent on $\hat{\Omega}$, converges in distribution to that of a "pseudo" variable, $\sqrt{T}(\bar{\alpha}-\alpha_0)$, whose distribution is given on the right hand side of (16).

In the normal linear model we have seen that 3SLS and FIML are asymptotically equivalent, and so 3SLS attains the Cramer Rao lower bound (CRLB) asymptotically under normality. Amemiya (1977) shows that in general for the

nonlinear model this is not the case. He considers the performance of NL3SLS with the most efficient choice of A. It can then be shown that NL3SLS only reaches the CRLB asymptotically under normality if

$$f_i(y_t, x_t, \alpha) = c_i(\alpha_i)'z(y_t, x_t) + K_i(\alpha_i, x_t),$$

where z_t is of the same dimension as y_t. This special case is of no practical interest because typically in econometric nonlinear SEM's of this class the dimension of $z(y_t, x_t)$ is greater than the number of endogenous variables due to the contemporaneous feedback between variables involving different functions of the variables in different equations. For the linear model it was argued that the computationally less burdensome 3SLS can be used to approximate FIML as it has the same asymptotic distribution. However the failure of NL3SLS to reach the CRLB asymptotically under normality for any practically useful cases means that it cannot be used similarly as an approximation to NLFIML.

3.3. Properties of NLFIML: Amemiya (1977) and Phillips (1982)

There has been some controversy in the literature about the properties of NLFIML in the general nonlinear static model. Amemiya (1977) implied the true distribution must be normal for NLFIML to be consistent and asymptotically normally distributed. Phillips (1982) has shown that this requirement is only sufficient and not necessary for consistency. Given consistency it can easily be established

that NLFIML is asymptotically normally distributed. The asymptotic efficiency of NLFIML when the distribution is correctly specified follows directly from standard likelihood theory. Before we explore the conditions for consistency and asymptotic normality of NLFIML in various situations it is necessary to outline how both Amemiya and Phillips came to their respective conclusions. This will serve to illustrate the complexity of the problem within this general framework and the limitations to existing analysis.

Conditions for a consistent root to the likelihood equation.

Let the log likelihood function, indexed by the parameter vector α, for a sample size T be denoted $L_T(\alpha)$. Expand $T^{-1}L_T(\alpha)$ around the true value α_0 using the second order mean value theorem:

$$T^{-1}L_T(\alpha) = T^{-1}L_T(\alpha_0) + T^{-1}\frac{\partial L_T}{\partial \alpha'}\bigg|_{\alpha_0}\cdot(\alpha-\alpha_0) + \frac{1}{2}(\alpha-\alpha_0)'T^{-1}\frac{\partial^2 L_T}{\partial\alpha\partial\alpha'}\bigg|_{\alpha_T^*}(\alpha-\alpha_0), \tag{17}$$

where α_T^* lies between α and α_0. Taking probability limits on both sides of (17) gives

$$\text{plim } T^{-1}L(\alpha) = \text{plim } T^{-1}L_T(\alpha_0) + \frac{1}{2}(\alpha-\alpha_0)'\text{plim } T^{-1}\frac{\partial^2 L}{\partial\alpha\partial\alpha'}\bigg|_{\alpha_T^*}(\alpha-\alpha_0).$$

From this we deduce that sufficient conditions for a consistent root to the likelihood equation are

$$\text{plim } T^{-1}\frac{\partial L_T}{\partial \alpha'}\bigg|_{\alpha_0} = 0,$$

$$\text{plim } T^{-1}\frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\bigg|_{\alpha_0} \quad \text{is negative definite.}$$

The second of these is the condition for identification of the parameters derived by Rothenberg (1971) and Bowden (1973).

The asymptotic normality of NLFIML comes from the first order mean value theorem applied to the score vector:

$$\frac{\partial L_T}{\partial \alpha}\bigg|_{\hat{\alpha}} = \frac{\partial L_T}{\partial \alpha}\bigg|_{\alpha_0} + \frac{\partial^2 L_T}{\partial \alpha \partial \alpha'}\bigg|_{\alpha_T^{**}}(\hat{\alpha}-\alpha_0).$$

If we can apply a Central Limit Theorem to $\partial L_T/\partial \alpha\big|_{\hat{\alpha}}$ with appropriate scaling then we have the desired result. To begin with we concentrate on the arguments involved in establishing consistency and deal with asymptotic normality later.

The model we consider is an m equation system

$$f_i(y_t, x_t, \alpha_i) = u_{it} \qquad \begin{matrix} i = 1,2,..,m, \\ t = 1,2,..,T \end{matrix}$$

where α_i are the parameters in i^{th} equation, we assume:

1) u_t is distributed independently and identically normal with covariance matrix Ω
2) there are no constraints amongst the α_i's

3) the mapping $f_i: y_t \to u_t$ is continuous one to one mapping from a subset of R^n onto the whole R^n, and the inverse function is continuous

4) all relevant derivatives exist and are continuous for a given x_t and almost all y_t in the neighborhood of the true value of α_i

5) $\partial f_i/\partial y_t'$ and $\sum_{t=1}^{T} f_t f_t'$ where $f_t' = (f_{1t},\ldots,f_{nt})'$ are nonsingular in the same domain as 4).

This model specification is essentially a generalisation of the linear model assumptions. It was remarked earlier that the implications of assumption 5 for the underlying variables may be unclear, but it should also be noted that assumption 3) is likely to be extremely restrictive. We return to the implications of 3) for the model in chapter 6, but for the present follow the established practice in the literature of assuming a unique inverse exists without considering the implied restrictions on the model.

The log likelihood is

$$L_T^* = -\frac{T}{2}\log|\Omega| + \sum_{t=1}^{T}\log||\partial f_t/\partial y_t'|| - 1/2\sum_{t=1}^{T} f_t'\Omega^{-1}f_t.$$

This can be concentrated with respect to Ω to give

$$L_T = \sum_t \log||\partial f_t/\partial y_t'|| - \frac{T}{2}|T^{-1}\Sigma_t f_t f_t'|.$$

and so the score vector is

$$\frac{\partial L_T}{\partial \alpha_i} = \Sigma \frac{\partial g_i}{\partial u_i} - T(\Sigma g_i f')(\Sigma f f')_i^{-1}$$

where t subscripts have been supressed, $g_i = \partial f/\partial \alpha_i'$, $\partial g_i/\partial u_j = (\partial g_i/\partial y')(\partial f/\partial y')^{-1}_j$ and $(A)^{-1}_j$ indicates the j^{th} row of A^{-1}.

We can rewrite the score vector as

$$\left.\frac{\partial L_T}{\partial \alpha_i}\right|_{\alpha_0} = \Sigma[\frac{\partial g_i}{\partial u_i} - g_i u'\sigma^i] - T^{-1}\Sigma g_i u'[(\frac{\Sigma uu'}{T})^{-1}_i - \sigma^i],$$

where σ^i is the i^{th} column of Ω.

To establish that $\left.\partial L_T/\partial \alpha_i\right|_{\alpha_0}$ has a zero probability limit Amemiya (1977) uses the following lemma: "if $u_1,..,u_n$ are jointly normal with mean 0 and covariance σ_{ij} and h(u) is such that Eh and $E\partial h/\partial u_i$ are finite then $E(\partial h/\partial u_i) = E(h\Sigma\sigma^{ij}u_j)$". This of course implies $\text{plim } \left.\partial L_T/\partial \alpha_1\right|_{\alpha_0}$ is zero. Amemiya concluded that normality was therefore crucial for consistency. His mistake was to assume that this was a property of normal random variables alone.

Phillips (1982) presents a "Possibility Theorem" which shows that whenever NLFIML is consistent when the assumed and true distributions are normal then it is also consistent when the true distribution is a particular discrete mixture of normals. His proof, which we outline below, is for a one equation one parameter model, for simplicity, but the arguments can be generalised.

The true p.d.f. of u_t is

$$\text{pdf}(u) = \int_0^\infty (2\pi w)^{-1/2}\tilde{\sigma}^{-1}\exp(-u^2/2w\tilde{\sigma}^2)dG(w),$$

where $\tilde{\sigma}^2 = \{\int_0^\infty w dG(w)\}^{-1}E(u^2)$ and G(w) is a distribution function supported on $[0,\infty)$.

The proof is concerned with showing

$$\text{plim } T^{-1}\Sigma(g'-gu\sigma^{-2}) = 0, \text{ (where } g' = \partial g/\partial u),$$

as elementary arguments show the remaining term of $\partial L/\partial \alpha_1 \Big|_{\alpha_0}$ to be zero. As $g(\cdot)$ is a function of x_t and u_t we need to consider expectations with respect to the joint distribution of (u_t, x_t). So we have

$$T^{-1}\sum_{t=1}^{T}(g'-\sigma^{-2}gu_t) \to \int_x dF(x;\theta)\int_u (g'-\sigma^{-2}gu)pdf(u)du, \qquad (18)$$

where θ are the parameters of the distribution of x_t. We need to show the RHS of (18) is zero and the argument rests on using the weights of the mixing distribution to offset the nonlinearities in the system.

Proof:

Since $g(u)pdf(u)$ is absolutely continuous it follows that

$$\int_a^b g'pdf(u)du + \int_a^b gpdf'(u)du = [gpdf(u)]_a^b.$$

Let $a \to -\infty$, $b \to -\infty$ then the right hand side can be shown to be zero. So,

$$\int_u g'pdf(u)du = -\int_u gpdf'(u)du$$

$$= \sigma^{-2}\int gu\int(2\pi w)^{-1/2}\,\tilde{\sigma}^{-1}\exp\{-u^2/2w\tilde{\sigma}^2\}m_w w^{-1}dG(w),$$

where $m_w = \int_0^\infty wdG(w)$. Therefore

$$T^{-1}\Sigma(g'-\sigma^{-2}gu) + \sigma^{-2}\int_x dF(x;\theta)\int gudu\int(2\pi w)^{-1/2}\,\tilde{\sigma}^{-1}\exp(-u^2/2w\tilde{\sigma}^2)$$

$$X\ [m_w w^{-1}-1]dG(w)$$

$$= \sigma^{-3}m_w^{1/2}\int_0^\infty (2\pi w)^{-1/2}[m_w w^{-1}-1]dG(w)\int_x dF(x;\theta)\int_u gu\exp(\frac{-u^2}{2w\tilde{\sigma}^2})du$$

$$= \int_0^\infty (2\pi w)^{-1/2}[m_w w^{-1}-1]h(w;\alpha,\sigma^2,\theta)dG(w). \qquad (19)$$

We now require a mixing distribution for which this is zero. In the normal case $G(w) = \begin{cases} 0 & w < 1 \\ 1 & w \geq 1 \end{cases}$ and $m_w = 1$. Phillips (1982) takes the following mixing distribution

$$G(w) = \begin{cases} 0, & w < w_1, \\ \alpha, & w_1 \leq w < w_2 \\ 1-\varepsilon+\alpha, & w_2 \leq w < w_3 \\ 1, & w_3 \leq w \end{cases}$$

with $0 < \alpha < \varepsilon$. As $\varepsilon \to 0$, $w_3-w_1 \to 0$ and the density approaches normality. We need to establish that we can move away from this case in a systematic way so that for every mixing distribution of the above form the limit function is zero.

Put $\eta > 0$ and assume:-

$$1-\eta < w_1 < 1,\ w_1 < w_2 < w_3,\ 1 < w_3 < 1+\eta.$$

Choose w_2 such that $m_w = \alpha w_1 + (1-\varepsilon)w_2 + (\varepsilon-\alpha)w_3 = 1$, which implies $w_2 = (1-[\alpha w_1 + (\varepsilon-\alpha)w_3])/(1-\varepsilon)$. The limit of (19) is now

$$(2\pi)^{-1/2}[\alpha h(w_1)(w_1^{-1}-1)w_1^{-1/2} + (1-\varepsilon)h(w_2)(w_2^{-1}-1)w_2^{-1/2}$$

$$+ (\varepsilon-\alpha)h(w_3)(w_3^{-1}-1)w_2^{-1/2}].$$

Of course if $h(w;\alpha,\sigma^2,\theta) = 0$ the result is easy to show, but we must consider what happens when $h(w;\alpha,\sigma^2,\theta) \neq 0$ for $1-\eta \leq w \leq 1+\eta$. For this limit to be zero we must choose α such that

$$\alpha[h(w_1)(w_1^{-1}-1)w_1^{-1/2}-h(w_3)(w_3^{-1}-1)w_3^{-1/2}] \qquad (20)$$

$$= -\varepsilon h(w_3)(w_3^{-1}-1)w_3^{-1/2}-(1-\varepsilon)h(w_2)(w_2^{-1}-1)w_2^{-1/2}.$$

The final point we need to check is that if α is the solution to (20) for certain $w_1 < 1$ and $w_3 > 1$ then $0 < \alpha < \varepsilon$, as required for $G(w)$. Phillips (1982) verifies this but, as it is not crucial to the intuition behind the argument, we do not reproduce it here.

There are two points worth noting at present about the result. Firstly the arguments can be generalised to multiequation multiparameter models using multiple mass points. We would then end up with a system of linear equations for the vector α. Secondly Phillips has established the existence of an infinite number of distributions for which NLFIML is consistent, formed as w_1, w_2 move away from unity. NLFIML is therefore always consistent when the true distribution moves away from normality in this fashion.

3.4 Nonlinear IV

The interpretation of LS as an IV can also be extended to the nonlinear setting, although the idea behind IV does not translate easily. In the linear model we had

$$y = Z\delta + u,$$

and the IV estimators were constructed as solutions to

$$W'y = W'Z\delta,$$

which is a system of equations in δ. To produce an analogous estimator in the nonlinear model

$$z_t = f(Z_t,\alpha) + u_t,$$

we need to linearise the system about the parameter value α_0. The IV estimator is therefore artificial as α_0 is unknown. Linearsation gives

$$y_{it} = f_i(z_{it},\alpha_i^0) + \sum_{j=1}^{p_i} f_{ijt}(\alpha_{ij}-\alpha_{ij}^0) + u_{it},\ i = 1,\ldots,m, \tag{21}$$

where $f_{ijt} = \partial f_i(z_{it},\alpha_i^0)/\partial\alpha_{ij}$ and the $p_i \times 1$ vector α_i consists of the coefficients in the i^{th} equation.

Putting,

$$F_1 = \begin{bmatrix} \left.\frac{\partial f_1}{\partial \alpha_1'}\right|_{\alpha_0} \\ \vdots \\ \left.\frac{\partial f_p}{\partial \alpha_p'}\right|_{\alpha_0} \end{bmatrix},$$

and $y_t^0 = f.(z_t-\alpha_0)$, the system in equation (21) can be written as

$$y-y^0 = F_1.(\hat{\alpha}-\alpha_0) + u,$$

and the IV estimator, $\hat{\alpha}_{IV}$, is defined as the solution to

$$W'(y-y^0) = W'F_1.(\hat{\alpha}-\alpha_0).$$

Using similar arguments to the NL3S case we can show that

$$\sqrt{T}(\hat{\alpha}_{IV}-\alpha_0) \overset{a}{\sim} N(0,\text{plim}[(\tfrac{1}{T}W'F_1)^{-1}(\tfrac{1}{T}W'(\Omega \otimes I)W)(\tfrac{1}{T}F_1'W)^{-1}]).$$

Clearly if we put $W = \bar{X}(\Omega \otimes X'X)\bar{X}'F_1$ and $\bar{X} = I \otimes X$, the $\hat{\alpha}_{IV}$ is asymptotically equivalent to the NL3SLS estimator. However in the nonlinear model construction of the most efficient instruments will run into problems. In the linear model the optimal set were based on a consistent estimator of the systematic part of the j^{th} reduced form equation, independent of the errors. By analogy, in the nonlinear model we seek the systematic part of $\partial f_i/\partial \alpha_i'\big|_{\alpha_0}$, due to the linearisation, and so the reduced form even if it were available will not provide the answer. Jorgenson and Laffont (1974) consider some possible solutions to this problem, but as these do not provide estimators asymptotically equivalent to NL3SLS we do not review them here.

The derivatives are easier to calculate for the nonlinear in variables but linear in parameters model

$$f(y_t, x_t)B + x_t C = u_t,$$

as they are not functions of the parameters. Hatanaka (1978) outlines a routine for constructing an IV estimator. Although the derivatives are simpler, the problem of calculating the systematic component of the variables remains. Hatanaka (1978) suggests estimating the structural equations by OLS to obtain consistent estimates of the parameters $\hat{\delta}_{OLS}$. The deterministic solution of the estimated model can then be obtained by numerical techniques to yield predicted values for the endogenous variables $\hat{y}_{OLS}$. These are then transformed to the appropriate functional forms for the structural equations, $f(\hat{y}_{OLS}, x_t)$ and used as instruments. Each equation is estimated separately by IV, to obtain consistent estimators of u_t. These can be used to estimate the covariance matrix of u_t, which is needed for the final step of estimating the equations simultaneously to give FIIV.

Using similar arguments as in section 3.2, it can be shown that the resulting estimator $\hat{\delta}_{IV}$ is consistent but asymptotically inefficient under normality. The intuition behind this fact is that $\hat{\delta}_{IV}$ uses the deterministic solution of the model and not the conditional expectation. Any nonlinear effects of u_t in the reduced form are completely ignored. This reduced information set is a cause of the inefficiency and of course is a problem for IV estimation of nonlinear models in general.

Whilst NL3SLS has an IV interpretation Amemiya (1977) has shown that NLFIML is not an IV estimator. He replicates the arguments used by Hausman (1974) in the linear model and

shows that the estimator is not FIML at each iteration. We can stack the score vector equations to give

$$[T^{-1}\Sigma\frac{\partial g_i}{\partial u'}F'-G_i']F(T^{-1}F'F)^{-1}_i = 0, \qquad (22)$$

where F is the m x T matrix whose i,t^{th} element is $f_i(y_t,x_t,\alpha_i)$ and G_i' is the matrix whose t^{th} column is $\partial f_i(y_t,x_t,\alpha_i)/\partial\alpha_i$. If we let $\hat{G}_i' = G_i'-T^{-1}\Sigma\frac{\partial g_i}{\partial u'}F'$ and

$$\hat{G}' = \begin{bmatrix} \hat{G}_1' & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \hat{G}_2' & & & & \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ 0 & & & & & \hat{G}_m' \end{bmatrix}.$$

Then, putting $\hat{\Omega} = T^{-1}F'F$, (22) can be rewritten as

$$\hat{G}'(\hat{\Omega}^{-1} \otimes I)\text{vec}F = 0.$$

Let $\hat{\alpha}_1$ be an initial estimator of α. By expanding $\text{vec}F(\hat{\alpha}_1)$ around α_0 using a first order Taylor series expansion we obtain an updated estimator, $\hat{\alpha}_2$, as the solution to

$$\hat{G}'(\hat{\Omega}^{-1} \otimes I)(\text{vec}F(\alpha_0) + G(\hat{\alpha}_2-\alpha_0)) = 0.$$

This gives

$$\hat{\alpha}_2 = \hat{\alpha}_1 - [\hat{G}'(\hat{\Omega}^{-1} \otimes I)G]^{-1}\hat{G}'(\hat{\Omega}^{-1} \otimes I)\text{vec}F. \qquad (23)$$

For this second stage estimator to be maximum likelihood its distribution must not depend on that of $\hat{\alpha}_1$. This is easily seen by considering the general class of iterative

shows that the estimator is not FIML at each iteration. We can stack the score vector equations to give

$$[T^{-1}\Sigma\frac{\partial g_i}{\partial u'}F'-G_i']F(T^{-1}F'F)^{-1}_i = 0, \tag{22}$$

where F is the m x T matrix whose i,tth element is $f_i(y_t,x_t,\alpha_i)$ and G_i' is the matrix whose tth column is $\partial f_i(y_t,x_t,\alpha_i)/\partial\alpha_i$. If we let $\hat{G}_i' = G_i'-T^{-1}\Sigma\frac{\partial g_i}{\partial u'}F'$ and

$$\hat{G}' = \begin{bmatrix} \hat{G}_1' & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \hat{G}_2' & & & & \\ \cdot & & \cdot & \cdot & \cdot & \\ 0 & & & & & \hat{G}_m' \end{bmatrix}.$$

Then, putting $\hat{\Omega} = T^{-1}F'F$, (22) can be rewritten as

$$\hat{G}'(\hat{\Omega}^{-1} \otimes I)\text{vec}F = 0.$$

Let $\hat{\alpha}_1$ be an initial estimator of α. By expanding $\text{vec}F(\hat{\alpha}_1)$ around α_0 using a first order Taylor series expansion we obtain an updated estimator, $\hat{\alpha}_2$, as the solution to

$$\hat{G}'(\hat{\Omega}^{-1} \otimes I)(\text{vec}F(\alpha_0) + G(\hat{\alpha}_2-\alpha_0)) = 0.$$

This gives

$$\hat{\alpha}_2 = \hat{\alpha}_1 - [\hat{G}'(\hat{\Omega}^{-1} \otimes I)G]^{-1}\hat{G}'(\hat{\Omega}^{-1} \otimes I)\text{vec}F. \tag{23}$$

For this second stage estimator to be maximum likelihood its distribution must not depend on that of $\hat{\alpha}_1$. This is easily seen by considering the general class of iterative

solutions:-

$$\hat{\alpha}_2 = \hat{\alpha}_1 - A\frac{\partial L}{\partial \alpha}\bigg|_{\hat{\alpha}_1},$$

where A is some matrix. Taking a Taylor series expansion of $\partial L/\partial \alpha|_{\hat{\alpha}_1}$ around α_0,

$$T^{1/2}(\hat{\alpha}_2 - \hat{\alpha}_0) = -T^{1/2}A\frac{\partial L}{\partial \alpha}\bigg|_{\alpha_0} + [I - A\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\bigg|_{\alpha^*}]T^{1/2}(\hat{\alpha}_1 - \alpha_0),$$

where α^* lies between $\hat{\alpha}_1$ and α_0. The condition for the distribution of $T^{1/2}(\hat{\alpha}_2 - \alpha_0)$ not to depend on $\hat{\alpha}_1$ is that

$$\text{plim}TA^{-1} = \text{plim}T^{-1}\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\bigg|_{\alpha_0}.$$

Our estimator falls into this general class with $A = [\hat{G}'(\hat{\Omega}^{-1} \otimes I)G]^{-1}$. However Amemiya (1977) demonstrates that this choice of A does not satisfy the condition, and so $\hat{\alpha}_2$ in (23) is not the maximum likelihood estimator in general. The linear model case discussed by Hausman (1974) is a special case for which the condition is satisfied.

In this chapter it has been demonstrated that the close relationship between the FI estimators, which provides the basis for the derivation of conditions for consistency and asymptotic normality of FIML in the linear model, does not persist to the nonlinear setting. From standard likelihood theory it is known that NLFIML has those properties if the model is correctly specified. Otherwise all that is known is that if the true distribution is a member of a particular family of discrete mixtures of normals, then NLFIML retains

these properties. Phillips' (1982) analysis establishes that the coincidence of assumed and true distribution is only sufficient for NLFIML to be consistent. It provides a starting point for an examination of the nature of the trade-off between the nonlinearities in the system and true distributions for which NLFIML is consistent. The approach taken in the Possibility theorem cannot be generalised to other true distributions, and so to pursue the question of the consistency of NLFIML we need an alternative type of analysis. This is explored in chapters 5 and 6. Before that, we set our analysis within the context of the quasi MLE theory developed by White (1982). This serves to provide the background for our subsequent analysis, in the course of which we are able to provide a more unified treatment of the conditions for consistency of NLFIML.

4. INFERENCE IN MISSPECIFIED MODELS.

4.1. Theory of the quasi MLE

Our main focus in the remaining chapters is behavior of NLFIML when the distributional assumption about the errors is the only misspecification. Our analysis is limited to the case common in practice in which ML estimation is carried out under the assumption of normality. White (1982) has considered the more general framework of maximum likelihood estimation for distributionally misspecified models for the i.i.d. case. This can be generalised to the i.n.i.d. case under consideration as follows.

Define the average information measure $I(g_t,h_t,\alpha)$ as

$$I(g_t,h_t,\alpha) = \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E(\log(g_t(y_t)/h_t(y_t,\alpha)),$$

where y_t are i.n.i.d. variables with true distribution $g_t(\cdot)$ in period t, but MLE is carried out assuming $h_t(\cdot,\alpha)$ to be the p.d.f. of y_t. $I(g_t,h_t,\alpha)$ is a generalisation of the Kullback Liebler (1951) Information Criterion (KLIC). Let α_* be the parameter vector that minimises the KLIC. Then under the following regularity conditions, which are a generalisation of White (1982) assumption A3: a) $E(\log g_t(y_t))$ exists and $|\log h_t(y_t,\alpha)| \leq m(y_t)$ for all α in A, where m is integrable with respect to the distribution function of y_t, b)$I(g_t:h_t, \alpha)$ has a unique minimum at α_* in A; the quasi maximum likelihood estimator, $\hat{\alpha}_T$, converges to α_* almost surely.

The first order conditions for KLIC minimisation are obtained by differentiating $I(\cdot)$ with respect to α. This gives

$$\frac{\partial I}{\partial \alpha} = -\lim T^{-1} \sum_{t=1}^{T} \int [\partial \log h_t(y_t, \alpha)/\partial \alpha] g_t(y_t) dy_t . \qquad (24)$$

The second order conditions are

$$\frac{\partial^2 I}{\partial \alpha \partial \alpha'} = -\lim T^{-1} \sum_{t=1}^{T} \int \frac{\partial^2 \log h_t(y_t, \alpha)}{\partial \alpha \partial \alpha'} \cdot g_t(y_t) dy_t .$$

The QMLE is obtained by setting $\sum_{t=1}^{T} \frac{\partial \log h_t(y_t, \alpha)}{\partial \alpha} = 0$ and solving for α. For α_* to be the KLIC minimising value it is sufficient that $\partial I/\partial \alpha\big|_{\alpha_*} = 0$ and $\partial^2 I/\partial \alpha \partial \alpha'\big|_{\alpha_*}$ is negative definite. These are the conditions derived earlier in the discussion of the Amemiya/Phillips debate on consistency, where of course expectations have to be taken with respect to the true distribution. In the context of White's analysis these represent conditions for the convergence of $\hat{\alpha}_T$ to α_*. He terms $\hat{\alpha}_T$ "consistent" for α_* if they are satisfied, but we shall not do so to avoid confusion. We refer to $\hat{\alpha}_T$ as being consistent for α if the conditions for KLIC minimisation are satisfied for $\alpha = \alpha_0$, the true value. We are concerned with the conditions under which $\hat{\alpha}_T$ converges to the particular value α_0, and not the general conditions for the existence of a KLIC minimising value α_*.

To check these conditions for a consistent root we need to examine the behavior of the quasi score and quasi hessian at the true parameter value. The second order conditions are very difficult to verify in general as the "sign" of the hessian is likely to depend on the properties of the exogenous variables and the unknown parameters. In the case of nonlinear in variables models, however we can use the fact that the second order condition for consistency is also the condition for identification of the parameters.

Rothenberg (1971) and Bowden (1973) considered this link between identification and the existence of a well defined maximum likelihood estimator. Rothenberg (1971) essentially uses the arguments outlined earlier to derive the second order conditions for consistency based on the expansion of the true likelihood. Bowden (1973) uses the KLIC minimising arguments restricted to the case where the family of distribution is correct, but it is desired to distinguish between two parameter vectors. White (1982) generalises this result for i.i.d. variables to cover situations in which the family is misspecified. His arguments revolve around taking mean value expansions of the quasi likelihood. Rothenberg (1971) shows that the familiar "distribution free" criterion based on the observational equivalence arguments in the linear model result from the requirement that the information matrix be negative definite when the true distribution is normal. The generalisation of White (1982)'s theorem 3.1 to the i.n.i.d. case explains this result as the structural parameters can be identified from those of the reduced form using knowledge of only the first two moments of the distribution. The observational equivalence criterion can therefore be derived from the quasi hessian condition for all distributions with the same first two moments as the normal.

The literature on misspecified models has largely been concerned with the conditions for convergence of the QMLE to the KLIC minimising value without relating these ideas to the more familiar concepts of consistency and identification. White (1982) notes that identification retains its importance in misspecified models and in the

subsequent chapters we consider the conditions for and the importance of consistency. We have already noted the difficulty of evaluating the "sign" of the hessian, and in the next section we examine the distribution free identification criteria developed by Brown (1983) for models nonlinear in the variables but linear in the parameters. This criteria is used to check second order conditions in some worked examples in chapter 5. Brown's arguments are reproduced fairly rigorously because we need to extend his arguments to dynamic models in chapter 7 and also to relate his assumptions about model specification to our discussion of the conditions for a unique reduced form in chapter 6.

4.2. Identification in nonlinear models

Brown (1983) has developed "distribution free" criteria for identification of nonlinear in variables models using arguments based on observationally equivalent structures. In the analysis of the linear model we used the nonstochastic restrictions, $A\phi = 0$, and the stochastic restriction that $E(u_t|x_t) = 0$ to provide sufficient information for the discrimination of one equation from linear combinations of the rest; necessarily this is the only class of transformations that need to be considered. In the nonlinear model a linear combination of nonlinear transformations may produce an equation observationally indistinguishable from the i^{th} equation of the system. This is best demonstrated by an example from Fisher (1966, p. 133). Consider the system

$$a_0 + a_1 y_1^2 + a_2 y_2^2 + a_3 y_1 y_2 + a_4 x = u_1 \quad (26)$$

$$b_1 y_1 + b_2 y_2 = u_2. \quad (27)$$

Put $u_1 u_2 = 0$, then if we construct a third equation by squaring (27) and adding (26) to it, then this equation is indistinguishable from (26). For this example above the set of possible transformations that need to be considered is larger than a system linear in both parameters and variables. Consequently, the identification criterion from the linear model has not used enough restrictions (information) on the system to be applicable. We use Brown's criteria later in our analysis and so outline the basis of his results.

Consider the system

$$Aq(y,x) = u, \quad (28)$$

where y is a m x 1 vector of endogenous variables, x is a k x 1 vector of exogenous variables, u is a m x 1 vector of disturbances, $q(\cdot)$ is a n x 1 vector of known functions of y,x and A is the m x n matrix of unknown coefficients. In this context a structure consists of a coefficient matrix, A, and a conditional distribution $f(u|x)$. Two structures are then observationally equivalent when they imply the same conditional distribution for y.

The procedure is similar to the linear model except that we increase the information set. For a structure to be model admissible it must satisfy the following,

1) nonstochastic restrictions: $A\phi = 0$,

2) stochastic restrictions: a) the mean of the error conditional distribution is zero, b) the conditional distribution is independent of the exogenous variables.

The assumption 2b) is essentially arbitrary, but justifiable as in this context exogenous means determined outside the system. Brown notes, however, that exactly the same conditions would be derived from replacing this with the restriction that $E(uu')$ be positive definite.

It is assumed that (28) defines a single relevant inverse relationship

$$y = G(u,x;A), \tag{29}$$

where $G(\cdot)$ is an (mx1) vector of continuous functions obtained by either analytic or numerical techniques. We discuss the implications of this assumption for the generality of the model in chapter 5. Therefore two structures (A^1,f^1), (A^0,f^0) are observationally equivalent if, and only if,

$$\tilde{u} = A^1 q(G(u,x;A^0),x)$$

has the conditional distribution f^1 when y follows the conditional distribution determined by (29) for u distributed as f^0.

First consider necessary and sufficient conditions for $\tilde{u}$ to be stochastically independent of x when u follows f^0. Let $\tilde{Q}'$ be an n row matrix which forms a basis for the space generated by

$$\tilde{q}(u,x) = \frac{\partial q(G(u,x;A^0),x)}{\partial x'}.$$

If $\det(A^0 \partial q/\partial y') \neq 0$ for all u,x then $A^1\tilde{Q}'$ spans the space $\partial\tilde{u}/\partial x'$. By analysing $A^1\tilde{Q}'$ we can derive our condition for stochastic independence. If $A^1\tilde{Q}' \neq 0$ then $\partial\tilde{u}/\partial x' \neq 0$ and $\tilde{u}$ depends on x. However if $A^1\tilde{Q}' = 0$ then $\partial\tilde{u}/\partial x' = 0$ and so $\tilde{u}$ is locally invariant with respect to x. As $\tilde{u}(u,x)$ is continuous with respect to x, $\tilde{u}$ must be a function of u alone. The stochastic independence of u and x therefore implies that of $\tilde{u}$ and x, and so the required condition is that $A^1Q' = 0$.

The next step is to derive conditions under which $E(\tilde{u}|x)$ is zero. Let

$$\bar{q}(x) = E(q(G(u,x;A^0),x|x),$$

where u is distributed f^0, then $E(u|x) = 0$ if and only if $A^1\bar{q}(x) = 0$. Taken together with the nonstochastic restrictions, this gives the conditions that if (A^1,f^1) is observationally equivalent to (A^0,f^0) then $A^1(\bar{q},\tilde{Q}',\Phi) = 0$. The condition for the i^{th} equation to be identifiable (up to a scalar multiple) and therefore to be the only structure satisfying the restrictions is

$$\text{rank}(\bar{q}:\tilde{Q}':\Phi_i) = n-1, \tag{30}$$

where Φ_i is the $m \times R_i$ matrix of restrictions on the coefficients of the i^{th} equation, α_i', the i^{th} row of A.

The sufficiency of this condition follows from the fact that if (30) holds then $\alpha_i'(\bar{q},\tilde{Q}',\Phi_i) = 0$ and so the coefficients of the i^{th} equation of every structure

observationally equivalent to (A^0,f^0) are unique up to a scalar multiple.

The condition is necessary because it ensures only (A^0,f^0) is model admissible. For if rank $(\bar{q},\bar{Q}',\phi_i) < n-1$, then it is possible to find α_i arbitrarily close to α_i^0 satisfying the restrictions where α_i is not a scalar multiple of α_i. If A^1 is composed of A^0 only with α_i replacing α_i^0, then it will satisfy the restrictions and so A^0 is not identifiable.

This condition has the drawback that it requires the specification of the higher moments of u to evaluate $\bar{q}(x)$. (The derivatives can be calculated from the implicit function theorem). This cancels out one of the chief advantages of this method namely the minimal assumptions about the error distribution. However this can be avoided by considering the "implied equations" of the system. These are equations linear in q(y,x) but independent of our original model and satisfying its stochastic restrictions. Their coefficients are related to the properties of the orginal model, and so offer an alternative source of information about the original parameters.

We have assumed that (A^0,f^0) is model admissible and so $A^0(\bar{q},\bar{Q}') = 0$. The m independent rows of A^0 therefore lie in the row kernel of $Q^{*}{}' = (\bar{q},\bar{Q}')$. The row kernel has dimension m^*, where

$$m^* = \dim Q^{*}{}' - \text{rank } Q^{*}{}' = n - \text{rank } Q^{*}{}'.$$

This implies we can find $m^* - m$ additional independent rows giving the $(m^*-m) \times n$ matric C such that

$$A^* = \begin{bmatrix} A^0 \\ C \end{bmatrix},$$

forms a basis for the row kernel of $Q^{*\prime}$. The matrix C contains the coefficients of the following implied equations of the system:

$$w = Cq(G(u,x;A^0);x) = h(u).$$

Note these have been constructed so that $E(w|x) = 0$ and $CQ^{*\prime} = 0$, and therefore the augmented matrix A^* automatically satisfies the stochastic restrictions. We can now derive an equivalent condition for identification. Consider the matrix

$$A^*(Q^{*\prime}:\phi_1) = (0:A^*\phi_1).$$

This matrix has rank equal to the number of independent (of each other and $Q^{*\prime}$) columns of ϕ_1. Therefore

$$\text{rank}(A^*\phi_1) = \text{rank}(Q^{*\prime}:\phi_1) - \text{rank}(Q^{*\prime}).$$

This implies that $\text{rank}(Q^*:\phi_1)$ equals N-1 if, and only if, $\text{rank}(A^*\phi_1)$ equals $m^* - 1$. The latter is therefore an alternative condition for identification of the i^{th} equation of the original system.

This approach has enabled us to replace the identification condition based on the nonlinearities in the system by a linear model type condition for an augmented system of equations, the additional information coming from

the coefficients of the implied equations which depend on the nonlinearities of the system. For this new condition to be workable we need to be able to find C. Brown shows that the rows of C may be chosen as those linearly independent rows such that $C(Q':\bar{Q}':\bar{q}) = 0$.* This still depends on $\bar{q}$, but if the constant term is unrestricted in the i^{th} equation then C may be chosen as those $m^* - m$ linearly independent rows, in addition to the (nx1) vector (0,..,0,1), such that $C(Q':\bar{Q}') = 0$. This revised condition would be applicable in most practical circumstances.

