Econometric Structural Models: a Model Selection Approach

Dr. Chen Pu

January 4, 2002

Using a model selection approach, this thesis proposes a constructive data-and-theory-combined procedure to identify model structures in the framework of a linear simultaneous equations system based on observed data. A model structure is characterized by restrictions on the structural parameters. To identify these restrictions two issues have to be taken into account: the first is the problem of observational equivalence, i.e. different models may have an identical density function, henceforth data cannot differentiate such observationally equivalent models; the second is the identification of the restrictions on structural parameters. For the first problem we classify models into different observationally equivalent classes and give necessary and sufficient conditions for the uniqueness of observationally equivalent models. For the second problem we take an approach based on the information criterion and give a (strong) consistent criterion to identify the restrictions on the structural parameters. We apply this model selection criterion to cointegration systems and provide a unified approach to analyzing linear simultaneous equations systems and cointegration systems. The model selection criterion is also used to identify the encompassing relations among different structural models under mis-specification. Through constructive use of the model selection criterion, we may get the most parsimonious structural model that is compatible to the data among the models under investigation. However, all conclusions conducted from the model selection criterion are valid only asymptotically. Nevertheless, the relevance for practical applications of this criterion is demonstrated by some simulation studies.

Contents

1 Introduction

One of the most important tasks of empirical modeling of economic data is to uncover the interpretable relations among variables that can either be used to verify existing economic theories or can provide empirical evidence for a new theory. In the context of an econometric model such relations are manifested in the parameters and the restrictions on the parameters of the model. Therefore, it is of great interest to construct an econometric model as a DGP that can generate data that will have the same characteristics as the observed data. In this way the observed data can be viewed as if they have been generated from this model and it follows that the relations among the variables

described in this model can be regarded as empirically verified. There are principally three approaches to constructing such an econometric model[1]: The traditional Cowles Commission approach or the theory conducted approach; the atheoretical VAR approach or data conducted approach; and the LSE approach or the data-and-theory combined approach.

1.1 Approaches to Building Econometric Models

1.1.1 The Theory Conducted Approach

The theory conducted approach was first developed by researchers of the Cowles Commission. It is also called the Cowles Commission method. The starting point of this approach is the theoretic foundation of a model. Usually an econometric model is seen as a linearized and estimable version of a comprehensive derived economic theoretical model.[2] The main focus of econometric work is on the estimation of parameters.

According to this approach, a structural model consists of correctly specified equations. The underlying premise is that suitable economic theoretical consideration should provide enough identification conditions to specify a structural model that can approximate the real data generating process (DGP). This approach enforces a model structure on a set of observed data and pays little attention to the question whether the restrictions on the DGP implied by the structural model are compatible with the data or not[3].

1.1.2 The Data Conducted Approach

Sims (1980) criticized the "incredible" identification restrictions of the structural models and showed vividly how serious this problem may be. He promotes therefore VAR (vector autoregressive) models without any restrictions on the density function of concerning variables. A VAR model provides here a general statistic framework to describe the observed data. The estimated model will describe the dynamic property of the DGP.

However, VAR models are usually much overparameterized[4]. Most of the estimated parameters are insignificant to zero. And VAR models do not provide intuitively interpretable relations among the variables. If VAR is a suitable instrument to catch the dynamic property of the variables, it is by no means a suitable instrument to understand the data i.e. to give the theoretic interpretation to the parameters.

[1] See Granger (1990) for detailed discussion

[2] See Fair (1984) Powell and Murphy (1997) and Klein (1983) for detailed discussion.

[3] See Spanos (1990) for more discussion.

[4] See Amisano and Giannini (1997) p.2 for more discussion.

1.1.3 The Data-and-Theory-Combined Approach

While the Cowles Commission method emphasizes the theoretic interpretation aspect of an empirical model, the VAR method focuses on the data conformity of the empirical model. Both aspects are essential to empirical economics. "Theory without empirics is empty. Empirics without theory is blind."[5] The theory and data combined approach developed by researchers at the London School of Economics (LSE) combines these two aspects. This approach starts from a general statistic model (it is usually a VAR model) and formulates the economic theories in a set of statistically testable hypotheses and tests these hypotheses within the statistical model.[6] If the test results support these hypotheses, a more restrictive model will be constructed. In this way, a specific structural model may be conducted from a general atheoretical model via a series of comprehensive statistic tests.

1.2 A Model Selection Approach to Structural Modeling

Economic theories do not usually provide enough unambiguous identification restrictions[7] from which we can conduct a unique structural model. This ambiguity in the economic theory leads to alternative structural models to the same economic phenomenon. Furthermore competing economic theories exist in many areas of economics at the same time.

The LSE approach tries to conduct an economic-theoretically founded econometric model by statistical tests. In case of a statistical test the null hypothesis and the alternative hypothesis are not symmetric. In the formulation of a null hypothesis one has to put a great degree of confidence in it. It is questionable whether one would have such confidence in such an economic-theoretically conducted hypothesis, while other competing theories exist.[8] Statistical tests are rather for confirmative study than for explorative study. Furthermore, this kind of test approach may result in contradictory conclusions that two or more rival models could be supported by observed data at the same time.

To overcome these difficulties we adopt the model selection approach to construct an econometric model, where all alternative economic hypotheses are treated equally.

As mentioned at the beginning, the task of empirical modeling is to uncover the real DGP. This is unfortunately an unsolvable problem because the real

[5] Immanuel Kant - German Philosopher (1724 - 1804)
[6] For details see Hendry (1995).
[7] Hendry (1995) p. 5-9
[8] For detailed discussion of this point see Granger and White (1995)

DGP of empirical data is usually too complex to be explicitly describable by a traceable model. Hence, we are forced to approximate the real DGP by a smaller well defined class of models and to develop a procedure to approximate the real GDP by a model in this class. This procedure should be able to identify the real DGP if the real DGP of the observed data were really within this class of models. In case the real DGP is not within this class this procedure should be able to choose the "best" one from this class as the closest approximation[9] to the real DGP.

In the context of linear simultaneous equations systems, this class will be the set of all possible linear simultaneous equations systems.

The basic idea in this thesis is to view the theory-conducted structural models as different set of restrictions on the unconstrained reduced form. Using a model selection approach we can identify which set of restrictions are true. Then we will choose the structural model from which this set of restrictions are derived. In this way, we can get a structural model that is both theoretically founded and compatible with the observed data. This kind of model provides empirical evidence for the economic theory and gives a theoretical understanding to the observed data.

1.2.1 Task of Structural Models and the Requirement

The statements of economic theories are mostly formulated as certain relations among economic variables in the structural form.[10] These relations are expressed by the parameters that link these variables. Hence, structural models provide a natural framework to present theories, to test theories and to interpret data. The estimated parameters in a structural model are usually interpreted to reveal some "behaviour constant". Some questions arise here: are there any alternative models that would describe the data equally well? If yes, do the corresponding parameters in the alternative models have the same value? If not, how should the parameters be interpreted?

It is well known that all exactly identified structural models will have the same reduced form and hence the same goodness of fit to the data. Therefore it is impossible to differentiate these models from the data. In case two rival theories would correspond to two exactly identified structural models, we would not be able to say which one is more appropriate based on the observed data.

According to the requirement for falsifiability of a scientific theory, an economic theory should be formulated as a testable hypothesis in a structural

[9] The measure of the closeness is Kullback-Leibler Information Criterion. See Chapter 8 for details.

[10] It means that these relations are among the interdependent variables.

model if it is going to be tested in an econometric model. In the context of model selection we rank alternative models. If we associate alternative models with alternative economic theories, a test of economic theories can be carried out by selection of models.

1.3 The Organization of this Thesis

Generally, the problem of identifying true structural models based on observed data rests on two levels. On one level, it is to identify the true restrictions on the parameter of the density function using the observed data. On the other level the problem is mapping the restrictions on the parameter of the density function to the restriction on the structural parameters. The former is a statistical issue, the latter is rather an algebraic issue.

In Chapter 2 we give a formal definition of structural models. We define a simultaneous equations system on the unconstrained reduced form. A structural model is taken as a possible representation of the simultaneous equations system, in which some specific properties of data are outstanding. In this context, building a structural model is realized by a proof of its appropriateness as an alternative representation of the unconstrained reduced form.

In Chapter 3 we discuss the problem of observational differentiability. The existence of observational equivalence in a simultaneous equations system is a well known problem in econometrics. This problem has two consequences for statistical inference. The first consequence is for the estimatability of the structural parameters. This is known as the identification problem of structural models. The solution is given by imposing a priori restrictions called identification conditions on the structural parameters. This condition guarantees the uniqueness of the mapping from the parameters of density function to structural parameters.[11] the second consequence is for the observational differentiability. Two identifiable models may still be observationally equivalent. In this case one cannot differentiate these two models from the observed data. This problem arises when one tries to identify true structural models from observed data. It is not yet well discussed in the literature of econometrics. We solve this problem by giving necessary and sufficient conditions for observational differentiability.

After solving the observational differentiability problem we turn in Chapter 4 to the problem of identifying the true structure. We start with the maximum likelihood (ML) principle. The ML-function can pick out false restrictions on parameters but it suffers from the problem of overfitting, i.e. the more parameters a model may have, the larger its likelihood function value will

[11] See Judge (1985) p. 574-581

be. The AIC solves the problem by applying the principle of maximization of the relative entropy. It leads to adding a penalty - the number of free parameters - to the ML function. However, the AIC does not solve this problem completely. It can be shown that the AIC is inconsistent[12]. We discuss then the issue of consistent selection criteria and provide a general condition for (weak) consistent selection criteria.

In Chapter 5 we develop a (strongly) consistent model selection criterion for structural models. It turns out that this criterion is formally identical to the Hannan-Quinn criterion for AR processes. This criterion will choose the true model with probability one asymptotically.

In Chapter 6 we provide an alternative representation of a cointegration system. Under this alternative representation, the consistent model selection criterion provides a unified approach to analyzing simultaneous equation systems and a cointegration system.

In Chapter 7 we look at the problem of misspecification. We adopt the pseudo-true value concept from White (1982) and Gourieroux and Monfort (1984), and the encompassing concept from Hendry and Richard (1988) and Gourieroux and Monfort (1996). We discuss the result of the consistent model selection criterion under misspecification. The model selection criterion provides an instrument to identify the parsimonious encompassing relation. Hence, the model selection criterion will choose the most parsimonious model that is the closest to the real DGP among all candidate models asymptotically.

In Chapter 8 we study the performance of the consistent model selection criterion for diverse constellation of model characters via simulation. We first look at the small sample size performance of the criterion to see when the asymptotical property prevails. Then we look at the sensitivity of this criterion to choose the most parsimonious model. We illustrate the performance of the criterion under nested, non-nested admissible models, as well as non-admissible models. We also study the performance of the model selection criterion for large scale models with up to 100 equations in a model.

[12]Shibata (1976)

2 Structural Models

2.1 The General Setting of a Simultaneous Equations Model

2.1.1 Reduced Form

We consider the following simultaneous equations system:

$$Y_t = \Pi X_t + V_t \qquad \text{for t=1,2,...T} \tag{2.1}$$

with the following assumptions:

- $Y_t \in \mathbb{R}^{G\times 1}$ is a random variable called the endogenous variable.
- $X_t \in \mathbb{R}^{K\times 1}$ is called the predetermined variable with :

  $$X_t' = (Y_{t-1}', Y_{t-2}', ...Y_{T-p}, \xi_t')$$

  $\xi_t \in \mathbb{R}^{K_e} \times 1$ is the exogeneous deterministic variable with:

  $$\operatorname*{plim}_{T\to\infty} \frac{\sum_{t=1}^{T} \xi_t \xi_t'}{T} = M_{\xi\xi'}.$$

  where $M_{\xi\xi}$ is a nonsingular constant matrix.
- $V_T \in \mathbb{R}^{G\times 1}$ is random disturbance. It is identically independently distributed as $N(0, \Omega)$
- X_t and V_t are uncorrelated.

  $$E(X_t V_t') = 0$$

- Π is a $G \times K$ matrix of parameters that satisfies the following stability condition. We rewrite the model explicitly in the lags of Y_t:

  $$Y_t = \Pi X_t + V_t = \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + ... + \Pi_p Y_{t-p} + \Pi_e \xi_t + V_t$$

  The stability condition is:

  - $\max\{|\lambda_i|; i = 1, 2, 3, ...GP\} > 1$
    λ_i is the i-th root of the following equation:

    $$|I - \Pi_1 \lambda_1 - \Pi_2 \lambda^2 - ... - \Pi_p \lambda^p| = 0$$

  - The initial value of Y_t is given.

The equations system (2.1) with the assumptions above is called the unconstrained reduced form of a simultaneous equations system. The conditional density function of the dependent variable Y_t given X_t is:

$$f(y_t|x_t;\Pi,\Omega) = (2\pi)^{-\frac{G}{2}}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2}(y_t-\Pi x_t)'\Omega^{-1}(y_t-\Pi x_t)}$$

We denote the realization of Y_t, X_t and V_t by y_t, x_t and v_t respectively. For a data set of T observations we stack all T observations into one equation:

$$\mathbf{y}_T - \mathbf{x}_T\Pi' = \mathbf{v}_T$$

where $\mathbf{y}_T = (y_1', y_2', ...y_T')$ and $\mathbf{x}_T = (x_1', x_2', ...x_T')$ and $\mathbf{v}_T = (v_1', v_2', ...v_T')$.

The log likelihood function for these T observations is:

$$\begin{aligned}&\log L_T(\Pi,\Omega;\mathbf{y}_T,\mathbf{x}_T)\\ =&-\frac{TG}{2}\log(2\pi)-\frac{T}{2}\log|\Omega|-\frac{1}{2}tr(\Omega^{-1}(\mathbf{y}_T-\mathbf{x}_T\Pi')'(\mathbf{y}_T-\mathbf{x}_T\Pi'))\end{aligned}$$

We know that the reduced form is seemingly unrelated (SUR)[13], the maximum likelihood estimate(MLE) can be obtained by applying ordinary least squares (OLS) to each single equation in (2.1)[14].

2.1.2 Structural Form

If there exists a nonsingular $G \times G$ matrix B and a $G \times K$ matrix Γ and a set of a priori restrictions (this will be explained below) on B and Γ, such that

$$B^{-1}\Gamma = -\Pi, \tag{2.2}$$

we can premultiply B to both sides of (2.1) and get:

$$BY_t + \Gamma X_t = U_t \qquad \text{for t=1,2,...T} \tag{2.3}$$

with $U_T = BV_T, E(U_t) = 0.E(U_t{U_t}') = B\Omega B'$. The equations system (2.3) is called the structural form of the simultaneous equations system. Often it

[13]See Theil (1971)

[14]See Hamilton (1994) p. 291-296

is also called a structural model. According to the assumption we made for the unconstrained reduced form (2.1), the structural equations system (2.3) fulfills the conventional assumptions of simultaneous equations system.[15]

If no restrictions are placed on the structural parameters matrix (B, Γ), (B, Γ) is unidentified in the sense that any matrix (B^*, Γ^*) which is similarly unconstrained will also satisfy the condition (2.2) where $B^* = AB$, $\Gamma^* = A\Gamma$ and A is an arbitrary nonsingular matrix:

$$B^{*-1}\Gamma^*(AB)^{-1}(A\Gamma) = B^{-1}\Gamma = -\Pi. \tag{2.4}$$

The number of parameters in (B, Γ) is $G(G + K)$, while we have only GK conditions in (2.2) to determine $G(G + K)$ parameters. Obviously we need a priori restrictions to identify (B, Γ). Identification conditions are fully discussed in Schmidt (1976) p. 128-145[16]. In this thesis we consider only identified structural models with zero restrictions and normalization restrictions[17].

2.1.3 Implication of Structural Representation on the Reduced Form

Exactly identified structural models have the same number of parameters as the reduced form, they do not impose any restrictions on Π. **Overidentified** structural models have less free parameters than the reduced form, they impose some restrictions on Π matrix. These restrictions can be demonstrated in the following example:

Exapmple:

$$\begin{pmatrix} 1 & \beta_{12} & \beta_{13} \\ 0 & 1 & \beta_{23} \\ \beta_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & 0 & \gamma_{24} \\ 0 & 0 & \gamma_{33} & \gamma_{34} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}.$$

For a simultaneous equations system as defined in (2.1) to have this structural presentation the following equation must be satisfied:

$$B\Pi = -\Gamma.$$

For the first row of the equation above we have:

[15] See Appendix A.1, Schmidt (1976) p. 120, Dhrymes (1993), and Theil (1971)

[16] See also Amemiya (1985) p. 231

[17] Normalization condition is that the diagonal elements of B are restricted to be unit; zero restriction on (B, Γ) is that some other elements in (B, Γ) are assumed to be zero

$$\begin{pmatrix}1 & \beta_{12} & \beta_{13}\end{pmatrix}\begin{pmatrix}\pi_{11} & \pi_{12} & \pi_{13} & \pi_{14}\\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24}\\ \pi_{31} & \pi_{32} & \pi_{33} & \pi_{34}\end{pmatrix} = -\begin{pmatrix}\gamma_{11} & 0 & 0 & 0\end{pmatrix}.$$

For $\begin{pmatrix}1 & \beta_{12} & \beta_{13}\end{pmatrix}$ to be uniquely identified the following sub matrix of Π must satisfy the rank condition[18]:

$$rank\begin{pmatrix}\pi_{12} & \pi_{13} & \pi_{14}\\ \pi_{22} & \pi_{23} & \pi_{24}\\ \pi_{32} & \pi_{33} & \pi_{34}\end{pmatrix} = 2$$

This rank condition implies that the elements in this submatrix cannot be estimated freely. Exactly one element is determined by the others. Similarly, the third equation is overidentified. It imposes also a restriction on the Π matrix. The second equation is exactly identified, we can determined the four structural parameters of the second equation from any Π. Hence, the second structural equation does not impose any restrictions on the Π matrix. □

Generally, zero restrictions of an overidentified structural model impose, according to their positions in the (B, Γ) matrix, such rank constrains on the corresponding sub matrix in Π. The number of restrictions imposed on Π by (B, Γ) is the number of overidentification conditions on (B, Γ)[19]. Because only zero restrictions on structural parameters are considered and the variance covariance matrix is assumed to be unconstrained, a structural model as defined in (2.3) is fully determined by the restrictions on (B, Γ). Hence we use (B, Γ) to represent a structural model in this thesis.

2.2 Model Selection Approach

2.2.1 Structural Form vs. Reduced Form

In the context of a theoretical framework one may derive the "behaviour" of economic agents and present these "behaviours" in the structural form of a simultaneous equations system (2.3). The attribute "structural" refers to an explicit description of how an action of economic agents - "Y_{it}" depends directly on other actions "Y_{jt}" and on the given information X_t. The reduced form (2.1) sums up all direct and indirect dependence of Y_t on all predetermined information X_t. "Generally, the structural form is more revealing of

[18] See Frohn (1995) p. 169

[19] The number of overidentification conditions is the difference between the number of zero restrictions in an equation and $G - 1$.

the manner in which an economic system is operating. The reduced form is less revealing."[20] In fact, the reduced form encompasses every structural model. It can always be estimated without referencing any structural form. An overidentified structural model imposes restrictions on the unconstrained reduced form; it corresponds only to a specific reduced form. If an overidentified structural model can encompass the DGP, the relation (2.2) will be satisfied. (2.2) implies then restrictions on the unconstrained reduced form Π. Hence, we can make a judgement about the appropriateness of a structural model by testing the associated restrictions on the reduced form.

2.2.2 Model Selection Approach

The reduced form as defined in (2.1) provides a general framework to study structural models. A structural form as described in (2.3) provides potentially a more parsimonious alternative presentation of the real DGP and may provide more interpretable facts of the observed data. If the real DGP can be encompassed by such a structural model, the observed data should reveal this property. *Identify such structural model from the study of data is the approach to structural modeling pursued in this thesis.*

If we know a potential candidate of the structural model, we can test the restrictions imposed on Π by this candidate model.

The question is how to find such candidates. There are principally two ways to get such candidates. One way is per permutation and an automatical search for such structural models[21]. Another way is to conduct alternative structural models by theoretical reasoning. Because the ultimate motivation of structural modeling is to understand and interpret the observed data, often to understand them in a specific way, we will take the latter approach to formulate possible potential candidates. [22] When we have a group of candidate models, we will evaluate each candidate model by a model selection criterion. The task is then to discover a parsimonious and interpretable structural model (2.3) for a given reduced form (2.1).

[20] See Dhrymes (1993) p. 13-14

[21] See Hendry and Krolzig (2001) for different strategies of search

[22] Of course there is no guarantee that a theoretical founded model will encompass the DGP. In case none of the candidates pass the proof, a new trail will be made, in this way our knowledge about the phenomena will be accumulated.

3 Observational Differentiability

In finding an overidentified structural model based on a given reduced form, one question rises naturally: is the identified structural model unique if we can identify it? In other words: are there overidentified models that will induce the same restriction on Π? We are going to answer this question in the following sections.

3.1 Definitions

We give at first a few definitions to formalize our discussion.

Definition 3.1 (Structure) *A structure is a complete specification of the parameters in the probability function of the variable concerned, say Y_t.*

We denote a structure by $(\bar{B}, \bar{\Gamma}, \bar{\Sigma})$. For a structural model as defined in (2.3), a structure is a point in the space of $\mathbb{R}^{G(G+K+(G+1)/2)}$, i.e. a numerically specified $(\bar{B}, \bar{\Gamma}, \bar{\Sigma})$. A structure corresponds uniquely to a numerically specified reduced form, i.e. $(-\bar{\Pi}, \bar{\Omega}) = (\bar{B}^{-1}\bar{\Gamma}, \bar{B}^{-1'}\bar{\Sigma}\bar{B}^{-1})$.

Definition 3.2 (True structure) *A structure is called true structure, if the data under investigation is generated by the density function specified by this structure.*

The corresponding reduced form is called true reduced form. We denote the true structure by $(\bar{B}_0, \bar{\Gamma}_0, \bar{\Sigma}_0)$ and the true reduce form by $(-\bar{\Pi}_0, \bar{\Omega}_0) = (\bar{B}_0^{-1}\bar{\Gamma}_0, \bar{B}_0^{-1'}\bar{\Sigma}_0\bar{B}_0^{-1})$.

Definition 3.3 (Model) *A model is a set of all possible structures. A model is characterized by the a priori restrictions on the parameter matrix (B, Γ).*

Throughout this thesis we consider only zero restrictions on parameters in the matrix (B, Γ). Different zero restrictions on the matrix (B, Γ) will be treated as different models. Because free varying parameters in (B, Γ) are complimentary to the zero restrictions (B, Γ), we define a model either by the free parameters in the matrix (B, Γ) or the zero restrictions on (B, Γ). The covariance matrix is considered to be unconstrained in this thesis. If the number of restrictions on a model is r, the parameter space of the model will be $R^{G(G+K+(G+1)/2)-r}$. Such a model with r restrictions can generate different structures that all fulfill the r restrictions.

Definition 3.4 (Admissible to a structure) *A model M_i is called admissible with respect to a structure, if the model can generate a density function that is identical to that specified by the structure.*

For example, the unconstrained reduced form, that is a model with zero restrictions on all off-diagonals of matrix B, is always an admissible model to any structural model, because the true reduced form is always one point in the set of all unconstrained reduced forms. If a model cannot generate a density function that is identical to the true density function, the model is said to be not admissible.

It is worthy to point out that it matters whether the parameter of the true reduced form is contained in the parameter space of the reduced form of a model M_i but not whether the true structural parameter is contained in the structural parameter space of M_i.

For example, the unconstrained reduce form is admissible to the true structure of an overidentified interdependent model. Obviously, this true structure must not be a point contained in the parameter space of the unconstrained reduced form, because some of its off-diagonal elements in the B matrix may not be zero.

Definition 3.5 (Admissible to a model) *A model M_i is called admissible with respect to another model M_j, if the model M_i is admissible to any structures that are contained in M_j.*

In this sense, M_i is admissible with respect to M_j, if the parameter space of the reduced form of M_j is contained in the parameter space of the reduced form of M_i.

Definition 3.6 (Observationally equivalent models) *Two models are called observationally equivalent, if they are admissible with respect to each other.*

Obviously, two observationally equivalent models can generate identical density functions. Therefore, for any set of data their maximum values of respective likelihood functions will be the same. That is why they are called observationally equivalent. Two observationally equivalent models will have the same number of zero restrictions, because they impose the same restriction on Π.

Definition 3.7 (True model) *A model M_0 is called a true model if it is admissible to the true structure and contains the same number of zero restrictions as the number of zeros in the true structure.*

According to this definition a true model must not be unique. If M_0 is a true model then the observationally equivalent models of M_0 will also be true models. (Compare Proposition 3.8) This definition is justified by the property that if a true model has observationally equivalent models, we cannot differentiate from which one of the observationally equivalent models the observed data may be generated. Hence they are all equally true if we judge them according to the data.

Admissible models with respect to M_0 may have a different number of free parameters. Among a set of admissible models with respect to M_0 there must be a model with a minimum number of free parameters. Models with minimum number of free parameters are called the most parismonious model within this admissible set.

M_0 is itself an admissible model with respect to M_0. It is also the most parsimonious model among all admissible models with respect to M_0[23].

3.2 observationally equivalent Models

To describe the property of observationally equivalent models we have the following propositions.

Proposition 3.8 *If M_i is admissible with respect to M_j and has the same number of free parameters as M_j, then M_j and M_i are observationally equivalent .*

Proof:

M_j imposes a set of restrictions on the unconstrained reduced form Π. The number of restrictions are $Z_j - G(G-1)$, where Z_j is the number of zero restrictions in model M_j. Because M_i has as many zero restrictions as M_j, it imposes also $Z_j - G(G-1)$ restrictions on the unconstrained reduced from.

Now M_i is admissible with respect to M_j: it implies the $Z_j - G(G-1)$ restrictions imposed by M_i are the same as those restrictions on Π imposed by M_j. In other words the derived reduced form of M_i and M_j are the same. Because both M_j and M_i are identifiable, there is a 1-1 mapping between (B_j, Γ_j) and Π_j, and between (B_i, Γ_i) and $\Pi_i = \Pi_j$. It follows that there exists a 1-1 mapping between (B_j, Γ_j) and (B_i, Γ_i). In other words, for any density function generated by (B_i, Γ_i) there exists a (B_j, Γ_j) that generates the same density function. This means M_j is admissible with respect to M_i. □

From the Proposition 3.8 above we have the following statements:

[23] See next section.

- Two structural models are observationally equivalent, if they induce the same restrictions on the unconstrained reduced form.
- All models in the most parsimonious admissible group with respect to M_0 are observationally equivalent to M_0.
- Observationally equivalent models have the same number of zero restrictions.
- Observationally equivalent models are admissible with respect to each other.

Proposition 3.9 (Exact identification and linear transformation) *If a model is exactly identified, there exists a linear transformation for each structure of the model, such that the number of zeros remains unchanged after this transformation.*

Proof:

Suppose we have a exactly identified structural model:

$$BY_t + \Gamma X_t + U_t$$

Premultiply the equation by B^{-1} so that we get the reduced form:

$$Y_t = -B^{-1}\Gamma X_t + B^{-1}U_t$$

According to the definition of exact identification the number of zero restriction in the structural form is $G(G-1)$ and the number of zero restrictions in the reduced form is also $G(G-1)$. Because B is a full rank matrix B^{-1} corresponds to a linear transformation. (In case the B is a unit matrix, then a linear transformation that eliminates an element in the Γ matrix will add one zero restriction into Γ matrix but at same time reduce one zero restriction in the B matrix. This linear transformation keeps the number of zeros unchanged after the transformations.) □

Corollary 3.10 *If one equation in a structural model is exactly identified, there exists a linear transformation of the model that transforms this equation into a new one and keeps the number of zero restrictions in this equation unchanged.*

Proposition 3.11 (Observational equivalence and linear transformation) *Two different models (B_i, Γ_i) and (B_j, Γ_j) are observationally equivalent , if and only if*

- *The number of zero restrictions in both model are equal:* $Z_i = Z_j$
- *For any structure* $(\bar{B}_i, \bar{\Gamma}_i)$ *in* (B_i, Γ_i) *there exists a structure* $(\bar{B}_j, \bar{\Gamma}_j)$ *in* (B_j, Γ_j) *such that* $(\bar{B}_j, \bar{\Gamma}_j) = A(\bar{B}_i, \bar{\Gamma}_i)$ *where* $A \neq I$.

Proof:

Necessity:

Because (B_i, Γ_i) and (B_j, Γ_j) are observationally equivalent, it follows that the number of zero restrictions are the same in both models and for any given structure $(\bar{B}_i, \bar{\Gamma}_i)$ there exists a $(\bar{B}_j, \bar{\Gamma}_j)$ such that their density function are identical. This implies that

$$\bar{B}_i^{-1}\bar{\Gamma}_i = \bar{B}_j^{-1}\bar{\Gamma}_j$$

We find a $A = \bar{B}_j^{-1}\bar{B}_i^{-1}$, Suppose that A would equal I we would have $\bar{B}_j = \bar{B}_i$ and $\bar{\Gamma}_j = \bar{\Gamma}_i$. This would contradict the assumption that (B_i, Γ_i) and (B_j, Γ_j) are different models. Hence $A \neq I$.

Sufficiency:

suppose that for $(\bar{B}_i, \bar{\Gamma}_i)$ there exists a structure $(\bar{B}_j, \bar{\Gamma}_j)$ such that $(\bar{B}_j, \bar{\Gamma}_j) = A(\bar{B}_i, \bar{\Gamma}_i)$ and $A \neq I$. It follows then that the density of the structure $(\bar{B}_j, \bar{\Gamma}_j)$ will be the same as that of the structure $(\bar{B}_i, \bar{\Gamma}_i)$:

$$\bar{B}_j^{-1}\bar{\Gamma}_j = (A\bar{B}_i)^{-1}(A\bar{\Gamma}_i) = \bar{B}_i^{-1}\bar{\Gamma}_i$$

This means that model (B_j, Γ_j) is admissible with respect to model (B_i, Γ_i). Since they have the same number of zero restrictions following proposition 3.8 they are observationally equivalent.

Corollary 3.12 *For a structural model, if we can always find a linear transformation to transform the structural model into another structural model and this transformation keeps the number of zeros unchanged, then these two models are observationally equivalent .*

Corollary 3.13 *For two exactly identified models there always exists a full rank linear transformation that transforms any given structure of one model into a structure of another model.*

Corollary 3.14 *Two observationally equivalent models have the same maximum likelihood function values for a given set of data.*

Proof: See Frohn (1995) p. 179 □

For example, all exactly identified models are observationally equivalent , because they correspond to the same unconstrained reduced form and hence have the same likelihood for any given set of data.

Following are a few more examples. These models all have 6 equations and 6 predetermined variables.

Example 1:

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & 0 & \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & 1 & \beta_{23} & 0 & 0 & 0 & \gamma_{21} & 0 & \gamma_{23} & 0 & 0 & 0 \\ \beta_{31} & 0 & 1 & 0 & 0 & 0 & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \beta_{45} & 0 & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ 0 & 0 & 0 & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ 0 & 0 & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

is observationally equivalent to :

$$\begin{pmatrix} 1 & 0 & \beta_{13} & 0 & 0 & 0 & \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ \beta_{21} & 1 & 0 & 0 & 0 & 0 & \gamma_{21} & 0 & \gamma_{23} & 0 & 0 & 0 \\ \beta_{31} & 0 & 1 & 0 & 0 & 0 & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \beta_{45} & 0 & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ 0 & 0 & 0 & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ 0 & 0 & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

These are two models with seemingly unrelated blocks. These two models are characterized by their partially exactly identified sub-block in the system, namely the first three equations; when we neglect the zero blocks in the first three equations we would have a three equations system of exactly identified equations. We know from proposition 3.9 and the Corollary of proposition 3.11 that for any given parameters in the first three equations there always exists a linear transformation that transforms these three equations into a structure of first three equations of the second model. Then, according to proposition 3.11 two models are observationally equivalent.

□

Example 2:

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & 0 & \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & 1 & \beta_{23} & 0 & 0 & 0 & \gamma_{21} & 0 & \gamma_{23} & 0 & 0 & 0 \\ \beta_{31} & 0 & 1 & 0 & 0 & 0 & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ 0 & 0 & \beta_{43} & 1 & \beta_{45} & 0 & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ \beta_{51} & 0 & 0 & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ 0 & \beta_{62} & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

is observationally equivalent to :

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & 0 & \gamma_{11} & 0 & \gamma_{13} & 0 & 0 & 0 \\ 0 & 1 & \beta_{23} & 0 & 0 & 0 & \gamma_{21} & \gamma_{22} & 0 & 0 & 0 & 0 \\ \beta_{31} & 0 & 1 & 0 & 0 & 0 & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ 0 & 0 & \beta_{43} & 1 & \beta_{45} & 0 & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ \beta_{51} & 0 & 0 & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ 0 & \beta_{62} & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

These are two models with recursive blocks. Also here we have a partially exactly identified sub-block: the first three equations. Similarly we can get the second model by a corresponding linear transformation within the three first equations of the first model.

□

Example 3:

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & \beta_{16} & \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & 1 & \beta_{23} & 0 & 0 & \beta_{26} & \gamma_{21} & 0 & \gamma_{23} & 0 & 0 & 0 \\ \beta_{31} & 0 & 1 & 0 & 0 & \beta_{36} & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ \beta_{41} & \beta_{42} & 0 & 1 & \beta_{45} & \beta_{46} & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ 0 & 0 & \beta_{53} & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ \beta_{61} & 0 & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

is observationally equivalent to :

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & \beta_{16} & \gamma_{11} & 0 & \beta_{13} & 0 & 0 & 0 \\ \beta_{21} & 1 & 0 & 0 & 0 & \beta_{26} & 0 & \gamma_{22} & \gamma_{23} & 0 & 0 & 0 \\ 0 & \beta_{32} & 1 & 0 & 0 & \beta_{36} & 0 & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ \beta_{41} & \beta_{42} & 0 & 1 & \beta_{45} & \beta_{46} & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ 0 & 0 & \beta_{53} & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ \beta_{61} & 0 & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

These two models are overidentified interdependent models. Also here, we have a partial exactly identified sub block in the first three equations. Similarly, we can always get the second model from a linear transformation in the first three equations of the first model. □

Doing model selection, we make the judgement: from which model a given set of observed data are generated. It is impossible to make such a judgement between two observationally equivalent models, because the likelihood of two observationally equivalent models are exactly the same. Hence, from observed data we can only identify the observationally equivalent group. If we identify a single model from the observed data, this model should not have any observationally equivalent models but itself. This is the motivation for the definition of observational differentiability.

3.3 Observational Differentiability

Definition 3.15 (Observational Differentiability) *A model is called observationally differentiable, if it has no observationally equivalent models but itself.*

An observational differentiable model can be identified from the observed data. Just as the concept of identification guarantees the uniqueness of parameter estimation, the concept of observational differentiability guarantees the uniqueness of a solution for model selection. It is of interest now to ask the question: what is the condition for model to have no observationally equivalent models?

To conduct a condition for observationally differentiable models, we introduce the concept of **partial identification**. In an identifiable structural model, not every variable would appear in one equation, otherwise this equation would not be identified. If we look at a part of a model, say $g \leq G$ equations, usually not every variable of the model appears in this partial model with g equations. If we apply the identification criterion i.e. the rank condition and the order condition to this partial model and take only those variables into account that appear in this partial model, we may assess whether each equation is underidentified, exactly identified, or overidentified within this partial model.

Example:

$$\begin{pmatrix} 1 & \beta_{12} & 0 & 0 & 0 & 0 & \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & 1 & \beta_{23} & 0 & \beta_{25} & 0 & \gamma_{21} & 0 & \gamma_{23} & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 1 & 0 & 0 & 0 & \gamma_{31} & \gamma_{32} & \gamma_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \beta_{45} & 0 & 0 & 0 & 0 & 0 & \gamma_{45} & \gamma_{46} \\ 0 & 0 & 0 & 0 & 1 & \beta_{56} & 0 & 0 & 0 & \gamma_{54} & 0 & \gamma_{56} \\ 0 & 0 & 0 & \beta_{64} & 0 & 1 & 0 & 0 & 0 & \gamma_{64} & \gamma_{65} & 0 \end{pmatrix}$$

The first three equations consist of a partial model. The variables $(y_{4t}, y_{6t}, x_{4t}, x_{5t}, x_{6t})$ do not appear in this partial model. The first equation is partial overidentified, the second is partially exactly identified and the third is partially underidentified.

Corresponding to a partial model, there is a zero block in the matrix (B, Γ). The number of rows of this zero block corresponds to the number of equations in the partial model, the number of columns of this zero block is the number of variables that are excluded from this partial model.

Theorem 3.16 (Conditions for the existence of observationally equivalent models) *If and only if there exists a partial exactly identified equation i in the model,*

the model has observationally equivalent models. Formally this condition can be stated as follows:

$$z_{gi} - m_{gi} = g - 1$$

z_{gi} *: the number of zeros in the i-th equation of a partial model.*

m_{gi} *: the number of columns of the zero block of the partial model.*

g *: the number of rows of the zero block of the partial model.*

Proof:

Necessity:

Suppose a Model $(\tilde{B}, \tilde{\Gamma})$ is observationally equivalent to (B, Γ)

According to 3.11 for any given structure of (B, Γ) there exists a full rank matrix $A \neq I$ such that

$$A(\tilde{B}, \tilde{\Gamma}) = (B, \Gamma),$$

where $(\tilde{B}, \tilde{\Gamma})$ is the observationally equivalent structure with respect to the structure (B, Γ).

Without loss of generality, we assume that the diagonal elements of A are not zero[24].

We denote the number of zeros in the i-th equation of the model (B, Γ) by Z_i and the number of zero of the i-th equation of $(\tilde{B}, \tilde{\Gamma})$ by $\tilde{Z}_i$.

Because the total number of zeros in the observationally equivalent models are equal, we can find some equation i, with: $Z_i \geq \tilde{Z}_i$.

For this i-th equation of the structure (B, Γ) we have:

$$A_i(\tilde{B}, \tilde{\Gamma}) = (B, \Gamma)_i$$

Denote the columns of the matrix $(\tilde{B}, \tilde{\Gamma})$ that correspond to the zero elements in $(B, \Gamma)_i$ by $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$.

It holds:

[24]Because A is of full rank, the determinant of A does not equal to zero, i.e. at least one product that consists of elements of A from different rows and columns of A is nonzero. We can rearrange the rows of A according to the order of the column index of the factors of this product and get a matrix whose diagonal element are not zero. The rearrangement of rows of A will not change its observationally equivalent property.

$$A_i(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i = 0 \tag{3.5}$$

where $(\tilde{B}, \tilde{\Gamma})^i$ is a $G \times Z_i$ matrix, and $Z_i > G-1$ (owing to the identification condition for (B, Γ)).

Because a matrix $A \neq I$ must exist, it follows that the equations system (3.5) must have a non-zero solution A_i. This implies that the following rank condition must be satisfied:

$$rank(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})_i \leq G - 1. \tag{3.6}$$

If $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ contains a pure zero row for all $i = 1, 2, ..., G$, then $(\tilde{B}, \tilde{\Gamma})$ and (B, Γ) are identical models, because they have the same number of zeros and at the same position. $(\tilde{B}, \tilde{\Gamma})$ and (B, Γ) may differ only in the order of equations in the system.

Because $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ is a $G \times Z_i$ matrix, the rank of $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ would be G if there were no zero block in it.

If $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ does not contain a pure zero row for some i, and the rank condition 3.6 must be satisfied, then $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ must contain a $(\tilde{l}_i \times \tilde{m}_i)$ zero block, such that

$$Rank(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i = (G - \tilde{l}_i) + (Z_i - \tilde{m}_i) \leq G - 1.$$

where $\tilde{l}_i$ is the number of rows and $\tilde{m}_i$ is the number of columns of the zero block in $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$.

This rank condition can equivalently be put as:

$$Z_i + 1 \leq \tilde{l}_i + \tilde{m}_i$$

We observe that in calculation of $A_i(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i = 0$ the rows in the $\tilde{l}_i \times \tilde{m}_i$ zero block of $(\tilde{B}^{\triangle\triangle}, \tilde{\Gamma}^{**})^i$ correspond to the nonzero elements in A_i, while the rows outside the zero block corresponds to the zero elements in A_i. Hence, the i-th equation must be within the zero block, because $a_{ii} \neq 0$.

Denote the number of rows of the zero block in $(\tilde{B}, \tilde{\Gamma})$ by $\tilde{l}$ and the number of columns by $\tilde{m}$.

Because $Z_i \geq \tilde{Z}_i$, it follows

$$\tilde{l} + \tilde{m} \geq \tilde{l}_i + \tilde{m}_i \geq Z_i + 1 > \tilde{Z}_i + 1.$$

Note that $\tilde{Z}_i > \tilde{m}$ there exists a $\tilde{g}$ with $0 < \tilde{g} < \tilde{l}$ such that it holds:

$$\tilde{g} + \tilde{m} = \tilde{Z}_i + 1,$$

or equivalently:

$$\tilde{Z}_i - \tilde{m} = \tilde{g} - 1.$$

This is the condition for partial exact identification.

Sufficiency: If there exists a partial system with a partial exactly identified equation in this system, then a linear transformation within this partial system that transfers this exactly identified equation into the reduced form (and then to other structural forms) will lead to an observationally equivalent model. □

Corollary 3.17 (Condition for observational differentiability) *If there is no partial exactly identified equation in a structural model, the model is observationally differentiable.*

The condition for observational differentiability makes it possible for us to check if we can identify a unique most parsimonious model from the observed data.

4 Model Selection Problems

4.1 Basic Assumptions of the Model Selection Problem

4.1.1 Model Selection without Misspecification

The model selection problem for structural models can be described as follows: given a set of well defined candidate models $\{M_i, i = 0, 1, 2, ...C\} = \mathcal{M}$ and a set of given data of exogeneous variable $\{\xi_t\}_{t=1}^T$ and a set of observed data of endogenous variable $\{y_t\}_{t=1}^T$ that is generated from one of these models, the problem is to find out the true model that generated the data.

We assume:

- The unconstrained reduced form is within the candidate models, so that we have always at least one admissible model.
- The unconstrained reduced form is correctly specified, i.e. the lags of the predetermined variables are correctly specified.
- The true model is within the set of candidate models under consideration.
- The data are infinite many. This assumption is because we are also interested in the asymptotic property of the model selection problem.

4.1.2 Model Selection with Misspecification

In the context of model selection for structural models misspecification may take two different forms. Firstly, the basic settings of the model are not correct, the lags of the predetermined variables may be incorrect, the distribution of the disturbance may be nonnormal, etc. Secondly, the true model may not be included in the set of alternative models or the restriction on the true structure may have a form other than zero restrictions.

We will discuss the misspecification problem in section 5.

4.2 Principles for Model Selection

4.2.1 The Maximum Likelihood Principle

Structural econometric models are defined as linear simultaneous equations models with normal disturbance. In this parametric setting, a natural approach to identifying the true model is using the maximum likelihood principle. We may calculate the maximum likelihood function value for each

alternative model. The large value of the likelihood function should provide evidence for appropriateness of the model.

In the case of iid observations Jensen's inequality and the law of large numbers provide a justification for the application of the maximum likelihood principle[25]. For the model selection problem of structural models we are actually dealing with dependent observations. Jensen's Inequality is not directly applicable. We have its asymptotic counterpart[26]:

$$\lim_{T\to\infty}\frac{1}{T}E\log L_T(\theta_0) > \lim_{T\to\infty}\frac{1}{T}E\log L_T(\theta) \qquad \text{for } \theta\neq\theta_0 \tag{4.7}$$

where $\log L_T(\theta_0)$ is the log likelihood function as defined in (B.25), evaluated at the true parameter $\theta_0 = (\bar{B}_0, \bar{\Gamma}_0, \bar{\Sigma}_0)$ and $\theta = (B, \Gamma, \Sigma)$.[27]

Under general assumptions of structural models (See A.1) the maximum likelihood estimate (MLE) is (strongly) consistent[28], and the law of large number (LLN) holds for the log likelihood function. This implies that for an admissible model $(B_i, \Gamma_i, \Sigma_i)$ we have:

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) = \lim_{T\to\infty} E\frac{1}{T}\log L_T(B_0, \Gamma_0, \Sigma_0)$$

For a nonadmissible model $(B_j, \Gamma_j, \Sigma_j)$, MLE will converge to the pseudo true parameter $(\bar{B}_j, \bar{\Gamma}_j, \bar{\Sigma}_j)$ that is different from the true parameter $(B_0, \Gamma_0, \Sigma_0)$.

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_j, \hat{\Gamma}_j, \hat{\Sigma}_j) = \lim_{T\to\infty} E\frac{1}{T}\log(L_T(\bar{B}_j, \bar{\Gamma}_j, \bar{\Sigma}_j)$$

It follows from (4.7):

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) > \operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_j, \hat{\Gamma}_j, \hat{\Sigma}_j)$$

or equivalently for $T > T_0$:

$$P\left(\frac{1}{T}\log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) > \frac{1}{T}\log L_T(\hat{B}_j, \hat{\Gamma}_j, \hat{\Sigma}_j)\right) \to 1 \tag{4.8}$$

[25] See Amemiya (1985) p. 115

[26] For proof see appendix Lemma 2.9

[27] This condition is a basic condition for the application of the maximum likelihood estimation. It is known as the asymptotically identifiable condition, or identifiable uniqueness. See Davidson and Mackinnon (1993a) p.259 and Pötscher and Prucha (1997) p.16 for details. For proof of this condition for structural models see Lemma 2.9.

[28] See (2.4) for proof.

Hence all nonadmissible models will ultimately have a smaller likelihood function value than admissible models. Under the assumption that the true model is in the set of candidate models, the reduced form (2.1) is an admissible model with respect to the true model. Therefore we can identify nonadmissible models by comparing their average likelihood with that of the unconstrained reduced form.

For all admissible models under investigation MLE will converge to the true parameter and their average likelihood function will converge to the same value. Thus we can identify all admissible models by comparing their likelihood to that of the unconstrained reduced form. In this way we can find the group of admissible models from the candidate set. But, we cannot use the average log likelihood function value to identify the true mode M_0. Because the average log likelihood function value cannot differ it from the other admissible ones. What is even worse is that the more overparameterized models will have larger average maximum likelihood than the parsimonious models in finite sample, because their maximum is chosen from a larger domain than the parsimonious ones.

We observed that both M_0 and the unconstrained reduced form are admissible models with respect to the true structure $(\bar{B}_0, \bar{\Gamma}_0)$; the difference is only that the number of free parameters of M_0 is not larger than that of the reduced form. This relation holds not only between M_0 and the unconstrained reduced form but also between M_0 and all other admissible models with respect to M_0. Hence we will find the true model M_0 by looking for the most parsimonious admissible model. If the solution is unique then we find the unique true model. If the solution is not unique, then we will have many true models which are indifferentialable from the observed data.

The well known Akaike information criterion[29] seems to provides a solution to this problem.

4.2.2 AIC Principle

To overcome the problem of model selection by maximum likelihood, AIC maximizes the relative entropy over all alternative models.

AIC results in a modification of the maximum likelihood criterion by subtracting the number of the free parameters of the model from the maximum of the log likelihood function[30].

$$AIC = \log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) - J_i$$

[29] See Akaike (1973)

[30] For an interesting derivation of AIC see Tong (1990)

It seems that this criterion may solve the problem of overfitting: if the first terms in AIC were equal, the second terms would become decisive and AIC would prefer the model with less parameters.

However, this intuition does not work asymptotically because the first term in the AIC converges to infinity with the growing sample size, while the second term remains constant. The difference in the first terms of the AIC may overwhelm the difference in the second terms. It is shown that the AIC is not consistent in the sense that the AIC will choose the overparameterized model with positive probability[31].

4.2.3 Consistent Criterion

A model selection criterion is defined as a function $\Phi : (\mathbb{R}^G)^{\mathbb{N}} \times \mathcal{M} \to \mathbb{R}$, $\mathcal{M}$ is the set of all candidate models. $(\mathbb{R}^G)^{\mathbb{N}}$ is the space of the random variable $\mathbf{Y}_T$, and a model M_i will be selected by the criterion if

$$\Phi(\mathbf{Y}_T, M_i) \geq \max_{M_j \in \mathcal{M}} \Phi(\mathbf{Y}_T, M_j)$$

A model selection criterion is called consistent, if it has the following property[32]:

$$\lim_{T \to \infty} P\left(\Phi(M_0, \mathbf{Y}_T) \geq \max_{M_j \in \mathcal{M}} \Phi(M_i, \mathbf{Y}_T)\right) = 1$$

The rationale behind this definition is that for a consistent criterion the probability to choose the true model will converge to 1 with growing sample size.

Based on the discussion in the last sections we know that the penalty added to the likelihood function in the AIC is too small from the point of view of a consistent criterion. Hence the AIC will choose overparameterized models with positive probability. We need to increase the penalty on the number of parameters to get a consistent criterion. The value of the maximum likelihood function values depends, on one hand, on the number of parameters k and, on the other hand, on the number of observations T. A penalty that may have a consistent property will depend on both T and k.

To stimulate the discussion we modify the penalty term in the AIC by a product of a function in T and k to see which kind of property the penalty should have so that we can have a consistent selection criterion. We denote this modified criterion as S.

[31] See Shibata (1976)

[32] Compare Schlittgen and Streitberg (1999)

$$S = \log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) - f(T)k_i$$

We look at the difference of S values between a model M_i with k_i parameters presented by $(B_i, \Gamma_i, \Sigma_i)$ and the true model M_0 with k_0 parameters presented by $(B_0, \Gamma_0, \Sigma_0)$ to see how can we get a consistent criterion:

$$\frac{1}{T}(S_0 - S_i) = \frac{1}{T}\log L_T(\hat{B}_0, \hat{\Gamma}_0, \hat{\Sigma}_0) - \frac{1}{T}\log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) - \frac{f(T)}{T}(k_0 - k_i)$$

If M_i is nonadmissible, the first difference on the RHS will converge to a positive number. If $\frac{f(T)}{T}$ converges to zero, the criterion S will choose the true model asymptotically. If M_i is admissible with respect to M_0, the first difference on the RHS will converge to zero. If $\frac{f(T)}{T}$ converges more "slowly" than the difference in likelihood, the second term will be dominant, the criterion will be positive, and it chooses also the true model asymptotically.

Theorem 4.1 (Consistent model selection criterion) *Suppose that M_0 is the true model with k_0 free parameter, $M_i \in \mathcal{M}$ is one of the candidate models with k_i free parameters. Suppose furthermore:*

- *A1: The true model is within the candidate set.*
- *A2: The likelihood function satisfied the condition given in (4.7)*
- *A3: The log likelihood ratio between M_0 and an admissible M_j has a well defined asymptotic distribution over $(0, +\infty)$: $\lim_{T\to\infty}(\log L_T(\theta_j) - \log L_T(\theta_0)) \to D(k_0, k_j)$ $D(k_0, k_j)$ is density function over $(0, +\infty)$.*

An information criterion:

$$\Phi(M_i, \mathbf{Y}_T) = L_T(\hat{\theta}_i) - f(T)k_i \tag{4.9}$$

is consistent if and only if it holds:

$$\lim_{t\to\infty} f(t) = +\infty \tag{4.10}$$

$$\lim_{t\to\infty} \frac{f(t)}{t} = 0. \tag{4.11}$$

Proof:

Sufficiency:

Supposed model M_i is not admissible with respect to the true M_0, we calculate the difference of the selection criterion between M_i and the true model M_0:

$$\begin{aligned} & \log L_T(\hat{\theta}_0) - f(T)k_0 - \log L_T(\hat{\theta}_i) + f(T)k_i \\ = \quad & T\left(\tfrac{L_T(\hat{\theta}_0)}{T} - \tfrac{L_T(\hat{\theta}_i)}{T} + \tfrac{f(T)}{T}(k_i - k_0)\right) \end{aligned}$$

$$P[\log L_T(\hat{\theta}_0) - f(T)k_0 > \log L_T(\hat{\theta}_i) - f(T)k_i] = P\left(\frac{L_T(\hat{\theta}_i)}{T} - \frac{L_T(\hat{\theta}_0)}{T} < \frac{f(T)}{T}(k_i - k_0)\right)$$

To show the consistence of the criterion (4.9), we need only to show that the probability of the RHS of the equation above converges to unit. Because M_i is not admissible, it follows from A2 and (4.8): for $T \to \infty$ and some $\delta > 0$:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T} \log L_T(\hat{\theta}_0) \geq \operatorname*{plim}_{T\to\infty} \frac{1}{T} \log L_T(\hat{\theta}_i) + \delta.$$

Using the condition (4.11) we have for $T \to \infty$:

$$\operatorname*{plim}_{T\to\infty} \left(\frac{1}{T} \log L_T(\hat{\theta}_0) - \frac{1}{T} \log L_T(\hat{\theta}_i)\right) \geq \delta > \lim_{T\to\infty} \frac{f(T)}{T} = 0.$$

It follows then:

$$\lim_{T\to\infty} P\left(\Phi(M_0, \mathbf{Y}_T > \Phi(M_i, \mathbf{Y}_T)\right) = 1$$

If now M_i is admissible with respect to M_0, it holds $k_1 > k_0$.

We look at the following event:

$$\log L_T(\hat{\theta}_i) - f(T)k_i > \log L_T(\hat{\theta}_0) - f(T)k_0 \iff \log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > f(T)(k_i - k_0)$$

Because A3 and $f(T) \to \infty$ for $T \to \infty$, for any $\epsilon > 0$ there exists a T_0 such that for $T > T_0$:

$$P[D(k_1, k_0) > f(T)(k_1 - K_0)] < \epsilon/2$$

and

$$P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > f(T)(k_1 - K_0)] - P[D(k_1, k_0) > f(T)(k_1 - K_0)] < \epsilon/2.$$

Combine the two inequalities above, we have:

$$P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > f(T)(k_i - k_0)] < \epsilon$$

or equivalently

$$\lim_{T\to\infty} P[\log L_T(\hat{\theta}_0) - f(T)k_0 > \log L_T(\hat{\theta}_i) - f(T)k_i] = 1.$$

Necessity:

For an admissible model M_i with $k_i > k_0$,

$$\lim_{T\to\infty} P[\log L_T(\hat{\theta}_0) - f(T)k_0 > \log L_T(\hat{\theta}_i) - f(T)k_i] = 1$$

implies:

$$\lim_{T\to\infty} P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > f(T)(k_i - k_0)] = 0. \tag{4.12}$$

From (A3) we have

$$\lim_{T\to\infty} P[2(\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0)) > J] > 0 \qquad \text{for any } J > 0. \tag{4.13}$$

If $f(t)$ would be bounded from above, we could find an L such that $f(t) \leq L = J/(k_i - k_0)$.

$$\begin{aligned}
& \lim_{T\to\infty} P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > f(T)(k_i - k_0)] \\
\geq\ & \lim_{T\to\infty} P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > L(k_i - k_0)] \\
=\ & \lim_{T\to\infty} P[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) > J] \\
=\ & P[D(k_1, k_0) > J] \\
>\ & 0
\end{aligned}$$

This contradicts (4.12). Hence it must hold $f(t) \to \infty$ for $t \to \infty$.

For any non-admissible model M_i, the consistent criterion:

$$\lim_{T\to\infty} P[\log L_T(\hat{\theta}_0) - f(T)k_0 > \log L_T(\hat{\theta}_i) - f(T)k_i] = 1.$$

implies:

$$P \lim_{T\to\infty} [\frac{1}{T}\log L_T(\hat{\theta}_0) - \frac{1}{T}\log L_T(\hat{\theta}_i) > -\frac{1}{T}f(T)(k_i - k_0)] = 1.$$

If $\frac{1}{T}f(T)$ would have a lower bound c with $\frac{1}{T}f(T) > c$, we could construct such a true structural $(\bar{B}, \bar{\Gamma}, \bar{\Sigma}) \in M_0$ by fixing a parameter in M_0, say β_k so close to zero such that the difference of the average likelihood between M_0 and M_i that is achieved by the set that β_k to zero is smaller than $c(k_i - k_0)$:

$$\frac{1}{T}\log L_T(\hat{\theta}_0) - \frac{1}{T}\log L_T(\hat{\theta}_i) < c(k_i - k_0).$$

This contradicts to the consistency of the criterion: it follows for $t \to \infty$:

$$\frac{1}{T}f(T) \to 0.$$

□

In the practical application the penalty function $f(T)$ has to be concretely specified. The BSC and HQ criteria for linear regression models and ARMA models are two examples of such consistent criterion.

The BSC criterion:

$$BSC = \log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) - k_i \log T$$

The HQ criterion:

$$HQ = \log L_T(\hat{B}_i, \hat{\Gamma}_i, \hat{\Sigma}_i) - 2 * C * k_i \log\log T \qquad \text{with } C > 1$$

The rate of convergence to zero of the penalty term $f(T)/T$ is essential for the property of the selection criterion. A penalty that does not converge to zero may choose the nonadmissible models, while a penalty that converges to zero too fast may choose overparameterized models. Slower rates of convergence give a bigger penalty to the number of parameters and hence tend to choose a model with less parameters, while fast rate of convergence give smaller penalties and hence tend to prefer models with more parameters.

4.2.4 Inconsistence of AIC

The penalty added to the AIC is a constant $f(T) = 1$. It follows from the theorem above that the AIC is inconsistent for ARMA model selection. The two conditions (4.10) and (4.11) can be interpreted as conditions of consistency in selection against non-admissible and against admissible models respectively. More precisely, we know that $1/T \to 0$ as $T \to \infty$, hence the penalty of the AIC does not violate the consistent condition in selection against non-admissible models i.e. the probability for the AIC to choose a model that is too short converges to zero; but it violates the consistent condition in selection against admissible models. Therefore the AIC tends to choose longer models[33].

4.3 Hypothesis Test vs. Model Selection Criterion

4.3.1 Two Aspects of One Stochastic Process

We consider a stochastic process, say, Brownian motion $\{W_t\}_1^\infty$. To study the property of W_t we may look at the distribution of the process at time t. The distribution of W_t at time t describes the distribution of the realizations of different paths of the stochastic process at time t.

$$W_t \sim N(0, t)$$
$$\frac{W_t^2}{t} \sim \chi^2(1)$$

Based on these hypothetical distributions and realizations of W_t, we may make statistical inferences on the underlying stochastic property of W_t.

Another way of studying the stochastic process is to follow one path of W_t and look at how the path can be described. The iterated law of logarithm is one such result:

$$\limsup_{T\to\infty} \frac{W_t}{(2T \log\log T)^{0.5}} = 1$$

$$\liminf_{T\to\infty} \frac{W_t}{(2T \log\log T)^{0.5}} = -1$$

In these two equations it is understood that ultimately the Brownian motion will be bounded within the area described by $(-(2T \log\log T)^{0.5}(1 + \epsilon), (2T \log\log T)^{0.5}(1 + \epsilon))$.

[33] See Schlittgen and Streitberg (1999) p. 335 -338 for more discussion.

Based on these hypothetical bounds and the realization of W_t we can also make statistical inferences on the underlying property of W_t.

For empirical studies, especially for econometric analysis, the second aspect is more relevant, because economic time series are usually not repeatable. Most economic data are a single realization path of a stochastic process. Hence, statement based on such a single realization is more relevant for the analysis.

4.3.2 χ^2 Test vs. Consistent Criteria

In the context of model selection, these two aspects of a stochastic process lead to two different approaches to model selection: the approach based on statistical tests and the approach of consistent criteria.