Given this resemblance to the linear model condition it is worth considering when these conditions will coincide as then the nonlinearity in the variables can be ignored, and the model treated as linear for identification purposes. From the nature of the conditions discussed this is the case when there are no implied equations. Brown shows that this can be established for the class of models subject to one condition for which

$$Aq(y,x) = A_1q_1(y_1,x_1) + A_2q_2(y_1,x_1,y_2,x_2)+A_3q_3(x_3)+a_0,$$

where $y' = (y_1',y_2')$, $x' = (x_1,x_2,x_3)'$ and the elements of $q_2(y_1,x_1,y_2,x_2)$ are functionally independent when (y_1,x_1) are taken as constants. The condition for there to be no implied equations is $\text{rank}(A_2^0:A_3^0) = m$, i.e. of full rank. If this is satisfied then the i^{th} equation is identifiable if and only if $\text{rank}(A^0\phi_1) = m - 1$, which is the condition derived in our discussion of the linear model. Applications of these techniques are presented later in our discussion of the conditions for consistency of NLFIML in particular nonlinear in variables models.

* Where Q' forms a basis for the columns of $[\partial q/\partial y':\partial q/\partial x']$ when $A^0q(y,x)=0$.

5. CONSISTENCY OF NLFIML

5.1 Nonlinear regression model

The properties of FIML have recently received attention in the literature on the nonlinear regression model. Gourieroux, Monfort and Trognon (1984), GMT, develop the idea of accepting that any distributional assumption is likely to be incorrect so that the choice made should be the one delivering the most robust estimator. They term the resulting estimator a pseudo MLE as the choice of distribution is not made through any desire to accurately model the error process but because it determines the optimand of the estimation routine, and a suitable choice can deliver an estimator with desirable properties. Such a scheme is more in the spirit of least squares than maximum likelihood, hence the prefix pseudo. It should be distinguished from the QMLE which is the MLE derived from a misspecified model. Although our attention has been focused on the case in which the distribution alone is incorrect, the term QMLE denotes the MLE derived when any aspect of the model is misspecified. GMT further show that the assumed distribution must come from this family for the pseudo MLE to be consistent for all possible choices of conditional expectation of y_t and true error distributions. We return to their arguments in chapter 8 where we examine the use of the information matrix test as a general test of misspecification in this type of model.

Burguette, Gallant and Souza (1983) consider the properties of various estimators, including FIML, when the nonlinear regression model is used to approximate the general model outlined earlier. They are concerned with the

asymptotic properties of the resulting estimator when both the functional form and distribution are misspecified, and so their analysis is more general than that of Amemiya (1977) and Phillips (1982). However both these approaches do not generalise easily to the more complicated general nonlinear model as the use of the nonlinear regression model considerably simplifies the analysis. This is demonstrated in section 5.2 where it is seen that their model format is crucial to the strength of their results.

5.2 Consistency of NLFIML in the general model

We now explore various attempts to establish a general result in the manner of Phillips (1982) for the general static nonlinear model. Consider the model discussed by Amemiya (1977) and Phillips (1982):-

$$f_i(y_t, x_t, \alpha_i) = u_{it}, \qquad i = 1,2,\ldots,m,$$

where y_t is a m x 1 vector of endogenous variables,

x_t is a k x 1 vector of exogenous variables,

α_i is a p_i x 1 vector of parameter in the i^{th} equation.

We consider the case in which MLE is carried out under the assumption that $u_t = (u_{1t},\ldots,u_{mt})'$ is independently and identically normally distributed with mean zero and covariance Ω. The aim is to examine the properties of the QMLE $\hat{\alpha}_i$, when the normality assumption alone is incorrect.

The score of the quasi likelihood under normality is

$$\frac{\partial LLF_N}{\partial \alpha_i} = \sum_{t=1}^{T} \frac{\partial \ln||J_t||}{\partial \alpha_i} - \sum_{t=1}^{T} \frac{\partial f_t'}{\partial \alpha_i} \Omega^{-1} u_t,$$

$$\frac{\partial LLF_N}{\partial \Omega^{-1}} = \frac{T}{2}\Omega - \frac{1}{2}\sum_{t=1}^{T} u_t u_t',$$

where $\frac{\partial f_t'}{\partial \alpha_i}$ is p_i x m and has j-kth element $\frac{\partial f_{kt}}{\partial \alpha_i}$ and we have put $f_t = (f_1(y_t, x_t, \alpha_1)..,f_m(y_t, x_t, \alpha_m))'$. Recall the first order condition for α_0 to be the KLIC minimising value is

$$E\frac{\partial LLF_N}{\partial \theta}\Big|_{\theta_0} = 0,$$

where $\theta = (\alpha, vec\Omega')$ and expectations are taken with respect to the true distribution. By weak law of large number arguments the expectation of the derivative with respect to Ω^{-1} is zero at θ_0.

The score with respect to α_i is less easy to evaluate. However if we were dealing with the nonlinear regression model the analysis is simplified. The Jacobian is the identity matrix and so the only non zero term in the score is

$$\frac{\partial LLF_N}{\partial \alpha_i} = -\sum_{t=1}^{T} h(x_t, \alpha)'\Omega^{-1} u_t,$$

which has zero expectation at θ_0. The second order conditions are similarly easily verified, and enable GMT (1984) to develop powerful results for this class of model.

For more complicated models the presence of the Jacobian causes considerable problems. Typically it is a nonlinear function of the parameters and variables. Amemiya (1977) avoided having to examine the nature of this function

by recourse to a lemma for normal random variables. This is, of course, no use when the model is misspecified. Another possible solution is to use the properties of the true distribution and score to evaluate the quasi score at θ_0. The Jacobian of the transformation from assumed and true error p.d.f. to that of y_t is the same. Letting TLLF be the true LLF we know from conventional ML theory

$$\left.\frac{\partial TLLF}{\partial \theta}\right|_{\theta_0} = \sum_{t=1}^{T} \left.\frac{\partial \ln||J_t||}{\partial \theta}\right|_{\theta_0} + \sum_{t=1}^{T} \left.\frac{\partial Tpdf(u_t)}{\theta}\right|_{\theta_0},$$

has zero expected value. Therefore if we can show

$$\text{plim} T^{-1} \Sigma \left.\frac{\partial Tpdf(u_t)}{\partial_\theta}\right|_{\theta_0} = \text{plim} T^{-1} \Sigma \left.\frac{\partial Qpdf(u_t)}{\partial \theta}\right|_{\theta_0}$$

then it must follow that $\text{plim} T^{-1} \Sigma \left.\frac{\partial QLLF}{\partial \theta}\right|_{\theta_0} = 0$.

Given that normality could be argued to be a specification aimed at capturing a symmetrical error distribution, a natural choice of true p.d.f. to use is when u_t is distributed as a member of the elliptically symmetric family. It might be considered disturbing if NLFIML is not robust in this case. We consider the case in which the true p.d.f. is a continuous mixture of normals:

$$pdf(u_t) = \int_0^\infty (2\pi)^{-m/2} w^{-m/2} |\tilde{\Omega}^{-1}| \exp[-u_t' \tilde{\Omega}^{-1} u_t / 2w] g(w) dw,$$

where g(w) is a p.d.f. supported on the positive real line. This means we must compare

$$\text{plim}T^{-1}\Sigma[\int_0^\infty (2\pi w)^{-m/2}|\tilde{\Omega}^{-1}|\exp[-u_t'\tilde{\Omega}^{-1}u_t/2w]g(w)dw]^{-1}$$

$$\times \int_0^\infty (2\pi w)^{-m/2}|\tilde{\Omega}^{-1}|\exp[-u_t'\tilde{\Omega}^{-1}u_t/2w]\frac{m_w}{w}g(w)dw \cdot \frac{\partial f_t'}{\partial \alpha_i}\Omega^{-1}u_t,$$

where

$$\tilde{\Omega} = m_w^{-1}\Omega,$$

$$m_w = \int_0^\infty wg(w)dw,$$

with $\text{plim}T^{-1}\Sigma\frac{\partial f_t'}{\partial \alpha_i}\Omega^{-1}u_t$, where both plims are evaluated at the true parameter value. Clearly a sufficient condition would be for the integrals in denominator and numerator to have the same value. In general there is no reason for this to be the case. To illustrate the problems we consider the case where w has a particular inverted gamma distribution, so that the true p.d.f. of u_t is a multivariate Student t. If we let the p.d.f. of w be

$$h(w|v) = \frac{2}{\Gamma(v/2)}\left(\frac{v}{2}\right)^{v/2}\frac{1}{w^{v+1}}e^{-v/2w^2}, \qquad 0 < w < \infty,$$

then the true p.d.f. of u_t is the MV Student t with v degrees of freedom:

$$p(u_t|v,\Omega) = \frac{\Gamma[(v+m)/2]\ |\Omega|^{-1/2}}{\pi^{m/2}\Gamma(v/2)(v-2)^{m/2}}[1+u_t'\Omega^{-1}u_t/v-2]^{-(m+v)/2}.$$

Therefore we need to compare

$$\text{plim}T^{-1}\Sigma\frac{\partial T\text{pdf}(u_t)}{\partial\alpha_i}\bigg|_{\alpha_0}$$

$$= \text{plim}T^{-1}\frac{(m+v)}{v-2}\sum_t\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_t[1+\text{tr}(\frac{\Omega^{-1}}{v-2}u_tu_t')]^{-1}\bigg|_{\alpha_0}, \tag{31}$$

with

$$\text{plim}T^{-1}\sum_t\frac{\partial Q\text{pdf}(u_t)}{\partial\alpha_i}\bigg|_{\alpha_0} = \text{plim}T^{-1}\Sigma\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_t\bigg|_{\alpha_0}.$$

We first consider the problem with the assumptions in Amemiya (1977) namely that all summations in the score and likelihood converge to finite limits. In this case the two sides of (31) are not in general equal for finite v. Consider the quasi p.d.f. term first,

$$E[\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_t] = \int\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_t\cdot p(u_t|v,\Omega)du_t,$$

which is the expected value of $(\partial f_t'/\partial\alpha_i)\Omega^{-1}u_t$ taken with respect to the MV Student t with v degrees of freedom. To evaluate the other plim we need a result from Prucha and Kelejian (1983), namely

$$[1 + u_t'\Omega^{-1}u_t/v-2]^{-1}p(u_t|v,\Omega) = \frac{v}{v+m}p(u_t|v+2,\frac{v-2}{v}\Omega).$$

Therefore

$$E\{\frac{m+v}{v-2}\cdot\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_t[1+\text{tr}\frac{\Omega^{-1}}{v-2}u_tu_t']^{-1}\}\bigg|_{\alpha_0}$$

$$= \int\frac{v}{v-2}\cdot\frac{\partial f_t'}{\partial\alpha_i}\Omega^{-1}u_tp(u_t|v+2,\frac{v-2}{v}\Omega)du_t,$$

which is equal to $\frac{v}{v-2}$ times the expected value of $(\partial f_t'/\partial \alpha_i)\Omega^{-1}u_t$ taken with respect to a MV Student t distribution with v+2 degrees of freedom. Now $\partial f_t'/\partial \alpha_i$ is a nonlinear function of u_t, and so the constant adjustment does not transform from expectations taken with respect to the two distributions. However if v is infinite then the two plims are the same, but this just replicates Amemiya (1977)'s result as both distributions are then normal. To establish a general result we need further information about the system. Due to the symmetry of the MV Student t distribution we know that odd functions of u_t have zero expectation, therefore if $\partial f_t'/\partial \alpha_i$ is an even function of u_t, then both terms in (31) have the same plim when evaluated at α_o. However this condition requires knowledge of the reduced form of the model, which in general we do not have.

The conclusion to be drawn from the above analysis is as follows: we cannot say that NLFIML is consistent when we maximise the normal likelihood but the errors are actually distributed multivariate Student t under the conditions on f(·) in Amemiya (1977). It is the case that NLFIML may be consistent but this requires further knowledge and/or restrictions on the model. The problem is that unlike the linear model these are not easily verifiable. We later consider some particular examples to illustrate the relationship between the nonlinearities in the system and the conditions on the true distribution for the QMLE to be consistent. Before doing so we consider the situation in which we can derive a general result by relaxing one of the assumptions of the Amemiya (1977) model.

5.3 Consistency of NLFIML when u_t is a weakly stationary process

One of the advantages of the normal specification is the equivalence of the assumptions that the errors are uncorrelated or statistically independent over time. This special property enables us to consider the likelihood of T observations on a m-dimensional vector u_t or of one

5.3 Consistency of NLFIML when u_t is a weakly stationary process

One of the advantages of the normal specification is the equivalence of the assumptions that the errors are uncorrelated or statistically independent over time. This special property enables us to consider the likelihood of T observations on a m-dimensional vector u_t or of one

observation on the mT dimensional vector vecU and obtain an identical estimator. This is not a general property of random variables, and when we relax the normality assumption we must consider exactly what the appropriate specification is, given our knowledge of the system. In this section we show that there are families of stationary processes satisfying the first two moment conditions on the error, for which NLFIML under normality is consistent.

Phillips' (1982) arguments used the mixing distribution to offset the nonlinearities in the system whilst retaining the independence of the errors. In each circumstance the appropriate mixing weights are different as they depend on the nonlinearity present. It is possible to achieve the desired result by sacrificing the independence assumption but leaving the true distribution unconstrained, and using the dependence structure of u_t to offset the nonlinearities. We are therefore focusing attention on the vecU framework and consider the case in which vecU was mistakenly assumed to be normal.

Maximisation of the quasi and true likelihood are both just optimisation problems and what we need to show is that their solution is the same. This would be the case if the quasi and true scores are proportional. For if

$$\frac{\partial L_1}{\partial \alpha} \propto \frac{\partial L_2}{\partial \alpha} \text{ then } E_1 \frac{\partial L_1}{\partial \alpha} = 0 \text{ implies } E_1 \frac{\partial L_2}{\partial \alpha} = 0.$$

We of course need regularity conditions to ensure the optimisation problem is properly defined and these are listed in Amemiya (1977), although all expectations must be taken relative to the true distribution.

If $\text{vec}U \sim N(0, I \otimes \Omega)$ then the log likelihood is

$$L_N = \text{constant} + \frac{T}{2}\ln|\Omega^{-1}| - \frac{1}{2}(\text{vec } U)'(\Omega^{-1} \otimes I)\text{vec } U + \ln||J||,$$

where J is the Jacobian of the transformation from vecU to vec Y. We are primarily interested in the structural coefficients α and so concentrate the likelihood with respect to Ω. As

$$\text{vec } U'(\Omega^{-1} \otimes I)\text{vec } U = \sum_{t=1}^{T} u_t'\Omega^{-1}u_t = \text{tr}\Omega^{-1}\sum_{t=1}^{T} u_t u_t'$$

it follows that

$$\frac{\partial L_N}{\partial \Omega^{-1}} = \frac{T}{2}\Omega - \frac{1}{2}\sum_{t=1}^{T} u_t u_t',$$

implying that the QMLE for Ω is $\hat{\Omega} = T^{-1}\Sigma u_t u_t'$. If we then substitute this back into L_N to derive the concentrated log likelihood L_N^c, we have:-

$$L_N^c = \text{const} + \frac{T}{2}\ln|\hat{\Omega}^{-1}| + \ln||J||.$$

The QMLE for α is obtained by minimising L_N^c.

We now consider the log likelihood if vecU has a MV Student t distribution with v degrees of freedom. In this case

$$L_{st} = \text{const} + \ln||J|| + \frac{T}{2}\ln|\Omega^{-1}| - \frac{(mT + v)}{2}\ln\left[v + \text{vec } U'(\Omega^{-1} \otimes I)\text{vec } U \cdot \frac{v}{v-2}\right],$$

which implies

$$\frac{\partial L_{st}}{\partial \Omega^{-1}} = \frac{T}{2}\Omega - \frac{(mT + v)}{2} \cdot \frac{\sum_{t=1}^{T} u_t u_t'}{[v + tr\Omega^{-1} \sum_{t=1}^{T} u_t u_t' \frac{v}{v-2}]} \cdot \frac{v}{v-2},$$

setting this derivative to zero gives the solution

$$\hat{\Omega} = T^{-1} \sum_{t=1}^{T} u_t u_t', \text{ as } tr\hat{\Omega}^{-1}\hat{\Omega} = m.$$

We can use $\hat{\Omega}$ to concentrate the likelihood giving

$$L_{st}^{c} = \text{const} + \frac{T}{2} \ln|\hat{\Omega}^{-1}| + \ln||J||.$$

L_{st}^{c} has identical first and second derivatives with respect to α as L_{N}^{c}. Therefore as we know

$$E_{st} \left. \frac{\partial L_{st}^{c}}{\partial \alpha} \right|_{\alpha_0} = 0 \text{ and } E_{st} \frac{\partial^2 L_{st}^{c}}{\partial \alpha \partial \alpha'} \text{ is negative definite,}$$

it follows that

$$E_{st} \left. \frac{\partial L_{N}^{c}}{\partial \alpha} \right|_{\alpha_0} = 0 \text{ and } E_{st} \frac{\partial^2 L_{N}^{c}}{\partial \alpha \partial \alpha'} \text{ is negative definite, as well.}$$

Of course this argument can be used in reverse with expectations taken with respect to the normal distribution to show the QMLE under Student t is consistent if the true distribution is normal. Both optimisation problems are the same, and converge to the same solution. We have therefore established the consistency of $\hat{\alpha}$ subject to all except one of the conditions in Amemiya (1977). It was remarked earlier that the vecU and i.i.d. u_t specifications are not the same for the MV Student t case. Since it is the u_t

specification that is more commonly made it is important to determine the implications for u_t of vecU having a Student t distribution.

From Zellner (1971) we know that if $z' = (z_1', z_2')$ has a multivariate Student t distribution then $p(z_1|z_2)$ and $p(z_2)$, the conditional and marginal distributions also have MVST form. The joint density is

$$p(z_1,z_2) = \frac{\Gamma[(v+m)/2]|H|^{1/2}}{\pi^{m/2}\Gamma(v/2)}[1+Q_1+Q_2]^{-(m+v)/2},$$

where v = degrees of freedom,

z is m x 1,

$Q_i = z_i' H_i z_i$,

$H_i^{-1} = E(z_i z_i')/(v-2)$,

$E(z) = 0$,

$E(z_i z_j') = 0$.

This can be factorised to give

$$p(z_1,z_2) = \left[\frac{K_1|H_{22}|^{1/2}}{(1+Q_2)^{(m_2+v)/2}}\right] \cdot \left[\frac{K_2(1+Q_2)^{-m_1/2}|H_{11}|^{1/2}}{[1+Q_1/(1+Q_2)]^{(m+v)/2}}\right]$$

where $K_1 = \frac{\Gamma[(v+m_2)/2]}{\pi^{m_2/2}\Gamma(v/2)}$, $K_2 = \frac{\Gamma[(m+v)/2]}{\pi^{m_1/2}\Gamma[(v+m_2)/2]}$

and $m_1 + m_2 = m$, which can be denoted

$$p(z_1,z_2) = p(z_2)p(z_1|z_2) \text{ (in that order).}$$

The marginal distribution for z_2, $p(z_2)$, is clearly MV Student t with v degrees of freedom.

The situation with which we are working is

$Q = \sum_{t=1}^{T} q_i$, $q_i = u_i' H_i u_i$, and so

$$p(\text{vec}U) = \frac{\Gamma[(v+mT)/2)]}{\pi^{mT/2}\Gamma(v/2)} |H|[1+\sum_{t=1}^{T} q_i]^{-(mT+v)/2},$$

with $|H| = \prod_{i=1}^{T} |H_i|$. We can clearly factorise p(vecU) into the marginal distribution for any u_t and the conditional distribution of the remaining elements of vecU given u_t.

$$p(\text{vec}U) = \frac{\Gamma[(v+m)/2]}{\pi^{m/2}\Gamma(v/2)}|H_1|[1+q_k]^{-(m+v)/2}$$

$$x\frac{\Gamma[(mT+v)/2]}{\pi^{(T-1)m/2}\Gamma((v+m)/2)} \cdot \prod_{\substack{i=1 \\ i\neq k}}^{T} |H_i|[1+a\sum_{\substack{i=1 \\ i\neq k}}^{T} q_i]^{-(mT+v)/2}$$

$$x\ [1+q_k]^{-m(T-1)/2}, \qquad (33)$$

where $a = (1+q_k)^{-1}$.

The first line of (33) is the marginal distribution of u_k and the last two lines are in the form of a MV Student t distribution with (m+v) degrees of freedom. Therefore if vecU ~ MV Student t(v), then the marginal distribution of u_t is also MV Student t(v), with $E(u_t) = 0$ and $E(u_t u_t') = V$, for all t, as $H_i = \bar{H}$ for all i. However the conditional distribution of z_1 given z_2, is not the marginal distribution for z_1. Therefore whilst the u_t are identically distributed MVST, they are not independent. We now explore the nature of this dependence.

The conditional distribution $p(z_1|z_2)$ can be factorised in a similar fashion to the joint density. If we put z_2 =

u_1 and $z_1' = (u_2', \ldots, u_T')$,

$$p(z_1|z_2) = \frac{\Gamma[(v+2m)/2]}{\pi^{m/2}\Gamma[(v+m)/2]} |H_2|[1+a_1q_2]^{-(2m+v)/2}(1+q_1)^{-m/2}$$

$$\times \frac{\Gamma[(mT+v)/2]}{\pi^{(T-2)m/2}\Gamma[(v+2m)/2]} \prod_{i=3}^{T} |H_i|^{1/2}[1+a_1a_2\sum_{i=3}^{T} q_i]^{-(mT+v)/2}$$

$$\times [1+q_2a_1]^{-m(T-2)/2}[1+q_1]^{-m(T-2)/2},$$

where $a_2 = [1+a_1q_2]^{-1}$ and $a_1a_2 = [1+q_1+q_2]$. The first term of $p(z_1|z_2)$ corresponds to $p(u_2|u_1)$ and the remainder to $p((u_3'..u_T')'|u_2,u_1)$. The distribution $p(u_2|u_1)$ is MV Student t with $(v+m)$ degrees of freedom and

$$E(u_2|u_1) = 0,$$

$$var(u_2|u_1) = a_1 H_2^{-1}/(m+v-2).$$

We can clearly continue to make these factorisations to give the result that $p(u_k|(u_{k-1}'..u_1')'$ is a MV Student t with $(v+(k-1)m)$ degrees of freedom and

$$E(u_k|(u_{k-1}'...u_1')') = 0,$$

$$var(u_k|u_{k-1}'..u_1')') = \frac{1}{km+v-2} \bar{H}^{-1}[1+\sum_{i=1}^{k-1} q_i]^{-1}$$

$$= \frac{1}{m+\frac{(v-2)}{k}} \bar{H}^{-1} \frac{1}{k}[1+\sum_{i=1}^{k-1} q_i]^{-1}.$$

As k increases the conditional distribution of u_k tends to the MV normal, but its marginal distribution is still MV Student t.

To summarise: If the process behaves to ensure the joint distribution of the u_t's is MV Student t, then although the u_t's are no longer independent, the marginal distribution of each u_t is the same, and u_t forms a weakly stationary series as its unconditional moments are constant over time. The process is also serially uncorrelated over time.

Can we learn anything about other distributions for which NLFIML derived under normality is consistent? The MV t-distribution is a continuous mixture of normal distributions with identical means and covariance wV where $w^{1/2}$ has an inverted gamma distribution. Consider the case where vecU has a general mixture of normals distribution with weighting function g(w). The log likelihood function is

$$L_m = \ln\int_0^{\infty}(2\pi)^{-mT/2}w^{-mT/2}|\Omega^{-1}|^{mT/2}$$

$$\times \exp(-\text{vec}U'(\Omega^{-1} \otimes I)\text{vec}U/2)g(w)dw\} + \ln||J||,$$

$$\frac{\partial L_m}{\partial \Omega^{-1}} = [I]^{-1}\int_0^{\infty}(2\pi)^{-mT/2}w^{-mT/2}[\Omega\frac{mT}{2} - \frac{m}{2}\Sigma u_t u_t' w^{-1}]|\Omega^{-1}|^{mT/2}$$

$$\times \exp(-\frac{1}{2}\text{vec}U'(\Omega^{-1} \otimes I)\text{vec}U/2)g(w)dw,$$

where

$$I = \int_0^{\infty}(2\pi)^{-mT/2}w^{-mT/2}|\Omega^{-1}|^{mT/2}$$

$$\times \exp(-\frac{1}{2}\text{vec}U'(\Omega^{-1} \otimes I)\text{vec}U/w)g(w)dw.$$

The solution to the score is clearly of the form $\hat{\Omega} = cT^{-1}\Sigma u_t u_t'$, where c is a constant depending on the ratio of the integrals and so the likelihood can be concentrated as before. The p.d.f. of vecU can be factorised, putting $vec^*U' = (u_1'..u_{t-1}', u_{t+1}'..u_T')$, as

$$pdf(vecU) = \int_0^\infty (2\pi)^{-m(T-1)/2} \cdot w^{-m(T-1)/2} |\hat{\Omega}^{-1}|^{m(T-1)/2}$$

$$\exp(-\frac{1}{2} vec^*U' \frac{(\Omega^{-1} \otimes I)}{w} vec^*U)$$

$$\times (2\pi)^{-m/2} w^{-m/2} |\hat{\Omega}^{-1}|^{m/2} \exp(-\frac{1}{2} u_t' \Omega^{-1} u_t / w) g(w) dw,$$

$$= p(vec^*U) p(u_t).$$

To obtain the marginal distribution for u_t we integrate out vec^*U, and as $p(vec^*U)$ is a normal distribution as a function of vec^*U, the marginal for u_t is

$$\int_0^\infty (2\pi)^{-m/2} w^{-m/2} |\Omega^{-1}|^{m/2} \exp(-\frac{1}{2} u_t' \Omega^{-1} u_t / w) g(w) dw.$$

The conditional distribution for vec^*U given u_t is the ratio of two integrals over w, and so in general does not equal the marginal distribution for vec^*U.

All the distributions mentioned above are members of the class of elliptically symmetric distributions

$$f(v) = \sigma^{-r} |\Omega|^{-1/2} \phi(v' \Omega^{-1} v / \sigma^2),$$

where v is a (rx1) vector and $\sigma^2\Omega$ is a positive definite matrix and $\phi(\cdot)$ is a function on $[0,\infty)$. If vecU has a

distribution from this class then the log likelihood of the sample is

$$LLF = \ln||J|| + \frac{T}{2}\ln|\Omega^{-1}| + \ln\phi[\text{vec}U'(\Omega^{-1} \otimes I)\text{vec } U/2],$$

where $\text{cov}(\text{vec}U) = (I \otimes \Omega)$. The ML estimator for Ω is the solution to

$$\frac{\partial LLF}{\partial \Omega^{-1}} = \frac{T}{2}\Omega^{-1} + \frac{1}{\phi[\text{vec}U'(\Omega^{-1} \otimes I)\text{vec}U/2]} \cdot \frac{\partial \phi}{\partial v}\frac{1}{2}\sum_{t=1}^{T} u_t u_t' = 0,$$

where $v = \text{vec } U'(\Omega^{-1} \otimes I)\text{vec}U$. It will again be of the form

$$\hat{\Omega} = c\frac{1}{T}\sum_{t=1}^{T} u_t u_t' \text{ where } c = \frac{1}{\phi(v)}\frac{\partial \phi(v)}{\partial v}\bigg|_{v=v^*},$$

where $v^* = \text{vec}U'(\hat{\Omega}^{-1} \otimes I)\text{vec}U = cmT$ (assuming we can solve for c). The concentrated log likelihood is therefore

$$LLF^c = \ln||J|| + \frac{T}{2}\ln|\hat{\Omega}^{-1}| + \text{const},$$

which when optimised with respect to α will give identical solutions to when the quasi normal likelihood is used. Kelker (1970) has considered the distribution theory of the elliptically symmetric family in detail. He shows that the marginal distribution of each u_t is the same but whilst the first moment of the conditional distribution is zero, for all t, the conditional covariance depends on the history of the series. Again u_t forms a weakly stationary series, as its unconditional moments are constant over time. At present we are only concerned with conditions for the consistency of NLFIML. In chapter 6 we consider the

distribution from this class then the log likelihood of the sample is

$$LLF = \ln||J|| + \frac{T}{2}\ln|\Omega^{-1}| + \ln\phi[vecU'(\Omega^{-1} \otimes I)vec\ U/2],$$

where $cov(vecU) = (I \otimes \Omega)$. The ML estimator for Ω is the solution to

$$\frac{\partial LLF}{\partial \Omega^{-1}} = \frac{T}{2}\Omega^{-1} + \frac{1}{\phi[vecU'(\Omega^{-1} \otimes I)vecU/2]} \cdot \frac{\partial \phi}{\partial v}\frac{1}{2}\sum_{t=1}^{T} u_t u_t' = 0,$$

where $v = vec\ U'(\Omega^{-1} \otimes I)vecU$. It will again be of the form

$$\hat{\Omega} = c\frac{1}{T}\sum_{t=1}^{T} u_t u_t' \text{ where } c = \frac{1}{\phi(v)}\frac{\partial \phi(v)}{\partial v}\bigg|_{v=v^*},$$

where $v^* = vecU'(\hat{\Omega}^{-1} \otimes I)vecU = cmT$ (assuming we can solve for c). The concentrated log likelihood is therefore

$$LLF^c = \ln||J|| + \frac{T}{2}\ln|\hat{\Omega}^{-1}| + const,$$

which when optimised with respect to α will give identical solutions to when the quasi normal likelihood is used. Kelker (1970) has considered the distribution theory of the elliptically symmetric family in detail. He shows that the marginal distribution of each u_t is the same but whilst the first moment of the conditional distribution is zero, for all t, the conditional covariance depends on the history of the series. Again u_t forms a weakly stationary series, as its unconditional moments are constant over time. At present we are only concerned with conditions for the consistency of NLFIML. In chapter 6 we consider the

arguments for the asymptotic normality of the estimator in the static model with the i.i.d. specification. We do not at present explore the conditions for the vecU framework. Essentially we need to find the appropriate assumptions for applying a central limit theorem to the quasi score, when the model is dynamic. This is examined in chapter 7.

We have explored possible ways of establishing classes of true distribution for which NLFIML under normality is consistent. Although we have not been explicit about the nonlinearities present, both methods are implicitly dependent on the functional form $f(\cdot)$. Within the i.i.d. specification Phillips (1982) showed that there is always a family of distributions for which NLFIML under normality is consistent, but the form of this distribution depends on the weights in the mixture and so therefore on $f(\cdot)$. Alternatively we can consider the elliptically symmetric family of distributions and show that the marginal distribution of u_t is constant over time and independent of $f(\cdot)$, but that the dependence structure between the u_t must take a particular form which depends on the nonlinearities. Therefore the only general result for the static nonlinear model is that there are true distributions for which NLFIML under normality is consistent. The nature of these distributions, however, depends on the nonlinearities in the model. In the next section we examine this relationship for specific examples.

5.4 Examples of models for which NLFIML is consistent

To derive more substantial results it is necessary to specify the problem in greater detail. In this section we

consider a series of examples containing nonlinearities common in econometric models. These illustrate the type of restrictions placed on the true distribution to ensure that NLFIML under normality is consistent.

5.4.1 Expenditure and cost share models

Mellander (1983) provides an algorithm for NLFIML in the following class of models:

$$By_t *_1 (z_t'p) + Cz_t *_2 (z_t'q) = u_t, \quad t = 1,..,T,$$

where y_t and z_t are vectors of endogenous and predetermined variables respectively and $*_i$ denotes either the multiplication or division operator. The unknown parameters are contained in B, C, p, q. One restriction on the applicability of the algorithm is that the scalars $(z_t'p)$, $(z_t'q)$ must appear in every equation. The model is of some interest as it contains such forms as

(i) the system of expenditure shares corresponding to the indirect translog utility function (Christensen, Jorgenson and Lau (1975):

$$w_{it} = \frac{\alpha_i + \sum_{j=1}^{n} \gamma_{ij} \log(P_{jt}/m_t)}{-1 + \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \log(P_{jt}/m_t)} + u_{it}$$

(ii) the system of cost shares corresponding to the generalised Leontief cost function: Diewert (1971).

To incorporate possible cross equation restrictions

Mellander (1983) considers the case where B, C, p, q are functions of unknown parameters θ.

Within such systems of equations the error covariance matrix is singular due to the adding up constraint on the budget shares (see Barten, 1971). The solution to this is to omit an equation. Berndt and Savin (1975) have shown that 2 step "Zellner type" estimators depend on which equation is omitted, whereas Barten (1971) has shown FIML to be invariant to this choice. Against this has to be set the robustness of the known minimum distance estimators for a wide class of distributions compared to the unknown properties of NLFIML. Below we consider the properties of NLFIML in this model and show it to be consistent provided the first two moments are correctly specified, whenever it is consistent under normality. Our analysis is concerned with the case where $*_1$ represents division, but the same conclusion would be derived for the case where it represents multiplication. Consider the model

$$By_t(z_t'p)^{-1} + Cz_t(z_t'q)^{-1} = u_t,\ t = 1,..,T,$$

where in Mellander (1983)'s treatment y_t and z_t are m and n component vectors of observations at time t on the endogenous and predetermined variables respectively. B, C, p, q contain the unknown parameters. For the present we restrict attention to exogenous z_t and consider conditions under which the result can be generalised to dynamic models in chapter 7. This model is a special case of a family of nonsingular transformations of the nonlinear regression model of the form

$$f_i(y_t, x_t, \alpha_i) = h(x_t, \alpha_i)y_t + g(x_t, \alpha_i) = u_{it},$$

where $h(x_t, \alpha_i)$ is assumed invertible. In a similar fashion to the nonlinear regression model discussed in sections 5.1 and 5.2, the consistency of NLFIML depends only on the first two moments of the model being correctly specified provided the regularity conditions, that ensure the problem is well defined, are satisfied. This result follows easily because we have maintained the linearity of y_t.

5.4.2 Logs and Levels Models

In his paper Phillips (1982) states that maximum likelihood estimators derived under normality for two particular models are consistent when the true distribution is in fact a member of the mixtures of normals family. To establish exactly what is going on we examine his models A and B in detail, and discover that model B fails the second order conditions for consistency but is easily amended to ensure that these are satisfied. We start with Phillips' Model B:

The model is

$$\begin{aligned} \ln y_{1t} + a_1 \ln y_{3t} + a_2 &= u_{1t} \\ y_{1t}y_{2t} + b_1 y_{1t} + b_2 x_t &= u_{2t} \\ \ln y_{3t} + c_1 \ln y_{1t} &= u_{3t} \\ y_{4t} + d_1 y_{2t} &= u_{4t}. \end{aligned}$$

The Jacobian is $J_t = (y_{3t}^{-1}(1-c_1 a_1)$. The concentrated log likelihood function is:

$$LLF = \text{const} - \frac{T}{2}\ln|T^{-1}\sum_{t=1}^{T} u_t u_t'| + T\ln(1-a_1 c_1).$$

Let $A_T = |T^{-1} \sum_{t=1}^{T} u_t u_t'|$ and $m_{ij} = T^{-1} \sum_{t=1}^{T} u_{it} u_{jt}$, $i,j = 1,..,4$. Clearly the score vector depends on derivatives of A_T. Before we calculate these it is necessary to briefly outline some results on permutations and determinants from Pollock (1979, p. 62-3).