Suppose that the likelihood ratio between an admissible model M_i and the true model M_0 is distributed as follows: (Here M_0 has k_0 parameters and M_i has k_i parameters and we take M_0 as M_i under r restrictions on the parameter.)

$$2(\log L_T(\theta_i) - \log L_T(\theta_0)) \sim W_{1T}^2/T + W_{2T}^2/T + ... + W_{rT}^2/T \tag{4.14}$$

where W_{it} is Brownian motion, r is the number of restrictions imposed by M_0 on M_i and T is the number of observations.

We may view this likelihood ratio as a stochastic process. If we look at the distribution of this likelihood ratio at time T, it is χ^2 distributed:

$$2[\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0)] \sim \chi^2(r)$$

The expression above is the well known result of the likelihood ratio test.

In the framework of the likelihood ratio test, we look at the realization of this test statistics and compare it with a critical value chosen at a significant level, say, $\alpha = 0.05$. We will reject M_0, if the realized statistic is larger than the critical value, otherwise we will not reject M_0 and chose M_0 as the true model.

If we look at the path of stochastic process, we have:

$$P\left(W_{iT} < (2T \log\log T(1+\epsilon))^{\frac{1}{2}}, \text{ for } T \text{ large enough}\right) = 1$$

$$P\left(W_{iT}^2/T < (2\log\log T(1+\epsilon)), \text{ for } T \text{ large enough}\right) = 1$$

We add all r Brownian motions together:

$$P\left(W_{1T}^2/T + W_{2T}^2/T + ... + W_{rT}^2/T < (2r \log\log T(1+\epsilon)),\ \text{for } T \text{ large enough}\right) = 1.$$

Using (4.14) we have:

$$P\left((\log L_T(\hat{\theta}_i) - \log L_T(\hat{\theta}_0) < r(1+\epsilon)\log\log T),\ \text{for } T \text{ large enough}\right) = 1$$

$$P\left((\log L_T(\hat{\theta}) - k_0 C \log\log T > \log L_T(\hat{\theta}_i) - k_i C \log\log T),\ \text{for } T \text{ large enough}\right) = 1$$

So we can define the model selection criterion:

$$\Phi(i) = \log L_T(\hat{\theta}_i) - Ck_i \log\log T$$

where $C > 1$. This criterion $\Phi(i)$ is strongly consistent. This means that if there are enough data on a single path, the criterion will choose the true model with probability 1. This strongly consistent criterion implies consistency in probability. Hence, if M_i is not admissible with respect to M_0 this criterion will not choose M_i.

Now we may compare the two approaches from both conceptual and practical aspects. Conceptually, the test approach looks at the distribution under H_0, and controls the first type error, i.e. the probability to reject the true model. Some problems are associated with this approach:

The conclusion from non-rejection of M_0 to the claim that M_0 is the true model is based on a very strong confidence in M_0. In the case of the existence of alternative models it is very questionable why one model should have such strong confidence.

Once this procedure is applied to alternative models, one may get conflicting results that more than one model can be claimed to be the true model.

Furthermore, this procedure will have a small probability to reject the true model, and it provides no information about the probability to choose a false models or an overparameterized model.

The approach of consistent criterion look at the likelihood function value with a suitable penalty. All models are treated equally, and ranked with the selection criterion. If we have enough data we will choose the true model with probability 1. The problem with this approach is that the conclusion hold only asymptotically.

Practically, we may compare the χ^2 test approach with the approach of consistent model selection criterion through comparing their conclusion with respect to the final conclusion on the model selection for practically relevant numbers of observations and numbers of restrictions.

The value of $2 \log \log T$ will converge to ∞, hence for sufficiently large T we always have:

$$2(\log L_T(\theta_i) - \log L_T(\theta_0)) < \chi^2(r) < 2r(1+\epsilon) \log \log T$$

This means that if the χ^2 test will accept M_0 the consistent criterion will also choose M_0. This can be seen as the preference of consistent criterion for parsimonious models against the χ^2 test.

This tendency prevails even for practical relevant numbers of observations. For $50 < T < 2000$ $2 \log \log T$ is between $(2.72, 4.05)$ The ratio of the critical value for a significance level of 0,05 to the degree of freedom of χ^2 distribution is $\chi^2(r)/r < 2.6$ for $r > 2$

Hence we have for most case ($r > 2$, $T > 50$):

$$2\left(\log L_T(\theta_i) - \log L_T(\theta_0)\right) < \chi^2(r)_{0.05} = r\left(\frac{\chi^2(J)_{0.05}}{r}\right) < 2r(1+\epsilon) \log \log(T),$$

or equivalently

$$\log L_T(\theta_i) - k_i(1+\epsilon) \log \log T < \log L_T(\theta_0) k_0(1+\epsilon) \log \log T.$$

This means that in most cases if χ^2 test will accept M_0 the consistent criterion will also choose M_0.

The reason is that to achieve consistency the model selection criterion puts a higher penalty on the likelihood function. Consequently, it prefers a more parsimonious model comparing to χ^2 test. The rejection of the false model is controlled by the $f(t)/t \to 0$.

5 A Model Selection Criterion for Structural Models

In the last chapter we have provided a general condition for consistent model selection criterion. In this chapter we are going to give a strongly[34] consistent model selection criterion for the model selection problem of structural models.

5.1 A Consistent Selection Criterion for Multiple Regression Models

We consider a regression model:

$$Y_t = X_t\beta + U_t$$

where $Y_t \in \mathbb{R}$ is a dependent variable called the regressant, $X_t \in \mathbb{R}^k$ is a deterministic independent variable called the regressor. U_t is random disturbance with $U_t iid \sim N(0, \sigma^2)$. We have T observations of Y_t and X_t denoted by a $T \times 1$ matrix y, and $T \times k$ matrix x respectively. Accordingly, T realizations of U_t is denoted by a $T \times 1$ matrix u.

There are many potential regression models characterized by different regressors. A regression model can be presented by its regressors X_{it}. $T \times k_i$ matrix x_i is the observations matrix of model M_i. The number of parameters is denoted by k_i. We assume that the real DGP can be described by a true model M_0 whose regressors denoted by X_{0t} is only a subset of X_t. The task is to find a strongly consistent model selection criterion to select the true model.

As discussed in Chapter 3, models that do not include all the true regressors will have a lower average value at the maximum of likelihood function according to Jensen's inequality. Hence $\Phi(i) = \frac{1}{T}\log L_T(\beta_i) + o(1)$ will not choose these models. The models that include the true regressors will have asymptotically the same average value at maximum likelihood and we have to look for a penalty that is of order $o(1)$ but is larger than the difference of the two average likelihood function values at upon comparison.

We look at the difference between the log likelihood of the true model M_0 with x_0 as regressor and an admissible model denoted as M_1 with $x_1 = (x_0, x_{11})$ as regressors. Obviously, it holds that $k_0 < k_1$. For T observations we have:

$$y = x_0\beta_0 + u \qquad \text{for } M_0 \tag{5.15}$$

[34]For the definition of strong consistency see Schlittgen and Streitberg (1999)

$$y = x_1\beta_1 + u = x_0\beta_{10} + x_{11}\beta_{11} + u \qquad \text{for } M_1 \tag{5.16}$$

For M_0 the log likelihood function evaluated at the maximum likelihood estimate $\hat{\beta}_0$ is:

$$\log L_T(\hat{\beta}_0) = -\frac{T}{2} - \frac{T}{2}\log(\frac{2\pi}{T}) - \frac{T}{2}\log(e_0'e_0),$$

with $e_0'e_0 = (y - x_0\hat{\beta}_0)'(y - x_0\hat{\beta}_0)$

For M_1 the log likelihood function is:

$$\log L_T(\hat{\beta}_1) = -\frac{T}{2} - \frac{T}{2}\log(\frac{2\pi}{T}) - \frac{T}{2}\log(e_1'e_1),$$

with $e_1'e_1 = (y - x_1\hat{\beta}_1)'(y - x_1\hat{\beta}_1)$.

The difference between the two likelihood functions is[35]:

$$\begin{aligned} 2(\log L_T(\hat{\beta}_1) - \log L_T(\hat{\beta}_0)) &= T(\log(e_0'e_0) - \log(e_1'e_1)) && (5.17)\\ &= T\frac{1}{\tilde{e}'\tilde{e}}(e_0'e_0 - e_1'e_1) && (5.18) \end{aligned}$$

with $\tilde{e}'\tilde{e} \in (e_0'e_0, e_1'e_1)$, such that the equality above holds.

According to the Kolmogorov strong law of large number we have: $\lim_{T\to\infty} e_1'e_1/T = \sigma^2$ and $\lim_{T\to\infty} e_0'e_0/T = \sigma^2$. It follows: $\lim_{T\to\infty} \tilde{e}'\tilde{e}/T = \sigma^2$.

Insert this into (5.18), we have:

$$2(\log L_T(\hat{\beta}_0) - \log L_T(\hat{\beta}_1)) \overset{a.a.s}{\sim} \frac{e_0'e_0 - e_1'e_1}{\sigma^2} \tag{5.19}$$

$\overset{a.a.s}{\sim}$ is defined in appendix (2.8). It reads asymptotically almost surely. We are now going to calculate the order of

$$\frac{e_0'e_0 - e_1'e_1}{\sigma^2}.$$

$$\begin{aligned} e_0'e_0 &= (y - x_0\hat{\beta}_0)'(y - x_0\hat{\beta}_0)\\ &= (y - x_0(x_0'x_0)^1x_0'y)'(y - x_0(x_0'x_0)^1x_0'y)\\ &= y'(I - x_0(x_0'x_0)^{-1}x_0)y\\ &= y'M_0y\\ &= (x_0\hat{\beta}_{10} + x_{11}\hat{\beta}_{11} + e_1)'M_0(x_0\hat{\beta}_{10} + x_{11}\hat{\beta}_{11} + e_1)\\ &= e_1'e_1 + (x_{11}\hat{\beta}_{11})'M_0(x_{11}\hat{\beta}_{11}) \end{aligned}$$

[35]The second equality is due to Taylor expansion

Under the assumption $\beta_{11} = 0$ it holds[36]:

$$\hat{\beta}_{11} = (x'_{11}M_0x_{11})^{-1}x'_{11}M_0u$$

Insert this into the equation above, we get:

$$e'_0e_0 - e'_1e_1 = (x_{11}\hat{\beta}_{11})'M_0(x_{11}\hat{\beta}_{11}) = u'M_0x_{11}(x'_{11}M_0x_{11})^{-1}x'_{11}M_0u = u'Au \tag{5.20}$$

Divide both sides of (5.20) by σ^2 we have:

$$\frac{e'_0e_0 - e'_1e_1}{\sigma^2} = (u/\sigma)'A(u\sigma). \tag{5.21}$$

The matrix $A = M_0x_{11}(x'_{11}M_0x_{11})^{-1}x'_{11}M_0$ is indempotent with rank k_{11}

$trA = tr(M_0x_{11}(x'_{11}M_0x_{11})^{-1}x'_{11}M_0) = tr((x'_{11}M_0x_{11})^{-1}(x'_{11}M_0x_{11})) = tr(I_{k11})$

Recall that u is the realization of $U \sim N(0, \sigma^2 I_T)$, $(U/\sigma)'A(U/\sigma)$ is $\chi^2(k_{11})$ distributed.[37] Insert this result into (5.19) we get the classic result:

$$2(\log L_T(\hat{\beta}_1) - \log L_T(\hat{\beta}_0)) \sim \chi^2(k_{11}) \tag{5.22}$$

The matrix $(x'_{11}M_0x_{11})^{-1}$ is positive definite; there exists an orthogonal matrix P, such that $(x'_{11}M_0x_{11})^{-1} = P'DP$, D being the diagonal $k_{11} \times k_{11}$ matrix of the eigenvalues of $(x'_{11}M_0x_{11})^{-1}$.

We have:

$$(x'_{11}M_0x_{11})^{-1} = P'D'^{\frac{1}{2}}D^{\frac{1}{2}}P$$

$$(x'_{11}M_0x_{11}) = P^{-1}D^{-\frac{1}{2}}D'^{-\frac{1}{2}}P'^{-1}$$

$$D^{\frac{1}{2}}P(x'_{11}M_0x_{11})P'D'^{\frac{1}{2}} = I \tag{5.23}$$

Insert (5.23) into the difference of the standardized residuals (5.21)we get:

$$\frac{e'_0e_0 - e'_1e_1}{\sigma^2} = (D^{\frac{1}{2}}Px'_{11}M_0(u/\sigma))'(D^{\frac{1}{2}}Px'_{11}M_0(u/\sigma)) \tag{5.24}$$

[36]See Frohn (1995) p. 30

[37]See Schmidt (1976) p. 11 for proof.

We observe that $D^{\frac{1}{2}}Px'_{11}M_0(U/\sigma)$ is a linear transformation of the iid standard normal distributed vector (U/σ), and hence it is a normally distributed random vector.

$$E(D^{\frac{1}{2}}Px'_{11}M_0(U/\sigma)) = 0$$

$$\begin{aligned} & Var(D^{\frac{1}{2}}Px'_{11}M_0(U/\sigma)) && (5.25)\\ = \; & D^{\frac{1}{2}}Px'_{11}M_0E(u/\sigma)(u/\sigma)'M_0x_{11}P'D'^{\frac{1}{2}} && (5.26)\\ = \; & D^{\frac{1}{2}}Px'_{11}M_0x_{11}P'D'^{\frac{1}{2}} && (5.27)\\ = \; & I && (5.28) \end{aligned}$$

We denote (U/σ) by V and the $k_{11} \times T$ matrix $D^{\frac{1}{2}}Px'_{11}M_0$ by Φ and insert them into (5.24):

$$\begin{aligned} & \frac{e'_0e_0 - e'_1e_1}{\sigma^2} && (5.29)\\ = \; & (D^{\frac{1}{2}}Px'_{11}M_0(u/\sigma))'(D^{\frac{1}{2}}Px'_{11}M_0(U/\sigma)) && (5.30)\\ = \; & (\Phi V)'(\Phi V) && (5.31)\\ = \; & \begin{pmatrix} \Phi_1 V \\ \Phi_2 V \\ \vdots \\ \Phi_{k_{11}} \end{pmatrix}' \begin{pmatrix} \Phi_1 V \\ \Phi_2 V \\ \vdots \\ \Phi_{k_{11}} \end{pmatrix} && (5.32)\\ = \; & \sum_{j=1}^{k_{11}} (\Phi_j V)^2 && (5.33)\\ = \; & \sum_{j=1}^{k_{11}} (\sum_{t=1}^{T} \Phi_{jt} V_t)^2 && (5.34) \end{aligned}$$

$\sum_{t=1}^{T} \Phi_{jt}V_t$ is the j-th component of vector ΦV and from (5.25) and (5.31)we know that $\sum_{t=1}^{T} \Phi_{jt}V \sim N(0,1)$.

Therefore,

$$\sum_{t=1}^{T} \Phi_{jt}V_t\sqrt{T} \sim N(0,T) \tag{5.35}$$

Now $\sum_{t=1}^{T} \Phi_{jt}V_t$ is a partial sum of a martingale difference sequence with independent normal increment. The law of iterated logarithm (LIL)[38] holds for this partial sum.

[38]Stautt, 1970, A Martingale Analogue of Kolmogorov's Law of Iterated Logarithm, Wahrscheinlichkeitstheorie und Verw. Gebiet, Vol. 15, p. 279-290.

$$\limsup_{T\to\infty} \frac{\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T}}{(2T \log\log T)^{\frac{1}{2}}} = 1$$

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T}}{(2T \log\log T)^{\frac{1}{2}}} = -1$$

These two equalities are interpreted: for any $\epsilon > 0$ the partial sum of the martingale difference sequence $\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T}$ lies with probability one infinitely often in the interval $(-2T \log\log T(1+\epsilon), 2T \log\log T(1+\epsilon))$ and only finitely often in the interval $(-\infty, -(2T \log\log T(1+\epsilon))^{0.5}) \cup ((2T \log\log T(1+\epsilon))^{0.5}, \infty)$.

We have:

$$P\left(\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T} < (2T \log\log T(1+\epsilon))^{\frac{1}{2}}, \text{ for } T \text{ large enough}\right) = 1$$

$$P\left(\left(\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T}\right)^2 < (2T \log\log T(1+\epsilon)), \text{ for } T \text{ large enough}\right) = 1$$

$$2(\log L_T(\hat{\beta}_0) - \log L_T(\hat{\beta}_1)) \overset{a.a.s}{\sim} \frac{1}{T} \sum_{j=1}^{k_{11}} \left(\sum_{t=1}^{T} \Phi_{jt} V_t \sqrt{T}\right)^2$$

It follows:

$$P\left((\log L_T(\hat{\beta}_0) - \log L_T(\hat{\beta}_1) < k_{11} C \log\log T, \text{ for } T \text{ large enough}\right) = 1$$

with $c = (1+\epsilon)$.

Note that $k_{11} = k_1 - k_0$, we have equivalently:

$$P\left(\log L_T(\hat{\beta}_0) - k_0 C \log\log T > \log L_T(\hat{\beta}_1) - k_1 C \log\log T, \text{ for } T \text{ large enough}\right) = 1.$$

So we can define

$$\Phi(i) = \log L_T(\hat{\beta}_i) - C k_i \log\log T$$

as the model selection criterion. This is exactly the same as the HQ criterion. For any $C > 1$ this criterion is strongly consistent.

5.2 A Consistent Selection Criterion for Structural Models

For the model selection problem for structural models as defined in Chapter 4. We have:

$$B_i Y_t + \Gamma_i X_t = U_t \qquad \text{i=0,1,2,...C}$$

$(B_i, \Gamma_i) \in \mathcal{M}$. $\mathcal{M}$ is the set of all candidate models under consideration. We assume the true model (B_0, Γ_0) is an element in this set. (For the uniqueness of selection we need to assume that the true model is observational differentiable.)

The procedure to find a consistent selection criterion is similar to the case of the multiple regression model.

We treat structural models as alternative representations of the restrictions imposed on the unconstrained reduced form. For all nonadmissible structural models the maximum of average log likelihood function values will be strictly smaller than that of admissible models, asymptotically. Hence, a model selection criterion that is dominated by (average) log likelihood function will not choose nonadmissible models. For all admissible models the maximum of their average log likelihood function will converge to the same value. Thus we have to find a penalty function $f(T)/T$ that will converge to zero as $T \to \infty$ and will converge to zero more slowly than the log likelihood ratio between an admissible model and the true model. (The difference in log likelihood is the same as log likelihood ratio.)

We assume that true model M_0 imposes only linear restrictions (zero restrictions) on the parameter of admissible models within the set of candidates. The likelihood ratio test theory says the likelihood ratio between M_0 and M_i is χ^2 distributed. This provides a hint that this difference can be seen as the quadratic form of some martingale sequences (MG)[39]. If we find such MG sequence we may apply the LIL to assess their rate of convergence. Based on the results of the LIL we may construct a consistent model selection criterion.

Without loss of generality we look at the relation between true model M_0 and an admissible model M_1. We will carry out the exposition in the following steps.

1. Treat true model M_0 as M_1 under linear restriction $R\theta = 0$ and apply Lagrange multiplier method to get the MLE under restriction.

2. Derive an asymptotically equivalent quadratic form for the log likelihood ratio between M_0 and M_1.

[39]See last section for example.

3. Show the components in the quadratic form are MG sequences
4. Apply the LIL to the MG sequences and to assess their rate of convergence
5. Construct a strongly consistent model selection criterion for structural models

The log likelihood function for model M_1 is:

$$\begin{aligned} & l_T(\theta) \\ = \; & \log L_T(B_1, \Gamma_1, \Sigma_1; \mathbf{y}_T, \mathbf{x}_T) \\ = \; & -\frac{TG}{2}\log(2\pi) + T\log\|B_1\| - \frac{T}{2}\log|\Sigma_1| - \frac{1}{2}tr(\Sigma_1^{-1}(\mathbf{y}_T B_1' + \mathbf{x}_T\Gamma_1')'(\mathbf{y}_T B_1' + \mathbf{x}_T\Gamma_1')) \end{aligned}$$

where $\theta = \text{vec}(B_1, \Gamma_1, \Sigma_1)$, θ is the vector of all unspecified unknown parameters in the matrix $(B_1, \Gamma_1, \Sigma_1)$. We denote the number of free parameters in M_0 and M_1 by k_0 and k_1 respectively. θ is a $k_1 \times 1$ vector. We assume that the parameters of M_0 can be obtained through a linear restriction on them:[40]:

$$R\theta = r. \tag{5.36}$$

We denote the partial derivative of the log likelihood function with respect to the parameter vector as follows:

$$d_T(\theta) = \frac{\partial l_T(\theta)}{\partial \theta}$$

$$D_T(\theta) = \frac{\partial^2 l_T(\theta)}{\partial \theta \partial \theta'}$$

Further we denote $\hat{\theta}$ the MLE under constraint (5.36) and $\tilde{\theta}$ the MLE without constraint. Hence, $\hat{\theta}$ and $\tilde{\theta}$ correspond to the MLE for M_0 and M_1 respectively.

[40]Generally the relation between parameters of an admissible model and parameters of a true model is not linear. Therefore the restriction on θ will be nonlinear. However, under some regularity conditions of the restricting function such as twice differentiable and uniformly continuous around the true parameter θ, we can get the same conclusion as in the case of a linear restriction. In the presentation here we consider only linear restriction for simplicity. the nonlinear case is considered in the appendix.

Now we look at the difference of the log likelihood functions by Taylor expansion of the log likelihood function at the unconstrained estimate:

$$\log L_T(\hat{\theta}) = \log L_T(\tilde{\theta}) + d_T(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) + \frac{1}{2}(\hat{\theta} - \tilde{\theta})' D_T(b(\hat{\theta}, \tilde{\theta}))(\hat{\theta} - \tilde{\theta})$$

for some $b(\hat{\theta}, \tilde{\theta}) \in (\hat{\theta}, \tilde{\theta})$, such that the equality above holds

Because $d_T(\tilde{\theta}) = 0$, it follows

$$2(\log L_T(\tilde{\theta}) - \log L_T(\hat{\theta})) = -(\hat{\theta} - \tilde{\theta})' D_T(b(\hat{\theta}, \tilde{\theta}))(\hat{\theta} - \tilde{\theta}) \tag{5.37}$$

From the Taylor expansion of the first derivative of the log likelihood function we have:

$$d_T(\hat{\theta}) = d_T(\tilde{\theta}) + D_T(b(\hat{\theta}, \tilde{\theta}))(\hat{\theta} - \tilde{\theta})$$

$b(\hat{\theta}, \tilde{\theta}) \in (\hat{\theta}, \tilde{\theta})$ in the last two equations above are not the same, but we are only interested in the asymptotic behaviour of $b(\hat{\theta}, \tilde{\theta})$s, which will converge to the same value. We use the same expression for each to keep notation simple.

Because $d_T(\tilde{\theta}) = 0$ we get:

$$(\hat{\theta} - \tilde{\theta}) = D_T(b(\hat{\theta}, \tilde{\theta}))^{-1} d_T(\hat{\theta})$$

Now we look at the Lagrange function for the constrained maximization problem:

$$\psi(\theta, \lambda) = l_T(\theta) + \lambda'(R\theta).$$

We derive the Lagrange function with respect to θ and λ and set them to zero:

$$R\theta = 0$$

$$\frac{\partial l_T(\theta)}{\partial \theta} + \frac{\partial \lambda' R\theta}{\partial \theta} = 0$$

Let $\hat{\lambda}$, $\hat{\theta}$ be the solution of the maximization problem. Then they satisfy the following equations:

$$R\hat{\theta} = 0 \tag{5.38}$$

$$d_T(\hat{\theta}) + R'\hat{\lambda} = 0 \tag{5.39}$$

Insert these two equations into the log likelihood ratio (5.37) we have:

$$2(\log L_T(\tilde{\theta}) - \log L_T(\hat{\theta})) = -\hat{\lambda}' R D_T(b(\hat{\theta},\tilde{\theta}))^{-1} D_T(b(\hat{\theta},\tilde{\theta})) D_T(b(\hat{\theta},\tilde{\theta}))^{-1} R' \hat{\lambda}$$

Under the assumption that the restriction (5.36) on the parameter vector θ is true, both $\hat{\theta}$ and $\tilde{\theta}$ will converge to the true parameter θ_0, and consequently $b(\hat{\theta},\tilde{\theta})$ will also converge to θ_0. Therefore

$$\left(\frac{1}{T} D_T(b(\hat{\theta},\tilde{\theta}))\right)^{-1} \left(\frac{1}{T} D_T(b(\hat{\theta},\tilde{\theta}))\right) \left(\frac{1}{T} D_T(b(\hat{\theta},\tilde{\theta}))\right)^{-1}$$

will converge to $\bar{D}(\theta_0)^{-1} = E(\frac{1}{T} D_T(\theta_0))^{-1}$, if the law of large number holds for $\frac{1}{T} D_T(b(\hat{\theta},\tilde{\theta}))$. Actually we have the following theorem:

Theorem 5.1 *Under the assumptions on the structural model as defined in Appendix (A.1) we have:*

$$2\big(l_T(\tilde{\theta}_T) - l_T(\hat{\theta}_T)\big) \overset{a.a.s}{\sim} -\frac{\hat{\lambda}'}{\sqrt{T}} R \bar{D}(\theta_0)^{-1} R' \frac{\hat{\lambda}}{\sqrt{T}} \tag{5.40}$$

Proof: See appendix theorem 2.8

The matrix $-R\bar{D}(\theta_0)^{-1}R'$ is an $r \times r$ symmetrical positive definite matrix with $r = k_1 - k_0$. There exists an orthogonal $r \times r$ matrix P, such that $-R\bar{D}(\theta_0)^{-1}R' = P'DP$, D being the $r \times r$ diagonal matrix of the eigenvalues of $-R\bar{D}(\theta_0)^{-1}R'$.

Insert this into the log likelihood ratio we get:

$$\begin{aligned}
& 2\big(l_T(\tilde{\theta}_T) - l_T(\hat{\theta}_T)\big) \\
\overset{a.a.s}{\sim} \; & -\frac{\hat{\lambda}'}{\sqrt{T}} R \bar{D}(\theta_0)^{-1} R' \frac{\hat{\lambda}}{\sqrt{T}} \\
= \; & \left(D^{\frac{1}{2}} P \frac{\hat{\lambda}}{\sqrt{T}}\right)' \left(D^{\frac{1}{2}} P \frac{\hat{\lambda}}{\sqrt{T}}\right) \\
= \; & \left(\begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \vdots \\ \vdots & \dots & \ddots & 0 \\ 0 & \dots & 0 & d_r \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_r \end{pmatrix} \frac{\hat{\lambda}}{\sqrt{T}}\right)' \left(\begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \vdots \\ \vdots & \dots & \ddots & 0 \\ 0 & \dots & 0 & d_r \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_r \end{pmatrix} \frac{\hat{\lambda}}{\sqrt{T}}\right) \\
= \; & \sum_{j=1}^{r} \left(d_j P_j \frac{\hat{\lambda}}{\sqrt{T}}\right)^2
\end{aligned}$$

According to Crowder (1976):

$$(-\bar{D}(\theta_0))^{-\frac{1}{2}} R' \frac{\hat{\lambda}}{\sqrt{T}} \to N(0, I_r)$$

$$-\frac{\hat{\lambda}}{\sqrt{T}}' R\bar{D}(\theta_0)^{-1} R' \frac{\hat{\lambda}}{\sqrt{T}} \to \chi^2(r)$$

$$\frac{\hat{\lambda}}{\sqrt{T}} \to N(0, -(R\bar{D}(\theta_0)^{-1} R')^{-1})$$

We have:

$$\lim_{T\to\infty} E(D^{\frac{1}{2}} P^{-1} \frac{\hat{\lambda}}{\sqrt{T}}) = 0$$

$$\begin{aligned} \lim_{T\to\infty} Var(D^{\frac{1}{2}} P \frac{\hat{\lambda}}{\sqrt{T}}) &= \lim_{T\to\infty} D^{\frac{1}{2}} P E(\frac{\hat{\lambda}}{\sqrt{T}} \frac{\hat{\lambda}}{\sqrt{T}}') P' D^{\frac{1}{2}} \\ &= D^{\frac{1}{2}} P (R\bar{D}(\theta_0)^{-1} R')^{-1} P' D'^{\frac{1}{2}} \\ &= I_r \end{aligned}$$

It follows:[41]

$$d_j P_j \frac{\hat{\lambda}}{\sqrt{T}} \to N(0, 1)$$

and

$$d_j P_j \hat{\lambda} \to N(0, T).$$

Now we look at $\hat{\lambda}$ and show this is a partial sum of the martingale difference sequence.