A permutation α defined on the set of integers $I_n = [1,..,n]$ is a one to one mapping of I_n onto itself. For every $i \in I_n$ there is a unique $\alpha(i) = j \in I_n$ and for every $i \in I_n$ there exists a unique $l \in I_n$ such that $\alpha(l) = i$. The sign of a permutation is either negative or positive depending on the number of transpositions in every factorisation. This is determined as follows: for the permutation $[\alpha(1),\alpha(2),..,\alpha(n)]$, let p be the number of pairs of elements $[\alpha(i),\alpha(j)]$; $i < j$ such that $\alpha(i) > \alpha(j)$. The sign of the permutation, denoted $\text{sgn}(\alpha)$, is $(-1)^p$. We can use this notation to define the determinant of a matrix. Recall A_T denotes the determinant of $T^{-1}\Sigma u_t u_t'$, and so

$$A_T = \sum_{\alpha} \text{sgn}(\alpha) m_{\alpha(1),1} m_{\alpha(2),2} m_{\alpha(3),3} m_{\alpha(4),4}$$

where the summation is over 4! terms. This implies

$$\frac{\partial A_T}{\partial \theta} = \sum_{\alpha} \text{sgn}(\alpha) \sum_{k} \left(\frac{\partial m_{\alpha(k),k}}{\partial \theta} \right) \cdot \prod_{\substack{j=1 \\ j \neq k}}^{4} m_{\alpha(j),j}.$$

The score of the concentrated log likelihood function is

$$\frac{\partial LLF}{\partial \theta} = (1-a_1 c_1)^{-1} \frac{\partial (1-a_1 c_1)}{\partial \theta} - \frac{T}{2} A_T^{-1} \frac{\partial A_T}{\partial \theta}.$$

The first order conditions for consistency require

$$\text{plim}T^{-1}\left.\frac{\partial LLF}{\partial\theta}\right|_{\theta_0} = 0.$$

From the above

$$\text{plim}T^{-1}\left.\frac{\partial LLF}{\partial\theta}\right|_{\theta_0} = \text{plim}(1-a_1c_1)^{-1}\frac{\partial(1-a_1c_1)}{\partial\theta} - \frac{1}{2}\text{plim}A_T^{-1}\text{plim}\frac{\partial A_T}{\partial\theta}.$$

Now $\text{plim}A_T^{-1} = |\Omega^{-1}|$ by the weak Law of Large numbers, and $\text{plim}T^{-1}\sum_{t=1}^{T}u_{it}u_{jt} = \sigma_{ij}$, we therefore need to turn our attention to $\text{plim}\partial m_{\alpha(k),k}/\partial\theta$.

i) Consider $\partial m_{ij}/\partial a_1$. The coefficient a_1 only appears in the first equation and so the only nonzero derivatives are $\partial m_{1k}/\partial a_1$, $k = 1,2,..,4$. So

$$\frac{\partial m_{11}}{\partial a_1} = 2T^{-1}\sum_{t=1}^{T}u_{1t}\ln y_{3t},$$

and

$$\frac{\partial m_{1j}}{\partial a_1} = T^{-1}\sum_{t=1}^{T}u_{jt}\ln y_{3t}.$$

To evaluate the plims of these terms we need the reduced form for y_{3t}, this gives

$$\ln y_{3t} = (u_{3t}-c_1u_{1t}-a_2)(1-a_1c_1)^{-1}.$$

We need only consider the stochastic part of the reduced form as by elementary arguments $\text{plim}T^{-1}\Sigma u_{it}$ is zero. So consider

$$E(u_{it}\ln y_{3t}) = E(u_{it}[u_{3t}-c_1u_{1t}](1-a_1c_1)^{-1}.$$

Again using σ_{ij} for the i-jth element of Ω, we have

$$E(u_{it}lny_{3t}) = (\sigma_{i3}-c_1\sigma_{i1})(1-a_1c_1)^{-1}.$$

It follows that

$$\text{plim}\frac{\partial A_T}{\partial a_1} = \Sigma_{\alpha}\text{sgn}(\alpha)(\sigma_{\alpha(1),3}-c_1\sigma_{\alpha(1),1})$$

$$\times\ \sigma_{\alpha(2),2}\sigma_{\alpha(3),3}\sigma_{\alpha(4),4} \cdot \frac{2}{(1-a_1c_1)}.$$

The factor of 2 arising because $\partial m_{ij}/\partial a_1 = \partial m_{ji}/\partial a_1$. So, putting Ω_{ij} equal to cofactor associated with the i-jth term of Ω,

$$\text{plim}T^{-1}\left.\frac{\partial LLF}{\partial a_1}\right|_{\theta_0} = \frac{-c_1}{1-a_1c_1} - A^{-1}[\sum_{i=1}^{4}\sigma_{i3}\Omega_{i1}-c_1\det|\Omega|](1-a_1c_1)^{-1}. \quad (34)$$

The first term of (34) is an expansion of Ω along its "wrong" cofactor and so is zero. Therefore as $A^{-1} = |\Omega|^{-1}$ we have $\text{plim}T^{-1}\partial LLF/\partial a_1|_{\theta_0} = 0$ as required.

ii) It is easy to establish that $\text{plim}T^{-1}\partial LLF/\partial a_2|_{\theta_0} = 0$ by similar arguments to those used to limit attention to the stochastic part of lny_{3t} above.

iii) The derivatives with respect to b_1 clearly involve y_{1t}, so we need to first calculate the reduced form expression for y_{1t}. This follows from

$$lny_{1t} = u_{1t}-a_2-a_1(u_{3t}-c_1lny_{1t}),$$

to yield

$$y_{1t} = \exp[(u_{1t}-a_2-a_1u_{3t})/(1-c_1a_1)].$$

Again we need only consider the stochastic part, namely $\exp(u_{1t}-a_1u_{3t})$.

$$E(u_{it}\exp(u_{1t}-a_1u_{3t})) = (\partial/\partial s_i)mgf((s))\Big|_{\substack{s_1=1\\ s_2=0\\ s_3=a_1\\ s_4=0}},$$

where mgf(s) is the moment generating function of u_t. Following Phillips (1982) we consider the case in which the true distribution is a member of the mixture of normals. Therefore

$$\frac{\partial mgf}{\partial s} = \{\int_0^\infty we^{ws'\tilde{\Omega}s/2}dG(w)\}\tilde{\Omega}s,$$

where $\tilde{\Omega}\{\int_0^\infty wdG(w)\} = \Omega$.

$$\Omega s = \Omega\begin{bmatrix}1\\0\\a_1\\0\end{bmatrix} = \begin{bmatrix}\sigma_{11}+a_1\sigma_{13}\\ \sigma_{12}+a_1\sigma_{23}\\ \sigma_{13}+a_1\sigma_{33}\\ \sigma_{14}+a_1\sigma_{43}\end{bmatrix},$$

which implies

$$\frac{\partial}{\partial s_i}mgf(s)\Big|_{\substack{s_1=1\\ s_2=0\\ s_3=a_1\\ s_4=0}} = \{\int_0^\infty we^{2s'\tilde{\Omega}s/2}dG(w)\}\frac{[\sigma_{1i}+a_1\sigma_{i3}]}{\{\int_0^\infty wdG(w)\}}.$$

So

$$\text{plim}\frac{\partial A_T}{\partial b_1}\Big|_{\theta_0} = \Sigma_\alpha \text{sgn}(\alpha)\sigma_{\alpha(1),1}[\sigma_{\alpha(2),1}+a_1\sigma_{\alpha(2),3}]\sigma_{\alpha(3),3}\sigma_{\alpha(4),4}k,$$

where $k = 2\int_0^\infty we^{ws'\tilde{\Omega}s/2}dG(w)/\int_0^\infty wdG(w)$. This gives

$$\text{plim}\frac{\partial A_T}{\partial b_1}\Big|_{\theta_0} = A^{-1}[\sum_{i=1}^{4}\sigma_{1i}\Omega_{2i}+a_1\sum_{i=1}^{4}\sigma_{3i}\Omega_{2i}] = 0,$$

as both summations are expansions of Ω along the wrong cofactor.

iv) By similar arguments to the above we can easily show $\text{plim}T^{-1}\partial LLF/\partial b_2 = 0$, as $\partial m_{i2}/\partial b_2 = T^{-1}\Sigma u_{it}x_t$.

v) Consider $\partial LLF/\partial c_1$. As $\partial m_{i3}/\partial c_1 = T^{-1}\sum_{t=1}^{T} u_{it}\ln y_{1t}$, we need the reduced form for y_{1t}, calculated earlier. This gives

$$\ln y_{1t} = (u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}.$$

We need only consider the stochastic part of $\ln y_{1t}$, namely $(u_{1t}-a_1u_{3t})(1-a_1c_1)^{-1}$. As

$$\text{plim } T^{-1}\Sigma u_{it}(u_{1t}-a_1u_{3t}) = \sigma_{1i}-a_1\sigma_{3i},$$

we have

$$\text{plim}\frac{\partial A_T}{\partial c_1} = \Sigma_\alpha \text{sgn}(\alpha)\sigma_{\alpha(1),1}\sigma_{\alpha(2),2}[\sigma_{\alpha(3),1}-a_1\sigma_{\alpha(3),3}]\sigma_{\alpha(4),4}$$

$$\times \frac{2}{(1-a_1c_1)}.$$

That gives

$$\text{plim}T^{-1}\left.\frac{\partial LLF}{\partial c_1}\right|_{\theta_0} = \frac{-a_1}{(1-a_1c_1)} - |\Omega|^{-1}[\sum_{i=1}^{4}\sigma_{1i}\Omega_{3i} - a_1|\Omega|](1-a_1c_1)^{-1}$$

$$= 0,$$

as again the summation is an expansion of Ω along the wrong cofactor.

vi) Finally we need to consider $\text{plim}T^{-1}\left.\frac{\partial LLF}{\partial d_1}\right|_{\theta_0}$.

To evaluate this we need the reduced form for y_{2t}:

$$y_{2t} = (u_{2t}-b_2x_t)y_{1t}^{-1}-b_1,$$

and

$$y_{1t}^{-1} = \exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}],$$

which implies

$$y_{2t} = (u_{2t}-b_2x_t)\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}]-b_1.$$

This gives

$$\frac{\partial m_{i4}}{\partial d_1} = T^{-1}\sum_{t=1}^{T}u_{it}y_{2t}$$

$$= T^{-1}\sum_{t=1}^{T}(u_{it}(u_{2t}-b_2x_t)\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}]-b_1u_{it}).$$

We need only consider the stochastic part of this expression, and clearly $\text{plim}T^{-1}\sum_{t=1}^{T}b_1u_{it} = 0$, so we need only evaluate

$$E[u_{it}(u_{2t}-b_2x_t)\exp[-(u_{1t}-a_1u_{3t})(1-a_1c_1)^{-1}]$$

$$= E[u_{it}u_{2t}\exp[-(u_{1t}-a_1u_{3t})(1-a_1c_1)^{-1}]$$

$$- b_2E(x_tu_{it}\exp[-(u_{1t}-a_1u_{3t})(1-a_1c_1)^{-1}]$$

$$= \left.\frac{\partial^2}{\partial s_i\partial s_2}(\text{mgf}(s))\right|_{\underline{s}} - b_2x_t\left.\frac{\partial}{\partial s_i}(\text{mgf}(s))\right|_{\underline{s}},$$

where $\bar{s}' = [-(1-a_1c_1)^{-1}, 0, a_1(1-a_1c_1)^{-1}, 0]$.

From Phillips (1982) we know

$$\frac{\partial^2 mgf(s)}{\partial s \partial s'} = \{\int_0^\infty w e^{ws'\tilde{\Omega}s/2} dG(w)\}\tilde{\Omega} + \{\int_0^\infty w^2 \exp[ws'\tilde{\Omega}s/2] dG(w)\}\tilde{\Omega}ss'\tilde{\Omega}$$

and as the i-jth element of $\Omega ss'\Omega\big|_{\bar{s}} = [\Omega ss'\Omega\big|_{\bar{s}}]_{ij}$

where $[\Omega ss'\Omega\big|_{\bar{s}}]_{ij} = (-\sigma_{1i}+a_1\sigma_{i3})(-\sigma_{1j}+a_1\sigma_{j3})$,

we have (omitting scaling factors),

$$\frac{\partial^2 mgf(s)}{\partial s_i \partial s_2} = [\sigma_{12}-a_1\sigma_{23}][\sigma_{1i}-a_1\sigma_{i3}] + \sigma_{i2},$$

and so

$$\text{plim}\frac{\partial A_T}{\partial d_1} = \Sigma \text{sgn}(\alpha)\sigma_{\alpha(1),1}\sigma_{\alpha(2),2}\sigma_{\alpha(3),3}$$

$$\times \{(\sigma_{12}-a_1\sigma_{23})(\sigma_{\alpha(4),1}-a_1\sigma_{\alpha(4),3}) + \sigma_{\alpha(4),2}$$

$$+ b\mu(\sigma_{\alpha(4),1}(1-a_1c_1)^{-1}(-1) + a_1(1-a_1c_1)\sigma_{\alpha(4),3}\},$$

where $\mu = \text{plim}T^{-1}\Sigma x_t$.

This gives

$$\text{plim}T^1 \left.\frac{\partial LLF}{\partial d_1}\right|_{\theta_0} = -A^{-1}\{(\sigma_{12}-a_1\sigma_{23})(\sum_{i=1}^4 \sigma_{1i}\Omega_{4i} - a_1\sum_{i=i}^4 \sigma_{3i}\Omega_{4i})$$

$$+ \sum_{i=1}^4 \sigma_{i2}\Omega_{4i} - b_2\mu[a_1\sum_{i=1}^4 \sigma_{3i}\Omega_{4i} - \sum_{i=1}^4 \sigma_{1i}\Omega_{4i}]\}$$

$$= 0,$$

as $\sum_{i=1}^4 \sigma_{ij}\Omega_{4i} = 0$ for $j = 1,2,3$. So the first order

conditions for consistency are satisfied.

We now need to consider the second order conditions for consistency. To do this we shall make use of the identification criterion given in Brown (1983) and outlined in section 4.2 above.

Our model falls into the following class:

$$u = A_1 q_1(y_1, x_1) + A_2 q_2(y_1, x_1, y_2, x_2) + A_3 q_3(x_3) + a_0,$$

where the elements of $q_2(y_1, x_1, y_2, x_2)$ are functionally independent when (y_1, x_1) are taken as constants. This means that any two elements of $q_2(\cdot)$ must not contain the same variable when the variables in q_1 are held constant.

Recall the model is

$$\begin{bmatrix} \ln y_{1t} + a_1 \ln y_{3t} + a_2 \\ y_{1t} y_{2t} + b_1 y_{1t} + b_2 x_t \\ \ln y_{3t} + c_1 \ln y_{1t} \\ y_{4t} + d_1 y_{2t} \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ u_{4t} \end{bmatrix}.$$

Put

$$q(y,x) = \begin{bmatrix} \ln y_{1t} \\ \ln y_{3t} \\ y_{1t} y_{2t} \\ y_{1t} \\ y_{4t} \\ y_{2t} \\ x_t \\ 1 \end{bmatrix} = \begin{bmatrix} (u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1} \\ (u_{3t} - c_1(u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1} \\ u_{2t} - b_2 x_t - b_1 \exp[(u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1}] \\ \exp[(u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1}] \\ u_{4t} - d_1[(u_{2t} - b_2 x_t)\exp[-(u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1}]] + b_1 d_1 \\ (u_{2t} - b_2 x_t)\exp[-(u_{1t} - a_2 - a_1 u_{3t})(1 - a_1 c_1)^{-1}] - b_1 \\ x_t \\ 1 \end{bmatrix}$$

Following Brown (1983) we have

$$q_1(y_1,x_1)' = [\ln y_{1t}, y_{1t}y_{2t}, y_{1t}, y_{2t}]$$
$$q_2(y_1x_1y_2,x_2)' = [\ln y_{3t}, y_{4t}]$$
$$q_3(x_3) = [x_t],$$

so:

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & b_1 & 0 \\ c_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 \\ b_2 \\ 0 \\ 0 \end{bmatrix}, \quad a_0 = \begin{bmatrix} a_2 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The condition for there being no implied equations in the system is that $\text{rank}(A_2:A_3) = 4$. Were this to be satisfied identification would be assessed using the familiar linear model criteria. However it is clearly not satisfied for the above model, so we have to consider the alternative criteria developed by Brown.

Brown shows that the i^{th} equation of the system is identifiable iff $\text{rank}_P(\bar{q}:\bar{Q}':\bar{\Phi}_i) = n-1$ where:

$\bar{q} = E[q(u,x)|x=0]$,

$\bar{Q}'$ is a n row matrix given by $\frac{\partial q(u,x)}{\partial x'}$,

n is the dimension of $q(\cdot)$,

and $A\Phi_i = 0$ are the parameter restrictions on the i^{th} equation.

For our model:

$$\bar{q} = E[q(u,x)|x=0]$$

$$= \begin{bmatrix} -a_2(1-a_1c_1)^{-1} \\ -c_1a_2(1-a_1c_1)^{-1} \\ E[-b_1\exp[(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}]] \\ E[\exp(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1} \\ E[-d_1u_{2t}\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}]+b_1d_1 \\ E[u_{2t}\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}]-b_1 \\ 0 \\ 1 \end{bmatrix}$$

$$\bar{Q} = \frac{\partial q(u,x)}{\partial x'}$$

$$= \begin{bmatrix} 0 \\ 0 \\ -b_2 \\ 0 \\ d_1b_2\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}] \\ -b_2\exp[-(u_{1t}-a_2-a_1u_{3t})(1-a_1c_1)^{-1}] \\ 1 \\ 0 \end{bmatrix}$$

(i) Consider the first equation:

$$\Phi_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We assume the relevant expectations to exist and using r_{ij} for the more complicated nonzero elements we can write:

$$(\bar{q}:\tilde{Q}:\Phi_1) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ -c_1a_2(1-a_1c_1)^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ r_{31} & -b_2 & 1 & 0 & 0 & 0 & 0 \\ r_{41} & 0 & 0 & 1 & 0 & 0 & 0 \\ r_{51} & r_{52} & 0 & 0 & 1 & 0 & 0 \\ r_{61} & r_{62} & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}\right] p \\ \\ \end{matrix}$$

For the first equation to be identified we need the rank of this matrix to be 7. Whilst the rows marked p form a linearly independent set the remaining three are multiples of each other and so the rank of $(\bar{q}:\tilde{Q}:\Phi_1)$ is 6. The first equation is not identified, and so the conditions for consistency are not satisfied.

ii) For the second equation:

$$(\bar{q}:\tilde{Q}:\Phi_2) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & 0 & 1 & 0 & 0 & 0 & 0 \\ -c_1a_2(1-a_1c_1)^{-1} & 0 & 0 & 1 & 0 & 0 & 0 \\ r_{31} & -b_2 & 0 & 0 & 0 & 0 & 0 \\ r_{41} & 0 & 0 & 0 & 0 & 0 & 0 \\ r_{51} & r_{52} & 0 & 0 & 1 & 0 & 0 \\ r_{61} & r_{62} & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

We can construct a set of 7 linearly independent rows by excluding the third row, and using the remainder. The rank of the matrix is therefore 7.

iii) For the third equation:

$$(\bar{q}:\bar{Q}:\Phi_3) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & ,0 & ,0,0,0,0,0,0 \\ -c_1a_2(1-a_1c_1)^{-1} & ,0 & ,0,0,0,0,0,0 \\ r_{31} & ,-b_2 & ,1,0,0,0,0,0 \\ r_{41} & ,0 & ,0,1,0,0,0,0 \\ r_{51} & ,r_{52} & ,0,0,1,0,0,0 \\ r_{61} & ,r_{62} & ,0,0,0,1,0,0 \\ 0 & ,1 & ,0,0,0,0,1,0 \\ 1 & ,0 & ,0,0,0,0,0,1 \end{bmatrix}.$$

The bottom seven rows of this matrix form a linearly independent set and so the rank of the matrix is 7.

iv) For the fourth equation

$$(\bar{q}:\bar{Q}:\Phi_4) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & ,0 & ,1,0,0,0,0,0 \\ -c_1a_2(1-a_1c_1)^{-1} & ,0 & ,0,1,0,0,0,0 \\ r_{31} & ,-b_2 & ,0,0,1,0,0,0 \\ r_{41} & ,0 & ,0,0,0,1,0,0 \\ r_{51} & ,r_{52} & ,0,0,0,0,0,0 \\ r_{61} & ,r_{62} & ,0,0,0,0,0,0 \\ 0 & ,1 & ,0,0,0,0,1,0 \\ 1 & ,0 & ,0,0,0,0,0,1 \end{bmatrix}$$

which is of rank 7 and so the equation is identified. The lack of identification of the

We can construct a set of 7 linearly independent rows by excluding the third row, and using the remainder. The rank of the matrix is therefore 7.

iii) For the third equation:

$$(\bar{q}:\tilde{Q}:\phi_3) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & ,0 & ,0,0,0,0,0,0 \\ -c_1a_2(1-a_1c_1)^{-1} & ,0 & ,0,0,0,0,0,0 \\ r_{31} & ,-b_2 & ,1,0,0,0,0,0 \\ r_{41} & ,0 & ,0,1,0,0,0,0 \\ r_{51} & ,r_{52} & ,0,0,1,0,0,0 \\ r_{61} & ,r_{62} & ,0,0,0,1,0,0 \\ 0 & ,1 & ,0,0,0,0,1,0 \\ 1 & ,0 & ,0,0,0,0,0,1 \end{bmatrix}.$$

The bottom seven rows of this matrix form a linearly independent set and so the rank of the matrix is 7.

iv) For the fourth equation

$$(\bar{q}:\tilde{Q}:\phi_4) = \begin{bmatrix} -a_2(1-a_1c_1)^{-1} & ,0 & ,1,0,0,0,0,0 \\ -c_1a_2(1-a_1c_1)^{-1} & ,0 & ,0,1,0,0,0,0 \\ r_{31} & ,-b_2 & ,0,0,1,0,0,0 \\ r_{41} & ,0 & ,0,0,0,1,0,0 \\ r_{51} & ,r_{52} & ,0,0,0,0,0,0 \\ r_{61} & ,r_{62} & ,0,0,0,0,0,0 \\ 0 & ,1 & ,0,0,0,0,1,0 \\ 1 & ,0 & ,0,0,0,0,0,1 \end{bmatrix}$$

which is of rank 7 and so the equation is identified. The lack of identification of the

first equation follows intuitively because this equation is indistinguishable from a linear combination of the first and third equations.

To ensure consistency we need to amend the model, and this could be done by introducing x_t into the third equation to give,

$$\begin{aligned} \ln y_{1t} + a_1 \ln y_{3t} + a_2 &= u_{1t} \\ y_{1t}y_{2t} + b_1 y_{1t} + b_2 x_t &= u_{2t} \\ \ln y_{3t} + c_1 \ln y_{1t} + c_2 x_t &= u_{3t} \\ y_{4t} + d_1 y_{2t} &= u_{4t} \end{aligned}$$

This alters the reduced forms:

$$y_{1t} = \exp[(u_{1t}-a_2-a_1u_{3t}+c_2x_t)(1-a_1c_1)^{-1}]$$

$$y_{2t} = (u_{2t}-b_2x_t)\exp[-(u_{1t}-a_2+c_2x_t-a_1u_{3t})(1-a_1c_1)^{-1}]-b_1$$

$$y_{3t} = \exp[(u_{3t}-c_1(u_{1t}-a_2-a_1u_{3t}-c_2x_t)(1-a_1c_1)^{-1}-c_2x_t)]$$

$$y_{4t} = u_{4t}-d_1[(u_{2t}-b_2x_t)\exp[-(u_{1t}-a_2-a_1u_{3t}+c_2x_t)(1-a_1c_1)^{-1}]+b_1d_1.$$

However due to the independence of errors and regressors this will not effect the arguments for the first order conditions but it will have an impact on the second order conditions.

The vector $\bar{q}$ is evaluated with $x = 0$, and so this remains unaltered when the model is amended. However $\bar{Q}$ will now take the form:

$$\tilde{Q}' = [\tilde{r}_{21}, \tilde{r}_{22}, \tilde{r}_{32}, \tilde{r}_{42}, \tilde{r}_{52}, \tilde{r}_{62}, 1, 0].$$

The effect of this on the identification arguments is that for the first equation ((i) above), the 8th and 2nd rows combined with p form an independent set giving $(\bar{q}:\tilde{Q}:\phi_i)$ the required rank of 7.

The arguments for the remaining equations are still the same and so the whole system is identified. The second order conditions for consistency are therefore satisfied in the amended model.

Phillips (1982) model A:

$$\ln y_{1t} + a_1 + a_2 x_t = u_{1t}$$

$$y_{2t} + b_1 y_{1t} + b_2 x_t = u_{2t}$$

i) First order conditions for consistency:

Using similar arguments to those above it is easily seen that

$$\text{plim} T^{-1} \partial LLF^c / \partial a_1 \big|_{\theta_0} = \text{plim} T^{-1} \partial LLF^c / \partial a_2 \big|_{\theta_0} = \text{plim} T^{-1} \partial LLF^c / \partial b_2 \big|_{\theta_0}$$

$$= 0,$$

for the class of true distributions considered above. We need only consider $\partial LLF^c / \partial b_1$:

From the first equation:

$$y_{1t} = \exp[u_{1t} - a_1 - a_2 x_t],$$

which combined with

$$|T^{-1}\sum_{t=1}^{T} u_t u_t'| = m_{11}m_{22}-m_{12}m_{21}.$$

implies

$$\frac{\partial LLF^c}{\partial b_1} = -TA_T^{-1}(m_{11}T^{-1}\sum_{t=1}^{T} u_{2t}\exp[u_{1t}-a_1-a_2x_t]$$

$$- m_{12}T^{-1}\sum_{t=1}^{T} u_{1t}\exp[u_{1t}-a_1-a_2x_t]).$$

Therefore assuming the true distribution to come from the class of continuous mixtures of normals:

$$\text{plim}T^{-1}\frac{\partial LLF^c}{\partial b_1}\bigg|_{\theta_0} = -A^{-1}(\sigma_{11}\partial mgf/\partial s_2\Big|_{\substack{s_1=1\\ s_2=0}}$$

$$- \sigma_{12}\partial mgf/\partial s_1\Big|_{\substack{s_1=1\\ s_2=0}})\text{plim}T^{-1}\sum_{t=1}^{T}\exp(-a_1-a_2x_t).$$

$$= -A^{-1}(\sigma_{11}\sigma_{12}-\sigma_{12}\sigma_{11})$$

$$\times\ \text{plim}T^{-1}\sum_{t=1}^{T}\exp(-a_1-a_2x_t)\{\int_0^{\infty} we^{ws'\tilde{\Omega}s/2}dG(w)\}$$

$$= 0.$$

ii) The second order conditions are easily verified

$$q'(y,x) = [\ln y_{1t}, y_{2t}, y_{1t}, x_t, 1],$$

$$q_1'(y_1,x_1) = [\ln y_{1t}, y_{1t}],\ q_2(y_1x_1y_2,x_2) = [y_{2t}],\ q_3 = [x_t],$$

and

$$[A_2:A_3] = \begin{bmatrix} 0 & a_2 \\ 1 & b_2 \end{bmatrix}, \quad \text{which is of rank 2}$$

implying the identification criteria from the linear model is appropriate. For the first equation

$$[A\Phi_1] = \begin{bmatrix} 0 & 0 \\ 1 & b_1 \end{bmatrix} \quad \text{and} \quad [A\Phi_2] = \begin{bmatrix} 1 & a_1 \\ 0 & 0 \end{bmatrix}$$

both of which are of rank m-1 = 1 and so the second order conditions for consistency are satisfied.

The conclusions from these examples about the properties of NLFIML in static logs and levels models for the situation in which the true distribution is mixture of normals but we have assumed normality, appear to be:

1) The model does not need to be recursive, for consistency but our arguments have relied on being able to write down the explicit reduced form.
2) NLFIML provides consistent estimators for recursive models satisfying the identification criterion. This can be established by considering the arguments in Phillips model A. Note that for a recursive model the Jacobian is independent of the parameters and so we need only consider the derivatives of $\ln|T^{-1}\sum_{t=1}^{T} u_t u_t'|$.

   For the expansion along the wrong cofactor arguments we required $\partial m_{ij}/\partial\theta_j$ to be a linear combination of the sample covariances of the residuals excluding m_{ij}. For this we require the endogenous explanatory variables in the j^{th} equation not to have a

reduced form that does not depend on u_j. This condition will always be satisfied in the recursive model.

Note that for logs and levels models we will only need the first derivative of the mgf.

3) We have required the mgf to exist and so logs and levels models not be considered with a MV Student t as the true distribution.

5.4.3 Further examples:

1) Brown (1983) considers the identification of the following system

$$u_{1t} = y_{1t} + a_1$$

$$u_{2t} = b_1 y_{1t}^2 + y_{2t} + b_2 x_t + b_3,$$

and shows that the second equation is unidentified. The second order conditions for consistency are therefore not satisfied.

One possible modification to overcome this is to introduce an additional exogenous variable consider

$$u_{1t} = y_{1t} + a_1 + a_2 z_t$$

$$u_{2t} = b_1 y_{1t}^2 + y_{2t} + b_2 x_t + b_3.$$

Using our earlier notation

$$\text{rank}(A_2 : A_3) = \text{rank}\begin{bmatrix} 0 & 0 & a_2 \\ 1 & b_2 & 0 \end{bmatrix} = 2,$$

and so the linear model criterion is appropriate.

$$A\Phi_1 = \begin{bmatrix} 0 & 0 & 0 \\ b_1 & 1 & b_2 \end{bmatrix} \text{ which is of rank 1,}$$

$$A\Phi_2 = \begin{bmatrix} 1 & a_2 \\ 0 & 0 \end{bmatrix} \text{ which is also of rank 1, and so the system is identified.}$$

We can now consider the first order conditions: the reduced form of the amended system is

$$y_{1t} = -a_1 - a_2 z_t + u_{1t}$$

$$y_{2t} = u_{2t} + b_1(u_{1t} - a_1 - a_2 z_t)^2 + b_2 x_t + b_3 \quad .$$

Recall the log likelihood is $LLF^c = \text{const} - \frac{T}{2} \ln |T^{-1} \sum_{t=1}^{T} u_t u_t'|$ and so by arguments already used to establish the results for the Phillips' models we have

$$\text{plim} T^{-1} \partial LLF^c / \partial \theta_i = 0, \text{ for } \theta_i = a_1, a_2, b_2, b_3,$$

provided the true distribution has mean zero.

We therefore need only consider $\text{plim} T^{-1} \partial LLF^c / \partial b_1$:

$$LLF^c = \text{const} - \frac{T}{2} \ln |T^{-1} \sum_{t=1}^{T} u_t u_t'|,$$

$$\frac{\partial LLF^c}{\partial b_1} = -\frac{T}{2} A_T^{-1} \{ m_{11} \frac{\partial m_{22}}{\partial b_1} - 2 \frac{\partial m_{12}}{\partial b_1} m_{12} \}$$

$$= -TA_T^{-1} \{ m_{11} T^{-1} \sum_{t=1}^{T} u_{2t} y_{1t}^2 - m_{12} T^{-1} \sum_{t=1}^{T} u_{1t} y_{1t}^2 \}.$$

As before,

$$\text{plim} A_T^{-1} = A = \det\Sigma,$$

$$\text{plim}\ m_{ij} = \sigma_{ij},$$

and so we need to calculate $\text{plim} T^{-1} \sum_{t=1}^{T} u_{it} y_{1t}^2$,

$$\text{plim} T^{-1} \sum_{t=1}^{T} u_{2t} y_{1t}^2 = \text{plim} T^{-1} \sum_{t=1}^{T} u_{2t}(u_{1t}-a_1-a_2 z_t)^2$$

$$= \text{plim} T^{-1} \sum_{t=1}^{T} u_{2t}(u_{1t}^2-2a_1 u_{1t}-2u_{1t} z_t)$$

$$= \text{plim} T^{-1} \sum_{t=1}^{T} u_{2t} u_{1t}^2 - 2a_1\sigma_{12}-2\sigma_{12}\text{plim} T^{-1} \sum_{t=1}^{T} z_t,$$

and also

$$\text{plim} T^{-1} \sum_{t=1}^{T} u_{1t} y_{1t}^2 = \text{plim} T^{-1} \sum_{t=1}^{T} u_{1t}(u_{1t}-a_1-a_2 z_t)^2$$

$$= \text{plim} T^{-1} \sum_{t=1}^{T} u_{1t}^3 - 2\sigma_{11}(a_1+\text{plim} T^{-1} \sum_{t=1}^{T} z_t).$$

If we assume the true distributions to be symmetric about zero and so $\text{plim} T^{-1} \sum_{t=1}^{T} u_{it} u_{jt} u_{kt} = 0$, then

$$\text{plim} T^{-1} \frac{\partial LLF^c}{\partial b_1}\bigg|_{\theta_0} = -A^{-1}\{-2(a_1+\text{plim} T^{-1} \sum_{t=1}^{T} z_t)(\sigma_{11}\sigma_{12}-\sigma_{11}\sigma_{12}) = 0.$$

For this model NLFIML under the assumption of normality is consistent for the true parameters vector provided the first and third moments of the true error process are zero. Again we can see that the recursive nature of the model is crucial. If z_t were replaced by y_{2t} then we quickly run into problems in trying to calculate the reduced form, and the arguments used above would not go through. This situation is dealt with in chapter 6.

In this chapter it has been shown that for any nonlinear model we can find classes of true distributions for which NLFIML is consistent by carefully structuring the correlation pattern of the residuals or the mixing distribution of the true p.d.f. (as in Phillips, 1982).

The examples presented illustrate the connection between the nonlinearities in the system and the properties required of the true distribution for NLFIML under normality to be consistent. Our analysis has relied on being able to write down an explicit reduced form for the endogenous variables. In this case there are always a set of moment restrictions on the true distribution which guarantee NLFIML is consistent. However in the majority of cases we are not going to be able to find an explicit reduced form. This raises two questions: (i) under what conditions is there an implicit reduced form and (ii) can we say anything about its functional form? In the next chapter we explore the answers to these problems and their implications both for our model specification and the consistency of NLFIML.

6. MODEL SPECIFICATION AND THE CONDITIONS FOR THE CONSISTENCY AND ASYMPTOTIC NORMALITY OF NLFIML

6.1. Model Coherency

It is frequently argued that the structural form of an econometric model,

$$f_i(y_t, x_t, \alpha) = u_{it};\ i = 1,..,m,$$

should be considered well specified if it implies a well defined reduced form for y_t. This is interpreted by Gourieroux, Laffont and Monfort (1982) (GLM) as the requirement that the model "must associate a unique value of y_t with any admissible value of x_t, u_t and α" (GLM p. 675). They term the conditions on α under which this is the case as "coherency conditions". Typically it is assumed that the model satisfies these restrictions provided the Jacobian of the transformation is nonsingular. However this is only a necessary condition, as noted by GLM, and so it is important to explore the nature of the restrictions placed on the model by this requirement.

In general attention is focused on three types of mapping. If we let y^m be the sample space of the endogenous variables and R^m the m-dimensional Euclidean space then f_i can be regarded as a mapping from y_t to u_t with domain Y^m and range R^m. The mapping f_i: $y_t \to u_t$ is injective (or "one to one") if $f_i(y) = f_i(y')$ implies $y = y'$. The mapping is surjective (or "onto") if for every element u of R^m there is at least one value of y such that $f(y) = u$. Finally f_i: $y_t \to u_t$ is bijective (or "1-1 correspondence") if the mapping is both injective and surjective.

The importance of bijective mappings is they permit the definition of an inverse mapping from u to y. For in this case there is a unique value of y_t such that $f(y_t) = u_t$, and so we can construct a mapping g: $u_t \rightarrow y_t$ such that $g(f(y_t)) = y_t$. Our earlier analysis of the properties of estimators has been restricted to the consideration of bijections. Amemiya (1977) assumes that "f_i: $y_t \rightarrow u_t$ is a continuous one to one mapping from a subset of R^m onto the whole R^m and the inverse function is also continuous" (p. 956). Brown (1983)'s derivation of identification criteria for nonlinear in variables models assumes that the structural equations implicitly define "a single relevant inverse relationship .. of continuous functions" (p. 177). However neither of these authors explore any functional restrictions entailed in such an assumption.

6.2. Coherency in Piecewise Linear Models.

GLM consider the case where f_i comprises a set of piecewise linear mappings. To illustrate that the examination of coherency conditions focuses attention on a different issue in this case, we outline the simplest case considered in their paper.

Let $a_1,..,a_n$ be independent linear forms defined on R^m. For each subset I of the set {1,2,..,n}, let C_I be the case defined by

$$C_I = \{x | x \epsilon R^n,\ a_i x \geq 0 \text{ if } i \in I \text{ and } a_i x < 0 \text{ if } i \notin I\}.$$

The invertible linear mapping A_I is associated with each case, and our function f is set equal to

$$f = \sum_I A_I J_I,$$

where J_I is the indicator variable, defined to be one if $x \in C_I$ and zero otherwise.

Given the invertibility of each mapping, the condition for the invertibility of the piecewise mapping is that the cones C_I partition R^n. GLM show that a necessary and sufficient condition for this to be the case is that all the determinants, $\det A_I$, $I\{1,2,..,n\}$, have the same sign. This is equivalent, in this case, to requiring $|\partial f_t/\partial y_t'|$ to be everywhere nonzero. The Jacobian is continuous and so if the $\det A_I$ are not of the same sign then there must be a crossover point between regimes for which the Jacobian is zero. The assumed invertibility within regime combined with linearity guarantees the existence of a unique inverse mapping given the partition.

GLM concentrate on establishing the conditions for an injective piecewise linear mapping. This approach does not generalise to other nonlinear models, although the question of coherency is still important. Below we consider the type of restriction placed on more general nonlinear models by coherency conditions.

6.3. The Implicit Function Theorem.

The nature of the mapping between y and u guaranteed by a nonsingular Jacobian is described by the implicit function theorem (see Goursat, 1959, p. 45). This states that if we have a system of equations

$$f_i(y_t, x_t, \alpha) = u_{it}, \quad i = 1,2,..,m,$$

where $f_i(\cdot)$ are continuous and possess continuous first partial derivatives in the neighborhood of $\bar{y}_t$, $\bar{u}_t$ then if the Jacobian of the transformation from y to u, $|\partial f_t/\partial y_t'|$, is nonzero for $\bar{y}_t$ and $\bar{u}_t$ then there exists one and only one system of continuous functions, $y_{it} = \phi(u_t)$, which satisfy the original equations and which reduce to $y_t = \bar{y}_t$ for $u_t = \bar{u}_t$.

This theorem establishes conditions for a local bijection. Provided the Jacobian is nonsingular, there is a unique local inverse. The analysis is only local and the functional form of $\phi(\cdot)$ need not remain constant as we move through the sample space. If the Jacobian condition holds everywhere in the sample space then this implies that for all y_t there is a value of u_t that maps onto it. Similarly as the mapping from u_t to y_t has a Jacobian that is the inverse of that for the mapping of y_t to u_t, this implies that for every value of u_t there is a value of y_t mapping onto it. In global terms, therefore both of these mappings are surjections if the Jacobian is nonsingular (almost) everywhere in the sample space.

The type of restrictions on the model implied by the Jacobian condition can be seen by considering the following two examples.

a) the logs and levels model,

$$\ln y_{1t} + a_1 y_{2t} + a_2 x_t = u_{1t}$$

$$\ln y_{2t} + b_1 y_{1t} + b_2 x_t = u_{2t},$$

does not possess an explicit reduced form. The Jacobian is $|(y_{1t}y_{2t})^{-1} - a_1b_1|$, and so for y_{it} to have a locally defined inverse mapping we must restrict y_{it} to be greater than zero and $y_{1t} \neq a_1b_1y_{2t}$.

b) In the case where there are ratios in the model,

$$y_{1t}/y_{2t} + a_2x_t/y_{2t} = u_{1t}$$

$$y_{2t} + b_1y_{1t} + b_2z_t = u_{2t},$$

then we must restrict attention to nonzero y_{it} that satisfy $|J| = |y_{2t}^{-1}(1+b_1(y_{1t}+a_2x_t))| \neq 0$.

This type of restriction is usually handled by assuming the inverse is locally defined "almost everywhere" meaning that the values of y_{it} that do not satisfy the Jacobian condition have been attached a zero probability of occurence.

That this condition does not guarantee that the mapping is a global bijection can be seen from the following example given by Gale and Nikaido (1968).

Consider the mapping

$$f_1(y_1,y_2) = e^{2y_1} - y_2^2 + 3$$

$$f_2(y_1,y_2) = 4e^{2y_1}y_2 - y_2^3,$$

$$|\partial f/\partial y'| = 2e^{2y_1}(4e^{2y_1}+5y_2^2) > 0 \text{ in } R^2.$$

The two points (0,2) and (0,-2) are both mapped onto the origin, and so although the Jacobian is everywhere nonsingular the mapping is not a bijection.

6.4. Gale and Nikaido Univalence Theorems.

Gale and Nikaido (1968) examine the conditions on the Jacobian that ensure the mapping is an injection. The basis for their results is a theorem specifying sufficient conditions on a matrix A for the equations $Ax \leq 0$ and $x \geq 0$ to have only the trivial solution. To understand the stringency of these conditions and to appreciate the complexity of the problem we outline the most relevant Gale and Nikaido's results below, but before we can do this the following definitions are required.

1) The principal submatrices of an (nxn) matrix $A = \{a_{ij}\}$ are matrices of the form:

$$\begin{bmatrix} a_{ii} & a_{ij} & \cdot & \cdot & \cdot a_{im} \\ a_{ji} & a_{jj} & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ a_{jm} & & & & a_{mm} \end{bmatrix}$$

where (i,j,..,m) is any permutation of m integers from the set of integers {1,2,..,n}.

2) The principal minors of A are the determinants of the principal submatrices, and the determinants of matrices formed by excluding any number of pairs of

rows and columns that both contain the same diagonal element.

3) A P-matrix is a matrix with all its principal minors positive.

4) Vector inequalities in the proof should be interpreted as follows:

$$x \geqq y \text{ if } x_i \geqq y_i,\ i = 1,..,n,$$

$$x \geq y \text{ if } x_i \geq y_i \text{ but } x \neq y,$$

$$x > y \text{ if } x_i > y_i,\ i = 1,..,n.$$

The results in Gale and Nikaido (1968) stem from the following theorem (Gale and Nikaido, 1968, Theorem 1, p. 82):

If A is a P-matrix, then the inequalities $Ax \geqq 0$, $x \leqq 0$ have only the trivial solution $x = 0$.

This result is trivial if A and x are scalars, and the proof for higher dimensions is by induction. The crucial property of a P-matrix that delivers the conclusion is that if we delete one of its rows and columns then the resulting matrix is itself a P-matrix.

The following two results can be derived fairly simply from this theorem.

Corollary 1: If A is a P-matrix, there is a number $\lambda > 0$ such that for all nonnegative vectors $x \geq 0$ of norm 1 ($||x|| = 1$) some component of Ax is as great as λ.

Corollary 2: If A is a P-matrix, the inequalities $Ax > 0$, $x > 0$ have a solution.

Nonlinear Models

The basis for the sufficient conditions for an injective mapping is a nonlinear analogue of theorem 1. For this we need to show not only that the inequalities locally imply only one solution but that there are no other solutions outside this and so it would be expected that the appropriate conditions would be more restrictive than those of the implicit function theorem. Gale and Nikaido's (1968) theorem 3 establishes that if the Jacobian of the mapping F is a P-matrix, then for any a, x in Ω, the inequalities

$$F(x) \leqq F(a), \quad x \geqq a,$$

have only the solution x = a. The proof is as follows:

If we assume that F is differentiable and set F(a) = 0 (without loss of generality) then

$$\lim_{x \to a} \{F(x)/||x-a|| - J(a)(x-a)/||x-a||\} = 0, \tag{35}$$

where for any vector v, $||v|| = (v'v)^{\frac{1}{2}}$. For $x \geq a$, then if J(a) is a matrix such that $J(a)(x-a) \geq 0$ then a is the only point in the neighborhood for which F(x) = 0. By corollary 1, it is sufficient that $J(\cdot)$ be a P-matrix.

This part of the analysis is similar to the implicit function theorem. For the existence of a locally defined unique inverse, we require $J(a)(x-a) \neq 0$ in equation (35) so that $F(x) \neq F(a)$ in a suitably defined neighborhood of a.

As $(x-a) > 0$, it is sufficient that $J(a)$ be nonsingular for if $J(a)(x-a) = b$ then $(x-a) = J(a)^{-1}b > 0$ which implies $b \neq 0$.

To establish this uniqueness in a rectangular region we need to show that if x is the set of all solutions to the inequalities, then $\hat{x} = x - \{a\}$ is the empty set. The set $\hat{x}$ is compact, and if it were not empty it must contain a minimal element $\bar{x}$ with the property that no other element x of $\hat{x}$ satisfies $x \leq \bar{x}$.

Gale and Nikaido (1968)'s arguments for the emptiness of $\hat{x}$ are based on considering two cases.

Case 1: $\bar{x} > a$

Assume $J(\bar{x})$ satisfies condition 1. By corollary 2 there is a vector $u < 0$ such that $J(\bar{x})u < 0$. Because $\bar{x} > a$ we can choose λ positive satisfying

$$x(\lambda) = \bar{x} + \lambda u > a.$$

Therefore $a < x(\lambda) < \bar{x}$ so $x(\lambda)$ lies in Ω. From the differentiability of F we have

$$F(x(\lambda)) = F(\bar{x}) + \lambda J(\bar{x})u + 0(\lambda||u||)$$

so that,

$$\frac{F(x(\lambda)) - F(\bar{x})}{\lambda||u||} - J(\bar{x})\frac{u}{||u||} = 0.$$

The left hand term can be made as small as necessary by suitably choosing λ. However this implies $F(x(\lambda)) < F(\bar{x}) \leq F(a)$ and $x(\lambda) \in \Omega$, for a sufficiently small positive

λ, contradicting the minimality of $\bar{x}$.

Case 2: Some component of $\bar{x} = \{x_i\}$ is equal to the corresponding component of $a = \{a_i\}$. Let this be the first element of x and a.

Gale and Nikaido (1968) establish that if $x_i = a_i$, for any i, then x = a if the Jacobian is a P-matrix. They define a new mapping $\hat{F}$: $\hat{\Omega} \to R^{n-1}$ by

$$\hat{f}_i(x_2,..,x_n) = f_i(a_1,x_2,...,x_n), \ (i = 2,..,n),$$

where

$$\hat{\Omega} = \{(x_2,..,x_n) | p_i \leq q_i, \ (i = 2,..,n)\}.$$

The Jacobian matrix of the new mapping is necessarily a P-matrix, and $\hat{f}_i(a_2,..,a_n) = 0 \geq \hat{f}_i(x_2,..,x_n)$, $(i = 2,..,n)$. Then by case 1 we have $\bar{x} = a$. Note that if x and a are assumed to have more than one element in common, the structure of the P-matrix ensures a similar proof goes through.

This theorem is the basis for the following univalence theorem (Gale and Nikaido, theorem 4, p. 86). (The proof is reproduced in appendix 1).

If F: $\Omega \to R^n$, where Ω is a closed rectangular region of R^n, is a differentiable mapping such that the Jacobian matrix J(x) is a P-matrix for all x in Ω, then F is univalent in Ω.

This condition on the Jacobian, whilst very stringent, is only sufficient for an injective mapping. However from

the structure of the arguments it can be seen that if we wish to work at this level of generality, then the requisite condition on the Jacobian must be of this type.

It is not necessary for A to be a P-matrix for $Ax \leq 0$, $x \geq 0$ to imply only $x = 0$. Another sufficient condition can be derived from Cramer's theorem. Let $Ax = b$, then $x_k = |A|^{-1} \sum_{j=1}^{n} b_j a^+_{kj}$, where a^+_{kj} is the k-j^{th} element of A^+, the adjoint matrix of A. For $x = 0$ to be the only solution it is sufficient that all the a^+_{kj} be of the same sign as the determinant, which is clearly not equivalent to A being a P-matrix. We could therefore replace the condition in theorem 1, corollaries 1 and 2 by this requirement. Case 1 of theorem 3 would follow through, but for case 2 we require the adjoints of all the principal submatrices to have all elements of the same sign as their determinants. In particular all the leading diagonal elements of J must be positive, and so implicit in this restriction is that the determinants of all the principal submatrices must be positive. The Jacobian must therefore be a P-matrix, but the adjoint condition also places restrictions on the other minors of the principal submatrices and so is more restrictive.

Gale and Nikaido (1968) establish a univalence theorem under slightly weaker conditions. If we define a <u>weak P-matrix</u> as one with positive determinant and nonnegative principal minors, then it can be shown by topological arguments that:

If $F: \Omega \to R^n$, where Ω is an open rectangular region of R^n is a differentiable mapping such that the Jacobian matrix $J(x)$ is a weak P-matrix for all x in Ω, then F is univalent.

This area is worthy of further research. The proofs outlined above suggest that working to this degree of generality is likely to require such a restrictive condition. It should be explicitly considered in the work of Amemiya (1977) and Brown (1983).

Recursive systems satisfy this coherency condition provided the leading diagonal elements of the Jacobian are positive. Earlier we considered the following case in which no explicit reduced form could be written down,

$$\begin{aligned} \ln y_{1t} + a_1 y_{2t} + a_2 x_t &= u_{1t} \\ \ln y_{2t} + b_1 y_{1t} + b_2 x_t &= u_{2t}. \end{aligned} \tag{36}$$

For the implicit function theorem to be valid we requried $|\partial f_t / \partial y_t'|$ to be nonzero. The Gale-Nikaido univalence condition requires this determinant to be positive.

An interesting example of where the more restrictive Jacobian condition supports our intuition is an augmented version of the quadratic model discussed in section 5.4.3. Consider

$$\begin{aligned} y_{1t} + a_1 y_{2t}^2 &= u_{1t} - a_1 x_t = c_{1t} \\ y_{2t} + b_1 y_{1t} &= u_{2t} - a_2 x_t = c_{2t}, \end{aligned}$$

which implies

$$a_1 y_{2t}^2 - b_1^{-1} y_{2t} + b_1^{-1} c_{2t} - c_{1t} = 0,$$

and so,

$$y_{2t} = (2a_1b_1)^{-1} \pm (2a_1)^{-1}\sqrt{b_1^{-2}-4a_1c},$$

where

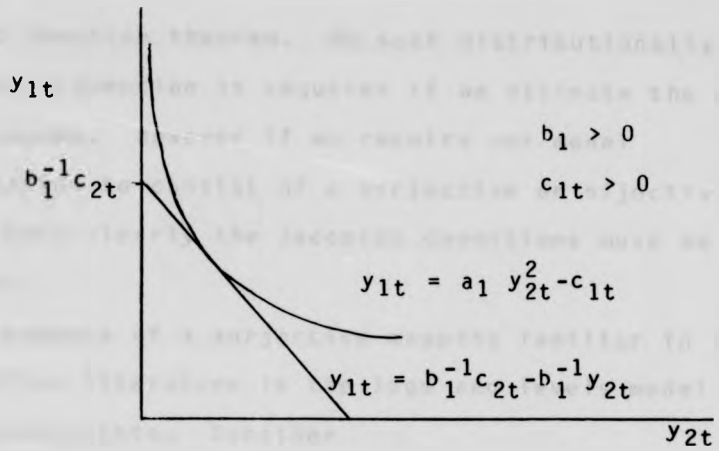
$$c = b_1^{-1}c_{2t}-c_{1t}.$$

Clearly there can be 0, 1 or 2 real solutions for y_{2t}. The Jacobian of the transformation is

$$J = \begin{matrix} 1 & 2a_1y_{2t} \\ b_1 & 1 \end{matrix}.$$

The requirement that J be nonsingular ensures $(2a_1b_1)^{-1} \neq y_{2t}$, and so eliminates the equal root case. Note that if the system has repeated roots then the line $y_{1t} = b_1^{-1}c_{2t}-b_1^{-1}y_{2t}$ is a tangent to the curve $y_{1t} = a_1y_{2t}^2-c_{2t}$. If y_{1t}^0, y_{2t}^0, c_{1t}^0, c_{2t}^0 are particular points satisfying those equations, then the inverse function $y_{it} = \phi(c_{1t},c_{2t})$ is not locally continuous about c_{1t}^0,c_{2t}^0 as this value is a boundary point of the set of feasible values of c_{1t}, c_{2t} that permit the system to have a solution. This demonstrated by figure 1,

Figure 1



For a given value of c_{1t}, if we reduce c_{2t} by any amount, no matter how small, then the equations become inconsistent.

This clearly does not restrict the mapping to be univalent as the system may have two roots. However if the Jacobian is a P-matrix then $y_{2t} < (2a_1b_1)^{-1}$, and attention is limited to one root of the quadratic.

6.5. Model Specification and Estimation

The nonsingularity of the Jacobian is required for the construction of the likelihood function. Throughout our analysis of ML we assumed the density of u_t was MV normal and so the p.d.f. of y_t is given by

$$\text{pdf}(y_t) = ||\partial u_t/\partial y_t'||\,\text{pdf}(u_t(y_t)).$$

Typically it is assumed that there exists a unique inverse of the mapping from u_t to y_t i.e. that it is a bijection. However, as noted by Pollock (1979), this procedure can still be employed if the mapping from u_t to y_t is a surjection, essentially due to the arguments of the

implicit function theorem. No such distributionally motivated assumption is required if we estimate the model by least squares. However if we require our model specification to consist of a surjective or bijective mapping then clearly the Jacobian conditions must be satisfied.

An example of a surjective mapping familiar in the econometrics literature is the logs and levels model with linear constraints. Consider

$$B_1 y_t + \Gamma_1 \log y_t + \Delta_1 z_t = u_{1t},$$

where y_t is of dimension m and u_{1t} is $m_1 \times 1$ with $m_1 < m$. To be able to transform from the p.d.f. of u_t to that of y_t we need to construct an invertible mapping from (u_t, v_t) to y_t where the v_t are $m-m_1$ dummy variables. The alternative is to use linear constraints on the variables to introduce additional information into the problem. If the constraints take the form

$$B_2 y_t + \Gamma_2 \log y_t + \Delta_2 z_t = 0,$$

a system of $m-m_1$ equations, then provided the conditions of the implicit function theorem are satisfied we now have a unique locally invertible mapping. Under normality the log likelihood function is

$$LLF = \sum_{t=1}^{T} \log|\det J_t| - \frac{T}{2}\log\det\Omega_1 - \frac{1}{2}\mathrm{tr}\Omega_1^{-1}A_1 X'XA_1 + \text{const},$$

where

$$A_1 = [B_1 : \Gamma_1 : \Delta_1],\quad X' = [x_1, .., x_T],\quad x_t' = (y_t', \log y_t', z_t')$$

and

$$J_t = \begin{bmatrix} B_1 + \Gamma_1 D_t^{-1} \\ B_2 + \Gamma_2 D_t^{-1} \end{bmatrix},\quad D_t = \operatorname{diag} y_t.$$

The concentrated LLF is

$$CLLF = \Sigma \log|\det J_t| - \frac{T}{2}\log\det\hat{\Omega}_1 + \text{const},$$

where

$$\hat{\Omega}_1 = T^{-1} A_1 X' X A_1',$$

$$\frac{\partial CLLF}{\partial A_1^c} = \Sigma (J_t^{-1})^r \frac{\partial J_t^c}{\partial A_1^c} - \frac{T}{2}\hat{\Omega}_1^r \frac{\partial \hat{\Sigma}_1^c}{\partial A_1^c}.$$

By the product rule

$$\frac{\partial (A_1 X' X A_1')^c}{\partial A_1^c} = (A_1 X' X \otimes I)\frac{\partial A_1^c}{\partial A_1^c} + (I \otimes A_1 X' X)\frac{\partial A_1'^c}{\partial A_1^c},$$

and

$$\frac{\partial J_t^c}{\partial B_1^c} = \begin{bmatrix} (I \otimes ii') \\ 0 \end{bmatrix}$$

$$\frac{\partial J_t^c}{\partial \Gamma_1^c} = \begin{bmatrix} D_t^{-1} \otimes ii' \\ 0 \end{bmatrix},$$

where i is a vector with every element equal to one.

To analyze the behavior of the score away from normality we need the reduced form of the system. In general this cannot be written down explicitly, the problem

is illustrated by the simple example used by Davidson (1981) for a simulation study

$$\log y_{1t} = a + b\log y_{2t} + u_t$$

$$y_{1t} = y_{2t} + z_t.$$

The log likelihood function is then

$$LLF = \sum_{t=1}^{T} \log[by_{2t}^{-1} - y_{1t}^{-1}] - \frac{T}{2}\log\hat{\sigma}^2,$$

$$\frac{\partial LLF}{\partial b} = \sum_{t=1}^{T} \frac{y_{2t}^{-1}}{(by_{2t}^{-1} - y_{1t}^{-1})} - T\hat{\sigma}^2 T^{-1} \sum_{t=1}^{T} u_{1t}\log y_{2t}.$$

$$\frac{\partial LLF}{\partial a} = -T\hat{\sigma}^{-2} \sum_{t=1}^{T} u_{1t}.$$

the latter having a zero plim provided the true distribution has mean zero. To examine the true distributions for which $\text{plim} T^{-1} \frac{\partial LLF}{\partial b}\Big|_{\theta_o} = 0$, we need the reduced form for y_{it}. For our example substituting into the identity gives:

$$\log(y_{1t}/y_{2t}^b) = a + u_t$$

$$y_{1t}/y_{2t}^b = e^{a+u_t},$$

and so

$$y_{1t} = e^{a+u_t} y_{2t}^b = y_{2t} + z_t.$$

Unless b = 1, the situation considered by Davidson, the reduced form cannot be written down explicitly. For b = 1

$$y_{1t} = \frac{-e^{a+u_t}}{1-e^{a+u_t}} z_t$$

$$y_{2t} = \frac{-1}{1-e^{a+u_t}} z_t,$$

but this case is of little interest for our purposes. The extent to which we can learn about the reduced form, and so the consistency of NLFIML, if $b \neq 1$ is explored in the next section.

Hatanaka (1978) considers nonlinear in variables models of the form,

$$f(y,x)B_1 + xc_1 = u_1$$

$$f(y,x)B_2 + xc_2 = 0.$$

The endogenous variables are partitioned into (y_1,y_2) with y_1 of the same dimension as u_1. The partition is arbitrary except that $(\partial f/\partial y_2')B_2$ must be nonsingular. Hatanaka argues that y_2 should be expressed as a function of y_1 and x from the identity, and this substituted into the stochastic equations. Estimation is then carried out on this problem of reduced dimension. This of course requires being able to solve for y_2, and from the implicit function theorem we know that the Jacobian condition does not guarantee an explicit solution.

6.6. The Implicit Function Theorem and Analytic Functions

The Jacobian conditions described above provide information on situations in which an implicit reduced form exists but as yet do not give any indication about its functional form. Goursat (1959, p. 402) shows that the implict function theorem can be extended in the following way:

If each of the functions $f_i(\cdot)$ (i) vanish when $y_j = u_i = 0$ (ii) is developable in a power series near that point and (iii) the Jacobian is nonsingular, then there exists one and only one system of solutions to the equations of the form $y_i = \phi(u_i)$ where $\phi_i(\cdot)$ are power series in u which vanish when $u = 0$.

This theorem is not directly applicable to the cases considered above due to the concentration on $y_i = u_i = 0$. Goursat (1959) considers power series of the form

$$y_i = \Sigma a_{ij\ldots r} v_1^i u_2^j \ldots u_p^r, \tag{37}$$

with $a_{00\ldots0} = 0$. However we can adapt his results so that the conditions are that the $\phi_i(\cdot)$ are developable around $y_i = y_i^0$, $u_j = u_j^0$ by considering power series of the form (37) with $a_{00\ldots0} \neq 0$. The convergence of the power series needs to be checked in each case.

The weights in the power series expansion can be calculated from repeated differentiation of the original equations. For instance consider the model in 6.4 (equation 36). Putting $z_{it} = \ln y_{it}$ to transform the equations into functions developable in power series we have

$$z_{1t} + a_1 e^{z_{2t}} = u_{1t} - a_2 x_t = v_{1t}$$

$$z_{2t} + b_1 e^{z_{1t}} = u_{2t} - b_2 x_t = v_{2t}.$$

This gives

$$z_{1t} + a_1 \exp[v_{2t} - b_1 \exp z_{1t}] = v_{2t}.$$

If the solution is of the form

$$z_{1t} = \sum_{i,j} c_{ij} v_{1t}^i v_{2t}^j,$$

then $c_{00} = z_{1t}\Big|_{v_t=0}$, where $v_t' = (v_{1t}, v_{2t})$. We therefore develop the power series about the point $z_{1t} = c_{00}$, $v = 0$. The coefficients c_{10} and c_{01} are given by

$$c_{10} = \frac{\partial z_{1t}}{\partial v_{1t}}\Big|_{v_t=0} = 1 + b_1 a_1 \exp[v_{2t} + z_{1t} - b_1 \exp z_{1t}] \frac{\partial z_{1t}}{\partial u_{1t}}\Big|_{v_t=0},$$

and so $c_{10} = -(1 - b_1 a_1 \exp[c_{00} - b \exp c_{00}])^{-1}$, similarly,

$$c_{01} = \frac{\partial z_{1t}}{\partial v_{2t}}\Big|_{v_t=0} = \frac{-a_1 \exp[-b_1 \exp c_{00}]}{1 - a_1 b_1 \exp[c_{00} - b_1 \exp c_{00}]}.$$

This method can be continued to give all the parameters of the power series. The next step would be to check the convergence. The above calculations give the flavour of what would be required to check this. Our subsequent arguments do not need it, and so we do not examine it for this example.

Having derived our power series solution for

$$z_{1t} = \Sigma c_{ij} v_{1t}^{i} v_{2t}^{j},$$

The same must be done for z_{2t}. We can then return to our original system to derive

$$y_{1t} = \exp(\Sigma c_{ij}(u_{1t}-a_2 x_t)^{i}(u_{2t}-b_2 x_t)^{j},$$

and a similar expression for y_{2t}.

The crucial point about the implicit function analysis is that it is only locally valid. Even if the functions $f(\cdot)$ are analytic the weights of the resulting power series are state dependent, being evaluated at a particular point. Our analysis of the behavior of y_{it} is considerably complicated by this fact. A similar observation was made by Bowden (1974) in the context of Taylor series expansions and locally linear models. If all the functions cannot be developed as power series and we cannot find a suitable transformation as in the example, then we can develop a power series approximation by omitting troublesome terms. This would amount to assuming their effect to be small and asymptotically negligible for consistency analysis. It is also worth noting that bilinear models have been suggested in the time series literature as a second order approximation to Volterra expansions. They could be justified in a static framework as a second order approximation to the power series for y_t. However this would require time varying parameter bilinear models as the assumption of constant coefficients is not justified by the theory.

6.7. Implicit Function Theorem and Consistency of NLFIML

What are the implications of these results for the original problem? Both Amemiya (1977) and Phillips (1982) restrict attention to $f_i(\cdot)$ satisfying the implicit function theorem, and their results only require the reduced form to exist. Phillips' "Possibility Theorem" is valid for implicit reduced forms, but to calculate the appropriate mixing distribution in a particular case in general requires an explicit reduced form for the calculations to be feasible.

If the reduced form cannot be written down explicitly, but only as a power series in the regressors and errors with time varying weights, then there is little that can be said about the consistency of NLFIML. For the case in which the reduced form was explicit then there were moment restrictions reflecting the nonlinearities in the system. When the reduced form is a power series, then it is not possible to identify these moment restrictions. In general they apply to all the moments of the distribution. We know that under normality NLFIML is consistent, but cannot specify any other classes of distribution explicitly for which consistency is guaranteed. The arguments for its asymptotic efficiency also require the distributions to be correctly specified. The above analysis does not rule out the possibility of other true distributions for which NLFIML under normality is consistent, but it does suggest that its robustness needs to be proved for particular cases rather than assumed. In the absence of an explicit reduced form this entails simulation studies, but the dependence of the estimators properties on the sequence of exogenous variables

would render the results of little general interest.

This contrasts with the properties of NL3SLS established by Jorgenson & Laffont (1974). Given the conditions for an asymptotic theory for nonlinear models are satisfied then NL3LS is consistent provided the mean of the error process is zero. A comparison of the covariances of NL3SLS & NLFIML for particular cases may be interesting but the calculation of the variance of NLFIML is more complicated in the misspecified case. Again results are model specific and likely to be dependent on the sequence of exogenous variables.

6.8. Asymptotic Normality of NLFIML

The foregoing analysis has concentrated on the point estimate properties of NLFIML. To complete our classical analysis of the estimator we must consider interval estimation using NLFIML, and so find appropriate conditions for it to have a well defined asymptotic distribution.

In a correctly specified model, the asymptotic normality of NLFIML is deduced from a mean value expansion of the score vector about the true parameter value,

$$\left.\frac{\partial L}{\partial \alpha}\right|_{\hat{\alpha}} = \left.\frac{\partial L}{\partial \alpha}\right|_{\alpha_0} + \left.\frac{\partial^2 L}{\partial \alpha \partial \alpha'}\right|_{\alpha^{**}} (\hat{\alpha}-\alpha_0),$$

where α^{**} lies between $\hat{\alpha}$ and α_0. Given a consistent root of the likelihood function the term on the left hand side is zero, and so by applying the central limit theorem to the score evaluated at α_0 we can deduce the asymptotic normality of $\sqrt{T}(\hat{\alpha}-\alpha_0)$. The arguments are similar to those in White (1982) used to establish the asymptotic normality of the

QMLE. The expansion then is about the KLIC minimising value α_*, and expectations are taken with respect to the true distribution. This leads to the conclusion that,

$$\sqrt{T}(\hat{\alpha}-\alpha_*) \overset{a}{\sim} N(0, A_*^{-1}B_*A_*^{-1}),$$

where

$$A_* = \lim T^{-1} \sum_{t=1}^{T} E\frac{\partial^2 L_t}{\partial\alpha\partial\alpha'}\Big|_{\alpha_*},$$

$$B_* = \lim T^{-1} \sum_{t=1}^{T} E\frac{\partial L_t}{\partial\alpha}\frac{\partial L_t}{\partial\alpha'}\Big|_{\alpha_*} - \lim T^{-1} \sum_{t=1}^{T} E\frac{\partial L_t}{\partial\alpha}\Big|_{\alpha_*} E\frac{\partial L_t}{\partial\alpha'}\Big|_{\alpha_*}$$

and L_t is the score associated with likelihood of the observation in period t, $L = \sum_{t=1}^{T} L_t$.

The arguments for the asymptotic normality of NLFIML therefore rely on the validity of the Central Limit Theorem to the quasi score. Amemiya (1977) shows that

$$T^{-1/2}\frac{\partial L}{\partial\alpha_i} = T^{-1/2}\Sigma[\frac{\partial g_i}{\partial u_i} - g_i u'\sigma^i] - T^{-1}\Sigma g_i u'.T^{-1/2}[(\Sigma\frac{uu'}{T})_i^{-1} - \sigma^i],$$

When evaluated at α_* the right hand side has zero expectation by definition. The function g can be considered as a function of u, x and α, and so the only stochastic elements are functions of u_t. If $\alpha_* = \alpha_0$ then the u_t form an i.i.d. sequence and we can apply the Central Limit Theorem provided we make the analogous regularity conditions to Amemiya (1977). Namely $E|g_{it}|^3$ and $E|\partial g_{it}/\partial u_{it}|^3$ are uniformly bounded for all t where u_t is evaluated at α_0 and expectations are taken with respect to the true distribution. For the arguments used in Amemiya (1977) to go through we

require the QMLE to be consistent, and not u_t to be normally distributed.

If $\alpha_* \neq \alpha_0$ then we must consider the behavior of $\partial L/\partial \alpha_i$ when u is evaluated at $u_t^* = f(y_t, x_t, \alpha_*)$. Now

$$u_t^* = u_t + f(y_t, x_t, \alpha_*) - f(y_t, x_t, \alpha_0)$$

$$= u_t + h(y_t, x_t, \alpha_*, \alpha_0)$$

$$= h^*(u_t, x_t, \alpha_*, \alpha_0).$$

Therefore $\partial L/\partial \alpha_i$ is a function of u_t, x_t, α_* and α_0 and as u_t is the only stochastic part of these we can use the same arguments as before. This, again, gives

$$\sqrt{T}(\hat{\alpha}-\alpha_*) \overset{a}{\sim} N(0, A_*^{-1} B_* A_*^{-1}).$$

We can therefore establish the asymptotic normality of NLFIML even when it is not consistent using the conventional assumptions.

In moving from the i.i.d. to the i.n.i.d. case, we encounter problems in consistently estimating the covariance of $\sqrt{T}(\hat{\alpha}-\alpha_*)$. In the i.i.d. case considered by White (1982), $E\partial L_t/\partial \alpha|_{\alpha_*} = 0$ and so the covariance can be estimated consistently by its sample analogue $A_T^{-1} B_T A_T^{-1}$ where

$$A_T = T^{-1} \sum_{t=1}^{T} \partial^2 L_t / \partial \alpha \partial \alpha' \Big|_{\hat{\alpha}},$$

$$B_T = T^{-1} \sum_{t=1}^{T} \frac{\partial L_t}{\partial \alpha} \frac{\partial L_t}{\partial \alpha'} \Big|_{\hat{\alpha}}.$$

However for our model $E\partial L_t/\partial\alpha|_{\alpha_*} \neq 0$ in general, and so $A_T^{-1}B_TA_T^{-1} - A_*^{-1}B_*A_*^{-1} \overset{a.s.}{\not\to} A_*^{-1}D_*A_*^{-1}$, where

$$D_* = \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E\frac{\partial L_t}{\partial \alpha_i} E\frac{\partial L_t}{\partial \alpha_j} \Big|_{\alpha_*}.$$

The matrix $A_T^{-1}B_TA_T^{-1}$ provides an estimator of the mean square error. White (1983) notes that this problem exists but incorrectly calculates the limit of $A_T^{-1}B_TA_T^{-1}$. The complications arise because B_T is not a consistent estimator of the covariance of the score in models of this generality. As White (1983) points out a consistent estimator of D is not available unless the true distribution is known. However as we have seen for regression models with errors assumed to be normal the QMLE is consistent provided the expected value of the true error process is zero. In this case we can consistently estimate the covariance matrix, as argued by GMT. Outside the regression framework, clearly consistent estimation of the first moment of the parameters is a precursor for consistent estimation of the second moments, and therefore for consistent inference using asymptotic tests. Although as argued by White (1983) one could undertake "conservative inference" using the sample moment matrices, as D is positive semi-definite. This underlines the importance of considering the conditions under which NLFIML is consistent. Further it provides another argument in favour of NL3SLS if we are to use the classical criterion to choose estimators. It was demonstrated earlier that NL3SLS is asymptotically normally distributed using the analogous assumptions to the linear model to construct an asymptotic theory.

We have considered the properties of the two most common estimators in the literature on systems of nonlinear static equations. The three stage least squares estimator only requires the assumption that the first moment of the error process be zero and its covariance be constant over time. Under these conditions it is consistent and asymptotically normally distributed, the desired properties for an estimator in classical statistics. The full information maximimum likelihood estimator shares these properties when the normality assumption is correct and is then asymptotically the most efficient estimator. The normality specification can be argued to be made as a way of capturing the symmetry of the errors in an analytically tractable fashion, and so to an extent should be considered arbitrary. Given this, it is desirable that the properties of our estimator should be robust when the distribution is misspecified. We have seen that in general it is not possible to explicitly write down the reduced form of the system, and so we cannot specify the class of true distributions for which NLFIML under normality is consistent. Although it is still asymptotically normally distributed, if NLFIML is not consistent then we cannot consistently estimate the covariance matrix of the QMLE. Therefore if we desire consistent estimators of the parameters and to be able to conduct inference about them using hypothesis tests, then our results suggest that least squares and not maximum likelihood should be used. This contrasts with the analogous result for linear models for which 3SLS and FIML converge in distributions.

7. CONDITIONS FOR GENERALISATION OF STATIC MODEL RESULTS TO DYNAMIC MODEL

7.1. Introduction

It might be anticipated that our conclusions about the properties of NLFIML in the static model can be extended to dynamic models for certain classes of stochastic processes. Our arguments required the convergence in probability of certain functions of random variables. Specifically we need

$$T^{-1}\sum_{t=1}^{T} f(y_t) \xrightarrow{p} \lim T^{-1}\Sigma Ef(y_t).$$

Finding the conditions on y_t under which this holds is refered to by Loeve (1978, p. 37) as the "central asymptotic problem". If y_t is i.i.d. then this result follows from the weak law of large numbers. For sequences of i.n.i.d. y_t, the result can be justified from Kolmogorov's first theorem (see Rao, 1973, p. 114), which states that

"If $\{X_i\}$ $i = 1,2,..$ is a sequence of independent random variables such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$ then $\sum_{i=1}^{\infty} (\sigma_i/i_i)^2 < \infty$ implies $\bar{X}_n \xrightarrow{a.s} \bar{\mu}_n$".

Such regularity conditions are implicit in our earlier analysis. However when considering the properties of NLFIML in dynamic models, we have moved into the world of neither independently nor identically distributed y_t, and so Kolmogorov's theorem is not applicable. The major problem is to find the conditions on y_t that allow law of large number type arguments to be used in dynamic models. Once this is done our earlier analysis easily extends for such processes. In this section we are concerned with the assumptions that must lie behind the construction of an

asymptotic theory of nonlinear dynamic models and so these results are relevant to both LS and ML estimators. The approach taken is quite rigorous as in the absence of such strong assumptions as independence, it is interesting to discover exactly what properties of the r.v's deliver the result. These conditions limit the processes that can be modelled using the theory and their identification is a necessary precursor to assessing whether economic time series satisfy these requirements.

An outline of the chapter is as follows. In section 7.2 we consider an extension of the work of Heijmans and Magnus (1983a) to show the QMLE to the KLIC minimising value in dynamic models. In section 7.3 we examine possible sets of regularity conditions that allow the development of a strong law of large numbers and central limit theorem for dynamic processes. In section 7.4 we show that our analysis of the robustness of NLFIML can be extended to particular dynamic models. The asymptotic normality of NLFIML is examined in section 7.5 and in section 7.6 we consider the plausibility of the mixing process assumption.

7.2. Convergence of QMLE.

7.2.1 Discussion of problem

Heijmans and Magnus (1983a) prove the consistency of the MLE of the parameter vector that indexes the joint density of a sequence of neither independent nor identically distributed random variables. The interest in their proof is the nature of the assumptions made about y_t which they "believe .. are weaker (and more readily applicable) than usual" (Heijmans & Magnus, 1983a, p. 1). Their conditions

do not require the derivatives of the likelihood, uniform convergence or the parameter space to be compact. The situation considered is, therefore, more general than our framework in which the behavior of derivatives is restricted. However their work is of interest for two reasons: as a basis of a more general proof about MLE, and as an example of the limitations of a particular form of analysis for our central question about the robustness of the MLE.

Heijmans and Magnus' (1983a) proof requires the joint p.d.f. of $y_1,..,y_n$ to be correctly specified. The last paragraph of their paper states:

"Finally, there is the problem of misspecification. We have assumed that the true distribution underlying the observations belongs to the parametric family defining the ML estimator. If this is not the case, can our proofs be modified to show that the ML estimator is still consistent?" (p. 26).