Following the notation of Amemiya (1985) [42], we derive the log likelihood function with respect to the unspecified elements of B, Γ and Σ respectively:

$$\frac{\partial \log L_T}{\partial \Gamma'} = -\mathbf{x}_T'(\mathbf{y}_T B' + \mathbf{x}_T \Gamma')\Sigma^{-1},$$

[41] Compare (5.35)

[42] Amemiya (1985) p. 233

$$\frac{\partial \log L_T}{\partial B'} = TB'^{-1} - \mathbf{y}_T'(\mathbf{y}_T B' + \mathbf{x}_T \Gamma')\Sigma^{-1},$$

and

$$\frac{\partial \log L_T}{\partial \Sigma} = -\frac{T}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(\mathbf{y}_T B' + \mathbf{x}_T \Gamma')'(\mathbf{y}_T B' + \mathbf{x}_T \Gamma')\Sigma^{-1}.$$

We evaluate the first derivative at the true parameter $(B_0, \Gamma_0, \Sigma_0)$

$$\frac{\partial \log L_T(\theta_0)}{\partial \Gamma'} = \mathbf{x}_T' \mathbf{u}_T \Sigma_0^{-1} = \mathbf{x}_{T-1}' \mathbf{u}_{T-1} \Sigma_0^{-1} + x_T u_T \Sigma_0^{-1},$$

$$\frac{\partial \log L_T(\theta_0)}{\partial B'} = TB_0'^{-1} - \mathbf{y}_T' \mathbf{u}_T \Sigma_0^{-1} = B_0'^{-1}(T-1)I - \mathbf{y}_{T-1}' \mathbf{u}_{T-1} \Sigma_0^{-1} + B_0'^{-1} - y_T' u_T \Sigma_0^{-1}$$

.

$$\begin{aligned} \frac{\partial \log L_T(\theta_0)}{\partial \Sigma} &= -\frac{T-1}{2}\Sigma_0^{-1} + \frac{1}{2}\Sigma_0^{-1}(\mathbf{y}_{T-1}B_0' + \mathbf{x}_{T-1}\Gamma_0')'(\mathbf{y}_{T-1}B_0' + \mathbf{x}_T\Gamma_0')\Sigma_0^{-1} \\ &\quad -\frac{1}{2}\Sigma_0^{-1} + \frac{1}{2}\Sigma_0^{-1}(y_T B_0' + x_T \Gamma_0')'(y_T B' + x_T \Gamma_0')\Sigma_0^{-1}. \end{aligned}$$

We calculate the conditional expectation.

$$\begin{aligned} & E_{T-1}\frac{\partial \log L_T(\theta_0)}{\partial \Gamma'} \\ = \;& E_{T-1}\mathbf{x}_{T_1}'(\mathbf{u}_{T-1})\Sigma_0^{-1} + E_{T-1}X_T' U_T \Sigma_0^{-1} \\ = \;& \mathbf{x}_{T-1}'(\mathbf{y}_{T-1}B_0{}' + \mathbf{x}_{T-1}\Gamma_0{}')\Sigma_0^{-1} \\ = \;& \frac{\partial \log L_{T-1}}{\partial \Gamma'}\Big|_{\theta_0}, \end{aligned}$$

$$\begin{aligned} & E_{T-1}\frac{\partial \log L_T(\theta_0)}{\partial B'} \\ = \;& B_0'^{-1}[E_{T-1}((T-1)I - (\mathbf{y}_{T-1}B_0)'\mathbf{u}_{T-1}\Sigma_0^{-1}) + E_{T-1}(I - (y_T B_0')' u_T \Sigma_0^{-1})] \\ = \;& (T-1)B_0'^{-1} - \mathbf{y}_{T-1}'\mathbf{u}_{T-1}\Sigma_0^{-1} + E_{T-1}(I - (u_T - x_T\Gamma_0')' u_T \Sigma_0^{-1}) \\ = \;& (T-1)B_0'^{-1} - \mathbf{y}_{T-1}'\mathbf{u}_{T-1}\Sigma_0^{-1} \\ = \;& \frac{\partial \log L_{T-1}}{\partial B}\Big|_{\theta_0}, \end{aligned}$$

and

$$\begin{aligned}
& E_{T-1}\frac{\partial \log L_T(\theta_0)}{\partial \Sigma} \\
= & -E_{t-1}\left(\frac{T-1}{2}\Sigma_0^{-1}+\frac{1}{2}\Sigma_0^{-1}(\mathbf{y}_{T-1}B'+\mathbf{x}_{T-1}\Gamma')'(\mathbf{y}_{T-1}B'+\mathbf{x}_T\Gamma')\Sigma^{-1}\right) \\
& +E_{T-1}(-\Sigma_0^{-1}+\frac{1}{2}\Sigma_0^{-1}(Y_TB'+X_T\Gamma')'(Y_TB'+X_T\Gamma')\Sigma^{-1}) \\
= & \frac{T-1}{2}\Sigma_0^{-1}+\frac{1}{2}\Sigma_0^{-1}(\mathbf{y}_{T-1}B_0'+\mathbf{x}_{T-1}\Gamma_0')'(\mathbf{y}_{T-1}B_0'+\mathbf{x}_T\Gamma_0')\Sigma_0^{-1} \\
= & \frac{\partial \log L_{T-1}}{\partial \Sigma}\Big|_{\theta_0}
\end{aligned}$$

Hence $d_T(\theta_0)$ is MG. We summarize this result in the following Theorem.

Theorem 5.2 (Martingale property of the first derivative) *For the structural model as defined in (A.1), the first derivative of the log likelihood function evaluated at the true parameter is a martingale.*

From the expansion of the first derivative of the log likelihood function at the constrained estimate θ_0 we get:

$$d_T(\hat{\theta}) = d_T(\theta_0) + D_T(b(\hat{\theta},\theta_0))(\hat{\theta}-\theta_0).$$

$$(\hat{\theta}-\theta_0) = D_T(b(\hat{\theta},\theta_0))^{-1}(d(\hat{\theta})-d(\theta_0))$$

Combine (5.36) and (5.38) we have:

$$R(\hat{\theta}-\theta_0) = RD_T(b(\hat{\theta},\theta_0))^{-1}(d_T(\hat{\theta})-d_T(\theta_0)) = 0.$$

$$RD_T(b(\hat{\theta},\theta_0))^{-1}d_T(\hat{\theta}) = RD(b(\hat{\theta},\theta_0))^{-1}d_T(\theta_0)$$

Insert (5.39):$R'\hat{\lambda} = -d_T(\hat{\theta})$ into the equation above and solve for $\hat{\lambda}$, we get:

$$\begin{aligned}
\hat{\lambda} &= -(RD_T(b(\hat{\theta},\theta_0))^{-1}R)^{-1}RD_T(b(\hat{\theta},\theta_0))^{-1}d_T(\theta_0) \\
&= \left(-R\left[\frac{1}{T}D_T(b(\hat{\theta},\theta_0))\right]^{-1}R\right)^{-1}R\left[\frac{1}{T}D_T(b(\hat{\theta},\theta_0))\right]^{-1}d_T(\theta_0)
\end{aligned}$$

It suggests that $\hat{\lambda}$ will converge to a linear combination of $d_T(\theta_0)$ and is itself a martingale. For this result we have the following theorem:

Theorem 5.3 (Martingale property of the Lagrange multiplier) *Under the assumption of the structural model as defined in Appendix (A.1), it holds*

$$\frac{\hat{\lambda}}{\sqrt{T}} \overset{a.a.s}{\sim} \big(- R\bar{D}(\theta_0)^{-1}R'\big)^{-1} R\bar{D}(\theta_0)^{-1}\frac{d_T(\theta_0)}{\sqrt{T}} \tag{5.41}$$

Proof: See Appendix Lemma 2.7.

□

Therefore $d_j P_j \hat{\lambda}$ is a MG sequence. The LIL can be applied to $d_j P_j \hat{\lambda}$. The log likelihood ratio is the sum of $d_j P_j \hat{\lambda}$, for $j = 1, 2, ...r$. We can find the rate of convergence of the log likelihood ratio. For this result we have the following theorem:

Theorem 5.4 (LIL for likelihood ratio) *Under the assumption of the structural model as defined in (A.1) with the condition of Lemma 2.14 in the appendix, we have*

$$P(\log L_T(\tilde{\theta}) - \log L_T(\hat{\theta}) < r(1+\epsilon)^2 \log\log T, \textit{for } T \textit{ large enough}) = 1,$$

Proof: See Appendix Theorem 2.3. □

Recall that $r = k_1 - k_0$. We can rearrange the equation above and get:

$$P(\log L_T(\tilde{\theta}) - k_1(1+\epsilon)^2 \log\log T < \log L_T(\hat{\theta}) - k_0(1+\epsilon)^2 \log\log T)) = 1.$$

So we can define

$$\Phi(M_i) = \log L_T(\hat{\theta}_i) - Ck_i \log\log T$$

with any $C = (1+\epsilon)^2 > 1$ as the model selection criterion. This criterion is strongly consistent. That means that we can identify the true model with probability 1, as long as we have enough observations along a single path of realization. This criterion is exactly the same as the HQ criterion. Combine the result above and the fact that this criterion will not choose the non-admissible models. We summarize this result in the following theorem.

Theorem 5.5 (Consistent model selection crtierion) *For structural models defined in Appendix A.1, the model selection criterion:*

$$\Phi(M_i) = \log L_T(\hat{\theta}_i) - Ck_i \log\log T \tag{5.42}$$

is strongly consistent for the model selection problem defined in Chapter 4.

6 Model Selection for Cointegration Systems

6.1 An Alternative Representation of Cointegration Systems

Cointegration systems have been used to handle the economic variables with stochastic trends in a structural model. The advantage of a cointegration system is that it models the long term relation and the short term adjustment mechanism simultaneously. For a more detailed discussion of cointegration system see Engle and Granger (1987), Hargreaves (1994), Johansen (1995) and Charemza (1997). In the error correction representation of cointegration system the cointegration relations are interpreted as long term equilibrium relations among the variables and the dynamic of I(0) variables are interpreted as short term adjustment mechanisms. Without loss of generality we consider only cointegration system with 1 lag in the short term dynamics.

$$\Delta y_t = \xi_1 \Delta y_{t-1} + \alpha + \xi_0 y_{t-1} + \epsilon_t$$

where y_t is $N \times 1$ vector ξ_1 is a $N \times N$ matrix and ϵ_t is iid $\sim N(0, \Omega)$. The log likelihood function of the cointegration system above is:

$$\begin{aligned} \log L &= -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log|\Omega| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}(\Delta y_t - \xi_0 y_{t-1} - \alpha - \xi_1 \Delta y_{t-1})'\Omega^{-1}(\Delta y_t - \xi_0 y_{t-1} - \alpha - \xi_1 \Delta y_{t-1}) \end{aligned}$$

In this error correction form of a cointegration system, we apply Johansen's auxiliary regressions and get[43]:

$$\hat{u}_t = \Delta y_t - \hat{\pi}_0 - \hat{\pi}_1 \Delta y_{t-1}$$

$$\hat{v}_t = y_t - \hat{\theta}_0 - \hat{\theta}_1 \Delta y_{t-1}.$$

We look at the concentrated likelihood function[44] where we take ξ_0 and Ω as known and maximize the log likelihood function with respect to (α, ξ_1).

[43] See Johansen (1991) and Hamilton (1994) p. 642

[44] For the motivation of the concentrated likelihood function see Koopmans and Hood (1953).

$$\begin{aligned} & \log L_T(\xi_1, \alpha|\xi_0, \Omega) \\ = & -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log|\Omega| \\ & -\frac{1}{2}\sum_{t=1}^{T}(\Delta y_t - \xi_0 y_{t-1} - \hat{\alpha}(\xi_0) + \hat{\xi}_1(\xi_0)\Delta y_{t-1})'\Omega^{-1}(\Delta y_t - \xi_0 y_{t-1} - \hat{\alpha}(\xi_0) - \hat{\xi}_1(\xi_0)\Delta y_{t-1}) \end{aligned}$$

Because the log likelihood function above can be seen as a SUR with $\Delta y_t - \xi_0 y_{t-1}$ as a dependent variable and $(1, \Delta y_{t-1})$ as regressors, the MLE is identical to the OLE that is characterized by the conditions that the following residual vector must have a sample mean of zero and be orthogonal to Δy_{t-1}:

$$[\Delta y_t - \xi_0 y_{t-1}] - \hat{\alpha}^*(\xi_0) - \hat{\xi}_1^*(\xi_0)\Delta y_{t-1}.$$

Notice that the OLS residuals $\hat{u}_t$ and $\hat{v}_t$ each satisfy these conditions and therefore the vector $\hat{u}_t - \xi_0\hat{v}_t$ also has a mean of zero and is orthogonal to Δy_{t-1}. Moreover, $\hat{u}_t - \xi_0\hat{v}_t$ is of the form:

$$\hat{u}_t - \xi_0\hat{v}_t = \Delta y_t - \hat{\pi}_0 - \hat{\pi}_1\Delta y_{t-1} - \xi_0(y_t - \hat{\theta}_0 - \hat{\theta}_1\Delta y_{t-1}) = -\hat{\alpha}^*(\xi_0) - \hat{\xi}_1^*(\xi_0)\Delta y_{t-1}$$

with $\hat{\alpha}^*(\xi_0) = \hat{\pi}_0 - \xi_0\hat{\theta}_0$ and $\hat{\xi}_1^*(\xi_0) = \hat{\pi}_1 - \xi_0\hat{\theta}_1$

The concentrated log likelihood function is found by replacing α by $\hat{\alpha}^*(\xi_0)$ and ξ_1 by $\hat{\xi}_1^*(\xi_0)$:

$$\log L_T(\Omega, \xi_0) = -\frac{NT}{2}\log(2\pi) - \frac{T}{2}log|\Omega| - \frac{1}{2}\sum_{t=1}^{T}[(\hat{u}_1 - \xi_0\hat{v}_t)'\Omega^{-1}(\hat{u}_1 - \xi_0\hat{v}_t)]$$

The task is now to find a ξ_0 and an Ω that maximize the concentrated log likelihood function above. This problem is equivalent to the following regression problem:

$$\hat{u}_t = \xi_0\hat{v}_t + \epsilon_t$$

with ϵ_t iid $N(0, \Omega)$.

The Johansen procedure treats a cointegration system by restricting $\xi_0 = BA'$ where A and B are $N \times h$ matrices and h is the number of cointegration

relations with $h < N$. Then the problem is solved by the calculation of the canonical correlation.

We can represent equivalently the cointegration system in an alternative way:

$$\xi_0 = B^* A^{*'}$$

$$B^* = \begin{pmatrix} b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ b_{N1} & b_{N2} & \dots & b_{NN} \end{pmatrix}$$

$$A^{*'} = \begin{pmatrix} 1 & a_{12} & \dots & \dots & a_{1N} \\ 0 & 1 & a_{23} & \dots & a_{2N} \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & a_{N-1N} \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

A cointegration system with $h < N$ cointegration relations can be represented by restricting the unspecified elements on the last $N - h$ columns of B^* and on the last $N - h$ rows of A^* to zero.

For instance a cointegration system with $h = N - 1$ cointegration relations can be represented through the following restriction:

$$H(N-1) : b_{NN} = 0$$

A cointegration system with $h = N-2$ cointegration relations can accordingly be represented through the following set of null restrictions:

$$H(N-2) : b_{NN} = 0, b_{N-1N-1} = 0, b_{N-1N} = 0, a_{N-1N} = 0.$$

The equivalence of the $\xi_0 = B^* A^{*'}$ and $\xi_0 = BA'$ can be shown as follows:

We partition A as

$$A' = [A_1', A_2']$$

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

Because A has rank h, A_1 is invertable. We have:

$$\xi_0 = BA' = B[A_1', A_2'] = BA_1'A_1^{-1'}[A_1', A_2'] = BA_1'[I, A_1^{-1'}A_2']$$

BA_1' is an $N \times h$ matrix with rank h. B_1A_1' is an $h \times h$ matrix with rank h. Hence, we can transform B_1A_1' into a lower triangular matrix. We denote this transformation matrix by C, C^{-1} then it is an upper triangular matrix.

$$\xi_0 = BA' = BA_1'CC^{-1}[I, A_1^{-1'}A_2'] = \begin{pmatrix} B_1A_1'C \\ B_2A_1'C \end{pmatrix} [C^{-1}, C^{-1}A_1^{-1'}A_2'] = B^*A^{*'}$$

The last equality is because $B_1A_1'C$ is a lower triangular matrix and C^{-1} is an upper triangular matrix with units on the diagonal. With $\xi_0 = B^*A^{*'}$ we look at a special cointegration relation represented by $A^{*'}y_t = [C^{-1}, C^{-1}A_1^{-1'}A_2']y_t$ and the corresponding adjustment matrix B^*.

6.2 Structural Models and Cointegration Systems

Recall the model selection problem for structural models:

$$Y_t = \Pi X_t + V_t$$

with V_t iid $N(0, \Omega)$ and $\Pi = -B^{-1}\Gamma$ where the B^{-1} and Γ are subject to some zero restrictions. The task is to find the true model that may have generated the observed data from a group of model candidates.

Comparing the model selection problem with the problem of cointegration analysis, we will find these two problems are extremely similar.

The cointegration analysis problem can be stated as follows:

$$\hat{u}_t = \xi_0\hat{v}_t + \epsilon_t$$

with ϵ_t iid $N(0, \Omega)$ and $\xi_0 = B^*A^{*'}$ where the B^* and $A^{*'}$ are subject to some null restrictions. The task is to find out the true model (i.e. identifying the number of cointegration relations h) that may have generated the observed data from a group of model candidates. These candidates are characterized by different set of null restrictions on the matrices B* and A* : H(N-1),H(N-2), ... H(1) and H(0). H(0) is the model without cointegration relations.

In this way we can translate the cointegration analysis problem into a model selection problem:

Suppose that the observed data are generated by one of the model candidate characterized by a set of null restrictions H(h). The task is to find a consistent model selection criterion that will identify the true cointegration model.

The difference between the structural model selection problem and the cointegration analysis problem is that in the cointegration analysis we have non-stationary variables as regressors, while in the structural model selection we have only stationary variables.

Due to the non-stationary regressors the likelihood ratio between a restricted model and that of the unrestricted model are no longer always asymptotically χ^2 distributed. However, according to Johansen, the likelihood ratio has a well defined distribution asymptotically. Hence, we can apply the weak convergence model selection criterion to the cointegration problem to identify the true model, i.e. to identify the number cointegration relations.

6.3 A (weak) Consistent Model Selection Criterion for Cointegration Systems

In Chapter 4 we have given a sufficient and necessary conditions for a (weak) consistent model selection criterion $\log L_T(\hat{\theta}_i) - f(T)k_i$:

$$\lim_{t\to\infty} f(t) = +\infty \tag{6.43}$$

$$\lim_{t\to\infty} \frac{f(t)}{t} = 0 \tag{6.44}$$

Now we verify the conditions of the consistent model selection criterion for the case of a cointegration system.

Sufficiency:

Supposing that the true number of cointegration relations is h, we calculate the difference between the selection criterion values for $H(h)$ and $H(h-i)$:

$$\begin{aligned} & \log L_T(\hat{B}_{h-i}\hat{A}'_{h-i}) - f(T)k_{h-i} - L(\hat{B}_h\hat{A}'_h) + f(T)k_h \\ = \;& T(\tfrac{1}{T}(\log L_T(\hat{B}_{h-i}\hat{A}'_{h-i}) - \tfrac{1}{T}L(\hat{B}_h\hat{A}'_h)) + \tfrac{f(T)}{T}(k_h - k_{h-i}) \end{aligned}$$

Because the MLE $(\hat{B}_h\hat{A}'_h)$ is consistent[45], the MLE for model $H(h)$ will converge to the true parameter of $H(h)$. $H(u-i)$ is not admissible; its corresponding average log likelihood function value will be smaller than that of $H(h)$[46].

[45] See Johansen (1995) p. 180.

[46] For proof compare Appendix Lemma 2.9.

$$\begin{aligned} & \operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_h\hat{A}_h) \\ = & \lim_{T\to\infty} E\frac{1}{T}L_T(B_hA_h) \\ > & \lim_{T\to\infty} E\frac{1}{T}L_T(B_{h-i}A_{h-i}) \\ = & \operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_{h-i}\hat{A}_{h-i}) \end{aligned}$$

$$P\left(\frac{1}{T}\log L_T(B_hA_h') > \frac{1}{T}\log L_T(B_{h-i}A_{h-i}')\right) \to 1$$

Using the condition (6.44) we have for $T \to \infty$:

$$P\left(\frac{1}{T}\log L_T(\hat{B}_h\hat{A}_h') - \frac{f(T)}{T}k_h > \frac{1}{T}\log L_T(\hat{B}_{h-1}\hat{A}_{h-i}\frac{f(T)}{T}k_{h-i})\right) \to 1.$$

It follows then

$$P\left(\log L_T(\hat{B}_h\hat{A}_h) - f(T)J_h > \log L_T(\hat{B}_{h-i}\hat{A}_{h-i}') - f(T)k_{h-i}\right) \to 1.$$

Now we compare $H(h)$ with $H(h+i)$,

According to Johansen (1995), we have:

$$2(\log L_T(\hat{B}_{h+i}\hat{A}_{h+i}') - \log L_T(\hat{B}_h\hat{A}_h')) \to d(.)$$

$d(.)$ has its mass over $(0, +\infty)$. Hence the conditions for the (weak) consistent criterion are satisfied. The model selection criterion:

$$\log L_T(\hat{\theta}_i) - f(T)k_i.$$

is consistent for the selection of the number of cointegration relations.

6.4 Calculation of the Consistent Selection Criterion for Cointegration System

Direct calculation of the criterion value for each candidate models (H(h); h = N-1, N-2, ... 0) is not trivial. It is therefore more convenient to use the calculation from the canonical correlation analysis to get the maximum likelihood function values for each model. For $f(T)$ we can use BSC or HQ criterion: $f(T) = \log(T)$ or $f(T) = \log\log(T)$ respectively.

7 Model Selection in the Case of Misspecification

7.1 Source of Misspecification

In the construction of a structural model we try to mimic the real DGP by an empirical model. Principally two kinds of misspecification may be involved: firstly, the real DGP is contained in the unconstrained reduced form, while the true model (and its observational equivalence) is not within the set of candidates. In this case the unconstrained reduced form provides a correctly specified model, and the structural model candidates consider only a part of the restrictions of the real DGP. Secondly, the DGP is not contained in the unconstrained reduced form, i.e. the reduced form is misspecified and henceforth all the structural models are misspecified. We discuss how the model selection criterion works in these two cases in the following subsections.

7.2 The Case of the Correctly Specified Reduced Form

It is a special feature of the structural model selection problem that all derived reduced forms of the structural models in the candidate set are nested in the unconstrained reduced form, in the sense that any derived reduced form of an overidentified structural model can be seen as the unconstrained reduced form under the corresponding restrictions imposed by the overidentified model, specifically, in the case of linear dependence in a sub-matrix of Π (see Chapter 2). All observationally equivalent structural models have the same derived reduced form. An overidentified structural model is then correctly specified if its associated restrictions on the true parameter of the reduced form Π_0 are correct. Therefore in order to check whether a structural model is correctly specified, we only need to check if Π_0 satisfies these restrictions imposed by the structural model.

Proposition 7.1 *For the case of the correctly specified unconstrained reduced form, if a structural model M_i characterized by (B_i, Γ_i) is chosen by the consistent model selection criterion:*

$$\log L_T(\hat{B}_i, \hat{\Gamma}_i) - k_i C \log\log T > \log L_T(\hat{\Pi}) - k_\pi C \log\log T \qquad \text{for all } T > T_0$$

then the restrictions imposed by this structural model are correct, where k_i and k_π are the number of the parameters in the structural model and the unconstrained reduced form respectively.

Proof:

We rearrange the inequality above and get:

$$0 < \frac{1}{T}\log L_T(\hat{\Pi}) - \frac{1}{T}\log L_T(\hat{B}_i^{-1}\hat{\Gamma}_i) < C(k_\pi - k_i)\frac{1}{T}\log\log T$$

The first inequality is because the maximum of the structural model is selected within the parameter range of the reduced form. Because the Σ_i matrix is unconstrained, the log likelihood function depends on (B_i, Γ_i) only through $(B_i^{-1}\Gamma_i)$. We write the argument $\hat{B}_i^{-1}\hat{\Gamma}_i$ in the log likelihood function of the structural model to make the comparison to the unconstrained reduced form more clear.

As $T \to \infty$ we have:

$$0 \leq \lim_{T\to\infty}\frac{1}{T}E\log L_T(\hat{\Pi}) - \lim_{T\to\infty}\frac{1}{T}E\log L_T(\hat{B}_i^{-1}\hat{\Gamma}_i) \leq 0$$

It follows

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{\Pi}) = \operatorname*{plim}_{T\to\infty}\frac{1}{T}\log L_T(\hat{B}_i^{-1}\hat{\Gamma}_i)$$

Because the unconstrained reduced form is correctly specified, the convergence of the log likelihood is uniformly in parameter and the limit of the average log likelihood function has a unique maximum at the true parameter, we have[47]

$$\Pi_0 = \operatorname*{plim}_{T\to\infty}\hat{\Pi} = \operatorname*{plim}_{T\to\infty}\hat{B}_i^{-1}\operatorname*{plim}_{T\to\infty}\hat{\Gamma}_i = B_0^{-1}\Gamma_0 \tag{7.45}$$

This means that the true parameter value Π_0 satisfies the restrictions imposed by the structural model (B_i, Γ_i). In other words the restrictions associated with the model M_i are true.

□

Corollary 7.2 *Assume that two models (B_i, Γ_i) and (B_j, Γ_j) are selected against the unconstrained reduced form by the model selection criterion for all $T > T_0$, if we can find a structural model denoted by $(B_{i\cup j}, \Gamma_{i\cup j})$ so constructed such that it unifies the restrictions of both models in it, it will also be selected by the model selection criterion.*

[47] For details see Davidson and Mackinnon (1993b) p. 255

Obviously, the restrictions associated with the model $(B_{i\cup j}, \Gamma_{i\cup j})$ are true and it has more restrictions than (B_i, Γ_i) or (B_j, Γ_j), hence the selection criterion will prefer $(B_{i\cup j}, \Gamma_{i\cup j})$ to (B_i, Γ_i) or (B_j, Γ_j).

This property of the model selection criterion provides us possible strategies to construct a structural model.

In the case we have several well specified structural models, we may use the model selection criterion to combine the features from different models to get a more parsimonious model which has all the properties of other models.

Because the restrictions associated with a structural model are the union of the restrictions imposed by each equation of the structural model, we may investigate the appropriateness of each structural equation separately.

In the case we have a structural model that is not supported by data according to the model selection criterion, we may detect the failure by substituting some potential "problem-making equations" by the corresponding reduced form equations and calculate the model selection criterion for this substituted model. If the model is now supported by data, then we have identified the problem.

A more constructive way of using this property is to specify the structural model step by step. We may specify one or more structural equations and check their data compatibility by comparing the values of the model selection criterion between the unconstrained reduced form and the structural model that consists of those specified structural equations and other equations of a reduced form. If this part of the structural model is shown to be data compatible then we may add another part of structural equations to the model by replacing the corresponding equations of the reduced form. In this way a structural model can be constructed step by step and at the end we can have a structural model that is theoretically well founded and conforms with the data.

7.3 The Case of the Misspecified Reduced Form

Generally, the DGP of empirical data is unknown. The determinants of the real DGP of economic data are nonlinear, dynamic, time varying and multidimensional. It is far too complex to be described completely in a structural model. A structural model is an empirical model that tries to capture the basic features of interest in the data. It is therefore not reasonable to assume that the DGP is contained in the unconstrained reduced from. In this case we are facing the issue of approximating the unknown real DGP by a parametric family of the structural models[48].

[48]For a detailed discussion of encompassing see Dhaene (1997)

Suppose that the real DGP can be described as a parametric family $\mathcal{W} = \{W_\alpha | \alpha \in \Omega_\mathcal{W} \subset R^m\}$ and a structural model by another parametric family $\mathcal{G} = \{G_\theta | \theta \in \Omega_\mathcal{G} \subset R^n\}$. What we are going to do is to approximate $\mathcal{W}$ with $\mathcal{G}$. The quality of the approximation can be described by the distance between the distribution of the real DGP and the distribution of the structural model. This distance can be described by the Kullback Leibler Information Criterion. The problem is to find a mapping of $\mathcal{W}$ into $\mathcal{G}$ that associates with each distribution $W_\alpha \in \mathcal{W}$ a distribution $G_\theta \in \mathcal{G}$ that is the closest to W_α within the family of $\mathcal{G}$.

7.3.1 Maximum Likelihood Estimation under Mis-specification

Consider a $T \times 1$ random vector Y_T taking its value in $A \in \mathbb{R}^{G \times T}$ with distribution W_α. Given another distribution G_θ over A, the distance between W_α and G_θ is provided by the Kullback-Leibler Information Criterion that is defined:

$$I(W_\alpha, G_\theta) = E_{W_\alpha} \log \frac{w(Y_T; \alpha)}{g(Y_T; \theta)}.$$

E_{W_α} denotes the expectation relative to W_α. $w(Y_T, \alpha)$ is the density of W_α, and $g(Y_T, \theta)$ is the density of G_θ.

$$I(W_\alpha, G_\theta) = E_{W_\alpha} \log w(Y_T, \alpha) - E_{W_\alpha} \log g(Y_T; \theta) \leq 0$$

The last inequality is according to Jensen's inequality. This distance depends on the entropy of the distribution of W_α and the relative entropy of G_θ with respect to W_α.

If we take W_α as the real DGP and G_θ as an empirical model, then the quality of the empirical model depends only on the expected log likelihood function value of the model relative to the real DGP. The larger the expected log likelihood function the better the model quality.

Now the number of the observations T may go to infinity. The distance between W_α and G_θ is given:

$$\bar{I}(W_\alpha, G_\theta) = \lim_{T \to \infty} E_{W_\alpha} \frac{1}{T} \log \frac{w(Y_T, \alpha)}{g(Y_T; \theta)}$$

We have:

$$\bar{I}(W_\alpha, G_\theta) = \lim_{T \to \infty} E_{W_\alpha} \frac{1}{T} \log w(Y_T, \alpha) - \lim_{T \to \infty} E_{W_\alpha} \frac{1}{T} \log g(Y_T, \alpha) \leq 0$$

Hence $\bar{I}(W_\alpha, G_\theta)$ can also be interpreted as the distance between $w(Y_\infty, \alpha)$ and $g(Y_\infty, \theta)$.