The answer to this question is yes and no. It is shown below that their arguments establish the convergence of the MLE to a particular value - the true value when the model is correctly specified. The majority of their proof* concentrates on the convergence property, and only in parts is the correct specification required. From our earlier analysis we would intuitively expect the QMLE to converge to the KLIC minimising value when the model is misspecified. This we establish below by using H&M's convergence arguments

*The original proof contains some errors, which I am indebted to Jan Magnus for bringing to my attention. Below we present a generalisation to the misspecified case of an amended version of their proof.

within the misspecified model, and the probability theory appropriate to this more general case.

As H&M note consistency proofs have taken two forms. We can either seek to establish that the score vector is zero when evaluated at the true parameter value, and so the likelihood has a consistent root, although not necessarily its maximum. This is the approach taken by Amemiya (1977), Phillips (1982) and in our chapters on static models. Alternatively we can examine the ratio $L_n(\gamma_0)/L_n(\gamma)$, where γ_0 is the true value, and show that it is almost everywhere greater than one and that accordingly the MLE must converge to this value. The latter is the approach taken by H&M. In the following analysis we also use this line of argument to show that the QMLE converges to the KLIC minimising value under similar regularity conditions to H&M. However it is seen that this line of argument cannot be used easily to establish the consistency of the QMLE when the model is misspecified. This requires further information on the model, and appears to be more easily handled within the consistent root framework.

To develop this second approach we need the y_t process to satisfy the mixing conditions outlined by White and Domowitz (1982) amongst others. These specify the rate at which the dependence between two observations in time decays as their distance in time increases. It can be shown that if the decay is fast enough we can establish a strong law of large numbers for such processes. In this chapter we outline the proof of this result and consider the applicability of these conditions to economic data. Before considering this work it is necessary to outline certain

definitions from topology, analysis and probability theory that can be found in texts such as Armstrong (1979), Apostol (1974) and Loeve (1962).

7.2.2 Definitions.

1) A random variable X is defined on the triple (Ω, F^*, P) where Ω is the sample space, F^* is a σ-field of subsets of Ω and P the probability density function of X.

2) G^* is a sub σ-field of F^* if it is a collection of subsets of F^* satisfying (i) ϕ and F^* belong to G^* (ii) if G belongs to G^* then so does G^c (iii) if $\{G_n\}$ is a sequence of sets in G^*, then $\bigcup_{n=1}^{\infty} G_n$ belongs to G^*.

3) The minimal σ-field over the class of all intervals from the real line, R, is the Borel field B in R and the elements of B are Borel sets in R.

4) Let $g \in G$, then a neighborhood of g, N(g), is the set $\{\phi : \phi \in G, ||\phi - g|| < r\}$ for some r.

5) Let G be a subset of F^* then G is open if it contains a neighborhood of each of its points.

6) Let p be a point of F^* and $G \subset F^*$ then p is a limit point of G if every neighborhood of p contains at least one point of $G - \{p\}$.

7) A set is closed if it contains all its limit points.

8) A topology on a set F^* is a nonempty collection of subsets of F^*, called open sets, such that any union of open sets is open, any finite intersection of open sets is open, and both F^* and the empty set are open. A set together with a topology on it, is called a topological space.

9) Let F^* be a topological space and let G^* be a family of open subsets of F^* whose union is all of F^*, such a family is called an open cover of F^*. If G' is a subfamily of G^* and if $UG' = F^*$, then G' is called a subcover of G^*.

10) A subset F^* of E^n is closed and bounded if and only if every open cover of F^* (with the induced topology) has a finite subcover.

11) A topological space X is compact if every open cover of F^* has a finite subcover.

12) To every set G there are assigned an open set G^0 and a closed set $\bar{G}$. (i) The interior G^0 of G is the maximal open set contained in G, i.e. the union of all open sets in G. G is open if $G^0 = G$. (ii) The adherence $\bar{G}$ of G is the minimal closed set containing G, that is the intersection of all closed sets containing G. If G is closed then $\bar{G} = G$. These two are related as follows:

$$(G^c)^0 = (\bar{G})^c \text{ and } (G^0)^c = (\overline{G^c}).$$

13) Every set containing a nonempty open set is a neighborhood of any point x of this open set. Let N(x) be the neighborhood of x, then (i) x is interior to G if G is a neighborhood of x (ii) x is adherent to G if no N(x) is disjoint from G.

14) The set G is said to be dense if F if $\bar{G} \supset F$.

7.2.3 Proof of convergence of QMLE to KLIC minimising value.

Notation

Let $y_{(n)} = (y_1, y_2, \ldots, y_n)$ be a set of continuous real

valued random variables, whose assumed joint density function $h_n(y,\gamma)$ is of known form except for the parameter vector $\gamma \in \Gamma \subset R^p$.

Denote the quasi likelihood function by $L_n(\gamma)$ and its log by $\Lambda_n(\gamma)$. The QMLE is the value $\hat{\gamma}(y) \in \Gamma$ such that

$$L_n(\hat{\gamma}_n) = \sup_{\gamma \in \Gamma} L_n(\gamma).$$

Our proof of the convergence of $\hat{\gamma}_n$ to the KLIC minimising value, γ_*, is in three stages. Firstly we show $\hat{\gamma}_n$ exists almost surely. Secondly given its existence we show $\hat{\gamma}_n$ converges to γ_* if a particular condition is satisfied. Finally it is shown that four assumptions guarantee that the condition holds and so the convergence is proved.

Existence of $\hat{\gamma}_n$

If Γ is a compact subset of R^p and $Y^{(n)}$ a measurable space then if

(i) for every γ, $h_n(y,\gamma)$ is a measurable function of y,

(ii) for every y, $L_n(\gamma)$ is a continuous function of γ,

then a QMLE for γ exists almost surely.

The proof is a simple adaptation of Jenrich (1969) lemma 2. Under the above conditions there exists a measurable function $\hat{\gamma}_n$ from $Y^{(n)}$ onto Γ such that for all y in Y:

$$L(\hat{\gamma}_n(y),y) = \sup_{\gamma} L(\gamma,y).$$

The proof depends on subsequence and continuity arguments. The situation is that we have parameter

estimators that are functions of the data. We therefore have a sequence of estimators each of which maximises the likelihood, for a particular sample size and which has a limit point as the sample size increases. We need to show that this limit is the optimum. Compactness ensures that the limit point of this sequence is a member of the parameter space.

Proof

Let (Γ_n) be an increasing sequence of finite subsets of Γ whose limit is dense in Γ.

For each n there is a measurable function $\bar{\gamma}_n$ from y into Γ_n such that

$$L(\bar{\gamma}_n(y),y) = \sup_{\gamma\in\Gamma_n} L(\gamma,y), \quad \text{for all } y \text{ in } Y.$$

Let $\bar{\gamma}_{n1}$ denote the first component of $\bar{\gamma}_n$. Let $\hat{\gamma}_1 = \lim_n \bar{\gamma}_{n1}$ and note $\hat{\gamma}_1$, is measurable as $\bar{\gamma}_{n1}$ is measurable for each n.

For each y in Y there exists a subsequence $(\bar{\gamma}_{n_1}(y))$ of $(\bar{\gamma}_n(y))$, which converges to a point $\tilde{\gamma}$ in Γ of the form $(\hat{\gamma}_1(y),\tilde{\gamma}_2,\ldots,\tilde{\gamma}_p)$.

$$\sup_{(\gamma_1,\ldots,\gamma_p)\in\Gamma} L((\hat{\gamma}_1(y),\gamma_2,\ldots,\gamma_p),y) \geq L(\tilde{\gamma},y) = \lim_i L(\bar{\gamma}_{n_i}(y),y)$$

$$= \lim_i \sup_{\gamma\in\Gamma_{n_i}} L(\gamma,y) = \sup_\gamma L(\gamma,y).$$

The inequality follows because we have enlarged the set over which the supremum is taken, and so it can only become smaller. The first equality is from the definition of $\tilde{\gamma}$ and uses the continuity of $L(\cdot)$. The last equality follows because the limit of Γ_n, a sequence of subsets of Γ, is

dense in Γ and so the minimal closed set containing the limit of Γ_n also contains Γ. Therefore the limit must be Γ itself. This implies

$$\sup_{(\gamma_1,\ldots,\gamma_p)\in\Gamma} L((\hat{\gamma}_1(y),\gamma_2,\ldots,\gamma_p,y) = \sup_{\gamma} L(\gamma,y), \text{ for all } y \in Y.$$

Let $L'(\gamma_1,\ldots,\gamma_p,y) = L(\hat{\gamma}_1(y),\gamma_2,\ldots,\gamma_p,y)$, then $L'(\gamma,y)$ is a continuous function of γ for all y in Y and a measurable function of y for all γ in Γ. Applying the same argument to L' as for L gives a measurable real valued function $\hat{\gamma}_2$ such that

$$\sup_{(\gamma_1,\ldots,\gamma_p)\in\Gamma} L((\hat{\gamma}_1(y),\hat{\gamma}_2(y),\gamma_3,\ldots,\gamma_p),y) = \sup_{\gamma} L(\gamma,y).$$

If we continue to use this argument we can deduce

$$L((\hat{\gamma}_1(y),\hat{\gamma}_2(y),\ldots,\hat{\gamma}_p(y)),y) = \sup_{\gamma} L(\gamma,y) \text{ for all } y \text{ in } Y.$$

Therefore $\hat{\gamma} = (\hat{\gamma}_1,\ldots,\hat{\gamma}_p)$ is a measurable function from y into Γ that maximises the quasi likelihood and so the proof is complete.

7.2.4 Convergence of $\hat{\gamma}_n$

Before considering the theorem and proof of convergence, it is necessary to list a couple of extra definitions from topology.

The parameter space is said to be an interval in R^p, as it is the Cartesian product of p one dimensional intervals. To prove convergence we require Γ to be compact or in other words the Cartesian product of p one dimensional

closed intervals. However if Γ is not compact we can overcome the problem by one-point compactification (see Ash, 1972, p. 388). By adding a point not in Γ to Γ we can construct a compact set, Γ^*, with the same topology as Γ. If we let $\Gamma^* = \Gamma \cup \{\infty\}$, where $\{\infty\}$ is a point not in Γ, and define Z to be an open set in Γ^* if and only if Z is open in Γ or Z is the complement in Γ^* of a compact subset of Γ then

(i) If $Z \subset \Gamma$, Z is open in Γ if and only if Z is open in Γ^*,

(ii) Γ^* is compact.

The property of a compact set that is crucial to our arguments is the existence of a finite subcover. This enables us to restrict attention to a particular finite subset of Γ^* with particular properties.

Let $\gamma_* \in \Gamma$ be the value of γ satisfying

$$S_n(\gamma_*, N(\gamma)) = \log(L_n(\gamma_*) / \sup_{\phi \in N(\gamma)} L_n(\phi)) > 0,$$

then if (i) $\hat{\gamma}_n$ exists a.s.,

(ii) for every $\gamma \neq \gamma_*$ there exists $N(\gamma)$ such that

$$\liminf_{n \to \infty} S_n(\gamma_*, N(\gamma)) > 0 \text{ a.s.},$$

the sequence $\{\hat{\gamma}_n\}$ converges a.s. to γ_* as $n \to \infty$.

This is theorem 1 of Heijmans and Magnus (1983a) except that we do not interpret γ_* as being equal to γ_0, the true value. H&M's result is in two parts. First they prove this theorem, and then establish a set of conditions that imply (ii). We inevitably follow this format. Their proof of theorem 1 is reproduced verbatim here as it does not require the model to be correctly specified. The conditions however

do need adaptation, and we develop the idea of γ_* as the KLIC minimising value when we examine that part of the proof.

Proof

Let $N_*(\gamma_*)$ be some neighborhood of γ_* and $N_*^c(\gamma_*)$ be its complement in Γ^*. For every point ϕ in $N_*^c(\gamma_*)$ there exists (by assumption) a neighborhood $N'(\phi)$ such that

$$\liminf S_n(\gamma_*, N'(\phi)) > 0 \text{ a.s. .}$$

The union of all such neighborhoods of points in ϕ in $N_*^c(\gamma_*)$ covers $N_*^c(\gamma_*)$,

$$\bigcup_{\phi \in N_*^c(\gamma_*)} N'(\phi) \supset N_*^c(\gamma_*).$$

Since Γ^* is compact, $N_*^c(\gamma_*)$ is compact as $N_*(\gamma_*)$ is an open set relative to Γ^* by definition. Therefore there exists a finite subcover of $N_*^c(\gamma_*)$. In other words we can find a finite number of points $\phi_1,..,\phi_r$ and $\phi_{r+1} = \{\infty\}$ (from the compactivisation), in $N_*^c(\gamma_*)$ with neighborhoods $N_h(\phi_h)$, $(h = 1,..,r+1)$ such that

$$\bigcup_{h=1}^{r+1} N_h(\phi_h) \supset N_*^c(\gamma_*),$$

and $\liminf_{n\to\infty} S_n(\gamma_*, N_h(\phi_h)) > 0$ a.s., $h = 1,..,r+1$. This implies

$$\sup_{\phi \in N_*^c(\gamma_*)} \Lambda_n(\phi) \leq \sup_{\phi \in \bigcup_{h=1}^{r+1} N_h(\phi_h)} \Lambda_n(\phi) = \max_{1 \leq h \leq r+1} \sup_{\phi \in N_h(\phi_h)} \Lambda(\phi).$$

The inequality follows because the union contains $N_*^c(\gamma_*)$,

and so the supremum over the larger set cannot be smaller.

In turn this gives

$$\Lambda_n(\gamma_*) - \sup_{\phi \in N_*^c(\gamma_*)} \Lambda_n(\phi) \geq \Lambda_n(\gamma_*) - \max_{1 \leq h \leq r+1} \sup_{\phi \in N_h(\phi_h)} \Lambda(\phi)$$

$$= \min_{1 \leq h \leq r+1} [\Lambda_n(\gamma_*) - \sup_{\phi \in N_h(\phi_h)} \Lambda_n(\phi)]$$

$$= \min_{1 \leq h \leq r+1} S_n(\gamma_*, N_h(\phi_h)).$$

From the basic definition of $N_h(\phi_h)$ we have

$$P[\lim_{n \to \infty} \inf\{ \min_{1 \leq h \leq r+1} S_n(\gamma_*, N_h(\phi_n))\} > 0] = 1,$$

and so

$$P[\lim_{n \to \infty} \inf\{\Lambda_n(\gamma_*) > \sup_{\phi \in N_*^c(\gamma_*)} \Lambda_n(\phi)\}] = 1.$$

As $\hat{\gamma}_n$ exists a.s., $\hat{\gamma}_n \in N_*(\gamma_*)$ a.s. as $n \to \infty$ and so $\hat{\gamma}_n$ converges almost surely to γ_*.

To establish a set of conditions that imply condition ii) of the theorem, we require a variant of the law of large numbers and less strict convergence properties of random variables. The arguments are formulated in terms of conditional expectations and so we are able to make use of the following results.

1) The monotone convergence theorem for conditional expectations: let X be a r.v defined on (Ω, F^*, p)

and G^* be a sub σ-field of F^* then if $X_n \to X$, and $E|X_n| < \infty$, it follows that $E\,X_n \to E\,X$.

2) The Stability Theorem (Loeve, 1978, p. 53):

If $\Sigma \frac{\sigma^2 X_n}{b_n^2} < \infty$ with $b_n \to \infty$, then

$$\frac{1}{b_n} \sum_{k=1}^{n} \{X_k - E(X_k | X_1, \ldots, X_{k-1})\} \overset{a.s.}{\to} 0.$$

The next step is to establish a set of conditions implying $\liminf S_n(\gamma_*, N(\gamma)) > 0$ a.s. This part of our proof, although relying heavily on H&M, differs from their result. Firstly because we need to correct their proof* for the original case they consider and secondly because for the misspecified case we take expectations with respect to the true density which is not $h(y, \gamma_*)$.

Define $g_n(\gamma) = L_n(\gamma)/L_{n-1}(\gamma)$ and $L_0(\gamma) = 1$, $g_n(\gamma)$ is just the conditional density of y_n given $y_{n-1}, \ldots, y_1$. For Φ a nonempty subset of Γ^*, define

$$T_n(\gamma_*, \Phi) = \log\{g_n(\gamma_*)/\sup_{\phi \in \Phi} g_n(\phi)\}.$$

To establish the desired result we make the following additional assumptions.

1) For theorem 1 we require the likelihood to be continuous and this in turn implies $g_n(\gamma)$ and $T_n(\cdot)$ are continuous. Also we assume $E[T_n | y_{n-1}, \ldots, y_1]$ is a continuous function of γ for all y.

---

*I am indebted to Jan Magnus for bringing the errors in the original to my attention.

2) $E(T_n|y_{n-1},..,y_1) > 0$ where $\gamma_* \neq \Phi$, for all n. This amounts to requiring γ to be identified for each conditional distribution $g_n(\cdot)$. Note this assumption also implies

(i) γ is asymptotically identifiable and so $\lim_{n\to\infty} \inf(1/n)E\log(L_n(\gamma_*)/L_n(\gamma)) > 0$ for every $\gamma \neq \gamma_*$,

(ii) $\lim_{n\to\infty} \inf(1/n)\Sigma\, E[T_n|y_{n-1},..,y_1] > 0$.

3) For every $\gamma \neq \gamma_*$, $ET_n(\gamma_*,\gamma)$ is uniformly bounded and $ET_n(\gamma_*,N(\gamma))$ is uniformly bounded for some neighborhood $N(\gamma)$ of γ.

4) For every $\gamma \neq \gamma_*$ there exists an $\alpha < 1$ such that $n^{-\alpha}ET_n^2(\gamma_*,\gamma)$ is uniformly bounded and $n^{-\alpha}ET_n^2(\gamma_*,N(\gamma))$ is uniformly bounded for some neighborhood $N(\gamma)$ of γ.

Assumptions 3) and 4) are just the regularity conditions for the stability theorem. Under these assumptions, the condition (ii) from theorem 1 is satisfied and so $\{\hat{\gamma}_n\} \to \gamma_*$ a.s. .

It is the identification condition that allows us to continue the KLIC minimising interpretation of γ_*. Recall the KLIC is

$$I(h,p:\gamma) = E\log[h(\gamma,y)/p(y)],$$

where $h(y,\gamma)$ is the assumed density and $p(y)$ the true p.d.f. of y.

Consider the ratio

$$R_{\gamma_*} = \frac{E[\log(h(y,\phi)/p(y))]}{E[\log(h(y,\gamma_*)/p(y))]},$$

and let γ_* be the KLIC minimising value. This implies

$$R_{\gamma_*} > 1 \text{ for all } \phi \neq \gamma_*.$$

In turn this gives

$$\text{Elogh}(y,\phi) - \text{Elogp}(y) > \text{Elogh}(y,\gamma_*) - \text{Elogp}(y),$$

and so

$$\text{Elogh}(y,\phi) - \text{Elogh}(y,\gamma_*) > 0.$$

Therefore if the vector γ_* is to satisfy the identification condition then it must be the KLIC minimising value. This proof consists of a generalisation to the misspecified case of one possible amended version of H&M's proof for the correctly specified model.

Let $\gamma' \neq \gamma_*$ be an arbitrary point of Γ^*. We need to find a neighborhood $N(\gamma')$ of γ' such that

$$\lim_{n\to\infty} \inf S_n(\gamma_*, N(\gamma')) > 0 \text{ a.s. .}$$

Now

$$S_n(\gamma_*, N(\gamma')) = \inf_{\phi \in N(\gamma')} \log(L_n(\gamma_*)/L_n(\phi))$$

$$= \inf_{\phi \in N(\gamma')} \sum_{t=1}^{n} \log(g_t(\gamma_*)/g_t(\phi))$$

$$\geq \sum_{t=1}^{n} \inf_{\phi \in N(\gamma')} \log(g_t(\gamma_*)/g_t(\phi))$$

$$= \sum_{t=1}^{n} T_t(\gamma_*, N(\gamma')) \text{ a.s. .}$$

Therefore our result follows if we can show our assumptions imply

$$\lim_{n\to\infty} \inf n^{-1} \sum_{t=1}^{n} T_t(\gamma_*, N(\gamma')) > 0 \text{ a.s. .}$$

The procedure is that we first establish that $T_t(\cdot)$ satisfies the conditions of the monotone convergence theorem. It is then established using this theorem and the identification condition that the conditional expectation of T_t lies between $(0,\delta)$. We then need to verify that the stability theorem applies to deduce that T_t is subject to the same bounds.

Condition 2) guarantees there exists a neighborhood $N^1(\gamma')$ of γ' such that

$ET_t(\gamma_*, N^1(\gamma'))$ is uniformly bounded.

Let $N^1(\gamma')$ be the first element of a sequence $\{N^i(\gamma'), i \in N, \text{ the set of natural numbers}\}$ of neighborhoods of γ' with property that

$$N^i(\gamma') \supset N^{i+1}(\gamma') \text{ and } \lim N^i(\gamma') = \gamma'.$$

For every $n \in N$

$$T_n(\gamma_*, N^i(\gamma')) \leq T_n(\gamma_*, N^{i+1}(\gamma')) \text{ a.s.,}$$

because we have shrunk the neighborhood under consideration. Therefore

$$T_n(\gamma_*, N^i(\gamma')) \to T_n(\gamma_*, \gamma') \text{ a.s. as } i \to \infty. \qquad (37)$$

This implies that for every $i \in N$ $ET_n(\gamma_*, N^i(\gamma'))$ is uniformly bounded in n.

Define

$$A_n^i(\gamma_*, \gamma') = E(T_n(\gamma_*, N^i(\gamma'))|y_1, .., y_{n-1}),$$

and

$$A_n(\gamma_*, \gamma') = E(T_n(\gamma_n, \gamma')|y_1, .., y_{n-1}).$$

As $ET_n(\gamma_*, N^i(\gamma'))$ and $ET_n(\gamma_*, \gamma')$ are uniformly bounded, and from (37), we can apply the monotone convergence theorem for conditional expectations to obtain

$$A_n^i(\gamma_*, \gamma') \to A_n(\gamma_*, \gamma') \text{ a.s., for } i \to \infty.$$

For the next stage of the proof we need to use this result to justify

$$0 \leq n^{-1} \sum_{t=1}^{n} A_t(\gamma_*, \gamma') - n^{-1} \sum_{t=1}^{n} A_t^i(\gamma_*, \gamma') < \varepsilon, \qquad (38)$$

for every ε satisfying $0 < \varepsilon < \sup_n\{A_n(\gamma_*, \gamma') - A_n^i(\gamma_*, \gamma')\}$, and for a neighborhood $N^i(\gamma')$ of γ'.

From the monotone convergence theorem we have

$$0 \leq A_n(\gamma_*,\gamma') - A_n^i(\gamma_*,\gamma') < \varepsilon \quad \text{a.s.},$$

for any i and some ε.

If we assume (i) T_n^i, A_n^i are continuous in i for all y and γ (ii) $A_n^i > 0$ for all n and i, then we must be able to find a finite i, i_* say, satisfying (38). Let i_* be the smallest i satisfying (38).

From assumption 2) we know $\liminf_{n\to\infty} n^{-1} \sum_{t=1}^{n} A_t(\gamma_*,\gamma') > 0$, and so from equation (38),

$$\liminf n^{-1} \sum_{t=1}^{n} A_t^i(\gamma_*,\gamma') > 0 \text{ a.s., for } i > i_*. \qquad (39)$$

We now need to show the conditions of Loeve's stability theorem are satisfied to use (39) to learn about $\liminf n^{-1}\Sigma T_t(\cdot)$.

Condition 4) ensures the existence of a neighborhood of γ', $\bar{N}(\gamma')$ such that

$$n^{-\alpha} E T_n^2(\gamma_*,\bar{N}(\gamma'))$$

is uniformly bounded. Let i_{**} be an index such that $N^i(\gamma') \subset \bar{N}(\gamma')$ for all $i \geq i_{**}$.

This implies

$$T_n(\gamma_*,\bar{N}(\gamma')) \leq T_n(\gamma_*,N^i(\gamma')) \leq T_n(\gamma_*,\gamma') \text{ a.s. } i \geq i_{**},$$

as the neighborhood shrinks as we move from left to right of this inequality.

Therefore we have

$$T_n^2(\gamma_*, N^i(\gamma')) \leq T_n^2(\gamma_*, \bar{N}(\gamma')) + T_n^2(\gamma_*, \gamma') \text{ a.s. .}$$

Taking expectations it follows that $n^{-\alpha}\text{var } T_n(\gamma_*, N^i(\gamma'))$ is uniformly bounded for every $i \geq i_{**}$. In particular i such that

$$\sum_{t=1}^{n} \frac{\text{var } T_t(\gamma_*, N^i(\gamma'))}{t^2} < \infty.$$

We can therefore use Loeve's stability theorem to deduce, for every $i > i_{**}$

$$n^{-1} \sum_{t=1}^{n} T_t(\gamma_*, N^i(\gamma')) - E(T_t(\gamma_*, N^i(\gamma^*)) | y_1, .., y_{t-1})$$

tends to zero a.s. for $n \to \infty$, and so

$$\lim \inf (1/n) \sum_{t=1}^{n} T_t(\gamma_*, N^i(\gamma')) > 0 \text{ a.s.,}$$

for every $i \geq \max(i_*, i_{**})$.

We have therefore answered H&M'S question. It has been shown that under a set of regularity conditions, the QMLE converges almost surely to the KLIC minimising value. What is apparent from contrasting this approach to our earlier analysis is that without further information about the model it is not possible to generalise this proof into one of consistency for the misspecified case. This problem is more easily handled by the other approach in ML theory. We return to the consistent root of the likelihood equation arguments later. To produce analogous arguments to the static model about the robustness of NLFIML we need to be

able to apply a version of the strong law of large numbers to functions of dependent random variables. The conditions that underlie this are explored in the next sections.

7.3 Assumptions underlying the strong Law of Large Numbers for dependent processes

The extension of our results on the asymptotic properties of NLFIML and NL3SLS requires the application of a strong law of large numbers and Central Limit theorem for dependent processes. In this section we examine possible sets of regularity conditions that deliver this result and that must be assumed to hold if we are to construct an asymptotic theory of nonlinear models. The main problem is going to be unravelling the implications of restrictions on functions of variables for the underlying raw series. This of course is a problem in the static model, but our analysis followed tradition and assumed it away. In the linear model the assumption that suitabley normalised cross product matrices converge to a limit has clear implications for the variables themselves. Typically in the nonlinear framework we make analogous regularity conditions to the linear model but the restriction that $T^{-1}\Sigma f_t f'_t$ tends to a limit has less easily interpretable implications for the variables.

Our analysis requires the convergence in probability of various functions of the variables. This can be handled by making regularity assumptions about each function as it becomes necessary, and so implicitly restricting the underlying variables. Alternatively we can seek assumptions about the variables that imply particular functions obey a SLLN. The latter approach would appear preferable, as it is

able to apply a version of the strong law of large numbers to functions of dependent random variables. The conditions that underlie this are explored in the next sections.

7.3 Assumptions underlying the strong Law of Large Numbers for dependent processes

The extension of our results on the asymptotic properties of NLFIML and NL3SLS requires the application of a strong law of large numbers and Central Limit theorem for dependent processes. In this section we examine possible sets of regularity conditions that deliver this result and that must be assumed to hold if we are to construct an asymptotic theory of nonlinear models. The main problem is going to be unravelling the implications of restrictions on functions of variables for the underlying raw series. This of course is a problem in the static model, but our analysis followed tradition and assumed it away. In the linear model the assumption that suitabley normalised cross product matrices converge to a limit has clear implications for the variables themselves. Typically in the nonlinear framework we make analogous regularity conditions to the linear model but the restriction that $T^{-1}\Sigma f_t f'_t$ tends to a limit has less easily interpretable implications for the variables.

Our analysis requires the convergence in probability of various functions of the variables. This can be handled by making regularity assumptions about each function as it becomes necessary, and so implicitly restricting the underlying variables. Alternatively we can seek assumptions about the variables that imply particular functions obey a SLLN. The latter approach would appear preferable, as it is

able to apply a version of the strong law of large numbers to functions of dependent random variables. The conditions that underlie this are explored in the next sections.

7.3 Assumptions underlying the strong Law of Large Numbers for dependent processes

The extension of our results on the asymptotic properties of NLFIML and NL3SLS requires the application of a strong law of large numbers and Central Limit theorem for dependent processes. In this section we examine possible sets of regularity conditions that deliver this result and that must be assumed to hold if we are to construct an asymptotic theory of nonlinear models. The main problem is going to be unravelling the implications of restrictions on functions of variables for the underlying raw series. This of course is a problem in the static model, but our analysis followed tradition and assumed it away. In the linear model the assumption that suitabley normalised cross product matrices converge to a limit has clear implications for the variables themselves. Typically in the nonlinear framework we make analogous regularity conditions to the linear model but the restriction that $T^{-1}\Sigma f_t f'_t$ tends to a limit has less easily interpretable implications for the variables.

Our analysis requires the convergence in probability of various functions of the variables. This can be handled by making regularity assumptions about each function as it becomes necessary, and so implicitly restricting the underlying variables. Alternatively we can seek assumptions about the variables that imply particular functions obey a SLLN. The latter approach would appear preferable, as it is

desirable to examine whether economic series in fact satisfy these requirements. It is this question that led White and Domowitz (1982) to suggest modelling economic series by mixing processes. We assess the arguments in favor of this practice in section 7.6. Before that, we consider other possible regularity conditions and their interrelationship.

7.3.1 Martingales

Martingale arguments are often invoked in the time series literature to apply a central limit theorem to the score vector and for the estimation of the covariance matrix of the MLE. It is therefore worth considering the extent to which they provide the solution to our problem.

A martingale is a sequence (X_n, F_n^*) defined on the probability triple (Ω, F_n^*, p) satisfying

i) F_n^* are an increasing sequence of σ fields i.e. $F_n^* \subseteq F_{n+1}^*$,

ii) $X_n \in L^1(\Omega, F_n^*, p)^*$, that is X is a r.v defined on (Ω, F_n^*, p) and $E|X_n| < \infty$ for all n.

iii) $X_n = E(X_{n+1}|F_n^*)$ a.s. for all n.

Our analysis is concerned with appropriately normalised summations of r.v's and so we define $S_n = \sum_{i=1}^{n} X_i$. If X_i does not have a zero mean than the arguments carry through by centering it about its conditional expectation. Put $z_i = X_i - E_{i-1}X_i$, then $EZ_i = 0$. Let X_i be a zero mean martingale then S_n is said to possess the martingale property as

---

*We write $X_n \in L^p$ if X_n is a r.v defined on (Ω, F^*, p) and $E|X_n|^p < \infty$.

$$E(S_n|S_1,..,S_{n-1}) = E(S_{n-1}+X_n|S_1,...,S_{n-1}) = S_{n-1}.$$

To establish the convergence of the summation we require a bound on some measure of the discrepency between S_i and S_j. For the independence case, laws of large numbers are based on the Chebyshev inequality,

$$P(|S_n|\geq\varepsilon) \leq \varepsilon^{-2}ES_n^2, \quad \text{for any } \varepsilon > 0,$$

and additional conditions on the probability of outliers and the order in probability of the first two moments of the process (see Loeve, 1977, p. 290).

Independence is only important in the derivation of the inequality because it implies the orthogonality of X_i and X_j. Hall and Heyde (1981) show that the assumption that $\{S_n,F_n^*\}$ is a zero mean martingale is also sufficient for this property as

$$E(X_iX_j) = E(X_jE(X_i|F_{i-1}^*)) = E(X_j[E(S_i|F_{i-1}^*)-S_{i-1}]) = 0.$$

The Chebyshev inequality can therefore be extended to such processes, and so to establish convergence we require a diminishing bound on $S_\infty-S_m$ as $m \to \infty$. Feller (1971, p. 242) uses these arguments to prove the martingale convergence theorem for S_n processes satisfying the above conditions and $ES_n^2 < c < \infty$.

The Chebyshev inequality implies

$$p(|S_n-S_m| > \varepsilon) \leq \varepsilon^{-2}var(S_n-S_m), \; n > m, \tag{40}$$

and so for convergence we need to show $var(S_\infty - S_m)$ tends to zero. As S_n has zero mean,

$$var(S_n - S_m) = E(S_n - S_m)^2 = ES_n^2 - 2ES_nS_m + ES_m^2.$$

By the tower property of conditional expectations,

$$ES_nS_m = ES_mE(S_n|S_m) = ES_m^2,$$

and so

$$var(S_n - S_m) = ES_n^2 - ES_m^2.$$

Under our assumptions (S_n^2, F_n^*) forms a submartingale sequence and so $\{E(S_n^2)\}$ is a monotonically increasing sequence that tends to a finite limit. Therefore if we set $n = \infty$, in equation (40), it has been shown that S_m converges in probability to S_∞.

As it stands this theorem does not give us the required interpretation of the limit in terms of an expectation. However the Kolmogorov strong law of large numbers* for i.n.i.d. r.v's cited earlier can be directly extended to martingales via Kronecker's lemma (see for instance Feller, 1971, p. 239). This states:

"Let $\{x_k\}$ be an arbitrary numerical sequence and $\{b_i\}$ a strictly increasing sequence of positive constants. If the series $\Sigma_1^\infty b_k x_k$ converges, then $\frac{x_1+x_2+..+x_n}{b_n} \to 0$."

---

*Rao (1973)'s proof on p. 114 and p. 142 is based on the Hajek Renyi inequality which only requires the X_i to be orthogonal.

PAGINATION ERROR

7.3.2 Mixingales

McLeish (1975) defines the sequence (X_n, F_n^*) as a mixingale if, for sequences of finite nonnegative constants c_n and ψ_m where $\psi_m \to 0$ as $m \to \infty$, we have for all $n \geq 1$, $m \geq 0$,

a) $||E_{n-m}X_n|| \leq \psi_m c_n$,

b) $||X_n - E_{n-m}X_n||_2 \leq \psi_{m+1} c_n$.

Mixingales can be considered as asymptotic martingales as the definition implies

c) $||E_{-\infty}X_i||_2 = 0$

d) $X_i - E_{+\infty}X_i = 0$ a.s. for all i.

This resemblance is sufficient for many of the martingale convergence theorems to be extended to these more general processes. Square integrable martingales are a special case of this definition obtained by putting $\psi_0 = 1$, $\psi_m = 0$ for $m \geq 1$ and $c_n = (EX_n^2)^{1/2}$.

Chebyshev's inequality can again be used as the basis for a convergence theorem, but the arguments need to be generalised. McLeish (1975) shows that if $\{X_n, F_n^*\}$ is a mixingale such that $\psi_n = O(n^{-1/2}(\log n)^{-2})^*$ as $n \to \infty$, then

$$E(\max_{i \leq n} S_i^2) \leq k \sum_{i=1}^{n} c_i^2,$$

where k is a constant depending only on ψ_n. This result can

*See appendix 3 for discussion/definition of size of ψ_n.

be used to bound the variance of the difference between S_n and S_m in our earlier analysis. We assume $\sum_{i=1}^{\infty} c_i^2 < \infty$ and $\psi_n = O(n^{-1/2}(\log n)^{-2})$ as $n \to \infty$, then

$$p\left(\max_{m<n\leq m'} |S_n - S_m| > \varepsilon\right) \leq \varepsilon^{-2} k \sum_{i=m}^{m'} c_i^2,$$

and so

$$p\left(\max_{n\leq m} |S_n - S_m| > \varepsilon\right) \leq \varepsilon^{-2} k \sum_{i=m}^{\infty} c_i^2 .$$

If we let $m \to \infty$, then the bound on the probability goes to zero and so S_n converges a.s.. As before we can use Kronecker's lemma to deduce the Kolmogorov strong law of large numbers for mixingale processes.

7.3.3 Mixing Processes

To define a mixing process it is necessary first to consider two measures of dependence between σ algebras F^* and G^*:

(i) a relative measure

$$\phi(F^*, G^*) = \sup_{\{F \in F^*, G \in G^* : p(F) > 0\}} |p(G|F) - p(G)|,$$

(ii) a "strong" measure

$$\alpha(F^*, G^*) = \sup_{\{F \in F^*, G \in G^*\}} |p(FG) - p(F)p(G)|.$$

The events in F^* and G^* are independent if and only if ϕ and α are zero. Let $\{X_t\}$ denote a sequence of random vectors defined on (Ω, F^*, p) and let $F_a^{*b} = \sigma(Z_t; a \leq t \leq b)$ - the Borel σ- algebra of events generated by $Z_a, Z_{a+1}, \ldots, Z_b$. Now define $\phi(m) = \sup_n \phi(F_{-\infty}^{*n}, F_{n+m}^{*\infty})$ and $\alpha(m) = \sup_n \alpha(F_{-\infty}^{*n}, F_{n+m}^{*\infty})$,

both of which measure the dependence that exists between events at least m periods apart. A sequence for which $\phi(m) \to 0$ as $m \to 0$ is called uniform mixing and one for which $\alpha(m) \to 0$ as $m \to 0$ is strong mixing. From the definition of conditional probability, ϕ mixing implies α mixing. Essentially mixing processes are sequences for which the dependence between two observations in time decreases as the distance between them grows larger. It would intuitively be expected that there would be connection between mixingales and mixing processes as both are defined in terms of a decaying dependence between observations over time. The relationship between convergence in absolute and conditional probability is translated into one between absolute and conditional moments by the following result due to McLeish (1975). This enables us to establish that mixing processes are mixingales and so the convergence theorems of the latter can be applied to the former.