For a given value α that describes the real DGP, denote the solution of the following maximization problem by θ_α:

$$\theta_\alpha = \arg\min_{\theta \in \Omega_\mathcal{G}} \bar{I}(W_\alpha, G_\theta)$$

θ_α is that parameter whose corresponding distribution is the closest to the real DGP within the family of $G(\theta)$ for $\theta \in \Omega_\mathcal{G}$. θ_α is defined as the pseudo true parameter.[49]Equivalently we have:

$$\theta_\alpha = \arg\max_{\theta \in \Omega_\mathcal{G}} \lim_{T\to\infty} E_{W_\alpha} \frac{1}{T} \log g(Y_T, \alpha)$$

Now we consider the maximum likelihood estimation in the model $\mathcal{G}$:

$$\hat{\theta} = \arg\max_{\theta \in \Omega_\mathcal{G}} \log g(\mathbf{y}_T, \theta).$$

$g(\mathbf{y}_T, \theta)$ corresponds to the likelihood function of model $\mathcal{G}$. Because logarithm is a monotone function, the MLE can be obtained by the maximization of the log likelihood function. We write it in the conventional way:

$$\hat{\theta} = \arg\max_{\theta \in \Omega_\mathcal{G}} \log L_T(\theta; \mathbf{y}_T, \mathbf{x}_T).$$

Under regularity assumptions, if the DGP fulfills the condition of ULLN and the condition of identifiable uniqueness, the pseudo-true maximum likelihood estimator of θ will converge to the pseudo-true parameter θ_α W_α-almost sure or in probability.[50] The pseudo-true parameter plays here the same role of true parameter in the theory of maximum likelihood estimation without misspecification. In the case of approximating the real DGP by a structural model we have:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty} \hat{\theta} &= \operatorname*{plim}_{T\to\infty} \arg\max_{\theta \in \Omega_g} \frac{1}{T} \sum_{t=1}^{T} \log L_T(\theta; Y_T, X_T) \\
&= \arg\max_{\theta \in \Omega_g} \lim_{T\to\infty} \frac{1}{T} E_{W_\alpha} \log L_T(\theta; Y_T, X_T) \\
&= \theta_\alpha
\end{aligned}$$

[49]This definition was first introduced in Sawa(1978), then further developed by White(1982)

[50]For details see Pötscher(1995)

7.3.2 Encompassing

In scientific fields, a more advanced theory should be able to account for or explain the results obtained by other, competing theories. This idea motivated the conception of encompassing for comparing two empirical models. If a model can account for the results of another it is said to encompass the latter model.[51] Following Hendry(1995) and Courieroux-Monfort(1995) we define for given $\Xi_T = \{\xi_t\}_{t=1}^{\infty}$ and for each real DGP D an encompassing relation as a binary relation on the candidate set $\mathcal{M}$ which involves the comparison of β_D and β_{α_D}[52].

Definition 7.3 (Encompassing) *For all* $(\mathcal{F}, \mathcal{G}) \in \mathcal{M} \times \mathcal{M}$ *with* $\mathcal{F} = \{F_\alpha |\ \alpha \in \Omega_\mathcal{F} \subset R^m\}$ *and* $\mathcal{G} = \{G_\theta |\ \theta \in \Omega_\mathcal{G} \subset R^n\}$, $\mathcal{F}$ *encompasses* $\mathcal{G}$ *relative to* D *and* Ξ_T, *denoted by* $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$, *if and only if*

$$\theta_D = \theta_{\alpha_D}.$$

If $\theta_D \neq \theta_{\alpha_D}$ we write $\mathcal{F}\,\mathcal{N}\mathcal{E}_D^{\Xi_T}\,\mathcal{G}$. If $\mathcal{F}$ is a subset of $\mathcal{G}$ and $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$, then $\mathcal{F}$ is said to parsimoniously encompasses $\mathcal{G}$.

The definition is justified by the properties that result from it. $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$ implies the property that the closest approximation to D by $\mathcal{G}$ coincides with the closest approximation by $\mathcal{G}$ to the closest approximation to D by $\mathcal{F}$. Encompassing is a formalization of the idea that, looking at $\mathcal{G}$ from D, one observes the same thing as looking at $\mathcal{G}$ from D through $\mathcal{F}$

7.3.3 The Properties of Encompassing

Property 1 $\{D\}\mathcal{E}_D^{\Xi_T}\mathcal{G}$.

Proof: This result follows directly from the definition of encompassing.

This is the fundamental property of a DGP that the performance of a model, no matter how poorly misspecified, is completely determined by the DGP.

Property 2 If $D \in \mathcal{F}$, then $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$.

Proof: If $D \in \mathcal{F}$, then $D = F_{\alpha_D}$. It follows $\theta_D = \theta_{\alpha_D}$.

This property implies that if a model is correctly specified it encompasses every model. In a correctly specified model the MLE will converge to the true parameter.

Property 3 If $D\bar{\in}\mathcal{F}$, then

[51] Hendry-Richard(1982,1983,1990), Mizon(1984), Mizon-Richard(1986) Smith(1993) White(1994) Courieroux-Monfort(1995)

[52] This and next two subsections are mainly due to the results in Dhaene (1997).

$$\exists \mathcal{G} : \mathcal{F} \subset \mathcal{G} \text{ and } D \in \mathcal{G} \rightarrow \mathcal{F} \mathcal{N} \mathcal{E}_D^{\Xi_T} \mathcal{G}$$

Proof: Let $D \bar{\in} \mathcal{F} \subset \mathcal{G}$ and $D \in \mathcal{G}$. Then $D = G_{\theta_D}$ and $F_{\alpha_D} = G_{\theta_{\alpha_D}} \neq D$ It follows that $G_{\theta_D} \neq G_{\theta_{\alpha_D}}$.

This property states that if $\mathcal{F}$ is misspecified, there are always models which $\mathcal{F}$ does not encompass. In other words, any correctly specified model contains enough information to invalidate any misspecified models.

Property 4 If $\mathcal{F} \subset \mathcal{G}$ then

$$\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G} \leftrightarrow F_{\alpha_D} = G_{\theta_D} \leftrightarrow \bar{I}(D, \mathcal{F}) = \bar{I}(D, \mathcal{G})$$

Proof:

If $\mathcal{F} \subset \mathcal{G}$, then $F_{\alpha_D} = G_{\alpha_D}$. Therefore $F_{\alpha_D} = G_{\theta_D}$ is equivalent to $G_{\theta_{\alpha_D}} = G_{\theta_D}$. That is $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$. Therefore $\bar{I}(D, \mathcal{G}) = \bar{I}(D, G_{\theta_D})$ iff $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$.

This property gears the principle of parsimony and the concept of encompassing. It tells us a more specific model will encompass a more general model if the specific one approximates the DGP equally well as the more general one. If a submodel has all the desired properties of the a comprehensive model, we only need to consider the submodel. Intuitively, if $\mathcal{G}$ is a valid reduction of the DGP and $\mathcal{F}$ parimoniously encompasses $\mathcal{G}$ then $\mathcal{F}$ is a valid reduction of the DGP.

Property 5 $\mathcal{F} \supset \mathcal{G}$ does not imply $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$.

Property 6 $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$ is not transitive.

Proof of property 5 and 6: This proof is an example given in Gourieroux and Monfort (1996).

7.4 Encompassing Relation and Model Selection Criterion

The concept of encompassing is designed to implement a progressive research strategy: a modeling strategy in which knowledge is gradually accumulated as codified. Following Hendry(1995) such a strategy will require:

- Reflexivity: $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{F}$.
- Anti-symmetry: $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$ and $\mathcal{G} \mathcal{E}_D^{\Xi_T} \mathcal{F}$ imply that $\mathcal{F}$ is equivalent to $\mathcal{G}$.
- Transitivity: $\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{G}$ and $\mathcal{G} \mathcal{E}_D^{\Xi_T} \mathcal{H}$ imply that t$\mathcal{F} \mathcal{E}_D^{\Xi_T} \mathcal{H}$.

Property 6 documents that encompassing does not generally have transitivity. However, parsimoniously encompassing does satisfy the transitivity requirement.

Proposition 7.4 *If $\mathcal{F} \subset \mathcal{G} \subset \mathcal{H}$, then $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$ and $\mathcal{G}\mathcal{E}_D^{\Xi_T}\mathcal{H}$ imply that $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{H}$.*

Proof: From $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{G}$ it follows according to property 4:

$$\bar{I}(D, \mathcal{F}) = \bar{I}(D, \mathcal{G}) = \bar{I}(D, \mathcal{H})$$

hence $\mathcal{F}\mathcal{E}_D^{\Xi_T}\mathcal{H}$.

The parsimonious encompassing implies that a model $\mathcal{F}$ with fewer parameters can explain equally well as a model with more parameters $\mathcal{G}$. This is a valid reduction. If the initial information set $\mathcal{G}$ is enlarged to $\mathcal{H}$, and the parsimonious encompassing still results, then model $\mathcal{F}$ explains more data information. This is the basis for a progressive research strategy.

7.5 The Consistent Model Selection Criterion and Parsimonious Encompassing

In the context of a model selection problem, each structural model corresponds to a derived reduced form that is a subset of the unconstrained reduced form. Hence, an overidentified structural model corresponds to a more parsimonious model relative to the reduced form. If the consistent model selection criterion will select a structural model for all $T > T_0$ the structural model will approximate the DGP equally well as the unconstrained reduced form. This implies a parsimonious encompassing.

Proposition 7.5 *We denote the unconstrained reduced form by M_{rd}. If a structural model denoted by $M_i = (B_i, \Gamma_i)$ is chosen by the consistent model selection criterion:*

$$\log L_T(\hat{B}_i^{-1}\hat{\Gamma}_i) - k_i C \log\log T > \log L_T(\hat{\Pi}) - k_\pi C \log\log T \qquad \text{for all } T > T_0$$

then the structural model encompasses the unconstrained reduced form parsimoniously, where k_i and k_π are the number of the parameters in the structural model and the unconstrained reduced form respectively.

Proof: Following (7.45) we have:

$$\bar{I}(D, M_i) = \bar{I}(D, M_{re}).$$

Because of $M_i \subset M_{rd}$ and Property 4, we have $M_i \mathcal{E}_D^{\Xi_T} M_{rd}$. $\square$

Intuitively, if the consistent criterion chooses a structural model, this structural model will approximate the real DGP equally well as the unconstrained reduced form. In addition, a structural model is usually overidentified; it contains fewer parameters than the unconstrained reduced form. Such a specific model that can explain the data information equally well with fewer parameters and provide a specific interpretation of the data means progress in understanding the data.

8 A Modeling Procedure to Construct a Structural Model

8.1 Encompassing in Structural Modeling

From the discussion in the last section we know that it is a specific feature of structural modeling that a structural model's derived reduced form is nested in the unconstrained reduced form. Two implications are involved in this specific feature. The unconstrained reduced form provides here a benchmark for the evaluation of the structural model. If the consistent model selection criterion chooses a structural model against the unconstrained reduced form, the structural model will encompass the unconstrained model parsimoniously. The structural model will account all the results obtained by the unconstrained reduced form. If the model selection criterion does not choose a structural model for large T, then the structural model may not encompass the unconstrained reduced form. It means some features of the DGP that can be explained by the unconstrained reduced form will not be explained by the structural model. This implies some deficits of the structural model in understanding the real DGP. Therefore the model selection criterion provides a useful instrument to construct/evaluate a structural model. The power of this instrument is limited, however, by the unconstrained reduced form, because the model selection criterion must compare the criterion value of a structural model with that of the unconstrained model. Hence if the unconstrained model is badly specified, the structural model will fit the data equally badly. Therefore a good reduced form is the necessary condition for a good structural model. This section gives instructions for the specification of a structural model; namely, we should start with the investigation of the reduced form. Only a satisfactory reduced form may lead to a satisfactory structural model.

8.2 A Modeling Procedure

In order to sum up what has been discussed in the previous sections and merge the information into a constructive way of model building, we suggest the following procedure:

- Describe the phenomena
- Construct economical theoretical models for the phenomena under investigation.

- Determine the unconstrained reduced form according to the theoretic models and check if the reduced form has captured all the features of interest in the data. If the reduced form is not satisfactory, rethink theoretical model and reconstruct the reduced form.
- Specify structural equations and compose a partial structural model through the substitution of the reduced form equations by specified structural equations.
- Calculate the model selection criterion values for the partially specified structural model and for the unconstrained reduced form respectively. Check the appropriateness of the restrictions implied by comparing the structural model by the criterion value of the structural model with that of the unconstrained reduced form.
- Repeat the last step until all the equations of the reduced form are replaced by structural equations and the whole structural model is completely specified.

If the procedure above can be completed, we will arrive at a structural model that is intuitively interpretable with respect to the economic hypotheses used to formulate the structural form and is supported by the empirical data.

9 Simulation Studies

9.1 Stationary Data

9.1.1 General Setting of Simulations

The model selection conclusions drawn from the consistent model selection criterion hold only asymptotically. In the empirical research we have always only a limited number of observations. To investigate the performance of the model selection criterion for empirical modeling in finite sample we conduct the following simulation studies.

The main focus of the simulation is to see when the asymptotic property prevails. The performance of the criterion in the following context will also be of interest:

- Number of equations in the model and its impact on the performance of model selection criterion
- Number of correctly specified restrictions and ability of the criterion to identify the most parsimonious model
- Ability to identify false restrictions
- Number of exogeneous variables and its impact on the criterion
- Model with non-stationary data

According to the targets listed above the following simulations are implemented. It is far from an exhaustive exploration of the performance of the consistent criterion for structural models but simply an illustrative demonstration of the performance of the model selection criterion for a few generated examples. The main feature of the simulated examples are listed in the following table.

No.	Number of Equations	Number of Ex.Variabl.	Number of Pred.Variabl.	Number of c. Restrict.	Number of f. Restrict.	Critical Point
1	9	9	18	126	0	31
2	9	9	18	127	1	120
3	9	9	18	127	1	630
4	9	9	18	126-121	0	280
5	9	9	18	125	0	
6	9	9	18	124-126	0	100
7	20	20	40	720	0	60
8	80	20	100	7520	0	90
9	3	20	100	1	1	25

The base case we look at is a structural model with 9 equations, i.e. $G = 9$.

We assume that the number of predetermined variable and the number of exogeneous variables increases proportionally with the number of equations in the simultaneous equation system.

$$\begin{aligned}
y_{1,t} &= \beta_{1,G} y_{G,t} + \gamma_{1,1} x_{1,t} + \gamma_{1,2} x_{2t} + \delta_{1,1} y_{1,t-1} + u_{1,t} \\
&\vdots \\
y_{i,t} &= \beta_{i,i-1} y_{i-1,t} + \gamma_{i,i} x_{i,t} + \gamma_{i,i+1} x_{i+1,t} + \delta_{i,i} y_{i,t-1} + u_{i,t} \\
&\vdots \\
y_{G,t} &= \beta_{G,G-1} y_{G-1,t} + \gamma_{G,G} x_{G,t} + \gamma_{G,1} x_{1,t} + \delta_{G,G} y_{G,t-1} + u_{G,t}
\end{aligned}$$

The we generate data by drawing iid random numbers from $N(0, 5)$ for u_t and and iid random numbers from N(0,5) for x_t. The starting value of y_0 is set to zero. The parameters are set as follows:

$$\beta(i,j) = \begin{cases} 0.3 & \text{for } i \neq j, \\ 1 & \text{for i=j.} \end{cases}$$

$$\gamma(i,j) = \begin{cases} 0.3 & \text{for } i \neq j, \\ 1 & \text{for i=j.} \end{cases}$$

$$\delta(i,j) = \begin{cases} 0 & \text{for } i \neq j, \\ 0.3664 & \text{for i=j.} \end{cases}$$

Then we calculate the value of the selection criterion respectively for different numbers of observations, namely for T = 19, 20, 21,...60,...1500 to see when in this parameter setting the asymptotic property prevails for the specified structural model and the reduced form.

The calculation of the criterion value is done by applying OLS to the unconstrained reduced form because the maximum likelihood estimate coincides in this case with the OLS estimate. For the structural model we apply iterative 3SLS to get the maximum likelihood estimate of the structural parameters, and the function value. The results are presented in the following graphs.

We demonstrate the results of simulation basically through plotting two pairs of variables. The first pair are the criterion values for the unconstrained reduced form (CCU) and for the structural form (CCR) respectively. The

second pair are the log likelihood function values of the unconstrained reduced form (LOGLKHU) and of the structural form (LOGLKHR) respectively. Through plotting them against the number of observations, we can see when the asymptotic property of the selection criterion becomes significant. In some cases we plot all four variables together in one graph to show how the penalty term works.

9.1.2 Simulation 1: True Structural Form vs. the Unconstrained Reduced Form

In simulation 1 we are interested in when the model selection criterion can identify the correctly specified structural mode against the unconstrained reduced form. The result is depicted in the following graphs.

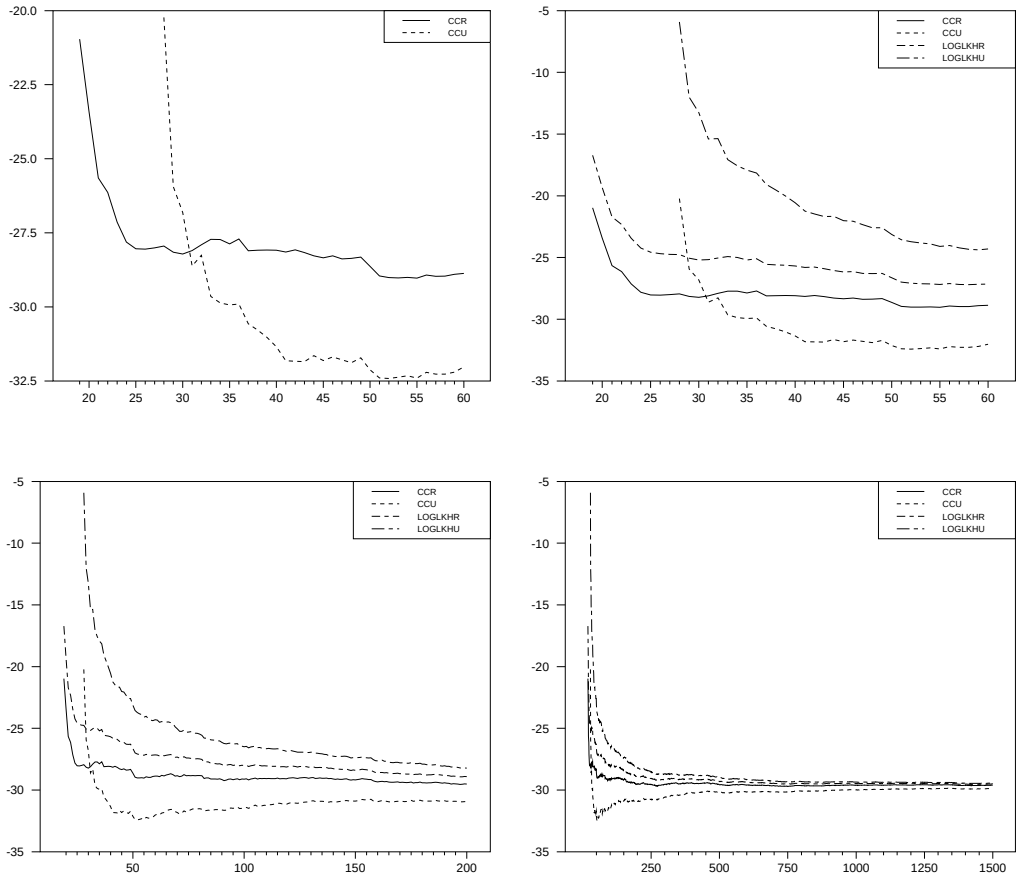


Figure 1: A True Structural Form vs. the Reduced Form (Sim. 1) CCU is the value of the model selection criterion of the unconstrained reduced form. CCR is the value of the model selection criterion of the structural form.

The first graph in the top left corner shows that for this 9 equations system the consistent property prevails already at 30 observations. The graph in the

to right corner shows that the log likelihood of an unconstrained model is significantly larger within this observation range.

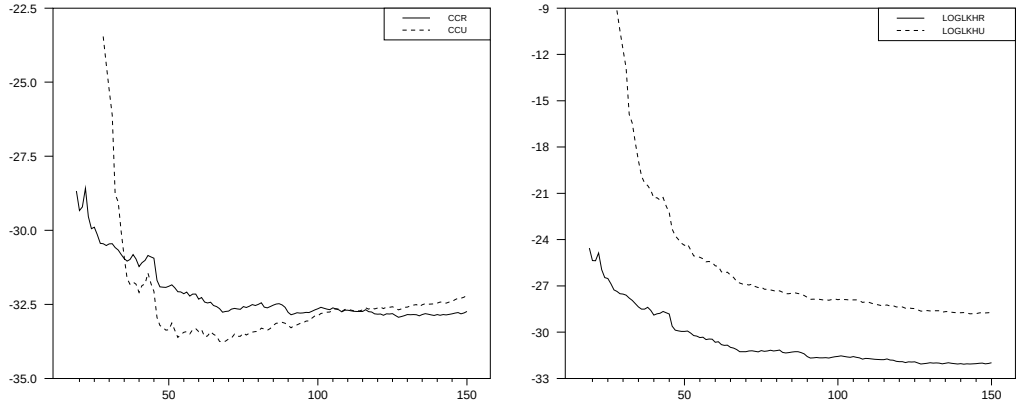
These next two graphs above show the asymptotic behaviour of the consistent criterion and the log likelihood function for this model.

9.1.3 Simulation 2,3: False Restrictions

While simulation 1 shows that the correctly specified structural model will be identified by the consistent criterion against the unconstrained reduced form, simulation 2 demonstrates that a misspecified model will be identified by the consistent criterion. The data used in simulation 2 are generated by the same DGP as in simulation 1 with $\gamma_{i,j} = 3.0$ for $i, j = 1, 2, ..G$. The structural model under investigation is obtained by restricting the coefficient of $x_{1,t}$ in the G-th equation to be zero: $\gamma_{G,1} = 0$. Hence we have a misspecified equation in the model:

$$y_{G,t} = \beta_{G,G-1} y_{G-1,t} + \gamma_{G,G} x_{G,t} + \gamma_{G,1} x_{1,t} + \delta_{G,G} y_{G,t-1} + u_{G,t}.$$

The simulation result is depicted in the following graphs.



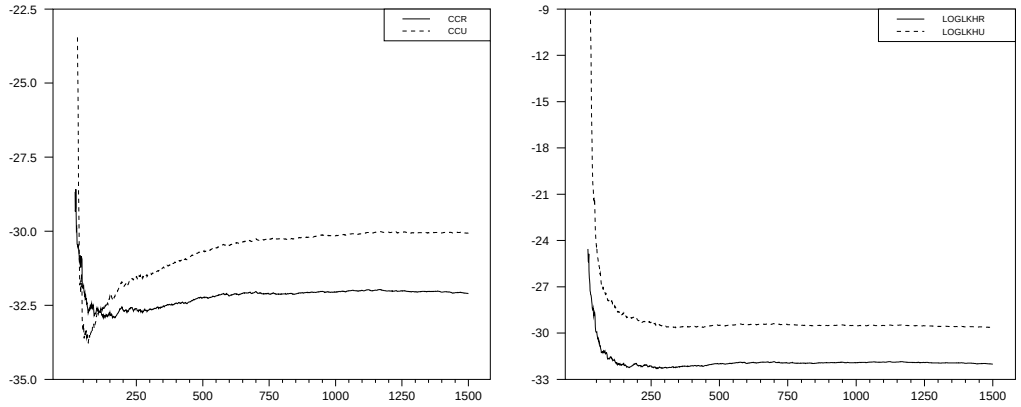


Figure 2: Misspecified Model vs. the Unconstrained Reduced Form (Sim. 2) CCU is the value of the model selection criterion of the unconstrained reduced form. CCR is the value of the model selection criterion of the misspecified structural model.

This case shows that even when only one true regressor is missing in the 9 equations system the consistent criterion can identify this misspecification already at a sample of 110 observations. The two graphs on the second row show the asymptotic behaviour of the criterion and the log likelihood. The average log likelihood of the misspecified model will converge to a lower level than that of the unconstrained reduced form that converges to the expected log likelihood of correctly specified models.

However, in some cases, especially when the misspecification is not very serious, for instance, if a parameter with a small value is set to be zero it will need a large number of observations to identify this kind of misspecification.

To see this effect we modify the DGP in simulation 2 by changing the value of $\gamma_{i,j} = 0.3$. In this case the true DGP is not greatly violated by the restriction $\gamma_{G,1} = 0.0$, and hence much more data are needed to identify this small false restriction. The result is shown in the following graph:

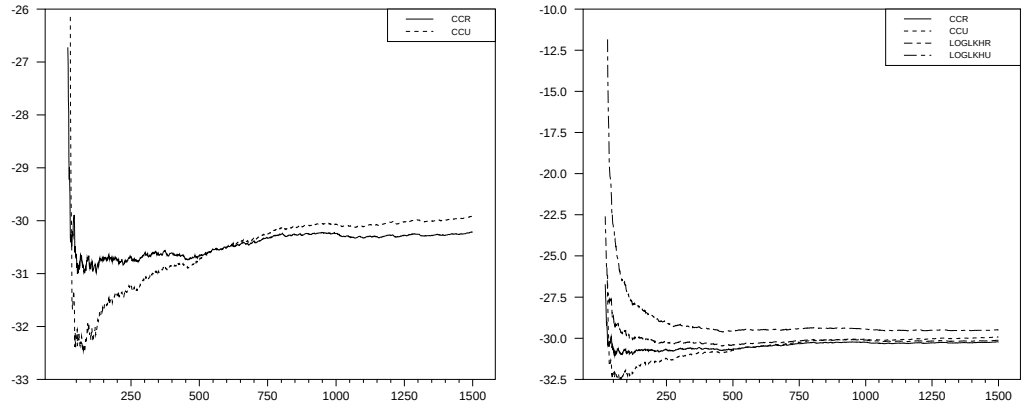


Figure 3: Misspecified Model vs. the Unconstrained Reduced Form (Sim.3) CCU is the value of the model selection criterion of the unconstrained reduced form. CCR is the value of the model selection criterion of the misspecified structural model. LOGLKHU is the value of the average log likelihood of the unconstrained reduced form. LOGLKHR is the average log likelihood function value of the structural model.

9.1.4 Simulation 4: Selection of the Most Parsimonious Model

Among a group of nested admissible models (including the unconstrained reduced form) the model selection criterion will choose the most parsimonious model i.e. that one that is nested in all other models. To see the behaviour of the model selection criterion in this case we construct the following nested admissible models.

Model 1 is the true model, i.e. the data that are used to estimate and evaluate all alternative models are generated from this model. For the generation of data, the parameter settings are the same as in the simulation 1.

Model 1:

$$\begin{aligned} y_{1,t} &= \beta_{1,G} y_{G,t} + \gamma_{1,1} x_{1,t} + \gamma_{1,2} x_{2t} + \delta_{1,1} y_{1,t-1} + u_{1,t} \\ &\vdots \\ y_{i,t} &= \beta_{i,i-1} y_{i-1,t} + \gamma_{i,i} x_{i,t} + \gamma_{i,i+1} x_{i+1,t} + \delta_{i,i} y_{i,t-1} + u_{i,t} \\ &\vdots \\ y_{G,t} &= \beta_{G,G-1} y_{G-1,t} + \gamma_{G,G} x_{G,t} + \gamma_{G,1} x_{1,t} + \delta_{G,G} y_{G,t-1} + u_{G,t} \end{aligned}$$

Model 2:

$$
\begin{aligned}
y_{1,t} &= \beta_{1,9}y_{9,t} + \gamma_{1,1}x_{1,t} + \gamma_{1,2}x_{2t} + \delta_{1,1}y_{1,t-1} + u_{1,t} \\
&\vdots \\
y_{i,t} &= \beta_{i,i-1}y_{i-1,t} + \gamma_{i,i}x_{i,t} + \gamma_{i,i+1}x_{i+1,t} + \delta_{i,i}y_{i,t-1} + u_{i,t} \\
&\vdots \\
y_{9,t} &= \beta_{9,8}y_{8,t} + \gamma_{9,9}x_{9,t} + \gamma_{9,1}x_{1,t} + \delta_{9,9}y_{9,t-1} + \gamma_{9,8}x_{8,t} + u_{9,t}
\end{aligned}
$$

Model 3:

$$
\begin{aligned}
y_{1,t} &= \beta_{1,9}y_{9,t} + \gamma_{1,1}x_{1,t} + \gamma_{1,2}x_{2t} + \delta_{1,1}y_{1,t-1} + u_{1,t} \\
&\vdots \\
y_{i,t} &= \beta_{i,i-1}y_{i-1,t} + \gamma_{i,i}x_{i,t} + \gamma_{i,i+1}x_{i+1,t} + \delta_{i,i}y_{i,t-1} + u_{i,t} \\
&\vdots \\
y_{8,t} &= \beta_{8,7}y_{7,t} + \gamma_{8,8}x_{8,t} + \gamma_{8,1}x_{1,t} + \delta_{8,8}y_{8,t-1} + \gamma_{8,7}x_{7,t} + u_{8,t} \\
y_{9,t} &= \beta_{9,8}y_{8,t} + \gamma_{9,9}x_{9,t} + \gamma_{9,1}x_{1,t} + \delta_{9,9}y_{9,t-1} + \gamma_{9,8}x_{8,t} + u_{9,t}
\end{aligned}
$$

...