Let X be a r.v. measurable with respect to F^* and $1 \leq p \leq r \leq \infty$. Then

a) $||E(X|F^*)-E(X)||_p \leq 2\phi(m)^{1-1/r}||X||_r$,

b) $||E(X|F^*)-E(X)||_p \leq 2(2^{1/p}+1)\alpha(m)^{1/p-1/r}||X_r||$.

Below we outline the proof of part a) and leave the proof of b) to appendix 4. Before we can do this it is necessary to introduce some extra definitions from measure theory (Loeve, 1962, p.82-84).

Definitions

1) A set function ψ is defined on a nonempty class Γ of sets in a space Ω by assigning to every set $A \in \Gamma$ a single number $\psi(A)$, finite or infinite, the value of ψ at A. If every set in Γ is a countable union of sets

in Γ at which ψ is finite, ψ is said to be σ-finite.

2) ψ is (countably or σ) <u>additive</u> if $\psi(\Sigma A_j) = \Sigma\psi(A_j)$ either for every countable or only for every finite class of disjoint sets respectively.

3) Let ψ be an additive function on a field Γ and define ψ^+ and ψ^- on Γ by

$$\psi^+(A) = \sup_{B\subset A}\psi(B),\quad \psi^-(A) = -\inf_{B\subset A}\psi(B),\quad A,B,\ \epsilon\ \Gamma$$

The set functions ψ^+, ψ^- and $\bar{\psi} = \psi^+ + \psi^-$ are called the <u>upper, lower and total variation</u> of ψ on Γ. Since $\psi(\phi) = 0$, these variations are nonnegative.

4) <u>Jordan Hahn decomposition</u>: If ψ is σ additive on the σ-field Λ, then there exists a set $D\ \epsilon\ \Lambda$ such that for every $A\ \epsilon\ \Lambda$

$$-\psi^-(A) = \psi(A\cap D),\quad \psi^+(A) = \psi(A\cap D^c).$$

ψ^+ and ψ^- are measures and $\psi = \psi^+ - \psi^-$ is a signed measure, as at least one of its components is finite.

We can now proceed with the proof.

<u>Proof</u>

a) This follows from theorem 2.2 of Serfling (1968). The argument is as follows:

We assume $E|X|^r < \infty$ for some $r > 1$. Let p denote the probability measure induced on $F^{*\infty}_{n+m}$ by (Ω,F^*,p) and $p(.|F^{*n}_{-\infty})$ denote a regular conditional probability measure on $F^{*\infty}_{n+m}$ given $F^{*n}_{-\infty}$. Let μ be the signed measure $p(\cdot|F^{*n}_{-\infty})-p(.)$. The space Ω corresponding to the r.v's $\{X_{n+m},X_{n+m+1},..\}$ has a Hahn decomposition $\Omega = \Omega^+ \cup \Omega^-$ with respect to μ, such that

for any measurable subset A of Ω, the sets $A \cap \Omega^+$ and $A \cap \Omega^-$ are measurable and $\mu(A\cap\Omega^+) \geq 0$, $\mu(A\cap\Omega^-) \leq 0$. As Ω is measurable so too are Ω^+ and Ω^-. Therefore as μ and $-\mu$ are measures on Ω^+ and Ω^- respectively

$$|E(X|F^{*n}_{-\infty})-E(X)| = |\int_\Omega X dP(\omega|F^{*n}_{-\infty})-dP(\omega)|$$

$$\leq |\int_{\Omega^+} X d\mu| + |\int_{\Omega^-} X d(-\mu)|$$

$$\leq \int_{\Omega^+}|X|d\mu + \int_{\Omega^-}|X|d(-\mu).$$

By Loeve's c_r inequalities*

$$|E(X|F^{*n}_{-\infty})-E(X)|^p \leq 2^{p-1}[\int_{\Omega^+}|X|d\mu]^p+2^{p-1}[\int_{\Omega^-}|X|d(-\mu)]^p. \quad (41)$$

From the definition of a ϕ mixing process we have

$$\int_{\Omega^+} d\mu = p(\Omega^+|F^{*n}_{-\infty}) - p(\Omega^+) \leq \phi(m).$$

By Holders inequality,

$$\int_{\Omega^+}|X|d\mu \leq (\int_{\Omega^+}|X|^p d\mu)^{1/p}(\int_{\Omega^+} d\mu)^{1/q}, \quad p^{-1}+q^{-1} = 1,$$

and so

---

*The bounds are manipulated to produce the required result using the following inequalities (see Loeve, 1962, p. 155-156).

i) c_r-inequality: $E|X+Y|^r < c_r E|X|^r + c_r E|Y|^r$, where

$c_r = 1$ or 2^{r-1} according to whether $r > 1$ or $r < 1$.

ii) Holder inequality: $E|XY| < E^{1/r}|X|^r E^{1/s}|Y|^s$ where $r > 1$.

and $s^{-1} + r^{-1} = 1$.

$$[\int_{\Omega^+}|X|d\mu]^p \leq [\phi(m)]^{p/q}\int_{\Omega^+}|X|^p d\mu.$$

Similar reasoning can be used for Ω^- and μ, to give

$$|E(X|F_{-\infty}^{*n})-E(X)|^p \leq 2^{p-1}[\phi(m)]^{p/q}[\int_{\Omega^+}|X|^p d\mu+\int_{\Omega^-}|X|^p d(-\mu)].$$

The RHS of this equation is $2^{p-1}[\phi(m)]^{p/q}[E(|X|^p|F_{-\infty}^{*n})-E|X|^p]$. Therefore just using the fact that we have taken the modulus of X we can rewrite it as,

$$|E(X|F_{-\infty}^{*n})-E(X)|^p \leq 2^{p-1}[\phi(m)]^{p/q}[E(|X|^p|F_{-\infty}^{*n}) + E|X|^p],$$

which implies

$$E|E(X|F_{-\infty}^{*n}) - E(X)|^p \leq 2^p(\phi(m))^{p/q}E|X|^p,$$

and so

$$||E(X|F_{-\infty}^{*n}) - E(X)||p \leq 2\phi(m)^{1-1/p}||X||p.$$

The final step in the proof follows from the fact that we could have used $r \geq p$ in the exponent of the c_r inequality in equation (41). $||\cdot||_k$ is a nondecreasing function of k and so the bound with r as exponent is also a bound when p is used. Therefore

$$||E(X|F_{-\infty}^{*n}) - E(X)||p \leq 2\phi(m)^{1-1/r}||X||_r.$$

Using these inequalities with $\phi(m)$ and $\alpha(m)$, it can be seen that mixing processes are indeed mixingales. For

instance a ϕ mixing process is a mixingale with $c_n = 2(EX_n^2)^{1/2}$ and $\psi_m = \phi^{1/2}(m)$. From our earlier definition of an L^2 martingale, which required $\psi_0 = 1$, $\psi_m = 0$ and $c_n = (EX_n^2)^{1/2}$, it can be seen that the L^2 martingales are not equivalent to ϕ mixing processes, although it would be anticipated that convergence theorems for the latter would apply to martingales due to the generosity of its bound. The one important property possessed by mixing processes is that functions of them are themselves mixing. The specification of the martingale in terms of conditional expectations meant it did not satisfy this requirement. However it would intuitively be expected that if a sequence has decaying dependence over time then, suitably restricted nonlinear transformations of the series would exhibit similar behavior. By defining mixing processes in terms of probabilities, we allow functions of the sequence to exhibit similar patterns of behavior.

McLeish (1975) establishes this property of mixing conditions. Let $\{\xi_n; -\infty<n<\infty\}$ be a ϕ mixing sequence*, and let $X_n = f_n\{\xi_j\}$ where f_n is a nonrandom function of the whole history, past and future, of the process, and $EX_i = 0$. Define

$$F_n^{*m} = \sigma(\xi_n,..,\xi_m) \text{ for } m \geq n,$$

$$\phi_m = \sup_n \phi(F_{-\infty}^{*n}, F_{n+m}^{*\infty}),$$

*A similar result can be shown for α mixing sequences, see McLeish (1975).

$$\nu_m = \sup_i ||E(X_i|F^{*i+m}_{i-m}) - X_i||_2,$$

where ν_m is $O[n^{1/2}\log n(\log\log n)^{1+\delta}]^{-1}$.

As $E(X_i|F^{*i+m}_{i-m}) = X_i$, we have

$$||E_{i-2m}X_i||_2 \leq ||E_{i-2m}E(X_i|F^{*i+m}_{i-m})||_2 + ||X_i - E(X_i|F^{*i+m}_{i-m})||_2$$

$$\leq 2\phi_m^{1-1/r}||E(X_i|F^{*i+m}_{i-m}||_2 + \nu_m,$$

from the definition of ϕ mixing and the fact that $EX_i = 0$. From Jensens inequality we can then show that

$$|E(X_i|F^{*i+m}_{i-m})||_r \leq ||X_i||_r,$$

and so

$$||E_{i-2m}X_i||_2 \leq 2\phi_m^{1-1/r}||X_i||_r + \nu_m.$$

The sequence X_i therefore satisfies part a) of the condition for a mixingale. To establish that it satisfies b) we need a lemma due to Billingsley (1968, p. 184). He shows that if F^* and G^* are two σ-fields with $F^* \subset G^*$ and $E(Y^2) < \infty$ then

$$E\{|Y - E(Y|G^*)|^2\} \leq E\{|\xi - E\{\xi|G^*\}|^2\}.$$

This can be proved by putting $\eta = Y - E(Y|G^*)$, which implies $Y - E(Y|G^*) = \eta - E(\eta|G^*)$, and so

$$E\{|\eta - E\{\eta|G^*\}|^2 G^*\} = E\{\eta^2|G^*\} - E^2\{\eta|G^*\} \leq E\{\eta^2|G^*\}.$$

The result follows by taking expectations of both sides of the inequality. This lemma can then be applied to our problem to show

$$||E_{i+2m}X_i - X_i||_2 \leq ||E(X_i|F^{*i+2m}_{i-2m}) - X_i|| \leq \nu_{2m}$$

and so X_i is a mixingale with $\psi_m = 2\phi^{1-1/r}_{[m/2]} + \nu_{[m/2]}$ (where the square brackets denote "the greatest integer contained in") and $c_i = \max(||X_i||_r, 1)$.

These theorems provide the basis for the strong law of large numbers presented by White and Domowitz (1982). The arguments used are similar to those of Heijmans and Magnus (1983) in their proof of the consistency of the MLE. Before we can outline their proof, we require their lemma A.1 which establishes conditions for the uniform convergence of $|T^{\gamma-1}\sum_{t=1}^{T} Z_t|$. This follows directly from the fact that mixing processes are mixingales, and is important in the proof because it provides bounds on functions of the data and parameters.

The lemma is as follows:

Let $\{Z_t\}$ have zero mean and suppose $\phi(m)$ is of size $r/(2r-1)$ (or $\alpha(m)$ of size $r/(r-1)$).

a) If there exists $\gamma \geq 0$ and p such that $r < p \leq 2r$ for which $\sum_{t=1}^{\infty}(E|Z_t|^p)^{1/r}T^{(\gamma-1)p/r} < \infty$, then

$$T^{\gamma-1}\sum_{t=1}^{T} Z_t \xrightarrow{a.s.} 0$$

b) If there exists $\Delta < \infty$ such that $E|Z_t|^p < \Delta$ for all t, then (i) $T^{\gamma-1}\sum_{t=1}^{T} Z_t \xrightarrow{a.s.} 0$ for $0 \leq \gamma < 1-r/p \leq 1/2$ and (ii) there exists T depending only on Δ and ϵ

such that for all $T > T(\Delta,\varepsilon)$, $|T^{\gamma-1}\sum_{t=1}^{n} Z_t| < \varepsilon$ a.s.

Parts a) and b(i) follow directly from the mixingale convergence theorem discussed above, and b(ii) can be derived from the Chebyshev inequality using the mixingale bound and $k\ \Sigma\ c_i^2$ which is independent of the sample size, due to the bounding condition.

Using this result, we are now in a position to establish the strong law of large numbers for mixing processes presented by White and Domowitz (1982). Their theorem 2.3 states:

Let $q_t(Z_t,\theta)$ be measurable for each θ in H, a compact subset of R^p, and continuous on H, uniformly in t a.s.. Suppose there exist measurable dominating functions $d_t(z_t)$ such that $|q_t(Z_t,\theta)| \leq d_t(Z_t)$ for all θ in H, and for some $r \geq 1$ and $0 < \delta \leq r$, $E|d_t(Z_t)|^{r+\delta} \leq \Delta < \infty$ for all t. If either a) $\phi(m) = O(m^{-\lambda})$, $\lambda > r/(2r-1)$ or b) $\alpha(m) = O(m^{-\lambda})$ for $\lambda > r/(r-1)$, $r > 1$ then

i) $E(q_t(Z_t,\theta))$ is continuous on H in t.

ii) $|T^{\gamma-1}\sum_{t=1}^{T}[q_t(Z_t,\theta) - Eq_t(Z_t,\theta)]| \overset{a.s.}{\to} 0$ uniformly in θ, for $0 \leq \gamma < \delta/(r+\delta) \leq 1/2$.

The proof is based on similar techiniques to that of Heijmans and Magnus (1983). We need to establish an upper and lower bound on $q_t(Z_t,\theta)$, and then to show that $q_t(Z_t,\theta) - E(q_t(Z_t,\theta))$ is bounded by the original bounds minus their respective expectations. The uniform convergence arguments from White and Domowitz (1982)'s lemma given above are used to show summation of both bounds minus their expectation goes to zero a.s.. From which it follows that

$$|T^{\gamma-1}\sum_{t=1}^{T}[q_t(Z_t,\theta)-Eq_t(Z_t,\theta)]| \overset{a.s.}{\to} 0.$$

Proof: Part (i) follows the continuity of $q_t(Z_t,\theta)$ and the uniform integrability of $q_t(Z_t,\theta)$ due to the existence of $d_t(Z_t)$.

Part (ii) Using part (i) we can set $E(q_t(Z_t,\theta)) = 0$ without loss of generality. Let

$$\bar{q}_t(Z_t,\theta,\rho) = \sup\{q_t(Z,\xi):||\xi-\theta|| \leq \rho\},$$

$$\underline{q}_t(Z,\theta,\rho) = \inf\{q_t(Z,\xi):||\xi-\theta|| \leq \rho\},$$

both of which are measurable functions. From the continuity of $q_t(Z,\theta)$ on H and the bound on $E|d_t(Z_t)|^{r+\delta}$ it follows that

$$\lim\bar{q}_t(Z,\theta,\rho) = q_t(Z_t,\theta),$$

and

$$\lim\underline{q}_t(Z,\theta,\rho) = q_t(Z_t,\theta) \text{ uniformly in } \rho \text{ a.s..}$$

This implies $E|\bar{q}_t(Z,\theta)|^{r+\delta} \leq \Delta$

and $E|\underline{q}_t(Z,\theta)|^{r+\delta} \leq \Delta.$

Therefore as the expected values are bounded it follows by definition that

$$\lim E|\bar{q}_t(Z,\theta,\rho)| = \lim E|q_t(Z,\theta,\rho)| = 0 \text{ as } \rho \to 0.$$

For each $\theta \in H$ there must exist $\rho_n(\theta)$ so small that

$$-\varepsilon n^{-\gamma} < E(\underline{q}_t(Z_t,\theta,\rho)) \leq E(\bar{q}_t(Z_t,\theta,\rho_n(\theta)) < \varepsilon n^{-\gamma} \qquad (42).$$

Define $\xi(\theta,\rho) = \{\xi : ||\xi-\theta|| < \rho\}$. This forms an open cover of the compact set H, and so there must exist a finite subcover: $\theta_{n_1},\ldots,\theta_{ng_n} \in H$ for which $H \subset \bigcup_{i=1}^{g_n} \xi(\theta_{n_i},\rho_n(\theta_{n_i}))$. From the definitions of $\bar{q}_t$ and $\underline{q}_t$ for all θ in H it must follow that

$$\min_{1\leq i\leq g_n} n^{\gamma-1} \sum_{t=1}^{n} \underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i})) \leq n^{\gamma-1} \sum_{t=1}^{n} q_t(Z_t,\theta)$$

$$- \max_{1\leq i\leq g_n} n^{\gamma-1} \sum_{t=1}^{n} \bar{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i})).$$

From (42)

$$\min_{1\leq i\leq g_n} n^{\gamma-1} \sum_{t=1}^{n} \underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i})) - E(\underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i}))$$

$$< \min_{1\leq i\leq g_n} n^{\gamma-1} \sum_{t=1}^{n} \underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i}))+\varepsilon \leq n^{\gamma-1} \sum_{t=1}^{n} q_t(Z_t,\theta)+\varepsilon.$$

From the uniform domination and lemma A.1, we know that for every i there exists $T(\Delta,\varepsilon)$ such that for all $T \geq T(\Delta,\varepsilon)$ and almost every sequence $\{Z_t\}$,

$$-\varepsilon < T^{\gamma-1} \sum_{t=1}^{T} \{\underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i})) - E(\underline{q}_t(Z_t,\theta_{n_i},\rho_n(\theta_{n_i}))\}, \qquad (43)$$

where $\gamma < \delta/(r+\delta)$.

Therefore $-2\varepsilon < T^{\gamma-1} \sum_{t=1}^{T} q_t(Z_t,\theta)$ for $T > T(\Delta,\varepsilon)$.

Similarly we can show

$$T^{\gamma-1} \sum_{t=1}^{T} q_t(Z_t,\theta) < 2\varepsilon \text{ a.s.}.$$

Since the set F_n of sequences $\{Z_t\}$ such that (43) fails to hold for any i has measure zero and since $G_n = U_{j=1}^{n} F_j$ is an increasing sequence of sets of measure zero, $P(\overset{\infty}{\underset{k=1}{U}} G_n) = \lim P(G_n) = 0$, we have

$$|T^{\gamma-1} \sum_{t=1}^{T} q_t(Z_t,\theta) - E(q_t(Z_t,\theta))| \xrightarrow{a.s.} 0$$

as $T \to \infty$, uniformly in θ, for $0 \leq \gamma < \delta/(r+\delta)$.

Our analysis of the consistency of NLFIML can be generalised to dynamic models under two types of regularity condition. We can use martingale or mixingale convergence theorems to the required functions of the variables such as the score and hessian, and in this way implicitly restrict the underlying variables of the system. Alternatively we can make the explicit assumption that the underlying variables are mixing processes, and then use the mixingale convergence theorem. In the remainder of this chapter we concentrate on the latter approach, but it is important to note martingale strong law of large numbers and central limit theorems could have been used to generate the results.

7.4 Relationship between robustness of NLFIML and the reduced form

Mixing processes are therefore ergodic and so provide an answer to the central asymptotic problem. If we are prepared to assume our series to be of this form then we can use similar analysis to that undertaken in the static case to examine the robustness of NLFIML. Furthermore if we wish to use the conventional estimators in dynamic models, it is necessary to make this type of assumption. This applies equally to LS and ML and has to be implicitly made in the analysis of Jorgenson and Laffont (1974).

The conditions for the consistency of NLFIML are the same as before and we illustrate below that the arguments used in the static model carry through by considering a two equation nonlinear in variables example. To avoid the difficulty of verification of the second order conditions we need to extend Brown's (1983) analysis. In his exposition Brown (1983) deals with contemporaneous nonlinear in variables models of the form $Aq(y,x) = u$. However his analysis is more general than it at first appears because of the assumptions made. The error process is assumed only to be distributed independently of the exogenous variables x with mean zero. It is these properties of u that are used to generate the identification criteria. The point to note is that there is no assumption about the serial independence of u because Brown wanted to allow for the more general case in which the structure of u is unknown. This is why he deals with conditions relative to exogenous variables x, whereas Fisher (1966) used the serial independence assumption and assumed x to consist of predetermined

variables. There is no reason not to use the criteria for Fisher's model as his mistake, corrected by Brown, is concerned with the number of implied equations and not his original assumptions. However once we consider Brown's disturbances with predetermined variables we increase the potential number of transformations that may produce observationally equivalent equations. In that case more information is needed, maybe, from covariance restrictions instead of the independence assumption.

We now consider the conditions for consistency of NLFIML in the following model from Howrey and Kelejian (1971).

Let

$$y_{1t} = b_1 x_t + u_{1t}$$

$$y_{2t} = b_2 y_{1t-1} + b_3 \exp y_{1t} + u_{2t},$$

where $u_t = (u_{1t}, u_{2t})'$ are $IIN(0,\Omega)$. The reduced form for y_{2t} is given by

$$y_{2t} = b_2 y_{t-1} + b_3 \exp(b_1 x_t + u_{1t}) + u_{2t}.$$

The concentrated quasi log likelihood is

$$LLF^c = \text{const} - \frac{T}{2} \ln \left| T^{-1} \sum_{t=1}^{T} u_t u_t' \right|.$$

a) First order conditions for consistency: From the arguments used in the static case $\text{plim}\ T^{-1} \left. \frac{\partial LLF^c}{\partial b_1} \right|_{\theta_0} = 0$ provided the true distribution has mean zero. Now

$$\frac{\partial LLF^c}{\partial b_2} = -\frac{T}{2}A_T^{-1}\{m_{11}\frac{\partial m_{22}}{\partial b_2} - 2m_{12}\frac{\partial m_{12}}{\partial b_2}\}$$

$$= -\frac{T}{2}A_T^{-1}\{2m_{11}T^{-1}\sum_{t=1}^{T}u_{2t}(b_1x_{t-1}+u_{1t-1}) - 2m_{12}T^{-1}\sum_{t=1}^{T}u_{1t}(b_1x_{t-1}+u_{1t-1})\},$$

where we have let $A_T = |T^{-1}\sum_{t=1}^{T}u_t u_t'|$ and $m_{ij} = T^{-1}\sum_{t=1}^{T}u_{it}u_{jt}$. Therefore given the serial independence of u_t,

$$\text{plim}T^{-1}\frac{\partial LLF^c}{\partial b_2}\Big|_{\theta_0} = 0.$$

Finally we need to check $\partial LLF^c/\partial b_3|_{\theta_0}$. For this we assume, as before, that the true distribution is a mixture of normals.

Now

$$\frac{\partial LLF^c}{\partial b_3} = -TA_T^{-1}\{m_{11}T^{-1}\sum_{t=1}^{T}u_{2t}\exp(b_1x_t+u_{1t}) - m_{12}T^{-1}\sum_{t=1}^{T}u_{1t}\exp(b_1x_t+u_{1t})\},$$

and as $\text{plim}T^{-1}\sum_{t=1}^{T}u_{it}\exp b_1x_t = 0$, we need only consider $\text{plim}T^{-1}\sum_{t=1}^{T}u_{it}\exp u_{1t}$.

For this family of distributions,

$$E(u_{it}\exp u_{1t}) = \frac{\partial}{\partial s_i}mgf(s)\Big|_{\substack{s_1=1\\ s_2=0}} = \sigma_{1i}.$$

This implies $\text{plim}T^{-1}\frac{\partial LLF^c}{\partial b_3}\Big|_{\theta_0} = 0$ as well. The first order conditions for consistency are therefore satisfied and we now need to check the second order conditions.

To verify the second order conditions for consistency using the identification criteria in dynamic models we need

to assume that the underlying variables are mixing processes.

If we put $q(y,z) = [y_{1t}, y_{2t}, \exp y_{1t}, y_{1t-1}, x_t, 1]$.

where z_t are predetermined variables, then the Brown's notation:

$$q_1(y_1,z_1)' = [y_{1t}, \exp y_{1t}],$$

$$q_2(y_1,z_1,y_2,z_2)' = [y_{2t}],$$

$$q_3(z_3) = [y_{1t-1}, x_t].$$

It follows that $\text{rank}(A_2:A_3) = \text{rank}\begin{bmatrix} 0 & -b & 0 \\ 1 & 0 & -b_2 \end{bmatrix} = 2$ and so identification is assessed using the linear model criteria.

The coefficient matrix is

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & -b_1 & 0 \\ 0 & 1 & -b_3 & -b_2 & 0 & 0 \end{bmatrix} \text{ and}$$

$\text{rank}(A\Phi_1) = \text{rank}(A\Phi_2) = 1$, which is the number of equations minus one and so the system is identified. NLFIML is therefore consistent in this model if the true distribution is a member of the mixture of normals.

To undertake this analysis we need to be able to write down an explicit reduced form. If this is possible then we are able to list a set of true distributions for which NLFIML is consistent. However if we are unable to do this the conclusion has to be that NLFIML is only definitely consistent when the model is correctly specified or the

error distribution is the mixture of normals considered by Phillips (1982). It is not possible to list conditions on a true nonnormal error process that guarantee the consistency of NLFIML. Again this is a marked contrast to NL3SLS which is consistent in this model provided u_t is i.i.d. with mean zero.

7.5 Asymptotic normality of an NLFIML in dynamic models.

White and Domowitz (1982) also present the following central limit theorem for mixing processes. (White and Domowitz, 1982, theorem 2.6, p. 10)

Let $\{Z_t\}$ be a sequence of random variables satisfying

a) $E(Z_t) = 0$,

b) there exists V, finite and nonzero such that $E(S_a(T)^2 - V) \to 0$ as $T \to \infty$, uniformly in a, where

$$S_a(T) = T^{-1/2} \sum_{t=1+a}^{a+T} Z_t,$$

c) $E|Z_t|^{2r} \leq \Delta < \infty$ for all t and some $r > 1$. If either $\phi(m)$ or $\alpha(m)$ is of size $r/(r-1)$ then

$$T^{-1/2} V^{-1/2} \left(\sum_{t=1}^{T} Z_t \right) \overset{a}{\sim} N(0,1).$$

This theorem can be used to establish the asymptotic normality of the score as a random vector is only multivariate normal if any linear combination of its elements is univariate normal, and functions of mixing processes are themselves mixing.

If we let L_t be the conditional quasi-log likelihood of the observation in period t and $L = \sum_{t=1}^{T} L_t$, then by

definition $\lim T^{-1}E\partial L/\partial\alpha_i\big|_{\alpha_*}$ is zero. Therefore if we centre the conditional quasi score about its expectation, $Z_t = \partial L_t/\partial\alpha_i - E\partial L_t/\partial\alpha_i\big|_{\alpha_*}$, the central limit theorem can now be applied to deduce

$$T^{-1/2}\partial L/\partial\alpha_i\Big|_{\alpha_*} \overset{a}{\sim} N(0, \lim T^{-1}\Sigma E\frac{\partial L_t}{\partial\alpha_i}\cdot\frac{\partial L_t}{\partial\alpha_i'}\Big|_{\alpha=\alpha_*} - \lim T^{-1}\Sigma E\frac{\partial L_t}{\partial\alpha_i}\Big|_{\alpha_*}\cdot E\frac{\partial L_t}{\partial\alpha_i'}\Big|_{\alpha_*}).$$

Using the mean value theorem applied to the score, as in the static model, it follows that

$$\sqrt{T}(\hat{\alpha}_T-\alpha_*) \overset{a}{\sim} N(0, A_*^{-1}B_*A_*^{-1}),$$

where

$$A_* = \lim T^{-1}E\frac{\partial^2 L}{\partial\alpha\partial\alpha'}\Big|_{\alpha=\alpha_*},$$

$$B_* = \lim T^{-1}E\Sigma\frac{\partial L_t}{\partial\alpha}\cdot\frac{\partial L_t}{\partial\alpha'}\Big|_{\alpha=\alpha_*} - \lim T^{-1}\Sigma E\frac{\partial L_t}{\partial\alpha}\Big|_{\alpha_*}E\frac{\partial L_t}{\partial\alpha'}\Big|_{\alpha_*}.$$

If we let A_T and B_T be the sample analogues evaluated at $\hat{\alpha}_T$ of A_*, B_* then as before

$$A_T^{-1}B_TA_T^{-1} - A_*^{-1}B_*A_*^{-1} \overset{a}{\sim} A_*^{-1}D_*A_*^{-1},$$

where D_* is the positive semi definite matrix

$$D_* = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E\frac{\partial L_t}{\partial\alpha_i}\Big|_{\alpha_*}\cdot E\frac{\partial L_t}{\partial\alpha_j}\Big|_{\alpha_*}.$$

The problem is that we cannot estimate D_* without knowledge of the true distribution. For our asymptotic tests to be

valid we must have a consistent estimator of α, as for instance would be the case for the nonlinear regression model. Otherwise we can only conduct what White (1983) termed "conservative inference" using conventional procedures in misspecified models.

This analysis has required the assumption that $E(T^{-1/2} \sum_{t=a+1}^{a+T} Z_t)^2 - V \to 0$ for some finite and nonzero V. Convergence of the variance of the process to a constant regardless of its index in time, a, is a limitation on the heterogeneity that can be covered by our model. Further White and Domowitz (1982) argue that the requirement $E(S_a(T)^2) - V_n \to 0$, uniformly in a, implies $V_n = V$. This again requires the covariance to converge to a value independent of the index a. White and Domowitz (1982) hypothesise that relaxing this condition may result in nonnormal limiting distributions, but Basawa, Feigin and Heyde (1976) and Hall and Heyde (1981) have shown that this need not be the case.

Following Hall and Heyde (1981) we consider a univariate one parameter model. They note that not all the results can be generalised, but this framework is sufficient to show that this alternative approach encounters the same problems.

Consider a sample $X_1, X_2, .., X_n$ of consecutive observations from some stochastic process whose distribution depends on a single parameter θ H. Hall and Heyde (1981) are concerned with the correctly specified case, and so, we first outline the assumptions that generate their result under these circumstances, as they provide a guide to those necessary for an extension to the misspecified case. (As it

happens, we also need to know potentially all the higher order moments of the true distribution for this to be possible). Let $L_n(\theta)$ be the likelihood associated with $X_1, X_2, \ldots, X_n$ and assume it to be twice differentiable with the expected value of the hessian being finite for each n. Denote the σ-field generated by $X_1, \ldots, X_k$ ($k \geq 1$) by F_k (F_0 = trivial σ-field).

Put

$$\frac{d\log L_n(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{d}{d\theta}[\log L_i(\theta) - \log L_{i-1}(\theta)] = \Sigma u_i(\theta),$$

and $E(u_i(\theta)|F_{i-1}) = 0$, so that $\{d\log L_n(\theta)/d\theta, F_n\}$ is a square integrable martingale. Also let

$$I_n(\theta) = \sum_{i=1}^{n} E_\theta(u_i^2(\theta)|F_{i-1}),$$

and

$$J_n(\theta) = \sum_{i=1}^{n} v_i(\theta) = \sum_{i=1}^{n} du_i(\theta)/d\theta.$$

The quantity $I_n(\theta)$ represents the conditional information and clearly varies over time. If we use this non constant normalisation of the MLE, $\hat{\theta}_n$, about the true value, instead of the constant V in the mixing CLT, then it can be shown that

$$I_n^{1/2}(\theta)(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1),$$

under the following assumptions:

i) $I_n(\theta) \overset{a.s.}{\to} \infty$, and so information is continually accruing.

ii) $I_n(\theta)/EI_n(\theta) \overset{p}{\to} n^2(\theta)$, some positive r.v. and $J_n(\theta)/I_n(\theta) \overset{p}{\to} -1$ as $n \to \infty$, the convergences being uniform.

iii) There exists some $\delta > 0$ such that

$$|\hat{\theta}_n - \theta| \leq \delta/EI_n(\theta)^{1/2}.$$

The result follows from a mean value expansion of the summations of the conditional scores in a similar fashion to the other CLT theorems presented above. The crucial point is that for models of this generality, we require random normalisation to induce the desired behavior on the MLE.

The problem in the misspecified case is two fold. Firstly, as remarked earlier, the score is not a martingale sequence. This can be overcome by centering $u_i(\theta)$ about its conditional expectation, and then, from $n^{-1} \lim EL_n/d\theta|_{\theta_*} = 0$, we have

$$c_n \sum_{i=1}^{n} u_u(\theta) \overset{d}{\to} N(0,1),$$

where $c_n = n^{-1} \sum_{i=1}^{n} E(u_i - E(u_i|F_{i-1})|F_{i-1})^2$. The problem is clearly going to be that when the model is misspecified we do not know this expectation. Therefore the extension of the theory to cover situations where the normalising factor of the summation is nonconstant still does not solve the problem of inference based on the QMLE.

It is worth noting that the situation is much easier to handle when we use the nonlinear regression model with

lagged dependent variables as regressors. In this case the QMLE is consistent, and so we can consistently estimate $A_*^{-1}B_*A_*^{-1}$ by its sample analogue. Heijmans and Magnus (1983b) present a proof of the asymptotic normality of the MLE under normality for this model, when it is correctly specified, using vector martingale arguments. Consistency of the QMLE is going to guarantee that their arguments can be generalised to produce similar results to those above. White and Domowitz (1983) present a series of specification tests for the regression model under the mixing assumptions, and of course these avoid the problems of the conventional tests in more complicated nonlinear dynamic models.

Therefore as White and Domowitz (1982) observe we can construct a complete asymptotic theory of inference for dynamic models on the basis of these assumptions. More correctly, given the conservative inference problems with our tests based on NLFIML, we can construct as much of a practically useful asymptotic theory as in the static model. The obvious question to turn to now is: are economic series mixing processes?

7.6 Verification and Suitability of the assumption that series are mixing processes.

Our conclusions on the asymptotic properties of NLFIML in dynamic models rely on the underlying variables being mixing processes. Not all series satisfy these requirements and so before we use the theory for economic modelling it is desirable to assess whether economic data obey these behavioral restrictions. White and Domowitz (1982) argue that

"although particular theoretical models can be demonstrated to yield ergodic or mixing processes, it is not possible to verify from a finite sample that a particular process is ergodic or mixing. Thus we adopt mixing as an operating assumption for economic processes on the basis of plausibility and convenience..." (p.5). There are clearly two issues at stake here. Firstly whether we can verify the mixing conditions from sample evidence and secondly whether we can verify them for a theoretical model.

As White and Domowitz (1982) observe it is not possible to use sample data to verify mixing assumptions. Quite simply they refer to limiting behaviour which cannot be assessed from a finite sample. We also face the problem that in the absence of information about the parameters, we would require a law of large numbers to apply for the substitution of parameter estimates to be valid.

White and Domowitz imply that the verification of the assumptions for theoretical models is quite straightforward. This does not appear to be the case as we demonstrate below for a simple nonlinear stationary process using the work of Jones (1976).

First consider the conditions of White and Domowitz (1982)'s theorem 2.3. These are

i) $q_t(Z_t,\theta)$ must be measurable for each θ in H,

ii) H must be compact,

iii) we require dominating functions $d_t(Z_t)$ to exist for all θ,

iv) moment restrictions: for some $r \geq 1$ and $0 < \delta \leq r$ $E|d_t(Z_t)|^{r+\delta} \leq \Delta < \infty$, for all t,

v) $\phi(m) = O(m^{-\lambda})$ for $\lambda > r/(2r-1)$,

or

$$\alpha(m) = O(m^{-\lambda}) \text{ for } \lambda > r/(r-1),\ r > 1.$$

Conditions (i) and (ii) are formalities that must hold for the model to be "well behaved". If the parameter space is not compact we can always make use of one point compactivisation as described earlier in the proof of the convergence of the QMLE to the KLIC minimising value.

The existence of a dominating function is not an unreasonable assumption given the physical real world constraints that exist on variables. This assumption, employed in the restriction of attention to Cesaro summable series, rules out models which make $Z_t = f(t)$, an increasing non converging sequence of random variables. For instance an AR(1) model with coefficient greater than one.

The mixing conditions themselves refer to probabilities and so to examine their validity we need the distribution of the process. To give an impression of the problems involved we outline some results due to Jones (1976) on the properties of nonlinear stationary Markov processes. Rosenblatt (1971) gives conditions for the ergodicity of stationary Markov process. Although these resemble mixing conditions in as much as they depend on probabilities, they are less stringent. The assumption of stationarity does not deliver the mixing result.

Consider the model,

$$X_{n+1} = \lambda(X_n) + Z_{n+1},\ n = (\ldots -1,0,1..),$$

where $\lambda(\cdot)$ is a fixed real function of a real argument and

Z_n is a sequence of i.i.d. random variables. The input series $\{Z_n\}$ has distribution F_Z. Typically this is our model specification without consideration of the implications for the distribution of X_{n+i}. The mixing conditions depend on this implied distribution, and so we must solve for the density of X_{n+i} as a function of $\lambda(\cdot)$ and F_Z.

If the autoregression function $\lambda(\cdot)$, is continuous everywhere then a sufficient condition for stationarity is the existence of constants ε, $\alpha > 0$ such that

$$E\{|\lambda(x)+Z|-|x|\} \leq -\varepsilon \quad (|x| > \alpha).$$

The distribution functions $F_{X,m}(\cdot;x_0)$,$(m\geq 1)$, of X_m conditional on a value $X_0 = x_0$ are given by

$$F_{X,1}(x;x_0) = F_Z(x-\lambda(x_0)),$$

$$F_{X,m}(x;x_0) = \int F_Z(x-\lambda(y))dF_{X,m-1}(y;x_0), \quad m = 2,3\ldots .$$

In most cases these equations can only be solved by numerical integration, the solution to which gives no idea of the properties of similar series. Jones (1976) instead considers the properties of the model

$$X_{n+1}(\beta) = a + bX_n(\beta) + \beta[\lambda\{X_n(\beta)\}-bX_n(\beta)-a],$$

as an expansion about the linear process for which $\beta = 0$. This leads to power series expansions in the parameter β about the known solution for the process $\{X_n(0)\}$.

For this part of his analysis Jones considers models with b = 0, and so putting

$$Y_n(\beta) = \beta[\lambda\{X_n(\beta)\}-a] \quad (n = ..-1,0,1,..),$$

and $Z_n^* = Z_n + a$, $\Lambda(x) = \lambda(x)-a$,

we have

$$X_{n+1}(\beta) = Y_n(\beta) + Z_{n+1}^*$$

$$Y_{n+1}(\beta) = \beta\Lambda\{Y_n(\beta)+Z_{n+1}^*\}.$$

It follows that Z_{n+1}^* must be independent of $(Y_n(\beta),Y_{n-1}(\beta),..)$. The approach Jones takes is to first find power expansions for the characteristic functions of the conditional distribution and to then find its implied Fourier inverse, the conditional density function.

The characteristic function (c.f.), $\phi_{Y,0}$, of $Y_n(\beta)$ given that $Y_n(\beta) = y$ is

$$\phi_{Y,0}(s;Y;\beta) = e^{isy} = L_{0,0}(s,y), \text{ say,}$$

because given information up to period n, Y_n can only take one value and is thus degenerate.

The c.f. of $Y_{n+1}(\beta)$ given that $Y_n(\beta) = y$ is

$$\phi_{Y,1}(s;y;\beta) = \int e^{is\beta\Lambda(y+z)} dF_{Z*}(z).$$

Expanding the exponential gives

$$\phi_{Y,1}(s;y;\beta) = \int \sum_{j=0}^{\infty} \frac{\{is\beta\Lambda(y+z)\}^j}{j!} dF_{Z^*}(z)$$

$$= \sum_{j=0}^{\infty} \beta^j (is)^j \int \frac{\Lambda^j(y+z)}{j!} dF_{Z^*}(z)$$

$$= \sum_{j=0}^{\infty} \beta^j L_{j,1}(s,y), \text{ say,}$$

where $L_{j,1}(s,y) = (is)^j p_j(y) = (is)^j \int \frac{\Lambda^j(y+z)}{j!} dF_{Z^*}(z)$, $j \geq 0$.

For $k \geq 0$ define $L_{j,1}^{(k)}(s,y) = \partial^k L_{j,1}(s,y)/\partial y^k$, then

$$\phi_{Y,2}(s;y;\beta) = \int \Sigma \beta^j L_{j,1}\{s,\beta\Lambda(y+z)\} dF_{Z^*}(z)$$

$$= \int \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \beta^{j+k} L_{j,1}^{(k)}(s,0) \frac{\Lambda^k(y+z)}{k!} dF_{Z^*}(z),$$

$$= \sum_{j=0}^{\infty} \beta^j \sum_{q=0}^{j} L_{j=q,1}^{(q)}(s,0) p_q(y),$$

$$= \sum_{j=0}^{\infty} \beta^j L_{j,2}(s,y).$$

Jones uses similar arguments to show that the characteristic function of $Y_{n+m+1}(\beta)$ given that $Y_n(\beta) = y$, is

$$\phi_{Y,m+1}(s;y;\beta) = \sum_{j=0}^{\infty} \beta^j L_{j,m+1}(s,y),$$

where

$$L_{j,m+1}(s,y) = \begin{cases} (is)^j p_j(y) & m = 0 \\ \sum_{q=0}^{j-m} L_{j-q,m}^{(q)}(s,0) p_q(y) & (j > m \geq 1) \\ L_{j,m}(s,0) = L_{j,j}(s,0) & (m \geq j). \end{cases}$$

If we define

$$p_j^{(n)} = \frac{d^n}{dy^n}[p_j(y)]\Big|_{y=0} = \frac{d^n}{dy^n}\{\int \frac{\Lambda^j(y+z)}{j!} dF_{Z^*}(z)\}\Big|_{y=0},$$

then it can be shown that

$$L_{N,m}(s,y) = \sum_{n=1}^{m-1} \Sigma(is)^{r_1} p_{r_1}^{(r_2)} p_{r_2}^{(r_3)} \cdots p_{r_{n-1}}^{(r_n)} p_{r_n}^{(0)}$$

$$+ \Sigma(is)^{r_1} p_{r_1}^{(r_2)} \cdots p_{r_{m-1}}^{(r_m)} p_{r_m}(y), \quad (1 \le m \le N),$$

where the first double summation is over all sets of $n \le m-1$ integers $r_1,..r_n$, $(r_i \ge 1)$, satisfying $r_1+r_2+\cdots+r_n = N$, and the second summation is over all sets of m integers $r_1,\ldots,r_m$, $(r_i \ge 1)$, satisfying $r_1+..+r_m = N$.

Let $T^{(N,j)}$ be given for $1 \le j \le N$ by

$$T^{(N,j)} = \begin{cases} p_j^{(0)} \\ \sum_{n=1}^{N-j} T^{(N-j,n)} p_j^{(n)}, \quad (1 \le j < N), \end{cases},$$

and so $T^{(N,j)} = \Sigma p_{r_1}^{(0)} p_{r_2}^{(r_1)} \cdots p_{r_k}^{r_{k-1}}$ where the summation is over all sets of k integers $(k \le N-j+1)$, $r_1,\ldots,r_k$, $(r_i \ge 1)$, satisfying $r_1+\cdots+r_k = N$ and $r_k = j$.

Define $S_m^{(N,j)}(y)$ for $(1 \le j \le N-m+1; m \ge 1)$ by

$$S_1^{(N,j)}(y) = \begin{cases} p_N(y) & (j=N) \\ 0 & (1 \le j < N) \end{cases},$$

and

$$S_m^{(N,j)}(y) = \sum_{n=1}^{N-j-m+2} S_{m-1}^{(N-j,n)}(y) p_j^{(n)}, \quad (1 \le j \le N-m+1;\ m=2,3,..),$$

then

$$S_m^{(N,j)}(y) = \Sigma p_{r_1}^{(r_2)} p_{r_2}^{(r_3)} \cdots p_{r_{m-1}}^{(r_m)} p_{r_m}(y), \quad (1 \leq j \leq N-m+1;\ m \geq 2).$$

Therefore we can write

$$L_{N,m}(s,y) = \sum_{j=N-m+2}^{N} (is)^j T^{(N,j)} + \sum_{j=1}^{N-m+1} (is)^j S_m^{(N,j)}(y), \quad 1 \leq m \leq N,$$

and

$$L_{N,m}(s,y) = \sum_{j=1}^{N} (is)^j T^{(N,j)}, \qquad (m > N \geq 1).$$

Finally, the characteristic function of $Y_{n+m}(\beta)$ given that $Y_n(\beta) = y$ is

$$\phi_{Y,m} = 1 + \sum_{j=1}^{\infty} \sum_{N=j}^{j+m-2} \beta^N (is)^j T^{(N,j)} + \sum_{j=1}^{\infty} \sum_{N=j+m-1}^{\infty} \beta^N (is)^j S_m^{(N,j)}(y).$$

(This relies on $p(\cdot)$ being continuously differentiable and the distribution of $Y_{n+m}(\beta)$ given $Y_n(\beta)$ having moments of all orders.)

We were originally concerned with the distribution of $X_n(\beta)$. Its characteristic function is the product of those of the independent random variables Z^*_{n+1} and $Y_n(\beta)$ and so

$$\phi_X(s;\beta) = \phi_{Z^*}(s)\phi_Y(s;\beta)$$

where $\phi_{Z^*}(s)$ is the c.f. of $Z_n + a$.

Therefore the c.f. of X_{n+1} given $X_n = x$ is

$$\phi_{X,1}(s;x;\beta) = \phi_{Z*}(s)\phi_{Y,0}(s;\beta\Lambda(x);\beta)$$

$$= \phi_{Z*}(s)\exp[is\beta\{\lambda(x)-a\}],$$

and

$$\phi_{X,m}(s;x;\beta) = \phi_{Z*}(s)\phi_{Y,m-1}(s;\beta\Lambda(x);\beta).$$

Jones (1976) notes that these summations only converge under conditions which are too restrictive for the result to be of practical use.

If the input distribution has a continuously differentiable density then we can invert the expressions for the characteristic equation for its Fourier transform, the density of X.

As the common density of $\{Z_n\}$ is $f_Z(z)$, that of $\{Z_n^*\}$ is $f_{Z*}(x) = f_Z(z-a)$. This gives the density of X as

$$f_X(x;\beta) = f_{Z*}(x) + \sum_{j=1}^{\infty}\sum_{N=j}^{\infty}\beta^N(-1)^j f_{Z*}^{(j)}(x)T^{(N,j)},$$

where

$$f_{Z*}^{(j)}(x) = \partial^j f_{Z*}(x)/\partial x^j.$$

We can truncate this at a fixed power of N,N^* say, and writing

$$h_j^{(N*)} = \begin{cases} \sum_{N=j}^{N*}\beta^N T(N,j) & (j\geq 1) \\ 1 & (j=0) \end{cases}$$

gives

$$f_X(x';\beta) \sim f_{Z^*}(x') + \sum_{j=1}^{N^*} h_j^{(N^*)}(-1)^j f_{Z^*}^{(j)}(x').$$

Similarly the conditional density of $X_{n+m}(\beta)$ given that $X_n(\beta) = x'$ is, for $m \geq 2$,

$$f_{X,m}(x;x';\beta) \sim f_{Z^*}(x) + \sum_{j=1}^{N^*} g_{j,m-1}^{(N^*)}\{\beta\Lambda(x')\}(-1)^j f_{Z^*}^{j}(x),$$

where

$$g_{j,m}^{(N^*)}(y) = \sum_{N=j}^{\min(N^*,j+m-2)} \beta^N T(N,j) + \sum_{N=j+m-1}^{N^*} \beta^N S_m(N,j)(y), \quad (j,m \geq 1).$$

All these expressions rely on being able to calculate $p_j^{(n)}$ and $p_j(y)$.

Recall $p_j^{(n)} = \frac{\partial^n}{\partial y^n}\{p_j(y)\}\Big|_{y=0}$,

$$p_j(y) = \int \frac{\Lambda^j(y+z)}{j!} dF_{Z^*}(z)$$

$$= \int \frac{\{\lambda(y+z+a)-a\}^j}{j!} dF_Z(z).$$

Jones (1976) considers these quantities for some simple functional forms with a normal input distribution. His analysis suggests that the development of analytically tractable solutions depends on the proportionality of the functional form to the probability density function.

For instance,

$$\lambda(x) = \lambda\exp\{-\frac{1}{2}w^2(x-d)^2\}, \qquad (-\infty<x<\infty),$$

then for $j \geq 1$ this implies

$$p_j[\lambda(x)](y) = \frac{\lambda^j}{j!}\frac{1}{\sqrt{2\pi}\sigma}\int \exp\{-\frac{1}{2}jw^2(z-d)^2 - \frac{1}{2\sigma^2}(z-a-y)^2\}dz$$

$$= \frac{\lambda^j}{j!}\frac{1}{(jw^2\sigma^2+1)^{1/2}}\exp\{\frac{jw^2(y+a-d)^2}{2(jw^2\sigma^2+1)}\}.$$

The derivatives $p_j^{(n)}$ can be calculated from p_j using the recurrence relations for Hermite polynomials.

Jones (1976) presents results for the joint density of $X_{n+m}(\beta)$ and $X_n(\beta)$ which involve similar types of calculations. These of course would be needed to verify if a process was strongly mixing. Rosenblatt (1971, p. 195) shows that the above calculations can be avoided if $\lambda(\cdot)$ has a particular form. He shows that when X_n has the same distribution as a nonlinear function of the input $f(Z_n, Z_{n-1}, ..)$, and so is purely non deterministic, then its stationarity implies it is strongly mixing. In which case if it is of the correct size White and Domowitz's (1982) results can be applied. However in general X_n does not have such a representation, and the assumption of stationarity is undesirable.

The verification of mixing conditions for theoretical models is by no means straightforward, especially as we have only considered the calculations for univariate processes. Jones (1976) outlines the extension of the analysis to vector prcesses which involves similar but more complicated expressions, as would be anticipated. If we are to proceed we clearly need to know the input distribution and correct functional form. The analysis of the properties of the series when there is misspecification, or in other words adjusting the relationship between $\lambda(\cdot)$ and $F_z(\cdot)$, entails nontrivial calculations. Strictly verification of the

mixing assumptions should procede any asymptotic analysis on their basis, but this is clearly not feasible for various combinations of $\lambda(\cdot)$ and $F_Z(\cdot)$.

The practical conclusion from this work seems to be that we must either decide to assume or not to assume the variables are mixing. As pointed out by White and Domowitz, (1983) their adoption has certain implications whicy may or may not be acceptable. For instance the covariance of mixing processes decays to zero as the distance between the observations increases and at a rate slower than that of ARMA models. The assumptions therefore allow the series to have more memory than conventional linear models. Rootzen (1974) has shown that if the process $\{Y_n\}$ is ϕ-mixing with limiting distribution G, then the range of $\{y_n(w)\}$ is dense in the support of G for almost all w. From the White and Domowitz (1983) central limit theorem we know

$$T^{-1/2}V^{-1/2} \sum_{t=a+1}^{a+T} Z_t \xrightarrow{d} N(0,1).$$

The support of the limiting distribution is the possible set of values that can be taken by a N(0,1) r.v. - in other words the real line. Recall that the set A is dense in B if B is a subset of the minimal closed set containing A. Therefore Rootzen (1974) has shown that the minimal closed set containing the range of Y_n contains the real line. For our purposes $Y_n(w) = n^{-1/2}V^{-1/2} \sum_{t=a+1}^{a+n} Z_t$, for n = 1,2,.., and so the result shows that mixing places no undesirable constraint on the possible values that can be taken by the variable Z_t, as it can take any value on the real line. Note that we can restrict the range of a mixing process by

allocating values of w that yield "improper" $Y_n(w)$ zero probability. Mixing does not therefore appear to be a particularly burdensome assumption.

Our analysis has shown that we can extend the results about the robustness of NLFIML and inference based on it from the static to the dynamic model under two types of regularity condition. Firstly we can implicitly bound the underlying variables by assuming various functions of them are martingales or mixingales. Alternatively we can assume the underlying variables to be mixing processess which, subject to size conditions, behave as mixingales. As functions of mixing processes are themselves mixing, we can then apply mixingale laws of large numbers to the appropriate functions. The analysis in section 6 suggests that the assumption that economic series are mixing is not particularly restrictive and can therefore be used as a basis of an asymptotic theory for nonlinear dynamic econometric models.

allocating values of w that yield "improper" $Y_n(w)$ zero probability. Mixing does not therefore appear to be a particularly burdensome assumption.

Our analysis has shown that we can extend the results about the robustness of NLFIML and inference based on it from the static to the dynamic model under two types of regularity condition. Firstly we can implicitly bound the underlying variables by assuming various functions of them are martingales or mixingales. Alternatively we can assume the underlying variables to be mixing processess which, subject to size conditions, behave as mixingales. As functions of mixing processes are themselves mixing, we can then apply mixingale laws of large numbers to the appropriate functions. The analysis in section 6 suggests that the assumption that economic series are mixing is not particularly restrictive and can therefore be used as a basis of an asymptotic theory for nonlinear dynamic econometric models.

8. THE INFORMATION MATRIX TEST AND THE EXPONENTIAL FAMILY.

8.1 Pseudo maximum likelihood estimators.

Gourieroux, Monfort and Trognon (1984a) consider the properties of MLE's in the nonlinear regression model. They argue that as the true distribution of the error process is unknown, the choice of assumed distribution should be one that ensures the resulting estimator has desireable statistical properties for a wide variety of true distributions. This leads them to discuss the idea of the pseudo MLE, which denotes the estimator derived by maximising what is acknowledged to be the wrong likelihood. In our earlier work, we have used the White (1982) terminology and refered to this estimator as a quasi MLE.

The model we consider here is of the following form

$$y_t = f(x_t,\theta) + u_t,$$

where y_t and u_t are G dimensional vectors and $f(x_t,\theta)$ represents the conditonal expectation of y_t. GMT (1984a) assume that the Cesaro summability conditions detailed by Burguette, Gallant and Souza (1983) (see section 3.1 above) are satisfied. GMT (1984a) establish the consistency and asymptotic normality of the PMLE for the situations in which a) we require estimates of the parameters of the mean and assume the distribution of u_t is a member of the linear exponential family, b) we require estimates of the parameters of the variance as well and the assumed distribution is a member of the quadratic exponential family. They further show that in each case it is necessary that the assumed distribution be a member of that particular

exponential family for the strong consistency of the PMLE for any true mean zero distribution of u_t. The arguments for restricting attention to such distributions are therefore quite strong.

GMT (1984a) show that

$$\sqrt{T}(\hat{\theta}_T-\theta_0) \xrightarrow{d} N(0,J^{-1}IJ^{-1}),$$

where

$$J = \lim_{T\to\infty} E_x E_0 T^{-1} \Sigma \frac{\partial^2 LLf_t}{\partial\theta\partial\theta'},$$

and

$$I = \lim_{T\to\infty} E_x E_0 T^{-1} \Sigma \frac{\partial LLF_t}{\partial\theta} \cdot \frac{\partial LLF_t}{\partial\theta'}.$$

The E_0 and E_x denoting expectations taken with respect to the true error process and regressors respectively. Therefore whilst the first order properties of the PMLE do not depend on the true distribution, the covariance matrix clearly does. In the absence of knowledge about the true distribution, it may be of interest to construct a specification test of the adequacy of the pseudo distribution as an approximation. It is argued below that the information matrix test suggested by White (1982) is a natural test of such a hypothesis. The analysis presented here examines the IMT and higher order likelihood derivative tests (see Chesher, 1983) for the linear and quadratic exponential families. They provide an alternative test of distribution to the goodness of fit type tests, although

each is based on different properties of the true distribution, and could similarly be used outside the regression context.

8.2 Linear exponential family.

The linear exponential family is a class of probability measures on R^G indexed by a parameter $m \in M \subset R^G$ that satisfy:

a) every element of the family has a density function with respect to a given measure $\nu(du)$

b) this density function can be written as

$$L(u,m) = \exp\{A(m)+B(u)+C(m)u\}: \quad u \in R^G,$$

where A(m), B(u) are scalars and C(m) is G-dimensional row vector.

c) m is the mean of the distribution whose density is L(u,m).

The reason for the necessity of using a member of this family to ensure strongly consistent estimators follows from the known properties of the true distribution. If all that we know is that its mean is m_0, and we require our estimator to be strongly consistent for all distributions with this property, then the quasi score must be a linear function of (u-m), if it is to have zero expectation under the true distribution. This is equivalent to requiring the log likelihood function to be linear in (u-m) and so the density to be of the form above. [Note that for this family $-\partial A/\partial m = (\partial C/\partial m)m$, see GMT 1984a].

The information matrix test is based on the fact that

if the model is correctly specified then

$$\lim T^{-1} \sum_t \{E\frac{\partial^2 LLF_t}{\partial\theta\partial\theta'} + E\frac{\partial LLF_t}{\partial\theta}\frac{\partial LLF_t}{\partial\theta'}\} = 0.$$

More precisely, this identity holds if the model is correctly specified to the second order. The error distribution may be misspecified, but its first k moments may coincide with the true distribution and so the mistake may be inconsequential for our inference.

In the linear exponential family,

$$\ln L(u,m) = A(m) + B(u) + C(m)u,$$

and so

$$\frac{\partial \ln L}{\partial m} = \frac{\partial A}{\partial m} + \frac{\partial C}{\partial m}u = \frac{\partial C}{\partial m}(u-m).$$

Therefore,

$$\frac{\partial^2 \ln L}{\partial m \partial m'} = \sum_{g=1}^{G} \frac{\partial^2 C_g}{\partial m \partial m'}(u-m) - \frac{\partial C}{\partial m},$$

where C_g is the g^{th} element of C. The information matrix test is therefore based on a comparison of zero with

$$\sum_{g=1}^{G} \frac{\partial^2 C_g}{\partial m \partial m'}(u-m) - \frac{\partial C}{\partial m} + \frac{\partial C}{\partial m}(u-m)(u-m)'\frac{\partial C}{\partial m'}.$$

GMT (1984a) show that $E(u-m)(u-m)' = [\partial C/\partial m]^{-1}$, and so the test examines whether the relationship between mean and variance implicit in the choice of assumed distribution is supported by the data. To illustrate this result for a

specific example we consider the Poisson models examined by GMT (1984b).

8.3 Poisson models

In this framework the endogenous variable is discrete, and may represent the frequency of a particular event in a fixed period of time. We consider the case where there is specification error, may be due to an omitted variable, and so

$$y_i \sim \text{Poisson}(\lambda_i) \text{ and } \lambda_i = \exp x_i'b + \varepsilon_i.$$

To obtain the conditional distribution of y_i, $L(y_i|x_i)$, it is necessary to integrate over ε_i, so

$$L(y_i|x_i) = \int \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} g(\varepsilon_i)d\varepsilon_i,$$

where $g(\varepsilon_i)$ is the p.d.f. of ε_i. This in general does not have a convenient form, although if $E(\exp\varepsilon_i) = 1$ and $V(\exp\varepsilon_i) = \eta^2$ then we know the first two moments of y_i:

$$E(y_i|x_i) = \exp x_i'b,$$

$$\text{var}(y_i|x_i) = \exp x_i'b + \eta^2 \exp 2x_i'b.$$

In spite of our ignorance of the true distribution, we can obtain consistent and asymptotically normal estimators of the parameters by using using PML methods provided y_i is assumed distributed linear exponential. GMT (1984b) show that the covariance of the PMLE is $J^{-1}IJ^{-1}$, where

$$J = \lim T^{-1} \sum_t E_x\left(\frac{\partial f}{\partial b}\Sigma_{\hat{0}}^{-1}\frac{\partial f'}{\partial b}\right),$$

$$I = \lim T^{-1} \sum_t E_x\left(\frac{\partial f}{\partial b}\Sigma_0^{-1}\Omega_0\Sigma_0^{-1}\frac{\partial f'}{\partial b}\right),$$

and $f(x_i'b) = \exp x_i'b$. In this context $\Omega_0 = \text{var}(y_i|x_i)$ in the true model (i.e. $\Omega_0 = \exp x_i'b + n^2\exp 2x_i'b$), and Σ_0 is the variance of the chosen linear exponential family.

GMT (1984b) point out that a specification test can be based on a comparison of J and I. This amounts to a test of whether the covariance matrix of the assumed and actual distributions are equal, and so can be regarded as a test of whether the estimated covariance of y_i in the assumed distribution is insignificantly different from the sample variance. Therefore rather than applying the information matrix test principle to J and I as given by GMT (1984b), it would appear computationally easier to apply the information matrix test to the pseudo distribution. GMT (1984b) consider four examples of the linear exponential families, and we calculate the appropriate information matrix test for each model. The covariances are left to an appendix.

8.3.1 Poisson Distribution

If we assume $y_i \sim$ Poisson $(\exp x_i'b)$, the the log likelihood function for one observation is

$$LLF = \lambda_i + y_i \log \lambda_i - \log(y_i!),$$

and so

$$\frac{\partial LLF}{\partial b} = x_i(y_i - \exp x_i'b),$$

$$\frac{\partial^2 LLF}{\partial b \partial b'} = -x_i x_i' \exp x_i'b.$$

Therefore the IMT compares the following quantity with zero,

$$\Sigma x_i x_i' \{(y_i - \exp x_i'b)^2 - \exp x_i'b\}.$$

If y_i does indeed have a Poisson distribution then $E(y_i) = var(y_i)$, and the statistic checks whether the parameter estimates support this restriction.

Typical elements of the covariance of the indicator vector are calculated in the appendix. Chesher (1983) and Lancaster (1984) have shown that under H_0 the IMT can be calculated from the auxilliary regression of a constant on the indicator vector and score vector. If $x_{1t} = 1$ for all t, then we can use the IMT principle to derive a simple test of the equality of mean and variance, based on,

$$\Sigma(y_i - \lambda_i)^2 - \Sigma\lambda_i = \Sigma\nu_i,$$

which can be calculated as nR^2 from the regression of a constant on ν_i and $\partial LLF_i/\partial b$ and is distributed χ_1^2 if the model is correctly specified.

8.3.2. Normal distribution

If we assume $y_t \sim N(\exp x_i'b, 1)$ then

$$LLF_i = \text{const} - \frac{1}{2}(y_t - \exp x_i'b)^2,$$

and

$$\frac{\partial LLF_i}{\partial b} = x_i(y_i - \exp x_i'b)\exp x_i'b,$$

$$\frac{\partial^2 LLF_i}{\partial b \partial b'} = x_i x_i' \exp x_i'b(y_i - 2\exp x_i'b).$$

The IM test therefore compares

$$\Sigma x_i x_i' \exp x_i'b[(y_i - 2\exp x_i'b) + (y_i - \exp x_i'b)^2 \exp x_i'b],$$

with zero, and is a test of whether the variance is unity. If $x_{1t} = 1$, for all t, then an asymptotically χ_1^2 test of whether y_i has this distribution can be calculated by regressing a constant on $\partial LLF_i/\partial b$ and v_i where

$$v_i = \exp x_i'b[(y_i - 2\exp x_i'b) + (y_i - \exp x_i'b)^2 \exp x_i'b].$$

8.3.3 Gamma Distribution

If $y_t \sim$ Gamma with a degrees of freedom then the LLF is

$$LLF_i = \text{const} - x_i'b - y_i \exp(-x_i'b),$$

$$\frac{\partial LLF_i}{\partial b} = x_i \exp(-x_i'b)(y_i - \exp x_i'b),$$

$$\frac{\partial^2 LLF_i}{\partial b \partial b'} = -x_i x_i' y_i \exp(-x_i'b),$$

which implies

$$\Sigma\left[\frac{\partial^2 LLF_i}{\partial b \partial b'} + \frac{\partial LLF_i}{\partial b}\frac{\partial LLF_i}{\partial b'}\right] = \Sigma x_i x_i'(-y_i + (y_i - \exp x_i'b)^2 \exp(-x_i'b)).$$

The moment generating function of the gamma distribution is $(\lambda/(\lambda-t))^r$ where we have set $\lambda = a\exp(-x_i'b)$ and $r = a$. This implies the variance is $r/\lambda^2 = a^{-1}\exp 2x_i'b$, and so the IM test examines whether var $y_i = \lambda^{-1}Ey_i$ as required.

As in the Poisson case if $x_{1t} = 1$, for all t, we can calculate a x_1^2 test of whether the distribution is Gamma by calculating nR^2 from the auxilliary regression of a constant on $\partial LLF_i/\partial b$ plus v_i where

$$v_i = (y_i - \exp x_i'b)^2\exp(-x_i'b) - y_i.$$

8.3.4 Negative binomial distribution

In this case the p.d.f. of y_i is

$$\frac{\Gamma(a^{-1}+y_i)}{\Gamma(a^{-1})\Gamma(y_i+1)}(1+a\exp x_i'b)^{-(a^{-1}+y_i)}(a\exp x_i'b)^{y_i}.$$

For a given value of a, the LLF of the i^{th} observation is

$$LLF_i = y_i x_i'b - (a^{-1}+y_i)\log(1+a\exp x_i'b),$$

and so,

$$\frac{\partial LLF_i}{\partial b} = \frac{x_i(y_i - \exp x_i'b)}{1+a\exp x_i'b},$$

$$\frac{\partial^2 LLF_i}{\partial b \partial b'} = -\frac{x_i x_i'(y_i+a^{-1})a\exp x_i'b}{(1+a\exp x_i'b)^2}.$$

(Note this second derivative is different from the result stated in GMT, 1984b,p. 706).

The IMT compares zero with

$$\Sigma\{\frac{x_i x_i'}{(1+a\exp x_i'b)^2} \cdot [(y_i-\exp x_i'b)^2-(y_i+a^{-1})a\exp x_i'b]\}.$$

The mean and variance of this family are: $E(y_i) = \exp x_i'b$ and var $y_i = \exp x_i'b(1+a\exp x_i'b)$. In the same fashion as before if $x_{1t} = 1$, for all t, we could construct a x_1^2 test of whether the distribution is negative binomial by calculating nR^2 from the regression of a constant on v_i and $\partial LLF_i/\partial b$, where

$$v_i = \frac{[(y_i-\exp x_i'b)^2-(y_i+a^{-1})a\exp x_i'b]}{(1+a\exp x_i'b)}.$$

8.4. Specification tests based on higher order derivatives of the likelihood

For the purposes of PML estimation the information matrix test is all that is required. However using the theory in Chesher (1983) we can develop specification tests based on the higher derivatives of the likelihood.

By differentiating $\int_{-\infty}^{\infty} f(y,\theta)dy = 1$, $f(\cdot)$ being the p.d.f. of y, we obtain

$$\int_{-\infty}^{\infty} \frac{\partial^i f(\theta_0)}{\partial\theta^i} \cdot \frac{f(y_i\theta_0)}{f(y,\theta_0)} = 0, \quad i = 1,2,3,\ldots .$$

For i = 1, this gives $E\frac{\partial \log f}{\partial\theta}\Big|_{\theta_0} = 0$, and by differentiating this identity we obtain the information matrix identity, but we could similarly base a test on third order derivative

which would give the indicator vector

$$d_3 = F_3(\theta_0) + 3F_2(\theta_0)F_1(\theta_0) + F_1(\theta_0)^3$$

where $F_j(\theta_0) = \partial^j \log f(y,\theta_0)/\partial\theta^j$.

From the nature of the linear exponential family,

$$F_3(\theta) = \frac{\partial}{\partial m}\left\{\frac{\partial^2 \ln L(u,m)}{\partial m \partial m'}\right\}^c$$

$$= \sum_{g=1}^{G} [(u-m)' \otimes I]\partial\left\{\frac{\partial^2 C(m)_g}{\partial m \partial m'}\right\}^c / \partial m - \sum_{g=1}^{G} (I_m \otimes \frac{\partial^2 C_g}{\partial m \partial m'})$$

$$- \frac{\partial(\partial C/\partial m)^c}{\partial m}.$$

A test based on this indicator vector examines the relationship between first, second and third moments of the distributions. Similarly, as in Chesher (1983), we can construct a test based on the j^{th} derivative of $E\partial \log f/\partial\theta|_{\theta_0} = 0$ and from the nature of the linear exponential family this compares the presented relationship between the $(j+1)^{th}$, j^{th},...1^{st} moments of the distribution.

Clearly for the information matrix and higher order derivative tets to have this interpretation in terms of the central moments of the distribution (up to the $j+1^{th}$) we require L(u,m) to be a member of the linear exponential family. In general if the k^{th} order, for all k, derivative test is to examine the relationship between the first $r(k+1)^{th}$ moments of the distribution then the log likelihood must be a polynomial of order r in (u-m).

As an example of these higher order tests we consider the standard normal distribution. The LLF of one observation is given by

$$LLF = \text{const} - \frac{1}{2}(u-m)^2,$$

$$\frac{\partial LLF}{\partial m} = u-m,$$

$$\frac{\partial^2 LLF}{\partial m^2} = -1,$$

$$\frac{\partial^k LLF}{\partial m^k} = 0, \quad k > 2.$$

Therefore

$$d_3 = (u-m)^3 - 3(u-m),$$

and

$$d_4 = F_4(\theta_0)+4F_3F_1(\theta_0)+3F_2(\theta_0)^2+6F_2(\theta_0)F_1(\theta_0)^2+F_1(\theta_0)^4$$

$$= 3-6(u-m)^2 + (u-m)^4,$$

$$d_5 =$$

$$F_5(\theta_0)+F_4(\theta_0)F_1(\theta_0)+4[F_4(\theta_0)F_1(\theta_0)+F_3(\theta_0)F_2(\theta_0)+F_3(\theta_0)F_1(\theta_0)^2]$$

$$+ 3[2F_2(\theta_0)F_3(\theta_0)+F_2(\theta_0)^2F_1(\theta_0)]$$

$$+ 6[F_3(\theta_0)F_1(\theta_0) + F_2(\theta_0)^2+F_2(\theta_0)F_1(\theta_0)^2]F_1(\theta_0)$$

$$+ 4F_2(\theta_0)F_1(\theta_0)^3+F_1(\theta_0)^5$$

$$= 3(u-m) + 6 - 6(u-m)^2 - 4(u-m)^3 + (u-m)^5$$

$$= (u-m)^5 - 4(u-m)^3 - 6(u-m)^2 + 3(u-m) + 6.$$

The first two tests involving d_3 and d_4 are identical to the LM tests for normality based on the Edgeworth expansion derived by Keifer and Salmon (1983). These two tests are independent under H_0 and a χ^2_2 test of normality can be derived by regressing a constant on d_3, d_4 and $\partial LLF/\partial m$. However the d_5 indicator vector is not interpretable as such an LM test, as it is not the sample estimate of the fifth cumulant of the distribution.

8.5 Quadratic exponential family

If we require estimates of the first and second conditional moments of the distribution of y_t, which are strongly consistent and asymptotically normally distributed for all possible true distributions with the same first two moments, then GMT (1984a) show that it is necessary for the assumed distribution to be a member of the quadratic exponential family.

This family is characterised as follows:

a) every element of the family has a density function with respect to a given measure $\nu(du)$, which can be written as

$$L(u,m,\Sigma) = \exp\{A(m,\Sigma)+B(u)+C(m,\Sigma)u+u'D(m,\Sigma)u\},$$

where $m \in M \subset R^G$, Σ is a p.d. matrix, $A(m,\Sigma)$, $B(u)$ are scalars, $C(m,\Sigma)$ is a row vector of size G and $D(m,\Sigma)$ is a G x G matrix.

b) m is the mean and Σ the covariance of the distribution $L(u,m,\Sigma)$.

The necessity of the assumed distribution to come from

this family for the strong consistency of the estimator follows easily when we note,

$$\frac{\partial A(m,\Sigma)}{\partial m} = -\frac{\partial C}{\partial m}m + \begin{bmatrix} \sigma_1' \partial D_1/\partial m \\ \sigma_2' \partial D_2/\partial m \\ \vdots \\ \sigma_G' \partial D_G/\partial m \end{bmatrix},$$

where D_i is the i^{th} column of D. Also

$$\frac{\partial A(m,\Sigma)}{\partial(\Sigma^{-1})^c} = (m' \otimes I)\frac{\partial C}{\partial \Sigma^{-1}{}^c} + \begin{bmatrix} \sigma_1' \partial D_1/\partial \Sigma^{-1}{}^c \\ \vdots \\ \sigma_G' \partial D_1/\partial \Sigma^{-1}{}^c \end{bmatrix}.$$

The score vector can therefore be written as

$$\frac{\partial \ln L}{\partial m} = \frac{\partial C}{\partial m}(u-m) + \begin{bmatrix} \sum_j (u_1 u_j - \sigma_{1j}) \partial D_{1j}/\partial m \\ \vdots \\ \sum_j (u_G u_j - \sigma_{Gj}) \partial D_{Gj}/\partial m \end{bmatrix},$$

$$\frac{\partial \ln L}{\partial \Sigma^{-1}{}^c} = [(u-m)' \otimes I]\frac{\partial C}{\partial(\Sigma^{-1})^c} + \begin{bmatrix} \sum_j (u_1 u_j - \sigma_{1j}) \partial D_{1j}/\partial \Sigma^{-1}{}^c \\ \vdots \\ \sum_j (u_G u_j - \sigma_{Gj}) \partial D_{Gj}/\partial \Sigma^{-1}{}^c \end{bmatrix}.$$

The pseudo likelihood will therefore always have a consistent root. However if all we know about the true distribution is its first two moments then for the PMLE to be strongly consistent for all true distributions with those moments, the pseudo score must be linear in $(uu'-\Sigma)$ and $(u-m)$.