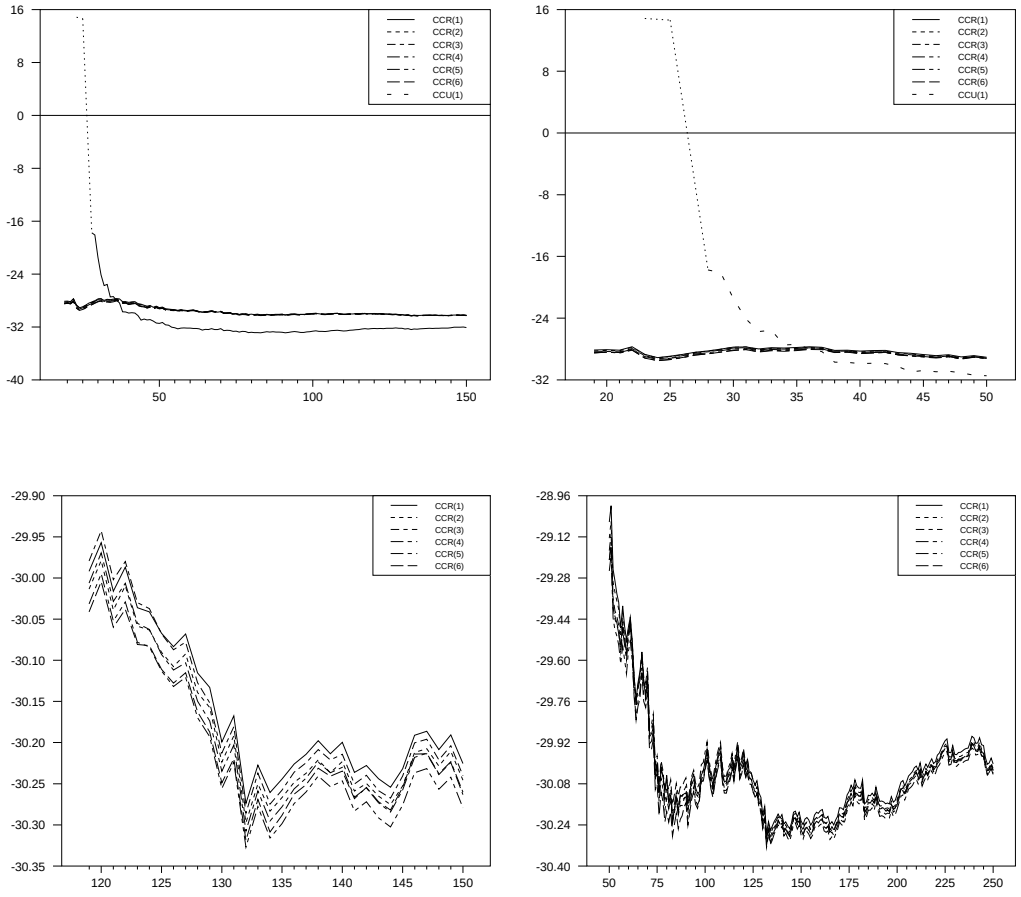
Model 6:

$$
\begin{aligned}
y_{1,t} &= \beta_{1,9}y_{9,t} + \gamma_{1,1}x_{1,t} + \gamma_{1,2}x_{2t} + \delta_{1,1}y_{1,t-1} + u_{1,t} \\
&\vdots \\
y_{i,t} &= \beta_{i,i-1}y_{i-1,t} + \gamma_{i,i}x_{i,t} + \gamma_{i,i+1}x_{i+1,t} + \delta_{i,i}y_{i,t-1} + u_{i,t} \\
&\vdots \\
y_{4,t} &= \beta_{4,3}y_{3,t} + \gamma_{4,3}x_{4,t} + \gamma_{4,1}x_{1,t} + \delta_{4,4}y_{4,4} + \gamma_{4,3}x_{3,t} + u_{4,t} \\
&\vdots \\
y_{8,t} &= \beta_{8,7}y_{7,t} + \gamma_{8,8}x_{8,t} + \gamma_{8,1}x_{1,t} + \delta_{8,8}y_{8,t-1} + \gamma_{8,7}x_{7,t} + u_{8,t} \\
y_{G,t} &= \beta_{G,8}y_{8,t} + \gamma_{G,G}x_{G,t} + \gamma_{G,1}x_{1,t} + \delta_{G,G}y_{G,t-1} + \gamma_{G,8}x_{8,t} + u_{G,t}
\end{aligned}
$$

Model 2, Model 3,... and Model 6 each add one more exogeneous variable to the structural model. Therefore, Model 1 is nested in Model 2 which is nested in Model 3, which, in turn, is nested in Model 4, and so on. Model 6 nested Model 1 to Model 5. In addition, we also consider the unconstrained reduced from that nests, by definition, every structural models.

The data are generated by Model 1. Hence all models listed above are admissible, because they all nest Model 1.

The estimation technique is the same as in the last subsection. We apply OLS to the unconstrained reduced form and iterated 3SLS to MLE for each models and then calculate the values of the selection criterion for each model respectively. The result of the simulation is shown in the following graphs:



These graphs show that in comparison to the unconstrained reduced form all 6 models are much better. Since T = 35 the model selection criterion will choose these admissible structural models. Among the structural models the differences are not very large. For $T > 125$ the selection criterion will clearly identify the most parsimonious model - the Model 1.

9.1.5 Simulation 5: Non-nested Admissible Models

While the selection criterion induces an ordering in the nested models, there is no such order in comparing between non-nested models. Simulation 5 should demonstrate this fact.

The data are generated by the model as in simulation 1 with $\gamma_{i,j} = 0.9$. 6 different admissible structural models are constructed. Model 1 is the true model. Model 2 is constructed by adding an additional regressor $x_{G-1,t}$ to the equation G of the true model. Model 3 is constructed by adding an additional regressor $x_{G-2,t}$ to the equation $G-1$ of the true model,... Model 6 is constructed by adding x_{G-5} to the equation $G-4$. Except for Model 1, each model has one unnecessary regressor, but they are not nested in each other. The following graphs show the value of the model selection criterion for the 5 structural models.

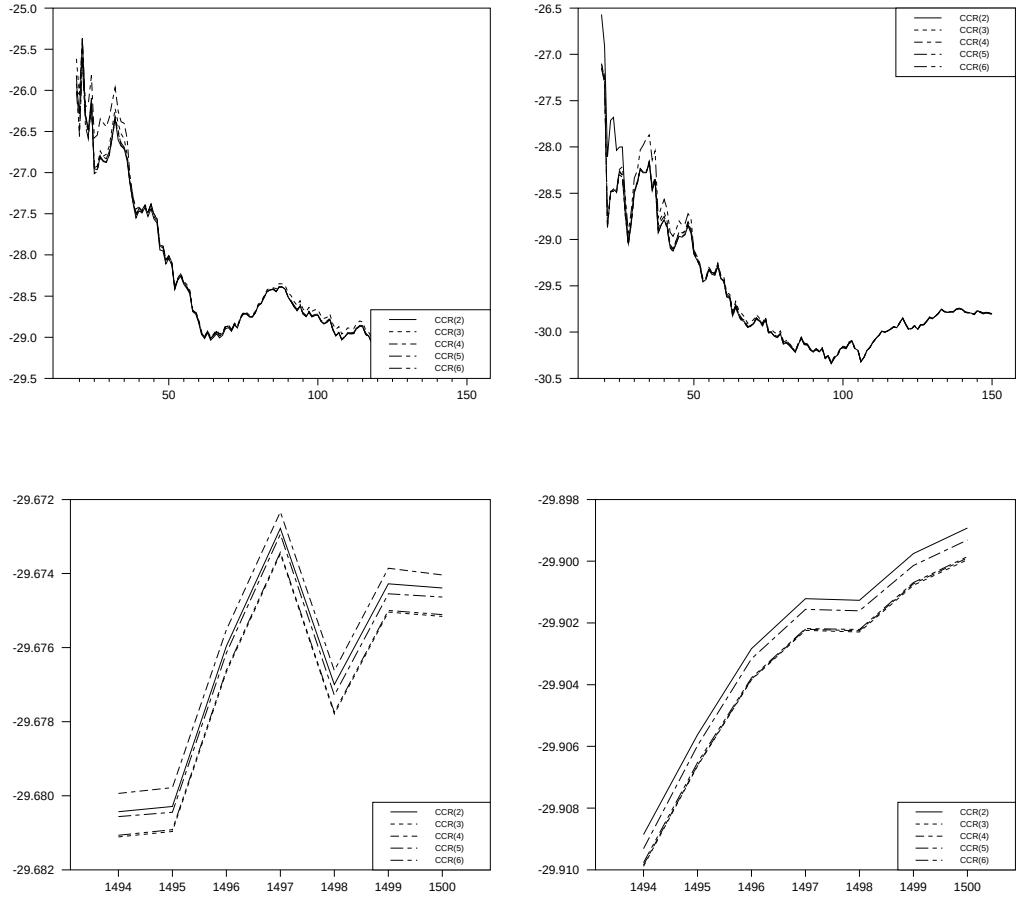
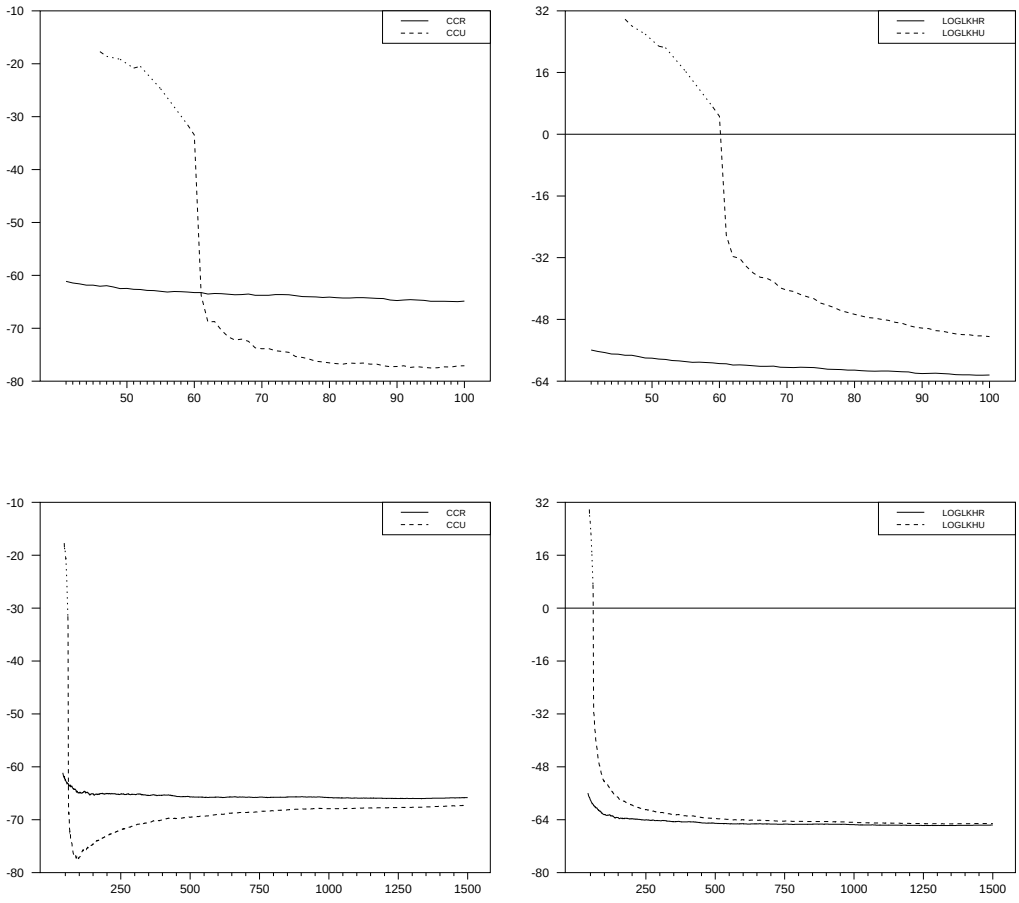


Figure 4: No Order in Non Nested Admissible Models (Sim. 5) CCR(i) is the value of the model selection criterion of Model i.

From the last two graphs we see that even for $T > 1450$ the order of the criterion value for each model changes. We look at the graph on the left side. The criterion value for model 2 is the second largest, while in the graph on the right side the criterion value for model 2 is the largest. The order of the values for other models changes too.

9.1.6 Simulation 7,8: Middle Scale Simultaneous Equations

To assess the impact of the model scale on the performance of the model selection criterion, we study a model with 20 equations. The data generation process is the same as the simulation 1 with $G = 20$. We compare the model selection criterion value of the correctly specified structural model with the unconstrained reduced form. The following graphs show the simulation result.



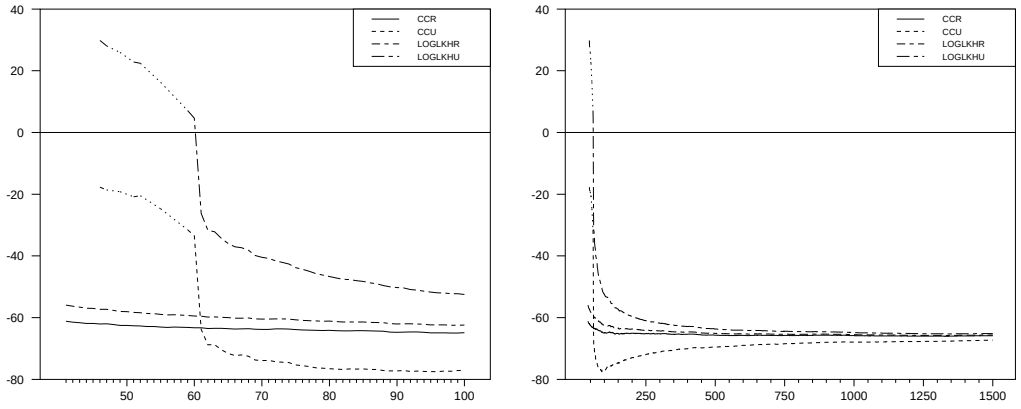


Figure 5: Model Selection Criterion in Case of Middle Scale Models (Sim. 7). CCU is the criterion value of the unconstrained reduced form. CCR is the criterion value of the structural model.

This simulation result shows that for $T > 60$ the model selection criterion will choose the correctly specified structural model. We observe that the average log likelihood of the structural model is almost constantly for $60 < T < 80$, while the log likelihood of the reduced form decreases very fast for $60 < T < 80$. The reason is that for a middle scale model with $G = 20$ and 40 predetermined variables, 60 observations may give a perfect fit. This overfitting decreases with the increase in the number of observations. Hence, the average log likelihood decreases very fast from $T = 60$ to $T = 80$.

For structural models with 4 explanatory variables in each equation, $T = 60$ provide already enough observations to obtain a stable result. Hence the average log likelihood of the structural model is almost constant.

For large scale structural models with 80 equations we do the same simulation. The data are generated by the model specified as in simulation 1 with $G = 80$ and the number of exogeneous variables is 20. Each structural equation has 4 explanatory variables. We compare the structural model with the unconstrained reduced form. The result is shown in the following graphs.

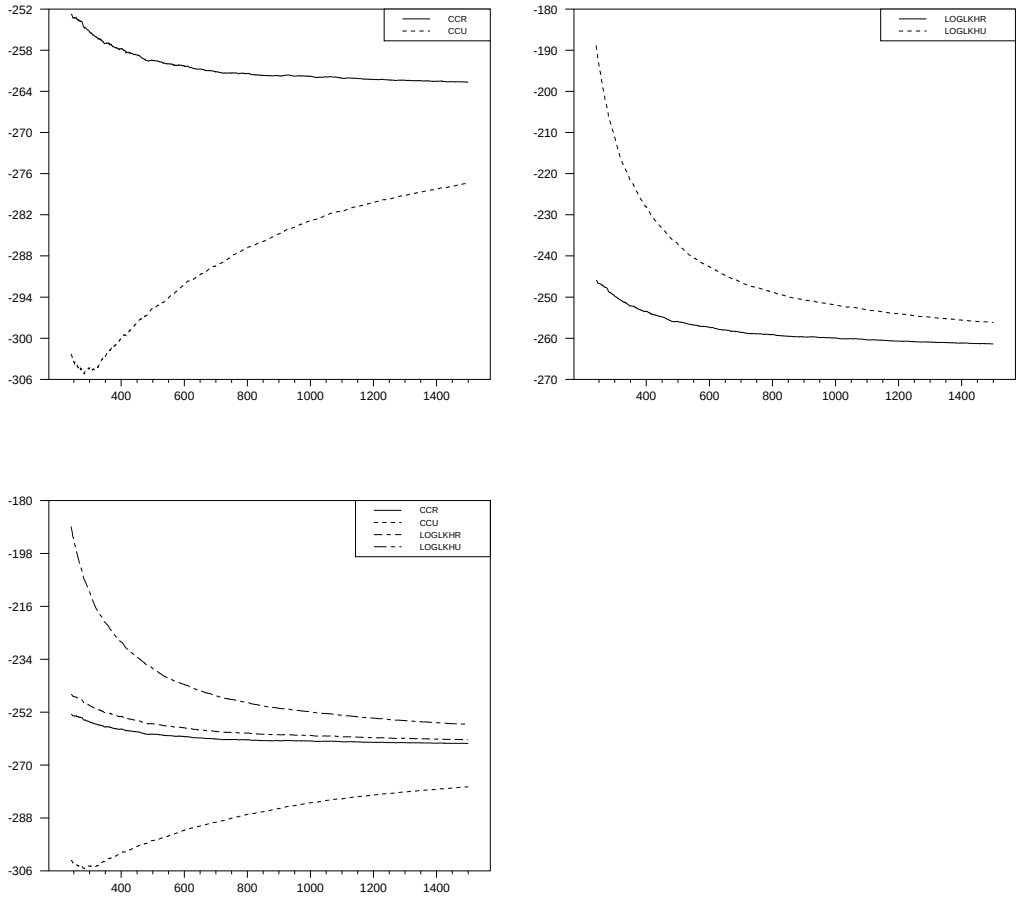


Figure 6: Selection Criterion in the Case of Large Scale Models with a Limited Number of Exogeneous Variables (Sim. 8)

The same phenomenon as in the case of a middle scale model is observed here. The unconstrained reduced form needs at least 200 observations to overcome the problem of overfitting, while the asymptotic property prevails already with 200 observations for the correctly specified structural model. Hence, the value of selection criterion for the structural model keeps almost constantly. The value of the selection criterion for the unconstrained reduced form will be very large at the beginning due to overfitting; it decreases sharply owing to the decrease of overfitting with the increase in the number of observations. The penalty is also large at the beginning. Due to the decrease of the penalty term, the log likelihood of the reduced form will take typically a "V" form, then converge to the limit of the average log likelihood.

9.2 Nonstationary Data

9.2.1 Cointegrated Systems

The DGP for the cointegration system is as follows

$$\Delta x_t = a_{11}x_t + a_{12}y_t + a_{13}z_t + c_{11}\Delta x_{t-1} + c_{12}\Delta y_{t-1} + c_{13}\Delta z_{t-1} + u_{1t}$$
$$\Delta y_t = a_{21}x_t + a_{22}y_t + a_{23}z_t + c_{21}\Delta x_{t-1} + c_{22}\Delta y_{t-1} + c_{23}\Delta z_{t-1} + u_{2t}$$
$$\Delta z_t = a_{31}x_t + a_{32}y_t + a_{33}z_t + c_{31}\Delta x_{t-1} + c_{32}\Delta y_{t-1} + c_{33}\Delta z_{t-1} + u_{3t}$$

with $a_{11} = -0.001, a_{12} = -0.001, a_{13} = -0.0005, \beta_1 = -.10, a_{21} = 0.15, a_{22} = -0.2, a_{23} = 0.2, \beta_2 = .10 a_{31} = \beta_1 a_{11} + \beta_2 a_{21}, a_{32} = \beta_1 a_{12} + \beta_2 a_{22}, a_{33} = \beta_1 a_{13} + \beta_2 * a_{23} c_{11} = 0.02, c_{12} = -.03, c_{13} = -0.01, c_{21} = 0.0, c_{22} = 0.0, c_{23} = 0.03, c_{31} = -0.03, c_{32} = 0.04, c_{33} = 0.05$

This is a cointegration system with one cointegration relation. In the cointegration analysis we investigate the system for 1,2 or no cointegration relations. The following graphs are the result of the simulation. CCN is the criterion value for no cointegration, CCU is the criterion value for 2 cointegration relations and CCR is the value of selections criterion for 1 cointegration relation.

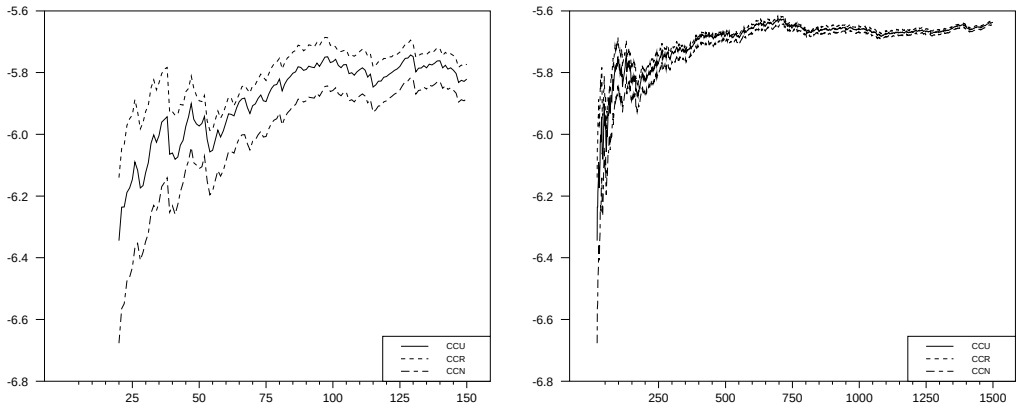


Figure 7: Selection Criterion in the Case of Cointegration Systems (Sim. 9).

In the graph we see that the curve CCR is always above CCN and CCU, hence the model selection criterion clearly identifies the single cointegration relation in the system.

10 Concluding Remarks

Structural models provide a natural framework to interpret economic data and to understand economic phenomenon. They were not popular for nothing during 50s and 60s. Owing to methodological weaknesses structural models as proposed by the Cowles Commission school were under fierce attack in the 70s and 80s[53]. This thesis proposes an alternative way to understand structural models, namely, a structural model is taken as an parsimonious specific representation of the general statistical model - the unconstrained reduced form.

Taking this approach, one has to solve the problem of observational equivalence and model selection. These two issues are the center pieces of this thesis. The existence of observationally equivalent models make the statistical inference on the structural models irresolute. The concept of identification was introduced to solve the problem of parameter estimation due to observational equivalence. We introduce the concept of observational differentiability to characterize the uniqueness of model selection. This thesis gives necessary and sufficient conditions of observational differentiability and answers the question, when a structural model is unique.

For model selection we use the information approach based on the log likelihood function and give a general condition for the (weak) consistent model selection criterion. This consistent criterion provides an alternative way to test the restriction imposed by structural models. It can be interpreted as a test against all possible alternative hypotheses. The probability of type II error for any alternative hypothesis will converge to zero, because the probability to choose a false model converges to zero. The probability of a type I error will also converge to zero. This is the consequence of a consistent criterion.

We further provide the strong consistent criteria for multiple regression models and for structural models. The strong consistency is of great importance for econometrics. Strong consistency is a statement based on a single realization path of a stochastic process. In the empirical economic research we have usually only a single realization path. Therefore, this kind of strong consistency is more relevant than the weak consistency.

Further research work along this line of structural modeling can be conducted in two ways. One way is to assess the practical applicability of the consistent model selection criteria and gain experience with testing the restrictions imposed by structural models; the other way is to develop strong consistence criterion for nonlinear restriction on parameters and variables, because all structural models are genuinely nonlinear. Either way there is still a much work to be done.

[53] See Leamer (1983), Lucas (1976)

A Structural Models

Conventional assumptions about a structural simultaneous equations system are as follows:[54]

$$BY_t + \Gamma X_t = U_t \qquad \text{for t=1,2,...T} \tag{A.1}$$

1: B is a nonsingular $G \times G$ parameter matrix and is normalized such that all the elements on the diagonal equal unit. Γ is a $G \times K$ parameter matrix.

2: $Y_t \in \mathbb{R}^{G\times 1}$ is a random variable called the endogenous variable.

3: $X_t \; in \mathbb{R}^{K\times 1}$ is called the predetermined variable with :

$$X_t' = (Y_{t-1}', Y_{t-2}', ...Y_{T-p}, \xi_t')$$

ξ_t is the exogenous deterministic variable with:

$$\underset{T\to\infty}{\text{plim}} \frac{\sum_{t=1}^{T} \xi_t \xi_t}{T} = M_{\xi\xi}.$$

4: U_T is identically independently distributed as $N(0, \Sigma)$

5: X_t and U_t are uncorrelated.

$$E(X_t U_t') = 0$$

6: We rewrite the model explicitly in the lags of Y_T:

$$BY_t + \Gamma_t X_t = BY_t + \Gamma_1 Y_{t-1} + \Gamma_2 Y_{t-2} + ... + \Gamma_p Y_{t-p} + \Gamma_\xi \xi_t + Ut$$

The stability condition is: $\max\{|\lambda_i| | i = 1, 2, 3, ...GP\} > 1$
λ_i is the i-th root of the following equation:

$$|B + \Gamma_1 L^1 + \Gamma_2 L^2 + ... + \Gamma_p L^p| = 0$$

6b: The initial value of Y_t is given.

[54] Compare Dhrymes (1993) p. 12

B Proof

B.1 Notations and Probability Space

Notations
r.v. = random variable
$\|A\| = \max_{ij} |a_{ij}|$, where $A = (a_{ij})_{ij}$

Variables
Let t represent the time and $t = 1, 2, \cdots$. Let the random variables $Y_t \in \mathbb{R}^{G\times 1}$ represent endogeneous variables at t. Let the vector ξ_t represent the exogeneous variables at t. Let $X_t = (Y_{t-1}, \cdots, Y_{t-p}, \xi_t)$ represent the predetermined variables. Let y_t and x_t be one realization for Y_t and X_t respectively. y_t and x_t can be observed. Throughout the text we will always write capital letters for r.v's and lower case for one realization.

Model
Let $\Theta \subset \mathbb{R}^M$ be the parameter set. Θ is compact.

We consider the model

$$y_t = F(x_t, \theta) + v_t, \tag{B.2}$$

where v_t is a realization for V_t. The random variable $V_t \in \mathbb{R}^{G\times 1}$ represents noises/innovations with a density function $\phi_\theta(v)$ depending on θ. Moreover, $V_t, t \in \mathbb{N}$ are i.i.d. and $\int v\,\phi_\theta(v)dv = 0,\ \forall\theta \in \Theta$. The initial conditions $Y_0 = y_0, \cdots, Y_{1-p} = y_{1-p}$ are known. We note $\mathbf{y}_0 = (y_0, \cdots, y_{1-p})$.

Probability Space
The stochastic in this model comes from V_t. We let $\Omega = (\mathbb{R}^G)^{\mathbb{N}}$ be the sample space for the model, because it describes all possible realizations of $V_t, t \in \mathbb{N}$. A sample point in this space is $\omega = (v_1, v_2, \cdots)$. The probability measure P_θ on Ω is determined by $\phi_\theta(v)$ the density function of V_t and their independence.

Maximum Likelihood Estimation (MLE)
We introduce, at first, likelihood functions. The likelihood function is defined as the joint density function of variables of relevance [55]. Here the endogenous variable Y_t is such a variable. Let $\mathbf{y}_T = (y_1, \cdots, y_T)$ be the observations of Y_t over T. According to the model (B.2) $\mathbf{y}_T$ is determined by a given sequence of exogeneous variables $\Xi_T = (\xi_1, \cdots, \xi_T)$ and a sample point $\omega_T = (v_1, \cdots, v_T)$. We denote $L_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0)$ as the likelihood function of $\mathbf{Y}_T = (Y_1, \cdots, Y_T)$ with respect to θ given Ξ_T and $\mathbf{y}_0$. We show

[55] See Hendry (1995) p.371.

Theorem 2.1

$$L_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0) = \prod_{t=1}^{T} \phi_\theta\big(y_t - F(x_t, \theta)\big). \tag{B.3}$$

Proof
We show it step by step.
We know $x_1 = (y_0, \cdots, y_{1-p}\, \xi_1)$ is given. So the density function of $Y_1|_{x_1}$ is $\phi_\theta\big(y_1 - F(x_1, \theta)\big)$ and therefore $L_1(\theta; \mathbf{y}_1|\Xi_1, \mathbf{y}_0) = \phi_\theta\big(y_1 - F(x_1, \theta)\big)$.
Let $\psi_2(x_2|\Xi_2, \mathbf{y}_0)$ be the density function of $X_2|_{\mathbf{y}_0, \Xi_2}$. Because in $X_2 = (Y_1, y_0, \cdots, y_{2-p}, \xi)$ only Y_1 is a random variable, then the density function of X_2 is the same as the density function of Y_1 $\psi_2(x_2|\Xi_2, \mathbf{y}_0) = \phi_\theta\big(y_1 - F(x_1, \theta)\big)$.
It is clear that the (conditional) density function of Y_2 given $X_2 = x_2$ is $\phi_\theta\big(y_2 - F(x_2, \theta)\big)$, $\forall x_2$. Therefore the joint density function of (Y_1, Y_2) given $\Xi_2, \mathbf{y}_0$ is given by

$$L_2(\theta; \mathbf{y}_2, |\Xi_2, \mathbf{y}_0) = \phi_\theta\big(y_2 - F(x_2, \theta)\big)\phi_\theta\big(y_1 - F(x_1, \theta)\big).$$

We continue in the same way and get (B.3).
□

By observation we obtain $\mathbf{y}_T$ and $(\Xi_T, \mathbf{y}_0)$. We assume $\mathbf{y}_T$ is one realization of $\mathbf{Y}_T$ which obeys the model (B.2) with some unknown $\theta_0 \in \Theta^\circ$, where Θ° represent the interior of Θ. We call θ_0 the true parameter. The maximum likelihood method is to infer θ_0 based on $\mathbf{y}_T$ (given $\Xi_T, \mathbf{y}_0$). The maximum likelihood estimator is defined by

$$\hat{\theta}(\mathbf{y}_T|\Xi_T, \mathbf{y}_0) = \arg\max_{\theta\in\Theta} L_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0).$$

Let $l_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0) = \ln L_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0)$. We assume that $l_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0)$ has a continuous third differential with respect to θ, $\forall T$ and $\forall \mathbf{y}_T, \Xi_T$. Let $d_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0) = \frac{\partial}{\partial\theta} l_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0)$ and $D_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0) = \frac{\partial^2}{\partial\theta\partial\theta'} l_T(\theta; \mathbf{y}_T|\Xi_T, \mathbf{y}_0)$.
In the following text we consider that the data are generated based on some fixed exogeneous variables Ξ_T and initial points $\mathbf{y}_0$. So sometimes we will omit them if it does cause confusion.