To calculate the information matrix test for this model, we require the hessian. To simplify the notation let

$$v_s = \begin{bmatrix} \sum_j (u_1 u_j - \sigma_{1j}) \partial D_{1j} / \partial \Sigma^{-1^c} \\ \vdots \\ \sum_j (u_G u_j - \sigma_{Gj}) \partial D_{Gj} / \partial \Sigma^{-1^c} \end{bmatrix},$$

and

$$v_m = \begin{bmatrix} \sum_j (u_1 u_j - \sigma_{1j}) \partial D_{1j} / \partial m \\ \vdots \\ \sum_j (u_G u_j - \sigma_{Gj}) \partial D_{Gj} / \partial m \end{bmatrix}.$$

Therefore

$$\frac{\partial^2 \ln L}{\partial m \partial m'} = -\frac{\partial C}{\partial m} + \sum_{g=1}^{G} \frac{\partial^2 C_g}{\partial m \partial m'} + \frac{\partial v_m}{\partial m},$$

$$\frac{\partial^2 \ln L}{\partial m \partial \Sigma^{-1^{c'}}} = \frac{\partial v_m}{\partial \Sigma^{-1^{c'}}} + ((u-m)' \otimes I) \frac{\partial [\partial C / \partial m]^c}{\partial \Sigma^{-1^{c'}}},$$

$$\frac{\partial^2 \ln L}{\partial \Sigma^{-1^c} \partial \Sigma^{-1^{c'}}} = (I \otimes [(u-m)' \otimes I] \partial [\partial C / \partial \Sigma^{-1^c}] / \partial \Sigma^{-1^{c'}} + \frac{\partial v_s}{\partial \Sigma^{-1^{c'}}}.$$

The information matrix test examines hypotheses about the first four moments of the distribution. In a similar fashion to Hall (1982), the indicator vector can be divided into three components. The first compares two estimater of the covariance of $\hat{m}_T$, $\{\partial^2 LLF / \partial m_i \partial m_j\}$ and $\{\frac{\partial LLF}{\partial m_i} \frac{\partial LLF}{\partial m_j}\}$, and is a test on a linear combination of the 2nd, 3rd and 4th moments of u. The second compares the estimator of the covariance of $\hat{m}_T$ and $\hat{\sigma}_{ij}$, and is a test on a linear combination of the first four moments of u. Finally the

vector comparing $\{\partial^2 LLF/\partial \Sigma_i^{-1^c} \partial \Sigma_j^{-1^c}\}$ with $\{\frac{\partial LLF}{\partial \Sigma_i^{-1^c}} \cdot \frac{\partial LLF}{\partial \Sigma_j^{-1^c}}\}$, is another test on a linear combination of the first four moments of the distribution.

Hall (1982) shows that in the normal linear fixed regressor model, the IMT decomposes asymptotically under H_0 into the sum of three independent tests: a test of homoscedasticity, a test of skewness and one of nonnormal kurtosis. This decomposition dependent on the symmetry of the distribution about zero and whilst it generalises to the nonlinear counterpart of this model, as we show below, it is clearly not going to be a general property of the quadratic exponential family.

We consider the case where G = 1, but our results generalise to higher order dimension vectors.

The LLF of one observation is

$$LLF = \text{const} - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y_t - f(x_t,\theta))^2,$$

and

$$\frac{\partial LLF}{\partial \theta} = \frac{1}{\sigma^2}(y_t - f(x_t,\theta))\frac{\partial f_t}{\partial \theta},$$

$$\frac{\partial LLF}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_t - f(x_t,\theta))^2,$$

$$\frac{\partial^2 LLF}{\partial \theta \partial \theta'} = \frac{1}{\sigma^2}\frac{\partial f_t}{\partial \theta}\frac{\partial f_t}{\partial \theta'},$$

$$\frac{\partial^2 LLF}{\partial \theta \partial \sigma^2} = -\frac{1}{\sigma^4}(y_t - f(x_t,\theta))\frac{\partial f_t}{\partial \theta},$$

$$\frac{\partial^2 LLF}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(y_t - f(x_t,\theta))^2.$$

The indicator vector is therefore identical to that derived by Hall (1982) once $\partial f_t/\partial\theta$ is substituted in for x_t, as

$$d' = [\Delta_1, \Delta_2, \Delta_3],$$

where Δ_1 has s^{th} element:

$$\Sigma \frac{1}{\sigma^2} \frac{\partial f_t}{\partial \theta_i} \cdot \frac{\partial f_t}{\partial \theta_j}(u_t^2 - \sigma^2),$$

Δ_2 has r^{th} element:

$$\Sigma \frac{\partial f_t}{\partial \theta_r}(u_t^3 - 3u_t\sigma^2)\frac{1}{\sigma^6},$$

and

$$\Delta_3 = \Sigma\sigma^{-8}\left(\frac{u_t^4}{4} - \frac{3u_t^2}{3}\sigma^2 + \frac{3}{4}\sigma^4\right).$$

For the decomposition we require the $\text{cov}(\Delta_i, \Delta_j) = 0$ for $i \neq j$. This follows immediately for $\text{cov}(\Delta_2, \Delta_j)$, as it is a linear function of the odd moments of u_t and so under H_0 is zero. The fact that $\text{cov}(\Delta_1, \Delta_3)$ is zero follows from

$$\Sigma \frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j}(u_t^2 - \sigma^2) = \Sigma\left(\frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j} - n^{-1}\Sigma \frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j}\right)(u_t^2 - \hat{\sigma}^2),$$

and so

$$\text{cov}(\Delta_1, \Delta_3) = n^{-1}\Sigma E\left[\frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j} - n^{-1}\Sigma \frac{\partial f_t}{\partial \theta_i} \frac{\partial f_t}{\partial \theta_j}\right](u_t^2 - \sigma^2)\Delta_3]$$

$$= 0.$$

In the nonlinear regression model with normally distributed errors, the IMT asymptotically decomposes into the sum of three statistics each of which tests homoscedasticity, skewness and zero mean, and nonnormal kurtosis alone. Further the tests are asymptotically independent, if the moments of the error process are finite up to and including order eight and with odd moments zero. Note the remaining part of the analysis in Hall (1982) concerning the power of the test cannot be generalised to the nonlinear model as it required Amemiya's residual decomposition.

8.6. Discussion

The PML procedure concentrates on modelling the first two moments of the process. The use of the exponential family guarantees that the estimator is always consistent and so we can construct consistent estimators of the covariance matrix of the PMLE.

It has been argued that the information matrix test is the natural procedure for assessing the validity up to the second moment of these models. In the example considered, the IM test examines whether the relationship between mean and variance implicitly assumed by the choice of distribution, is supported by the data. These tests can be used outside the regression framework, and as higher order derivative tests based on the linear exponential family are tests on linear combinations of the central moments, they provide useful tests of the distribution. In this context, the IM test can refute the hypothesis that the data were generated by a binomial distribution, say, but does not

necessarily confirm this if an insignificant statistic is recorded as more than one distribution may have this mean-variance relationship.

For the case in which y_t is scalar, then an alternative method of checking the validity of the distributional assumptions is to use the goodness of fit test outlined by Heckman (1984). This involves dividing the range of y up into more than two model admissible intervals, and comparing the expected and actual frequencies in each interval. It represents an inefficient test of a composite hypothesis about all the moments of the distribution, but its advantage compared to the IMT is that it examines the shape of the distribution and uses information on what range of values should have been observed.

However such a test is not wholly appropriate for the PML framework in which we are concerned with the first two moments of the process alone. Furthermore whilst the goodness of fit test uses information on the "shape" of the distribution, which may be of interest when forecasting, the construction of goodness of fit tests when y_t is a vector is a nontrivial exercise and so the method would not appear easily implementable for models of the generality discussed in this chapter.

The IM test is not without its problems as well. Firstly, although we can construct consistent estimators of the covariance if the model is misspecified, the interpretation of a significant statistic is only that the assumed distribution is not correct. We could conduct a succession of IM tests, given consistent estimates, to assess which distribution is most in keeping with the

data. The sequence is not independent and so we can only place a bound on the size of the test. Interpretation of the results may also be difficult if more than one test is insignificant.

9. CONCLUSIONS

In this thesis we have examined the properties of NLFIML in both static and dynamic models, and the parameters restrictions implicit in the requirement that our specification be coherent. It has been shown that the class of true distributions for which NLFIML is consistent depends on the nonlinearities present in the system. If it is possible to write down an explicit reduced form for the system, then we can find families of true distributions for which consistency is guaranteed. For instance in logs and levels models NLFIML is consistent for true distributions with a particular moment generating function. However if the reduced form is implicit then the only analytical results available at this degree of generality are two fold. Firstly, it is possible to find distributions either by suitably choosing the correlation structure or mixing distribution of the errors, but the exact choice in any particular case depends on the nature of the underlying reduced form. Secondly, and more generally, apart from these special cases, NLFIML is only guaranteed to be consistent if the model is correctly specified. This does not rule out the possibility of there being nonnormal true distributions in any particular case for which NLFIML is consistent, but it does suggest that its robustness can by no means be assumed.

Our analysis has frequently used the statistical nomenclature introduced by White in his development of QMLE theory. However we have focused attention on an issue that has not received proper treatment in this literature. White (1982) concentrates on the conditions for the convergence of

the QMLE to the KLIC minimising value, whereas we have extended this line of analysis by examining the conditions under which this is the true value. The latter is an important question because once we leave the i.i.d. framework with which White (1982) worked, the implications for our test procedures based on the QMLE are no longer the same regardless of whether the KLIC minimising value is or is not the true value. This point is acknowledged in White (1983), although the extent of its implications are not examined. In the class of models we have considered, if it is not possible to obtain consistent estimates of the first moments then we cannot obtain consistent estimates of the covariance. The result is that we are reduced to what White (1983) has termed conservative inference based on the QMLE. It is because of these implications for test procedures that we have rejected the White (1982) nomenclature and reserved the term consistency for convergence to the true value, which is its conventional meaning.

Typically it is argued that our nonlinear econometric model has a structural equation interpretation. Consistent estimation then has an intrinsic appeal of its own if the parameters themselves are of concern. However consistency retains its importance as a criteria for estimator selection even if we require only a forecasting model. For any specification searches for a more parsimonious representation or the calculation of forecast intervals require consistent estimation of the covariance matrix of the QMLE. The structural equation interpretation entails stringent parameter restrictions, in most cases, for there to be a unique reduced form. In the course of our work we

have noted that various authors (Gourieroux, Monfort and Trognon, (1984), Burguette, Gallant and Souza (1983)) have obtained powerful results for the nonlinear regression model. The interpretation of such models is open to question. In most situations they cannot be regarded as the reduced form equations associated with a nonlinear model. If we have a functionally constant nonlinear structural equation, then from the implicit function theorem we know that in the majority of cases in economic modelling, the reduced form will not remain functionally constant over the sample space. Whilst the regression model can be argued to be an approximation, in some sense, to the reduced form, its "accuracy" will vary from case to case and depend on what may be an arbitrary choice of functional form. Within these models the question of consistency, examined by GMT (1984a), retains its importance, as do specification tests even in the PMLE framework where it is explicitly acknowledged that the error distribution is incorrect. Specification tests provide evidence of when the existing choice of functional form may be inadequate, and improved forecasts can be gained from an alternative formulation. It has been argued in chapter 8 that the information matrix test (White, 1982) would be a natural choice for this purpose.

The performance of NLFIML has been contrasted with that of NL3SLS, which is consistent and asymptotically normally distributed for mean zero error processes under conditions analogous to the linear model. It has the drawback of being asymptotically inefficient, in general, if the true error distribution is normal. However it may be argued that normality is to some extent arbitrary as it is just an

analytically tractable way of capturing three basic properties of the error process, u, i) it can take on any value in R^m, ii) $p(0<u<a) = p(-a<u<0)$ for some point, a, iii) $p(|u|>a)$ monotonically decreases making extreme values unlikely.

Once we consider the case of other symmetric errors then NL3SLS has guaranteed desirable asymptotic properties, and NLFIML may no longer have an efficiency advantage. The central importance of testing procedures in econometrics and the analytical difficulties of comparing the efficiency of NL3SLS and NLFIML, suggest that the robustness property should be given more weight in our choice of estimator. Our analysis shows that the incorrect imposition of the normality assumption is likely to bias inference in nonlinear models, and so NL3SLS would appear to be the prefered estimator.

Although NL3SLS results from an optimisation routine that does not take explicit account of the Jacobian restrictions, if our model specification is to have a structural interpretation the estimators should satisfy those conditions. Evidence of their violation calls into question this interpretaton of the model. Alternatively our general nonlinear model can be considered as one of an infinite number of approximations to the dgp, and so should be interpreted like the regression model as a method of obtaining forecasts. From the implicit function theorem we know the choice of a regression or more general nonlinear model is not interchangeable in terms of the correlations exploited because a functionally constant system of nonlinear equations does not usually solve for a

functionally constant regression model. The advantage of the $f(y,x,\alpha) = u$ framework is that it can be used for policy simulation (in the spirit of Sims, 1982, see chapter 1 above), whereas the regression model relegates the simultaneity to the error process.

Estimators and forecasts should always then be interpreted conditional on the chosen functional form and loss function employed in estimation. This returns us to the problems of choosing a model and assessing its adequacy discussed in chapter 1. It was argued there that the strategy of assuming a linear model unless diagnostics suggested its inadequacy was undesirable because of blinkered interpretations of such tests. This problem appears in our choice of diagnostic for any model, as our conclusions result from the imposition of subjective opinions. At present, outside the linear framework, there exist limited methods of discriminating between two models that can be considered of the same functional form, for instance because they are both bilinear. Whearas what is ideally required is some method of selecting the appropriate class of functional form to be considered. The problems of statistical dependence between a sequence of tests and the choice of the appropriate correlations to examine suggest easily interpretable procedures for such class identification are going to be very difficult, if not impossible, to obtain for nonlinear models. In their absence it is important to be aware of the limitations of econometric models and to be very cautious in attaching any structural or dgp interpretation to them. Instead these models should be more properly regarded as an approximation

whose interpretation is conditional on the chosen functional form and estimation loss function. For a given general nonlinear model we have demonstrated that the loss function implicit in NL3SLS is more appropriate, given conventional requirements, than that of NLFIML. The development of more sophisticated methods of discriminating between nonlinear models, after estimation, remains an area worthy of further research.

APPENDICES

Appendix 1: Proof of Gale and Nikaido's Univalence Theorem

In this appendix we present Gale and Nikaido's (1968) proof of their univalence theorem (theorem 4 p. 86):

If $F: \Omega \to R^n$, where Ω is a closed rectangular region of R^n, is a differentiable mapping such that the Jacobian matrix $J(x)$ is a P-matrix for all x in Ω, then F is univalent in Ω.

Proof

Suppose $a,b \in \Omega$ and $F(a) = F(b)$. We need to show that $a = b$. Let a_i, b_i be the i^{th} elements of a and b respectively. Suppose, maybe after reordering that, $a_i \leq b_i$ $(i \leq k)$, $a \geq b_i$ $(i > k)$. We need to consider 3 cases.

Case 1: $k = n$, then $F(a) = F(b)$ and $a \leq b$, and so by the amended theorem 3, $a = b$.

Case 2: $k = 0$, $a = b$ by similar reasoning to case 1.

Case 3: $0 < k < n$. Define the mapping $D: R^n \to R^n$ by

$$D(x_1,..,x_n) = (x_1,...,x_k,-x_{k+1},...,-x_n).$$

D is univalent on R^n and $D^{-1} = D$. Further $D(\Omega)$ is again a closed rectangular region. Let $D(a) = a^*$ and $D(b) = b^*$. Let $H: D(\Omega) \to R^n$ be the composite mapping given by $H = D \circ F \circ D$. This implies $H(a^*) = H(b^*)$ and $a^* \leq b^*$. As the Jacobian matrix of the product of two transformations is the product of the Jacobian of each, we have the Jacobians of H and F are identical. Therefore the Jacobian of H is a P-matrix and by the amended theorem 3 applied to H we have $a^* = b^*$ which implies $a = b$

Appendix 2

Hajek-Renyi inequality:

Let $X_1, X_2 \ldots$ be independent r.v's such that $EX_i = 0$ and $V(X_i) = \sigma_i^2 < \infty$. If $c_1, c_2 \ldots$ is a nonincreasing sequence of positive constants, then for any positive integers m,n with $m < n$ and arbitrary $\varepsilon > 0$

$$p\left(\max_{m \leq k \leq n} c_k |X_1 + \ldots + X_k| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2}\left(c_m^2 \sum_{1}^{m} \sigma_i^2 + \sum_{m+1}^{n} c_i^2 \sigma_i^2\right).$$

The proof of this result only requires the independence of the X_i to imply their orthogonality. (See Rao, 1973, p. 142.)

Appendix 3

Order of Mixingale sequences

Our discussion of the convergence of a mixingale sequence is restricted to processes with ψ_n exhibiting a particular rate of convergence to zero. These conditions ensure that the various summations under consideration do in fact converge.

McLeish (1975) defines $\{\psi_k\}$ to be of size -p if there exists a positive sequence $\{L(k)\}$ such that

a) $\Sigma_n n^{-1} L(n) < \infty$,

b) $L_n - L_{n-1} = O(L(n)/n)$,

c) L_n is eventually nondecreasing,

d) $\psi_n = O[1/n^{1/2} L(n)]^{2p}$.

This can be translated into a single order condition, as any sequence which is $O[n^{1/2} \log n (\log\log n)^{1+\delta}]^{-p}$ with $\delta > 0$ is of size -p/2.

The mixingale convergence theorem requires ψ_n to be of

size -1/2, and subsequent order restrictions on $\phi(m)$ and $\alpha(m)$ guarantee the mixing processes are mixingales of size -1/2.

Appendix 4: Proof that α-mixing processes are mixingales

We need to show

$$||E(X|F^*)-E(X)||_p \leq 2(2^{1/p}+1)\alpha(F^*,G^*)^{1/p-1/r}||X||_r.$$

This is proved in lemma 2.1 of McLeish (1975).

Let $c = \alpha(m)^{-1/r}||X||_r$ and $X_1 = XI(|X|\leq c)$, where $I(\cdot)$ is the indicator function, and $X_2 = X-X_1$. We neglect $\alpha = 0$ case for which the result is trivial and put $\alpha > 0$.

By Minkowski's inequality, namely if $r \geq 1$ then

$$E^{1/r}|X+Y|^r \leq E^{1/r}|X|^r + E^{1/r}|Y|^r,$$

and the fact that $X = X_1 + X_2$ we have

$$||E(X|F^*)-E(X)||_p \leq ||E(X_1|F^*)-E(X_1)||_p + ||E(X_2|F^*)-EX_2||_p.$$

From the definition of $X_1 = XI(|X|\leq c)$ it follows that $X_1 \leq c$ and so $E(X_1|F^*)$ and $E(X_1)$ must lie between $[-c,c]$ and so the maximum discrepency between $E(X_1|F^*)$ and EX_1 is $2c$.

Therefore

$$||E(X_1|F^*)-EX_1||_p \leq E^{1/p}[2c]^{p-1}|E(X_1|F^*)-EX_1|$$
$$= (2c)^{\frac{p-1}{p}}E^{1/p}|E(X_1|F^*)-EX_1|.$$

From part a) and putting p = r we have

$$||E(X_2|F^*)-EX_2|| \leq 2\phi(m)^{1-1/p}||X_2||_p \leq 2||X_2||_p,$$

as $\phi(m)$ lies between 0 and 1 by its definition and $p > 1$. Taken together this implies

$$||E(X|F^*)-E(X)||_p \leq (2c)^{\frac{p-1}{p}} E^{1/p}|E(X_1|\ )-EX_1| + 2||X_2||_p, (1).$$

To develop the next part of the proof we need two results from Dvoretsky (1972): lemmas 5.1 and 5.2. Using the Jordan decomposition arguments of part a) we can establish that if x and y are two r.v's satisfying $|X| \leq 1$, $|Y| \leq 1$ and putting $\Delta = \sup_B |P(X\epsilon B)-P(y\epsilon B)|$ where the sup is over all Borel sets B, then $|EX-EY| \leq 2\Delta$. This can be shown as follows:

Let $\nu(B) = P(X\epsilon B)-P(y\epsilon B)$ for all Borel sets B, and so it is signed measure. Now

$$|EX-EY| = |\int t\nu(dt)| \leq \int |t||\nu|(dt) \leq \int |\nu|(dt),$$

where $|\nu|(B)$ is the total variation of λ on B. Let B^+ and B^- be a Jordan decomposition of [-1,1] corresponding to ν, then

$$|\nu|([-1,1]) = 2\nu(B^+) = 2\Delta.$$

We are interested in r.v's bounded by an arbitrary constant, c. The above argument carries through when we standardise the bound to give $|X/c| \leq 1$ and $|Y/c| \leq 1$, the resulting bound on $E|X-EY| \leq 2\Delta c$.

Using this result Dvoretsky (1972) shows that if

$|X| \leq c$ and if $F^* = B(x)$ with G^* any σ-field in the probability space then $E|E(X|G^*)-EX| \leq 4ac$ where α is the strong measure of dependence between two sets defined earlier. This can be proved as follows:

Let G denote the set where $E(X|G^*) \geq EX$, and G' its complement. From the tower property of conditional expectations we have

$$0 = E[E(X|G^*)-EX]$$

$$= E[E(X|G^*)-EX|G]P(G) + E[E(X|G^*)-EX|G']P(G'). \qquad (2)$$

Also

$$E|E(X|G^*)-EX| = E[E(X|G^*)-E(X)|G]P(G)-E[E(X|G^*)-EX|G']P(G'), \qquad (3)$$

and combining (2) and (3) we have

$$E|E(X|G^*)-EX| = 2E[E(X|G^*)-EX|G]P(G).$$

If we let $\tilde{X}$ be a r.v with the same distribution as X and independent of G^* then

$$E[E(X|G^*)-EX|G] = E(X|G)-E(\tilde{X}|G) \leq 2\sup_B|P(X\epsilon B|G)-P(\tilde{X}\epsilon B|G)|,$$

and as $P[(\tilde{X}\epsilon B)\ G] = P(X\epsilon B)P(G)$ the bound becomes $4\alpha c$.

This means we can rewrite equation (1) as

$$||E(X|G^*)-EX||_p \leq (2c)^{p-1/p}(4\alpha c)^{1/p} + \frac{2||X_2||_r^{r/p}}{c^{(r-p)/p}}, \qquad (4)$$

where the second term comes from

$$E|X|^p I(|X|>c) \leq \frac{1}{c^{r-p}} E|X|^r I(|X|>c).$$

From the Minkowski inequality it follows that $||X_2||_r \leq ||X||$ and so the upper bound in (4) becomes $2(2^{1/p+1})\alpha^{1/p-1/r}||X||_r$ which is the required result.

Appendix 5: Covariance matrices of PMLE's in Poisson model example.

Let $d(y_t,b)$ be the indicator vector consisting of the lower triangular elements of the matrix

$$\frac{d^2 LLF_t}{dbdb'} + \frac{dLLF_t}{db} \cdot \frac{dLLF_t}{db'}.$$

From Lancaster (1984), the covariance of d is

$$E(dd') - E(d\frac{dLLF_t}{db'}) \quad E[\frac{dLLF_t}{db} \cdot \frac{dLLF_t}{db'}] \quad ^{-1} E(\frac{dLLF_t}{db} \cdot d')$$

where E denotes $\lim n^{-1}\Sigma E(\cdot)$.

1) For the case where we assume $y_i \sim$ Poisson($\exp x_i'b$), typical elements of the component matrices of the covariance are as follows: where $\lambda_i = \exp x_i'b$.

$$Ed_i d_j = n^{-1} \Sigma x_{rt} x_{st} x_{mt} x_{nt} (3\lambda_t^2 + \lambda_t),$$

$$Ed_i \frac{dLLF_t}{db_j} = n^{-1} \Sigma x_{rt} x_{st} x_{kt} (-3\lambda_t^2),$$

$$E\frac{dLLF_t}{db_i} \frac{dLLF_t}{db_j} = n^{-1} \Sigma x_{it} x_{jt} \lambda_t.$$

2) $y_i \sim N(\exp x_i'b, 1)$

$$Ed_i d_j = n^{-1}\Sigma x_{rt}x_{st}x_{pt}x_{mt}\lambda_t^2(1-5\lambda_t^2+6\lambda_t^4),$$

$$Ed_i\frac{dLLF_t}{db_j} = n^{-1}\Sigma x_{rt}x_{st}x_{pt}\lambda_t^2,$$

$$E\frac{dLLF_t}{db_i}\frac{dLLF_t}{db_j} = n^{-1}\Sigma x_{it}x_{jt}\lambda_t^2.$$

3) $y_i \sim$ Gamma(λ=aexp($-x_i'b$), r=a) where p.d.f. of Gamma is "$\frac{\lambda^r}{\sqrt{(r)}}x^{r-1}e^{-\lambda x}$".

$$Ed_i d_j = n^{-1}\Sigma x_{rt}x_{st}x_{mt}x_{kt}\exp(-2x_i'b)[3a^2+8a+6]$$

$$Ed_i\frac{dLLF_t}{db_j} = n^{-1}\Sigma x_{rt}x_{st}x_{kt}[\frac{2a-1}{a}]$$

$$E\frac{dLLF_t}{db_i}\frac{dLLF_t}{db_j} = n^{-1}\Sigma x_{it}x_{jt}a^{-1}.$$

4) $y_i \sim$ negative binomial, and so with p.d.f.

$$\frac{\Gamma(a^{-1}+y_i)}{\Gamma(a^{-1})\Gamma(y_i+1)}(1+a\exp x_i'b)^{-(a^{-1}+y_i)} \cdot (a\exp x_i'b)^{y_i},$$

$Ed_i d_j =$

$$n^{-1}\Sigma x_{pt}x_{st}x_{rt}x_{mt}[\lambda_t+(4a^2+3a)\lambda_t^2+(9a^3+8a^2+2a)\lambda_t^3+(5a^4+5a^3+a^2)\lambda_t^4],$$

$$Ed_i\frac{dLLF_t}{db_j} = n^{-1}\Sigma\frac{x_{pt}x_{st}x_{rt}}{(1+a\lambda_t)}(\lambda_t+a\lambda_t^2+a^2\lambda_t^2+a^3\lambda_t^3),$$

$$E\frac{dLLF_t}{db_i}\frac{\partial LLF_t}{db_j} = n^{-1}\Sigma x_{it}x_{jt}\frac{\lambda_t}{(1+a\lambda_t)},$$

where $\lambda_t = \exp x_t'b$.

REFERENCES

Amemiya, T. (1977), The maximum likelihood and the nonlinear three stage least squares in the general nonlinear simultaneous equations model, Econometrica, 45, pp. 955-968.

Apostol, T.M, (1974), Mathematical Analysis, 2nd ed., Addison-Wesley, Reading.

Armstrong, M.A., (1979), Basic Topology, Springer Verlag, Berlin.

Ash, R.B. (1972), Real Analysis and Probability, Academic Press, New York.

Barten, A.P. (1969), Maximum Likelihood Estimation of a complete System of Demand Equations, European Economic Review, Fall, pp 7-73.

Basawa, I.V., Feigin, P.D. and Heyde, C.C. (1976), Asymptotic properties of Maximum Likelihood estimators for stochastic processes, Sankhya, 38, Series A, pp 259-270.

Billingsley, P. (1972), Convergence of Probability measures, Wiley, New York.

Bowden, R. (1973), The theory of parametric identification, Econometrica, 41, pp 1069-1074.

Bowden, R. (1974), Nonlinearity and nonstationarity in dynamic econometric models, Review of Economic Studies, 41, pp 173-179.

Brown, B.M. (1971), Martingale Central Limit Theorems, Annals of Mathematical Statistics, 42, pp 59-60.

Brown, B.W. (1983), The identification problem in systems nonlinear in the variables, Econometrica, 51, pp 75-196.

Brundy, J. and Jorgenson, D.W. (1971) Efficient Estimation of Simultaneous Equations Systems by Instrumental Variables, Review of Economics and Statistics, 53, pp 207-224.

Burguette, J.F., Gallant, A.R. and Souza, G. (1983), On the unification of the asymptotic theory of nonlinear econometric models, Econometric Reviews, pp 151-211.

Chesher, A. (1983), the Information Matrix Test, Extensions and Omnilous tests of Specification, Center for Econometrics and Decision Sciences, University of Florida, discussion paper No. 98.

Christensen, L.R., Jorgenson, D.W., and Lau, L.J., (1975), Transcendental Logarithmic Utility functions, American Economic Review, 65, No. 3 pp

REFERENCES

Amemiya, T. (1977), The maximum likelihood and the nonlinear three stage least squares in the general nonlinear simultaneous equations model, Econometrica, 45, pp. 955-968.

Apostol, T.M, (1974), Mathematical Analysis, 2nd ed., Addison-Wesley, Reading.

Armstrong, M.A., (1979), Basic Topology, Springer Verlag, Berlin.

Ash, R.B. (1972), Real Analysis and Probability, Academic Press, New York.

Barten, A.P. (1969), Maximum Likelihood Estimation of a complete System of Demand Equations, European Economic Review, Fall, pp 7-73.

Basawa, I.V., Feigin, P.D. and Heyde, C.C. (1976), Asymptotic properties of Maximum Likelihood estimators for stochastic processes, Sankhya, 38, Series A, pp 259-270.

Billingsley, P. (1972), Convergence of Probability measures, Wiley, New York.

Bowden, R. (1973), The theory of parametric identification, Econometrica, 41, pp 1069-1074.

Bowden, R. (1974), Nonlinearity and nonstationarity in dynamic econometric models, Review of Economic Studies, 41, pp 173-179.

Brown, B.M. (1971), Martingale Central Limit Theorems, Annals of Mathematical Statistics, 42, pp 59-60.

Brown, B.W. (1983), The identification problem in systems nonlinear in the variables, Econometrica, 51, pp 75-196.

Brundy, J. and Jorgenson, D.W. (1971) Efficient Estimation of Simultaneous Equations Systems by Instrumental Variables, Review of Economics and Statistics, 53, pp 207-224.

Burguette, J.F., Gallant, A.R. and Souza, G. (1983), On the unification of the asymptotic theory of nonlinear econometric models, Econometric Reviews, pp 151-211.

Chesher, A. (1983), the Information Matrix Test, Extensions and Omnilous tests of Specification, Center for Econometrics and Decision Sciences, University of Florida, discussion paper No. 98.

Christensen, L.R., Jorgenson, D.W., and Lau, L.J., (1975), Transcendental Logarithmic Utility functions, American Economic Review, 65, No. 3 pp

Davidson, J. (1981), Alternative Estimators for systems with log-linear stochastic equations and linear identities, LSE Econometrics Discussion Paper No. 81/29.

Diewert, W.E., (1971), An application of the Shephard Duality Theorem: A generalised Leontief Cost Function, Journal of Political Economy. 79, pp 481-507.

Dvoretsky, A. (1972), Asymptotic normality for sums of dependent random variables, Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability, vol II, Univ. of California Press, Berkeley.

Engle, R.F., Hendry, D.F. and Richard, J.F. (1983), Exogeneity, Econometrica, 51, pp. 277-304.

Feller, W. (1971), an introduction to probability theory and its applictions, Vol II., Wiley, New York.

Fisher, F. (1966), The identification problem in econometrics, McGraw Hill, New York.

Gale, D. and Nikaido, H. (1968) The Jacobian Matrix and Global Univalence of Mappings, in Readings in Mathematical Economics Volume 1 ed. Peter Newman, John Hopkins Baltimore Press.

Gallant, A.R. and Holly, A. (1980), Statistical inference in an implicit, nonlinear, simultaneous equation model in the context of maximum likelihood estimation, Econometrica, 48, pp 697-720.

Gourieroux, C., Laffont, J.J. and Monfort, A. (1980), Coherency conditions in simultaneous linear equation models with endogenous switching regimes, Econometrica, 48, pp 675-694.

Gourieroux, C., Monfort, A. and Trognon, A. (1984a): Pseudo Maximum Likelihood Methods: Theory, Econometrica, 52, pp 681-700.

Gourieroux, C., Monfort, A. and Trognon, A. (1984b): Pseudo Maximum Likelihood Methods: Applications to Poisson Models, Econometrica, 52, pp 701-720.

Goursat, E. (1969), A course in Mathematical Analysis: Volume 1. New York, Dover Publications.

Granger, C.W.J. and Andersen, A.P. (1978), An Introduction to Bilinear Time Series Models, Gothenberg, Vandenhoeck and Ruprecht.

Hall, A.R. (1982), The Information Matrix Test for the General linear model, unpublished MSc dissertation, University of Southampton.

Hall, P. and Heyde, C.C. (1981), Martingale limit theory and its application, Academic Press, New York.

Harrison, P.J. and Stevens, C.F. (1976), Bayesian Forecasting, J.R. Statist. Soc. B, 38, pp 205-227.

Harvey, A.C. (1981), Time series models, Oxford, Phillip Allen.

Hatanaka, M. (1978), On the efficient estimation methods for the macroeconomic models nonlinear in variables, Journal of Econometrics. 8, pp 323-356.

Hausman, J.A. (1974), Full information instrumental variables estimation of simultaneous equations sytrems, Annals of Economic and Social Measurement. 3/4, pp 641-652.

Heckman, J.J. (1984), The x^2 Goodness of Fit Statistic for Models with Parameters Estimated from Microdata, Econometrica. 52, pp 1543-1548

Heijmans, R. and Magnus, J. (1983a), On the consistency of the Maximum Likelihood estimator with dependent observations, Paper presented at the European meeting of the Econometric Society 1983.

Heijmans, R. and Magnus, J. (1983b), Asymptotic Normality of the Maximum Likelihood estimator in the nonlinear regression model with normal errors, LSE Econometrics discussion paper 83/83.

Hendry, D.F. (1976), The structure of simultaneous equations estimators, Journal of Econometrics, 4, pp. 51-88.

Howrey, E.P., and Kelejian, H.H. (1971) Simulation versus analytical solutions, The Case of Econometric Models In, Computer Simulation Experiments with Models of Economic Systems, (T.H. Naylor, ed.) New York: Wiley.

Jenrich, R.I. (1969), Asymptotic properties of nonlinear least squares estimators, Ann. of Math Stat, 40, pp 633-643.

Jones, D.A. (1978), Nonlinear autoregressive processes, Proc. R. Soc. London A, pp 71-95.

Jorgenson, D.W. and Laffont, J.F. (1974), Efficient estimation of simultaneous equations with additive disturbances, Annals of Economic and Social Measurement, 3/4, pp 615-640.

Keifer, N.M. and Salmon, M., (1983), Testing normality in econometric models, Economics Letters, 11, pp 123-127.

Kelker, D. (1970), Distribution Theory of Spherical distributions and a location-scale parameter generalisation, Sankhya, Series A, 32 pp 419-430.

Kullback, S. and Leibler, R.A. (1951) On information and sufficiency, Annals of Mathematical Statistics, 22, pp 79-86.

Lancaster, T., (1984), A covariance matrix of the information matrix test, Econometrica. 52, 1051-0154.

Loeve, M. (1962), Probability Theory 3rd ed., Nostrand, Princeton.

Loeve, M. (1978), Probability Theory II 4th edn, Springer Verlag, Berlin.

Lucas, R.E. (1976), Econometric policy evaluation: a critique, in "The Phillips Curve and Labour Markets" (K. Brunner and A.H. Meltzer eds.), Amsterdam, North Holland. (Carnegie-Rochester Conference Series on Public Policy No. 1, supplement to Journal of Monetary Economics. January 1976), pp 19-46.

McLeish, D.M. (1975), A maximal inequality and dependent strong laws, Annals of Probability. 3, pp 829-839.

Phillips, P.C.B. (1982), On the consistency of nonlinear FIML, Econometrica. 50, pp 1307-1324.

Pollock, D.S.G., (1979), The Algebra of Econometrics, Chichester, John Wiley and Sons.

Prucha, I.R. and Kelejian, H.H. (1983), The structure of simultaneous equation estimators: A generalisation towards nonnormal disturbances, University of Maryland Department of Economics Discussion Paper 1982-12.

Rao, C.R. (1973), Linear Statistical inference and its applications 2nd edn., Wiley, New York.

Rootzen, H. (1974), Some properties of convergence in distribution of sums and maxima of dependent r.v's., Z. Wahr. verw. Gebiete, 29. pp 295-307.

Rosenblatt, M. (1971), Markov Processes. Structure and asymptotic behaviour, Springer Verlag Berlin.

Rothenberg, T. (1971), Identification in parametric models, Econometrica, 39, pp 577-591.

Sims, C. (1980), Macroeconomics and Reality, Econometrica. 48, pp 1-48.

Sims, C. (1982), Policy Analysis with econometric models, Brookings Papers on Economic Activity (1982), No. 1, pp 107-164.

Tong, H. and Lim, K.S. (1980), Threshold autoregression, limit cycles and cyclical data, Journal of the Royal Statistical Society Series B. Vol. 42, pp. 245-268.

White, H. (1980), Using Least Squares to aproximate unknown regression functions, International Economic Review, 21, pp 149-170.

White, H. (1982), Maximum Likelihood in Misspecified Models, Econometrica, 50, pp 1-25.

White. H. (1983). Corrigendum. Econometrica, 51, p 513.

White, H. and Domowitz, I. (1982), Nonlinear regression with dependend observations, University of California Discussion Paper.

Zellner, A. (1971), An Introduction to Bayesian Econometrics, Wiley, London.

Additional References

Sargan, J.D., (1975), Asymptotic Theory and Large Models, International Economic Review, 16, pp. 75-91.

Serfling, R.J., (1968), Contributions to Central Limit Theory for Depdendent Variables, Annals of Mathematical Statistics, 39, pp. 1158-1175.