B.2 The Law of Iterated Logarithm for Martingales

Let $\mathcal{F}_t$ be the σ-algebra generated by $U_1, \cdots, U_t$. Let $(Z_t)_{t\in\mathbb{N}}$ be a (multidimensional) $(P, (\mathcal{F}_t)_{t\in\mathbb{N}})$ martingale and $\Delta Z_t = Z_t - Z_{t-1}$. The *variation* of a martingale $(Z_t)_{t\in\mathbb{N}}$ is defined by

$$V(Z_T) = \sum_{t=1}^{T} E[\Delta Z_t \cdot \Delta Z_t' | \mathcal{F}_{t-1}].$$

Theorem 2.2 *(Corollary 4, p.220, Wang, (1993)) Let $(Z_t)_{t\in\mathbb{N}}$ be now a one dimensional martingale. Let $\sigma_t^2 = E[\Delta Z_t^2|\mathcal{F}_{t-1}]$ and $V(Z_t) = \sum_{t=1}^{T}\sigma_t^2$. If*

(W1)

$$\lim_{T\to\infty} V(Z_T) = \infty \qquad a.s., \tag{B.4}$$

(W2) there is a $\delta > 0$, such that

$$\sup_t \frac{E\big[|\Delta Z_t|^{2+\delta}|\mathcal{F}_{t-1}\big]}{\sigma_t^{2+\delta}} < \infty \qquad a.s., \tag{B.5}$$

(W3) and for a α with $0 < \alpha < 1$

$$\sigma_T^2 = o(V(Z_T)^\alpha), \tag{B.6}$$

then

$$P\left(\overline{\lim_{T\to\infty}} \frac{Z_T}{\sqrt{2V(Z_T)\ln\ln V(Z_T)}} = 1\right) = 1. \tag{B.7}$$

B.3 The Asymptotically Behavior of Likelihood Ratios

Here we test the null restrictions on the parameters

$$H_0 : \{R_1(\theta) = 0, \cdots, R_r(\theta) = 0.\}$$

Let $R(\theta) = \big(R_1(\theta), \cdots, R_r(\theta)\big)$. Let $\hat{\theta}(\mathbf{y}_T)$ be the ML-estimators with respect to the restriction $R(\theta) = 0$. According to the Lagrange multiplicator method we will maximize the object function

$$\Psi(\theta, \lambda) = l_T(\theta; \mathbf{y}_T) + R(\theta)\lambda,$$

where $\lambda = (\lambda_1, \cdots, \lambda_r)'$. Let $\hat{\lambda}$ be the maximizer. Then the following equations are satisfied:

$$R\big(\hat{\theta}(\mathbf{y}_T)\big) = 0 \tag{B.8}$$

$$d_T\big(\hat{\theta}(\mathbf{y}_T)\big) + \dot{R}\big(\hat{\theta}(\mathbf{y}_T)\big)\hat{\lambda} = 0 \tag{B.9}$$

In order to denote the subspace of the parameter space satisfying H_0 we reparameterize θ with

$$\theta = h(\eta),$$

where the dimension of η equals $M - r$. η represents the free parameters under the restriction $R(\theta) = 0$. Thus $R(h(\eta)) = 0$, $\forall\eta$. Let η_0 denote

the true parameter under the reparametrization $h(\eta_0) = \theta_0$. Moreover, we assume that $\frac{\mathrm{d}h}{\mathrm{d}\eta}$ exists and is continuous in some neighborhood of η_0.

A natural choice of η is a subset of θ and let the rest of the parameters be functions of η. To obtain this we simply need to find r parameters, say, under rearrangement of the parameters, $\theta^{(M-r+1)}, \cdots, \theta^{(M)}$, such that $\frac{\partial R}{\partial \theta^{(M-r+j)}}(\theta_0) \neq 0, \forall j = 1, \cdots r$. Then, following the implicit function theorem, the parameter $\theta^{(M-r+j)}$ can be represented as a function of $\eta = \{\theta^{(1)}, \cdots \theta^{(M-r)}\}$ $\forall j = 1, \cdots, r$ and $\frac{\partial \theta^{(M-r+j)}}{\partial \eta}$ exists and is continuous in some neighborhood of θ_0.

The constrained likelihood function can be rewritten as $l_T(\theta; \mathbf{y}_t) = l_T(h(\eta); \mathbf{y}_T)$. We let $g_T(\eta; \mathbf{y}_T) = \frac{\partial}{\partial \eta} l_T(h(\eta); \mathbf{y}_T)$ and $G_T(\eta; \mathbf{y}_T) = \frac{\partial^2}{\partial \eta \partial \eta'} l_T(h(\eta); \mathbf{y}_T)$ be the differentials of the log likelihood function subject to restrictions. Let $\hat{\eta}$ be the maximizer of $l_T(h(\eta); \mathbf{y}_T)$. Then $g_T(\hat{\eta}; \mathbf{y}_T) = 0$ and $h(\hat{\eta}) = \hat{\theta}(\mathbf{y}_T)$ satisfies the equations (B.8) and (B.9).

The theorem about likelihood ratios (e.g. Godfrey (1988)) states that the limit distribution of the likelihood ratio $l_T\big(\tilde{\theta}(\mathbf{Y}_T)\big) - l_T\big(\hat{\theta}(\mathbf{Y}_T)\big)$ is χ^2-distribution under some conditions. The following theorem characterizes furthermore the limit barrier of the likelihood ratios. As proof for likelihood ratios we need the law of large numbers for averages of likelihood functions. It is worthy to note that the limits of the averages in our theorem depend on $\Xi = (\mathbf{y}_0, \xi_1, \xi_2, \cdots)$. The expectations in the following equations are taken with respect to P_{θ_0} and "a.s." means $P_{\theta_0}-$a.s. Now we state at first the conditions of the theorem:

The conditions

(A1) $\forall \theta \in \Theta, \exists\, \bar{l}(\theta, \Xi)$, such that

$$\lim_{T\to\infty} \frac{1}{T} E l_T(\theta; \mathbf{Y}_T|\Xi) = \bar{l}(\theta, \Xi)$$

and $\bar{l}(\theta, \Xi)$ has one unique maximum at θ_0 over Θ.

(A2) The convergence in (A1) holds locally uniformly, i.e. $\forall \theta \in \Theta^\circ, \exists \mathcal{N}(\theta) \subset \Theta^\circ$, such that

$$\lim_{T\to\infty} \sup_{\theta' \in \mathcal{N}(\theta)} \big| \frac{1}{T} E l_T((\theta; \mathbf{Y}_T|\Xi) - \bar{l}(\theta', \Xi) \big| = 0.$$

(A3) The strong law of large number holds locally uniformly, i.e. $\forall \theta \in \Theta^\circ, \exists \mathcal{N}(\theta) \subset \Theta^\circ$, such that

$$\lim_{T\to\infty} \sup_{\theta' \in \mathcal{N}(\theta)} \big| \frac{1}{T} l_T(\theta; \mathbf{Y}_T|\Xi) - \frac{1}{T} E l_T(\theta; \mathbf{Y}_T|\Xi) \big| = 0 \qquad \text{a.s. .}$$

(A4) $\forall\theta \in \Theta, \exists\bar{D}(\theta, \Xi)$, such that

$$\lim_{T\to\infty} \frac{1}{T} D_T(\theta; \mathbf{Y}_T|\Xi) = \bar{D}(\theta, \Xi) \qquad \text{a.s. .}$$

$\bar{D}(\theta, \Xi)$ is continuous in θ. $\bar{D}(\theta_0, \Xi)$ is invertible.

(A4') $\forall\eta, \exists\bar{G}(\eta, \Xi)$, such that

$$\lim_{T\to\infty} \frac{1}{T} G_T(\eta; \mathbf{Y}_T|\Xi) = \bar{G}(\eta, \Xi) \qquad \text{a.s. .}$$

$\bar{G}(\eta, \Xi)$ is continuous in η. $\bar{G}(\eta_0, \Xi)$ is invertible.

(A5) The convergence in (A4) and (A4') holds locally uniformly, i.e. $\forall\theta \in \Theta, \exists\mathcal{N}(\theta)$, such that

$$\lim_{T\to\infty} \sup_{\theta'\in\mathcal{N}(\theta)} \left\| \frac{1}{T} D_T(\theta'; \mathbf{Y}_T|\Xi) - \bar{D}(\theta', \Xi) \right\| = 0 \qquad \text{a.s. .}$$

(A6) $E[d_t(\theta_0; \mathbf{Y}_t)] = 0, \forall t$ and moreover $(d_t(\theta_0; \mathbf{Y}_t))_{t\in\mathbb{N}}$ is a martingale. It exists $\bar{V}_d(\Xi) \in \mathbb{R}^{M\times M}$, such that

$$\lim_{T\to\infty} \frac{V(d_T(\theta; \mathbf{Y}_T|\Xi))}{T} = \bar{V}_d(\Xi) \qquad \text{a.s.,}$$

where V denote the variation of a martingale.

(A6') $E[g_t(\eta_0; \mathbf{Y}_t, \mathbf{X}_t)] = 0, \forall t$ and moreover $(g_t(\eta_0); \mathbf{Y}_t, \mathbf{X}_t)_{t\in\mathbb{N}}$ is a martingale. It exists $\bar{V}_g(\Xi) \in \mathbb{R}^{M\times M}$, such that

$$\lim_{T\to\infty} \frac{V(g_T(\eta; \mathbf{Y}_T|\Xi))}{T} = \bar{V}_g(\Xi) \qquad \text{a.s..}$$

(A7) For any $\alpha \in \mathbb{R}^{M\times 1}$, the weighted martingale $\big(\alpha'(d_t(\theta_0; \mathbf{Y}_T|\Xi))\big), t \in \mathbb{N}$ satisfies the assumptions (W2) and (W3).

(A7') For any $\alpha \in \mathbb{R}^{M\times 1}$, the weighted martingale $\big(\alpha'(g_t(\eta_0; \mathbf{Y}_T|\Xi))\big), t \in \mathbb{N}$ satisfies the assumptions (W2) and (W3).

(A8)

$$\bar{V}_d(\Xi) = -\bar{D}(\theta_0, \Xi).$$

(A9) $\bar{D}(\theta, \Xi)$ is differentiable and

$$\sup_{\theta\in\Theta} \left\| \frac{\partial}{\partial\theta} \bar{D}(\theta, \Xi) \right\| < \infty$$

(A10) The convergence in (A5) is "fast" in the sense that there exists a constant c, such that

$$\overline{\lim_{T\to\infty}} \frac{\sup_{\theta'\in\mathcal{N}(\theta)} \left\| \frac{1}{T} D_T(\theta'; \mathbf{Y}_T|\Xi) - \bar{D}(\theta', \Xi) \right\|}{\sqrt{\ln\ln T/T}} \leq c \qquad \text{a.s.}$$

(A11) The second differential $\frac{\mathrm{d}^2 R}{\mathrm{d}\theta \mathrm{d}\theta'}$ exists and is continuous in the neighborhood of θ_0.

Theorem 2.3 *Let r be the number of the hypothesis H_0. If the hypothesis H_0 is true and all the conditions (A1) - (A11) and (A4') - (A7')are satisfied, then for such Ξ we have*

$$\overline{\lim_{T\to\infty}} \frac{\left(l_T(\tilde{\theta}_T; \mathbf{Y}_T|\Xi) - l_T(\hat{\theta}_T; \mathbf{Y}_T|\Xi)\right)}{r \ln\ln T} \leq 1 \qquad \textit{a.s.}\ .$$

Remark. We compare our conditions with those in Godfrey(1988) p.6-13. Our idea of how to represent the log likelihood ratio is exactly the same as Godfrey's. The difference is that he uses the central limit theorem (for martingales) to obtain the limit distribution of the log likelihood ratio, while we use the law of iterated logarithm (LIL) (for martingales) to obtain the barrier for the log likelihood ratio. Thus we need the conditions (A6) and (A7) instead of his conditions (viii).

Since the LIL requires a.s.-convergence, whereas the central limit theorem requires only convergence in distribution, we must add conditions such that all convergences in probability in his proof can be formulated as a.s.-convergences. This is why we require the a.s.-convergence in (A3) and (A4) and additionally the speed of the convergences in (A9) and (A10).

The condition (A8) corresponds Godfrey's (4.4) on p.6 which say something about the relationship between the first and second differentials of l_T

In the proofs we abbreviate $\tilde{\theta}(\mathbf{Y}_T)$ $\hat{\theta}(\mathbf{Y}_T)$ as $\tilde{\theta}_T, \hat{\theta}_T$.

Lemma 2.4 *If the assumptions (A1), (A2) and (A3) are satisfied, then the ML-estimators are strongly consistent:*

$$\lim_{T\to\infty} \tilde{\theta}_T = \theta_0 \qquad \textit{a.s.},$$
$$\lim_{T\to\infty} \hat{\theta}_T = \theta_0 \qquad \textit{a.s.}.$$

For the proof see Frydman (1980).

Lemma 2.5 *If the assumptions (A1) - (A7) and (A4') - (A7') and (A11) are satisfied, then there exist c_1 and c_2, such that*

$$\overline{\lim_{T\to\infty}} \frac{\left\|\sqrt{T}(\tilde{\theta}_T - \theta_0)\right\|}{\sqrt{\ln\ln T}} \leq c_1 \,, \tag{B.10}$$

$$\overline{\lim_{T\to\infty}} \frac{\left\|\sqrt{T}(\hat{\theta}_T - \theta_0)\right\|}{\sqrt{\ln\ln T}} \leq c_2 \,. \tag{B.11}$$

Lemma 2.6 *If (A1)-(A7), (A4') - (A7'),(A9) and (A10)- (A11) are satisfied, then there exists c_3 and c_4, such that*

$$\overline{\lim_{T\to\infty}} \frac{\left\|\frac{1}{T}D_T\big(b(\tilde{\theta},\hat{\theta})\big) - \bar{D}(\theta_0)\right\|}{\sqrt{\ln\ln T/T}} \leq c_3 \qquad a.s. \,, \tag{B.12}$$

$$\overline{\lim_{T\to\infty}} \frac{\left\|\frac{1}{T}D_T\big(b(\tilde{\theta},\theta_0)\big) - \bar{D}(\theta_0)\right\|}{\sqrt{\ln\ln T/T}} \leq c_4 \qquad a.s. \,, \tag{B.13}$$

We introduce a new notation. Let $A_T \overset{a.a.s}{\sim} B_T$ mean $\lim_{T\to\infty}\left\|A_T - B_T\right\| = 0$ P-a.s., where "a.a.s" represents "asymptotically almost surely".

We abbreviate $\frac{\mathrm{d}_R}{\mathrm{d}\theta}(\theta)$ as $\dot{R}(\theta)$, $\frac{\mathrm{d}_R}{\mathrm{d}\theta}(\theta_0)$ as $\dot{R}_0$, and also $\bar{D}(\theta_0)$ as $\bar{D}_0$.

Theorem 2.7

$$\frac{\hat{\lambda}}{\sqrt{T}} \overset{a.a.s}{\sim} \big(-\dot{R}_0'\bar{D}_0^{-1}\dot{R}_0\big)^{-1}\dot{R}_0'\bar{D}_0^{-1}\frac{d_T(\theta_0)}{\sqrt{T}} \tag{B.14}$$

Proof
Using the Taylor expansion we have

$$\frac{1}{\sqrt{T}}d_T(\hat{\theta}_T) = \frac{1}{\sqrt{T}}d_T(\theta_0) + \bar{D}(\theta_0)\sqrt{T}(\hat{\theta}_T - \theta_0) + \delta_{1,T}, \tag{B.15}$$

where

$$\delta_{1,T} = \big(\frac{1}{T}D_T(b(\hat{\theta}_T,\theta_0) - \bar{D}(\theta_0)\big)\sqrt{T}(\hat{\theta}_T - \theta_0).$$

Using (B.20) and (B.12) we can have

$$\overline{\lim_{T\to\infty}}\left\|\delta_{1,T}\right\| \leq \overline{\lim_{T\to\infty}} c\frac{(\ln\ln T)}{\sqrt{T}} = 0 \qquad \text{a.s..}$$

Thus we have

$$\frac{1}{\sqrt{T}}d_T(\hat{\theta}_T) \overset{a.a.s}{\sim} \frac{1}{\sqrt{T}}d_T(\theta_0) + \bar{D}(\theta_0)\sqrt{T}(\hat{\theta}_T - \theta_0). \tag{B.16}$$

Multiplying the both sides of the equation above with $\dot{R}_0'\bar{D}_0^{-1}$ and using (B.8) we have

$$\frac{1}{\sqrt{T}}\dot{R}_0'\bar{D}_0^{-1}d_T(\hat{\theta}_T) \overset{a.a.s}{\sim} \frac{1}{\sqrt{T}}\dot{R}_0'\bar{D}_0^{-1}d_T(\theta_0) + \sqrt{T}\dot{R}_0'(\hat{\theta}_T - \theta_0). \tag{B.17}$$

We show here the last term above $\lim_{T\to\infty}\sqrt{T}\dot{R}_0'(\hat{\theta}_T - \theta_0) = 0$ P-a.s.. Since $R(\hat{\theta}_T) = 0, R(\theta_0) = 0$, we can find a θ_T^* between $\hat{\theta}_T$ and θ_0 such that

$$0 = R(\hat{\theta}_T) - R(\theta_0) = \dot{R}(\theta_T^*)'(\hat{\theta}_T - \theta_0).$$

Then for $i = 1, 2, ...r$

$$\sqrt{T}\frac{\mathrm{d}R_i}{\mathrm{d}\theta'}(\theta_0)(\hat{\theta}_T - \theta_0) = \big(\frac{\mathrm{d}R_i}{\mathrm{d}\theta'}(\theta_0) - \frac{\mathrm{d}R_i}{\mathrm{d}\theta'}(\theta_T^*)\big)\sqrt{T}(\hat{\theta}_T - \theta_0)$$

$$= \underbrace{(\theta_T^* - \theta_0)'}_{\leq c\sqrt{\ln\ln T/T}\ P\text{-a.s.}} \underbrace{\frac{\mathrm{d}^2R_i}{\mathrm{d}\theta\mathrm{d}\theta'}(\theta_T^{**})}_{\leq K\ P\text{-a.s.}} \underbrace{\sqrt{T}(\hat{\theta}_T - \theta_0)}_{\leq \tilde{c}\sqrt{\ln\ln T}\ P\text{-a.s.}} \ \to\ 0 \qquad \text{a.s..} \tag{B.18}$$

Replacing $d_T(\hat{\theta}_T)$ in the equation (B.16) using (B.8) we have

$$-\frac{1}{\sqrt{T}}\dot{R}_0'\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T)\hat{\lambda} \overset{a.a.s}{\sim} \frac{1}{\sqrt{T}}\dot{R}_0'\bar{D}_0^{-1}d_T(\theta_0).$$

Because of the strong consistency of $\hat{\theta}_T$ and continuity of $\dot{R}(\theta)$, it follows that $\lim_{T\to\infty}\dot{R}_0'\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T) = \dot{R}_0'\bar{D}_0^{-1}\dot{R}(\theta_0)$ P-a.s.. It follows, therefore[56]

$$\frac{\hat{\lambda}}{\sqrt{T}} \overset{a.a.s}{\sim} \big(-\dot{R}_0'\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T)\big)^{-1}\dot{R}_0\bar{D}_0^{-1}\frac{d_T(\theta_0)}{\sqrt{T}}.$$

From (A6) and (A7) we have $\frac{d_T(\theta_0)}{\sqrt{T}} \leq c\sqrt{\ln\ln T}$. Moreover, with the same reason as in the equation (B.18) we have the P-a.s. convergence $\dot{R}_0'\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T) \to \dot{R}_0'\bar{D}_0^{-1}\dot{R}_0$ with the speed $\sqrt{\ln\ln T/T}$. Therefore

$$\big(\dot{R}_0'\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T)\big)^{-1}\dot{R}_0\bar{D}_0^{-1}\frac{d_T(\theta_0)}{\sqrt{T}} \overset{a.a.s}{\sim} \big(\dot{R}_0'\bar{D}_0^{-1}\dot{R}_0\big)^{-1}\dot{R}_0\bar{D}_0^{-1}\frac{d_T(\theta_0)}{\sqrt{T}}.$$

□

[56]Let A_T and B_T be one-dimensional r.v.'s for $T \in \mathbb{N}$. Let a and b be two finite constants. If $\lim_{T\to\infty} A_T = a$ P-a.s. and $\lim_{T\to\infty} B_T = b$ P-a.s., then $\lim_{T\to\infty} A_T B_T = ab$ P-a.s.. Therefore

$$\big(\dot{R}(\theta_0)'\bar{D}(\theta_0)^{-1}\dot{R}(\hat{\theta}_T)\big)^{-1}\Big(\dot{R}(\theta_0)'\bar{D}(\theta_0)^{-1}\dot{R}(\hat{\theta}_T)\frac{\hat{\lambda}}{\sqrt{T}} - \dot{R}(\theta_0)'\bar{D}(\theta_0)^{-1}\frac{d_T(\theta_0)}{\sqrt{T}}\Big)$$
$$\to S\cdot 0 = 0 \qquad \text{a.s..}$$

Theorem 2.8

$$2\big(l_T(\tilde{\theta}_T) - l_T(\hat{\theta}_T)\big) \overset{a.a.s}{\sim} -\frac{\hat{\lambda}'}{\sqrt{T}}\dot{R}_0'\bar{D}_0^{-1}\dot{R}_0\frac{\hat{\lambda}}{\sqrt{T}} \tag{B.19}$$

Proof of Theorem 2.8
From lemma 2.5 we have

$$\overline{\lim_{T\to\infty}} \frac{\big\|\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T)\big\|}{\sqrt{\ln\ln T}} \;\leq\; c_1 + c_2 \tag{B.20}$$

Using Taylor expansion we obtain

$$\begin{aligned} 2\big(l_T(\hat{\theta}) - l_T(\tilde{\theta})\big) =& 2\underbrace{d_T(\tilde{\theta}_T)}_{=0}(\hat{\theta}_T - \tilde{\theta}_T) \\ &+ \sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T)'\bar{D}_0\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) + \delta_{2,T}\ , \end{aligned} \tag{B.21}$$

where

$$\delta_{2,T} = \sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T)'\big(\frac{1}{T}D_T(b(\tilde{\theta},\hat{\theta})) - \bar{D}(\theta_0)\big)\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T).$$

Using (B.20) and (B.12)

$$\overline{\lim_{T\to\infty}} \big\|\delta_{2,T}\big\| \leq \overline{\lim_{T\to\infty}} c\frac{(\ln\ln T)^{\frac{3}{2}}}{\sqrt{T}} = 0 \qquad \text{a.s..}$$

Now consider another expansion

$$\frac{1}{\sqrt{T}}d_T(\hat{\theta}_T) = \frac{1}{\sqrt{T}}\underbrace{d_T(\tilde{\theta}_T)}_{=0} + \bar{D}(\theta_0)\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) + \delta_{3,T}\ , \tag{B.22}$$

where

$$\delta_{3,T} = \big(\frac{1}{T}D_T(b(\hat{\theta}_T,\tilde{\theta}_T) - \bar{D}(\theta_0)\big)\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T).$$

Using again (B.20) and (B.12) we have also

$$\overline{\lim_{T\to\infty}} \big\|\delta_{3,t}\big\| \leq \overline{\lim_{T\to\infty}} c\frac{(\ln\ln T)}{\sqrt{T}} = 0 \qquad \text{a.s..}$$

Multiplying both sides of the equation (B.22) with $(-\bar{D}_0)^{-1/2}$ and using (B.9) it follows that

$$(-\bar{D}_0)^{\frac{1}{2}}\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) \overset{a.a.s}{\sim} \frac{1}{\sqrt{T}}(-\bar{D}_0)^{-\frac{1}{2}}d_T(\hat{\theta}_T) = -\frac{1}{\sqrt{T}}(-\bar{D}_0)^{-\frac{1}{2}}\dot{R}(\hat{\theta}_T)'\hat{\lambda}. \tag{B.23}$$

Following Theorem 2.7 we have

$$\frac{\hat{\lambda}}{\sqrt{T}} \leq \tilde{c}\sqrt{\ln\ln T},$$

for some $\tilde{c}$ and with the same reason as in (B.18) the speed of the convergence $\dot{R}(\hat{\theta}_T) \to \dot{R}_0$ is $\sqrt{\ln\ln T/T}$. Therefore, it follows that

$$\bar{D}_0^{-1}\dot{R}(\hat{\theta}_T)\frac{\lambda}{\sqrt{T}} \overset{a.a.s}{\sim} \bar{D}_0^{-1}\dot{R}_0\frac{\lambda}{\sqrt{T}}. \tag{B.24}$$

Using (B.21) (B.23) and (B.24) we can obtain

$$\begin{aligned} 2\big(l_T(\tilde{\theta}_T) - l_T(\hat{\theta}_T)\big) &\overset{a.a.s}{\sim} -\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T)'\bar{D}_0\sqrt{T}(\hat{\theta}_T - \tilde{\theta}_T) \\ &\overset{a.a.s}{\sim} \frac{\hat{\lambda}'}{\sqrt{T}}\dot{R}_0'(-\bar{D}_0)^{-\frac{1}{2}}(-\bar{D}_0)^{-\frac{1}{2}}\dot{R}_0\frac{\hat{\lambda}}{\sqrt{T}}. \end{aligned}$$

□

Proof of Theorem 2.3
Following (A8), $-\bar{D}(\theta_0) = \bar{V}_d$ is positive definite. Therefore $\bar{V}_d^{\frac{1}{2}}$ is well-defined.

Let $S = -\dot{R}_0'\bar{D}_0^{-1}\dot{R}_0 \in \mathbb{R}^{r\times r}$. Then S is positive definite and we have $S^{\frac{1}{2}}$ well-defined.

Using Theorem 2.7 and Theorem 2.8 we can get

$$2\big(l_T(\tilde{\theta}_T) - l_T(\hat{\theta}_T)\big) \overset{a.a.s}{\sim} \frac{\hat{\lambda}'}{\sqrt{T}}\, S\, \frac{\hat{\lambda}'}{\sqrt{T}} \overset{a.a.s}{\sim} \big(S^{-\frac{1}{2}}\dot{R}_0'\bar{V}_d^{-1}\frac{d_T(\theta_0)}{\sqrt{T}}\big)'\big(S^{-\frac{1}{2}}\dot{R}_0'\bar{V}_d^{-1}\frac{d_T(\theta_0)}{\sqrt{T}}\big).$$

Let

$$Z_T = S^{-\frac{1}{2}}\dot{R}_0'\bar{V}_d^{-1}d_T(\theta_0)$$

and $Z_T^{(i)}$ be the i-th component of Z_T. It is clear that $(Z_T), T \in \mathbb{N}$ is a martingale and

$$\lim_{T\to\infty}\frac{V(Z_T)}{T} = S^{-\frac{1}{2}}\dot{R}_0'\bar{V}_d^{-1}\lim_{T\to\infty}\frac{V(d_T(\theta_0))}{T}\bar{V}_d^{-1}\dot{R}_0S^{-\frac{1}{2}} = S^{-\frac{1}{2}}SS^{-\frac{1}{2}} = I_{r\times r} \quad a.s. .$$

Therefore

$$\lim_{T\to\infty}\frac{V(Z_T^{(i)})}{T} = \lim_{T\to\infty}\frac{(V(Z_T))^{(ii)}}{T} = 1 \qquad \text{a.s. .}$$

Together with (A7), $(Z_T^{(i)}), T \in \mathbb{N}$ satisfies the conditions in Theorem 2.2, then applying the theorem we can have

$$1 = \overline{\lim_{T\to\infty}}\frac{Z_T^{(i)}/\sqrt{T}}{\sqrt{2\frac{V(Z_T^{(i)})}{T}\ln\ln\frac{V(Z_T^{(i)})}{T}\,T}} = \overline{\lim_{T\to\infty}}\frac{Z_T^{(i)}/\sqrt{T}}{\sqrt{2\ln\ln T}} \qquad \text{a.s. .}$$

Thus, for $\epsilon > 0$, for P-a.s. ω such that for $T \geq T_i(\omega)$,

$$\frac{\left(Z_T^{(i)}(\omega)\right)^2}{T} < 2(1+\epsilon)\ln\ln T,$$

for all $i = 1, \cdots r$. Therefore for P-a.s. ω we have

$$2\big(l_T(\tilde{\theta}_T(\omega)) - l_T(\hat{\theta}_T(\omega))\big) \overset{a.a.s}{\lesssim} \sum_{i=1}^{r} \frac{\left(Z_T^{(i)}(\omega)\right)^2}{T} < 2r(1+\epsilon)\ln\ln T,$$

for $T \geq \max_{i=1,\cdots,r} T_i(\omega)$.
□

B.4 Likelihood Ratios for Structural Models

Now we consider structural models. We have an unconstrained reduced form as defined in (2.1):

$$Y_t = \Pi X_t + V_t$$

The parameter of the reduced from satisfies a set of restrictions imposed by the following overidentified structural form as defined in (2.3):

$$BY_t + \Gamma X_t = U_t,$$

where
$\Pi \in \mathbb{R}^{G\times K}$ are parameters of the unconstrained reduced form.
$B \in \mathbb{R}^{G\times G}, \Gamma \in \mathbb{R}^{G\times K}$ are structural parameters and it holds that $B\Pi_0 = -\Gamma$,
$Y_t \in \mathbb{R}^{G\times 1}$ represents the endogeneous variables,
ξ_t are exogeneous variables,
$X_t' = (Y_{t-1}', \cdots, Y_{t-p}', \xi_t') \in \mathbb{R}^{1\times K}$ represent the predetermined variables,
V_t is the disturbance of the reduced form and is i.i.d. $N(0, \Omega_0)$-distributed,
$U_t = BV_t$ represents the disturbances of the structural form and is i.i.d. $N(0, B\Omega_0 B')$-distributed, and
$\theta = (\Pi, \Omega)$. $B\Pi_0 = -\Gamma$ implies restrictions on $\theta : \{R_1(\theta) = 0, \cdots, R_r(\theta) = 0.\}$

The conditional density function is

$$\begin{aligned} l(\theta, y_t|x_t) = &-\frac{G}{2}\ln 2\pi - \frac{1}{2}\ln\det\Omega \\ &-\frac{1}{2}(y_t - \Pi x_t)'\Omega^{-1}(y_t - \Pi x_t). \end{aligned} \tag{B.25}$$

We also assume that the reduced form has the following dynamics:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \phi_e \xi_t + B^{-1} U_t$$

and $\phi_1, \cdots, \phi_p$ are given such that the autoregressive part is stationary [57]. Let L represent the lag operator. We rewrite the dynamics as

$$\phi(L) Y_t = \phi_e \xi_t + B^{-1} U_t.$$

Because of the stationarity we have the inverse of $\phi(L)$, say $\psi(L)$, and we get

$$Y_t = \underbrace{\psi(L)\phi_e \xi_t}_{\tilde{\xi}_t} + \underbrace{\psi(L) B^{-1} U_t}_{\tilde{Y}_t}. \tag{B.26}$$

That means that we can divide our process linearly into two parts: $\tilde{\xi}_t$ represents the total historical effect of the exogenous variable at t and $\tilde{Y}_t$ represents the autoregressive part of the process. For technical reasons we assume $\xi_t = 0$ for $t \leq 0$.

In order to prove that the likelihood ratios for structural models have a limit barrier $r \ln \ln T$ we must check the conditions (A1)-(A10) for the structural models.

Let $\Delta\Pi = \Pi - \Pi_0$. We rewrite the log likelihood function as[58]

$$l(\theta, y_t | x_t) = -\frac{G}{2} \ln 2\pi + \frac{1}{2} \ln \det \Omega - \frac{1}{2} tr \left(\Omega^{-1} (v_t - \Delta\Pi x_t)(v_t - \Delta\Pi x_t)' \right). \tag{B.27}$$

Let $A = \Omega^{-\frac{1}{2}} \Delta\Pi$. Following Theorem 2.1 we obtain the likelihood function and we take the average

$$\begin{aligned} \frac{1}{T} \ln L_T(\theta; \mathbf{y}_t | \Xi) &= \frac{1}{T} \sum_{t=1}^{T} l(\theta, y_t | x_t) \\ = &- \frac{G}{2} \ln 2\pi - \frac{1}{2} \ln \det \Omega \\ &- \frac{1}{2} \text{ tr } (\Omega^{-1} \frac{\sum_{t=1}^{T} v_t v_t'}{T}) + \text{ tr } (\Omega^{-1} \Delta\Pi \frac{\sum_{t=1}^{T} x_t v_t'}{T}) \\ &- \frac{1}{2} \text{ tr } (A \frac{\sum_{t=1}^{T} x_t x_t'}{T} A'). \end{aligned} \tag{B.28}$$

[57] we mean that the roots of

$$|I_G \lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_p| = 0$$

satisfy $|\lambda| < 1$, see Hamilton (1994)

[58] $\text{tr } (AB) = \text{ tr } (BA)$, if AB and BA are well defined.

First we want to consider the convergence in (A1). We take the expectation in both sides of the equation above:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} El(\theta, Y_t|X_t) = &-\frac{G}{2}\ln 2\pi - \frac{1}{2}\ln\det\Omega \\ &-\frac{1}{2}\,\mathrm{tr}\,(\Omega^{-1}\frac{1}{T}\sum_{t=1}^{T}\underbrace{E[\,V_tV_t']}_{\text{constant w.r.t } t}\quad) + \mathrm{tr}\,(\Omega^{-1}\Delta\Pi\frac{1}{T}\sum_{t=1}^{T}\underbrace{E[X_tV_t']}_{=0}) \\ &-\frac{1}{2}\,\mathrm{tr}\,(A\frac{1}{T}\sum_{t=1}^{T}E[X_tX_t']\;A').\end{aligned} \tag{B.29}$$

For the convergence we need only to consider the last term. If there exists a constant matrix $\overline{X^2}$ such that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}X_tX_t' = \overline{X^2} \qquad \text{a.s.},$$

then using Jensen's inequality [59] we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E[X_tX_t'] = \overline{X^2}.$$

Therefore we have a well-defined limit of average log likelihood

$$\bar{l}(\theta,\Xi) = c - \frac{1}{2}\ln\det\Omega - \frac{1}{2}\,\mathrm{tr}\,(\Omega^{-1}\Omega_0) - \frac{1}{2}\,\mathrm{tr}\,(A\overline{X^2}A'), \tag{B.30}$$

where $c = -\frac{G}{2}\ln 2\pi$ and we can show

Lemma 2.9 *$\bar{l}(\theta,\Xi)$ has a unique maximum at θ_0.*

Proof is given in the next section.

Now we consider the locally uniform convergence w.r.t parameter in the assumption (A2). It is easy to obtain because the term $-\frac{1}{2}\,\mathrm{tr}(A\,\frac{\sum_{t=1}^{T}X_tX_t'}{T}\,A')$ is a product of the parameter A and the average $\frac{\sum_{t=1}^{T}X_tX_t'}{T}$.

Below, we want to consider the convergence in the assumption (A3). To do this is to prove

$$\lim_{T\to\infty}\frac{\sum_{t=1}^{T}V_tV_t'}{T} = \Omega_0 \qquad \text{a.s.}$$

[59] $E|\frac{1}{T}\sum X_tX_t' - \overline{X^2}| \geq |E[\frac{1}{T}\sum X_tX_t'] - \overline{X^2}|$

and

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} X_t V_t'}{T} = 0 \qquad \text{a.s. .}$$

The first limit is simply the result from the strong law of large number. For the second limit we need the strong law of large number of Mcleish(1975).

Lemma 2.10 *If Y_t is stationary and there is a constant matrix $\overline{\Xi^2}$ such that the following moment condition*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \xi_t \xi_t' = \overline{\Xi^2} \tag{B.31}$$

is satisfied, then

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} X_t V_t'}{T} = 0 \qquad \textit{a.s. ,}$$

by using the strong law of large number for mixingale in Mcleish (1975).

The discussion of the locally uniform convergence w.r.t parameter in the assumption (A3) is also because of the product form of the parameter and the average in (B.28).

Now we want to consider the convergence of the second differentials of the log likelihood in the assumptions (A4) and (A5). Let ω^{ij} be the component at ij position in Ω^{-1} and $\omega^{i\cdot}$ be the i-th row in Ω^{-1}. Let β^{ij} and $\beta^{i\cdot}$ be those in B^{-1}. It is easy to calculate the first differentials of l:

$$\frac{\partial l(\theta, y_t, x_t)}{\partial \pi_{ij}} = x_{jt} \omega^{i\cdot} (y_t - \Pi x_t), \tag{B.32}$$

and the second differentials:

$$\frac{\partial^2 l(\theta, y_t, x_t)}{\partial \pi_{ij} \partial \pi_{kn}} = -\omega^{ik} x_{jt} x_{nt} \tag{B.33}$$

Therefore if $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} \begin{pmatrix} Y_t' & X_t' \end{pmatrix}$ can converge to a constant matrix P-a.s., then we can have the P-a.s. convergences for the averaged second differentials as stated in the assumption (A4).

Lemma 2.11 *If the conditions in the lemma 2.10 are satisfied, then there exists a constant matrix $Z(\Xi) \in \mathbb{R}^{(G+K)\times(G+K)}$ depending on $\Xi = (\xi_1, \xi_2, \cdots)$ such that*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} \begin{pmatrix} Y_t' & X_t' \end{pmatrix} = Z(\Xi) \qquad \textit{a.s. .} \tag{B.34}$$

Now we show that $\big(\frac{\partial l(\theta,Y_t,X_t)}{\partial \gamma_{ij}}\big|_{\theta_0}\big)_{t\in\mathbb{N}}$ is martingale difference process $\forall i,j$. Using (B.32) it is easy to see that

$$E_{t-1}\big[\frac{\partial l(\theta,Y_t,X_t)}{\partial \pi_{ij}}\big|_{\theta_0}\big] = X_{jt}E_{t-1}[\omega^{i\cdot}V_t] = 0.$$

Therefore $d_t(\theta_0)_{t\in\mathbb{N}}$ is a martingale.

Now we want to show that

$$\lim_{T\to\infty}\frac{V(d_T(\theta_0))}{T} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E_{t-1}\big[d_T(\theta_0)d_T(\theta_0)'\big] = \bar{V}_d \qquad \text{a.s. .} \tag{B.35}$$

$$E_{t-1}\big[\frac{\partial l}{\partial \omega_{ij}}(\theta_0)\,\frac{\partial l}{\partial \omega_{kn}}(\theta_0)\big] = X_{jt}X_{nt}\omega^{ik}\ .$$

Therefore, if (B.34) is satisfied, then we can have the convergence (B.35).

Lemma 2.12 *The assumption (A7) (for LIL Theorem of Wang) is satisfied for the stationary structural model.*

Lemma 2.13 *If the assumptions in the lemma 2.10 are satisfied, then the assumption (A8) holds for the structural model.*

Lemma 2.14 *Assume that the $\tilde{Y}_t$ is stationary.*
If $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t' = \overline{\Xi^2}$, and moreover there exists a constant $\tilde{c}$ such that

$$\overline{\lim_{T\to\infty}}\ \frac{\big\|\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t' - \overline{\Xi^2}\big\|}{\sqrt{\ln\ln T/T}} \ \le \tilde{c}\quad , \tag{B.36}$$

then for the structural model there exists a constant c such that

$$\overline{\lim_{T\to\infty}}\ \frac{\big\|\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} Y_t \\ X_t\end{pmatrix}\begin{pmatrix} Y_t' & X_t'\end{pmatrix} - Z(\Xi)\big\|}{\sqrt{\ln\ln T/T}} \ \le\ c \qquad \textit{a.s. .} \tag{B.37}$$

To summarize the discussions above, we have the following theorem:

Theorem 2.15 *If the assumptions in the lemma 2.14 are satisfied, then the conditions (A1) - (A10) hold.*

B.5 Proofs

Proof of the lemma 2.5:
Using Taylor development we have

$$\frac{1}{\sqrt{T}}d_T(\tilde{\theta}_T) = \frac{1}{\sqrt{T}}d_T(\theta_0) + \frac{1}{T}D_T\big(b(\tilde{\theta}_T,\theta_0)\big)\sqrt{T}(\tilde{\theta}_T - \theta_0), \tag{B.38}$$

where $b(\tilde{\theta}_T,\theta_0)$ is a vector between $\tilde{\theta}_T$ and θ_0 satisfying the equation above.

The first step is to prove that

$$\lim_{T\to\infty}\frac{1}{T}D_T\big(b(\tilde{\theta}_T,\theta_0)\big) = \bar{D}(\theta_0) \qquad \text{a.s. .}$$

Because of the continuity of $\bar{D}(\theta_0)$, we can find a neighborhood $\mathcal{N}(\theta_0)$ such that

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\big\|\bar{D}(\theta) - \bar{D}(\theta_0)\big\| < \epsilon/2.$$

Following the strong consistency in Lemma 2.4 there exists $T_1(\omega)$ for P-a.s. ω such that

$$\tilde{\theta}_T \in \mathcal{N}(\theta_0) \qquad \forall T \geq T_1(\omega).$$

According to (A5) we can find $\mathcal{N}'(\theta_0)$ such that

$$\overline{\lim_{T\to\infty}}\sup_{\theta\in\mathcal{N}'(\theta_0)'}\big\|\frac{1}{T}D_T(\theta) - \bar{D}(\theta)\big\| \;=\; 0.$$

Therefore, for this ϵ, we can find $T_2(\omega)$ for P-a.s ω such that for $T \geq T_2(\omega)$

$$\sup_{\theta\in\mathcal{N}'(\theta_0)}\big\|\frac{1}{T}D_T(\theta,\omega) - \bar{D}(\theta)\big\| \;\leq\; \epsilon/2.$$

As a result, for this ϵ for P-a.s ω we can find $T \geq T_1(\omega)\vee T_2(\omega)$ such that

$$\begin{aligned}&\big\|\frac{1}{T}D_T(b(\tilde{\theta}_T,\theta_0)) - \bar{D}(\theta_0)\big\|\\ \leq&\big\|\frac{1}{T}D_T(b(\tilde{\theta}_T,\theta_0)) - \bar{D}(b(\tilde{\theta}_T,\theta_0))\big\| + \big\|\bar{D}(b(\tilde{\theta}_T,\theta_0)) - \bar{D}(\theta_0)\big\| \;\leq\; \epsilon.\end{aligned}$$

The first step is proved.

Using (B.38), the fact that $d_T(\tilde{\theta}_T) = 0$, the invertibility of $\bar{D}(\theta_0)$, and the strong convergence in step 1, we can have

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) = \big[-\frac{1}{T}D_T(b(\tilde{\theta}_T,\theta_0))\big]^{-1}\frac{1}{\sqrt{T}}d_T(\theta_0). \tag{B.39}$$

Under the conditions (A6) and (A7) we can apply the Theorem 2.2 for $(d_t(\theta_0))_{t\in\mathbb{N}}$ to obtain

$$P\left(\overline{\lim_{T\to\infty}} \frac{d_T^{(j)}(\theta_0)}{\sqrt{2V(d_T^{(j)}(\theta_0))\ln\ln V(d_T^{(j)}(\theta_0))}} = 1\right) = 1. \tag{B.40}$$

where $d_T^{(j)}(\theta_0)$ is the j-th element of $d_T(\theta)$.

Using (A6) and let $v_{jj} = \left(\bar{V}\right)_{jj}$, then we have

$$\begin{aligned} 1 &= \overline{\lim_{T\to\infty}} \frac{d_T^{(j)}(\theta_0)/\sqrt{T}}{\sqrt{2\frac{V(d_T^{(j)}(\theta_0))}{T}\ln\ln\frac{V(d_T^{(j)}(\theta_0))}{T}T}} \\ &= \overline{\lim_{T\to\infty}} \frac{d_T^{(j)}(\theta)/\sqrt{T}}{\sqrt{2v_{jj}\ln\ln T}} \qquad \text{a.s. ,} \end{aligned}$$

which means for some ϵ, for P-a.s. ω

$$\frac{d_T^{(j)}(\theta_0(\omega))}{\sqrt{T}} \leq (1+\epsilon)\sqrt{2v_{jj}\ln\ln T}, \tag{B.41}$$

for T great enough (depending on ω).

Now using step 1 and (B.41) we have

$$\left\|\sqrt{T}(\tilde{\theta}_T - \theta_0)\right\| \leq M\big(\left\| - \bar{D}(\theta_0)^{-1}\right\| + \epsilon\big)(1+\epsilon)\sqrt{2\max_j v_{jj}\ln\ln T}.$$

So we proved the first statement of the lemma. With the same argument we can prove

$$\frac{\left\|\eta_T - \eta_0\right\|}{\sqrt{\ln\ln T/T}} \leq \tilde{c}.$$

Using

$$\hat{\theta}_T - \theta_0 = h(\hat{\eta}_T) - h(\eta_0) = \frac{\mathrm{d}h}{\mathrm{d}\eta'}(\eta_t^*)(\hat{\eta}_T - \eta_0),$$

and also the strong consistency of $\hat{\eta}_T$ (h is invertible following the implicit function theorem) and continuity of $\frac{\mathrm{d}_h}{\mathrm{d}_\eta}$ in neighborhoods of η_0, the second statement is also proved.

□

Proof of Lemma 2.6:

Using the triangle inequality for the norm we have

$$\begin{aligned}
&\frac{\left\|\frac{1}{T}D_T b(\tilde{\theta}_T,\theta_0)-\bar{D}_0\right\|}{\sqrt{\ln\ln T/T}} \\
&\leq \underbrace{\frac{\left\|\frac{1}{T}D_T(b(\tilde{\theta}_T,\theta_0))-\bar{D}(b(\tilde{\theta}_T,\theta_0))\right\|}{\sqrt{\ln\ln T/T}}}_{\leq K \text{ , because of the condition (A10)}} + \frac{\overbrace{\left\|\frac{\mathrm{d}\bar{D}}{\mathrm{d}\theta}(\theta^*)\right\|}^{\leq}\overbrace{\left\|b(\tilde{\theta}_T,\theta_0)-\theta_0\right\|}}{\sqrt{\ln\ln T}} \\
&\leq \tilde{K},
\end{aligned}$$

because $\left\|\frac{\mathrm{d}\bar{D}}{\mathrm{d}\theta}(\theta^*)\right\| \leq K_1$ for T large enough and $\frac{\left\|b(\tilde{\theta}_T,\theta_0)-\theta_0\right\|}{\sqrt{\ln\ln T}} \leq c_2$ from Lemma 2.5.
□

Proof of Lemma 2.9:
For any B, we choose $\Gamma = B(B_0^{-1}\Gamma_0)$, then $\Delta\Pi = (-B^{-1}\Gamma + B_0^{-1}\Gamma_0) = 0$, also $A = 0$. That means, for any B, we can choose Γ such that the third term on the right hand side of the equation (B.30) equals zero.[60] Since the Ω is unconstrained, for the maximum of $\bar{l}$ in (B.30) we need only to decide Ω.

Let Q be the probability measure of $N(0,\Omega)$ and $q : \mathbb{R}^G \to \mathbb{R}_+$ be the density function of Q. Let Q_0 be the P-measure of $N(0,\Omega_0)$. Then

$$E_{Q_0}[\ln q(\cdot)] = \frac{1}{2}\ln\det\Omega - \frac{1}{2}\,\mathrm{tr}\,(\Omega\Omega_0^{-1}).$$

Now using Jensens inequality, $E_{Q_0}[\ln q(\cdot)]$ has it's unique maximum at Ω_0.
□

In order to prove Lemma 2.10, We quote at first the definition of mixingale in Mcleish (1975).

Definition 2.16 *A sequence of one-dimensional r.v. is called* mixingale*, if there exist sequences of positive constants* $(c_n)_{n\in\mathbb{N}}$ *and* $(\varphi_m)_{m\geq 0}$ *with* $\lim\varphi_m = 0$*, such that*

$$\begin{aligned}
&E\big[E_{n-m}[X_n]^2\big] \leq \varphi_m c_n \\
&E\big[(X_n - E_{n+m}[X_n])^2\big] \leq \varphi_{m+1} c_n.
\end{aligned}$$

[60] At this point the problem of observational differentiability can be seen clearly: the MLE is uniquely identified only by Π_0. From Π_0 to (B,Γ) is the problem we have discussed in Chapter 3.

Theorem 2.17 (Strong law of large number for mixingales) *((1.6) Theorem, p.831, Mcleish(1975) and (1.9) Corollary p.832)*
Let $X_t, t \in \mathbb{N}$ be a mixingale with constants $(c_n)_{n\in\mathbb{N}}$ and $(\varphi_m)_{m\geq 0}$ defined above. If (i) $\sum_{i=1}^{\infty} c_i^2/i^2 < \infty$ (ii) $\exists a_k > 0, k \in \mathbb{N}$ such that $\sum_{k=1}^{\infty} a_k < \infty$ and (iii) $\sum_{k=1}^{\infty} \varphi_k^2(a_k^{-1} - a_{k-1}^{-1}) < \infty$, then

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} X_t}{T} = 0 \qquad a.s. .$$

□

Proof of Lemma 2.10
Recall that there are two kinds of variables collected in $X_t = (Y_{t-1}, \cdots, Y_{t-p}, \xi_t)$. Without loss of generality we consider here only one-dimensional processes.

Consider the convergence of $\frac{1}{T}\sum_{t=1}^{T} \xi_t V_t$. This is a case of the strong law of large number (SLLN) for independent r.v's. Using Theorem 5.4.1 p.124 and Corollary p.125 in Chung (1974) we know for independent r.v's X_t

$$\sum_{t=1}^{\infty} E[X_t^2]/t^2 < \infty \quad \Rightarrow \quad \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} X_t = 0 \qquad \text{a.s..} \qquad (B.42)$$

We know that $\xi_t V_t$ are independent under t and $E[(\xi_t V_t)^2] = \xi_t^2\sigma^2$. We will use the following lemma.

Lemma 2.18 *Let $q_t, t \in \mathbb{N}$ be a sequence.*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} q_t^2 < \infty \quad \Rightarrow \quad \sum_{t=1}^{\infty} \frac{q_t^2}{t^2} < \infty \ .$$

(Without proof).

Because $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \xi_t^2 = \overline{\Xi^2}$, then
$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{E[(\xi_t V_t)^2]}{t^2} = \lim_{T\to\infty} \sum_{t=1}^{T} \sigma^2 \frac{\xi_t^2}{t^2} < \infty$. Using the theorem of Chung (B.42) it follows $\frac{1}{T}\xi_t V_t \to 0$ a.s..

Consider now the convergence of $\frac{1}{T}\sum_{t=1}^{T} Y_{t-k}V_t$, $k = 1, \cdots, p$. We apply at first the linear decomposition of Y_t in (B.26) and consider $\frac{\sum \tilde{Y}_{t-k}V_t}{T}$ and $\frac{\sum \tilde{\xi}_{t-k}V_t}{T}$ separately, where

$$\tilde{Y}_t = \sum_{i=0}^{\infty} \psi_i V_{t-i}, \qquad \tilde{\xi}_t = \sum_{i=0}^{\infty} \psi_i \xi_{t-i}$$

For the convergence of $\frac{\sum \tilde{\xi}_{t-k}V_t}{T}$, use the following lemma.

Lemma 2.19 *Assuming that the dynamic of $\tilde{Y}_t$ is stationary. Then*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\xi_t^2=\overline{\Xi^2}\quad\Rightarrow\quad\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\tilde{\xi}_t^2<\infty.$$

(Without proof.)

The condition $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\tilde{\xi}_t^2<\infty$ leads to
$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\xi_tV_t=0$ a.s.. The reason is the same as in the discussion about the convergence of $\frac{1}{T}\sum_{t=1}^{T}\xi_tV_t$ above.

We consider now $\sum\tilde{Y}_{t-k}V_t$. We note at first that $\tilde{Y}_t$ is mixingale with $c_n^2=(EV_t^2)^2/(1-\lambda^2)$ and $\varphi_m^2=\lambda^{2m}$, where λ maximal absolute eigenvalue of dynamics [61] is a value $0<\lambda<1$. We choose $a_k=k^{-2}$, then it can be checked that these c_n,φ_m,a_k satisfy the conditions in the theorem. Therefore $\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k}V_t=0$ a.s. .
□

Proof of Lemma 2.11
Following the linear decomposition for Y_t, we rewrite the terms in the r.h.s. of (B.34) as the following

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}Y_{t-k_1}Y_{t-k_2}=&\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2}+\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k_1}\tilde{\xi}_{t-k_2}\\&+\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k_2}\tilde{\xi}_{t-k_1}+\frac{1}{T}\sum_{t=1}^{T}\tilde{\xi}_{t-k_1}\tilde{\xi}_{t-k_2}\ ,\end{aligned}$$

where $k_1=0,\cdots,p$, $k_2=0,\cdots p$.

$$\frac{1}{T}\sum_{t=1}^{T}Y_{t-k}\xi_t=\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k}\xi_t+\frac{1}{T}\sum_{t=1}^{T}\tilde{\xi}_{t-k}\xi_t\ ,$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t\ .$$

Therefore we should discuss three kinds of convergences

(i) $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2}=c$, *a.s.* because $(\tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2})_{t\in\mathbb{N}}$ is strictly stationary [62] and is therefore ergodic by the ergodic theorem (Theorem 6.21) on p.113 in Breimen(1992).

[61] See Hamilton (1994), p.259
[62] See Hamilton (1994) p.46

(ii) $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\tilde{Y}_{t-k}\tilde{\xi}_t = c$, $a.s.$ by using the theorem of Mcleish with $\varphi_m^2 = \lambda^{2m}$ and $c_n^2 = \xi_n^2(EV_n^2)^2/(1-\lambda^2)$.

The convergence of the series of the third kind is the assumption.
□

Proof of Lemma 2.12
at first we want to show that the linear combination of the first differentials $\alpha' d_t(\theta_0)$ satisfies (W2).
Recall

$$\frac{\partial l(\theta)}{\partial \beta_{ij}}\Big|_{\theta_0} = -\bar{Y}_{jt}W_{it}(\theta_0) + \big(\beta^{ji} - V_{jt}W_{it}(\theta_0)\big)$$

where the second term on the r.h.s has an expectation of zero.

We observe that the term $\alpha' d_t(\theta_0)$ has the structure

$$\sum_{i=1}^{I} X_{it}\epsilon_{it} \tag{B.43}$$

where X_t is $\mathcal{F}_{t-1}$-measurable, $(\epsilon_{it})_{i=1,\cdots,I}$ are i.i.d with $E\epsilon_{it} = 0$ and $E|\epsilon_{it}|^{2+\delta} < \infty$. Let $Z_t = \sum_{i=1}^{I} X_{it}\epsilon_{it}$ and $\sigma_t^2 = E_{t-1}Z_t^2$, then

$$\sup_t \frac{E_{t-1}|Z_t|^{2+\delta}}{\sigma_t^{2+\delta}} < \infty.$$

First we "orthogonize" Z_t, i.e. we can find $\tilde{X}_{it}$, $\tilde{\epsilon}_{it}$ such that $\sum_{i=1}^{I} X_{it}\epsilon_{it} = \sum_{i=1}^{I}\tilde{X}_{it}\tilde{\epsilon}_{it}$ and $\tilde{X}_{it}$ are still $\mathcal{F}_{t-1}$ measurable, and $\tilde{\epsilon}_{it}$ are uncorrelated under i and also $E\tilde{\epsilon}_{it} = 0$, $E|\tilde{\epsilon}_{it}|^{2+\delta} < \infty$. Then

$$\frac{E_{t-1}|Z_t|^{2+\delta}}{\sigma_t^{2+\delta}} = \frac{E_{t-1}\big[|\sum_{i=1}^{I}\tilde{X}_{it}\tilde{\epsilon}_{it}|^{2+\delta}\big]}{\big(E_{t-1}\big[(\sum_{i=1}^{I}\tilde{X}_{it}\tilde{\epsilon}_{it})^2\big]\big)^{\frac{2+\delta}{2}}} \le \frac{I^{2+\delta}\sum_{i=1}^{I}E_{t-1}[|\tilde{X}_{it}\tilde{\epsilon}_{it}|^{2+\delta}]}{\big(\sum_{i=1}^{I}E_{t-1}[(\tilde{X}_{it}\tilde{\epsilon}_{it})^2]\big)^{\frac{2+\delta}{2}}}$$

$$\le I^{2+\delta}\frac{\sum_{i=1}^{I}|\tilde{X}_{it}|^{2+\delta}E|\tilde{\epsilon}_{it}|^{2+\delta}}{\sum_{i=1}^{I}|\tilde{X}_{it}|^{2+\delta}\big(E[\tilde{\epsilon}_{it}^2]\big)^{\frac{2+\delta}{2}}} \le I^{2+\delta}\max_i\frac{E|\tilde{\epsilon}_{it}|^{2+\delta}}{\big(E[\tilde{\epsilon}_{it}^2]\big)^{\frac{2+\delta}{2}}} < \infty.$$

□

Proof of Lemma 2.13
We already showed in (B.35) that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E_{t-1}\big[\frac{\partial l(\theta_0;Y_t,X_t)}{\partial\theta_i}\frac{\partial l(\theta_0;Y_t,X_t)}{\partial\theta_j}\big] = (\bar{V}_d)_{ij}.$$

Because the expectation of the l.h.s. converges to the same limit, it follows that

$$\begin{aligned}(\bar{V}_d)_{ij} &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E\big[\frac{\partial l(\theta_0;Y_t,X_t)}{\partial\theta_i}\frac{\partial l(\theta_0;Y_t,X_t)}{\partial\theta_j}\big] \\ &= -\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E[\frac{\partial^2 l(\theta_0;Y_t,X_t)}{\partial\theta_i\partial\theta_j}] = -D(\theta_0)_{ij}\end{aligned}$$

□

Proof of Lemma 2.14
Analogous to the proof of Lemma 2.11 we simply need to consider the convergence speed for the three types of series $\sum\tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2}/T$, $\sum\tilde{Y}_{t-k}\xi_t/T$ and $\sum\xi_t\xi_t/T$. Since the convergence speed of the third type is assumed, we only need to consider the first and second types.

Let Z_t represent $\tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2}$ or $\tilde{Y}_{t-k}\xi_t$. The basic idea of this proof is to represent $Z_t - E[Z_t]$ as

$$Z_t - E[Z_t] = \lim_{N\to\infty} Z_t - E_{t-N}[Z_t] = \lim_{N\to\infty}\sum_{n=0}^{N}\big(E_{t-n}[Z_t] - E_{t-n-1}[Z_t]\big),$$

where $E[Z_t] = \lim_{n\to\infty} E_{t-N}[Z_t]$. For any fixed n, $\big(E_{t-n}[Z_t] - E_{t-n-1}[Z_t]\big)_{t\in\mathbb{N}}$ is a martingale difference process. Then we can apply Theorem 2.2 to control the convergence speed.

At first let $Z_t = \tilde{Y}_{t-k_1}\tilde{Y}_{t-k_2}$ and $\varphi_t^{(n)} = \big(E_{t-n}[Z_t] - E_{t-n-1}[Z_t]\big)$. We consider the case $n \geq k_1 \vee k_2$. After some calculation we get

$$\begin{aligned}\varphi_t^{(n)} = \;& \psi_{n-k_1}\psi_{n-k_2}(V_{t-n}^2 - E[V_{t-n}^2]) \\ &+ \big(\psi_{n-k_2}E_{t-n-1}\tilde{Y}_{n-k_1} + \psi_{n-k_1}E_{t-n-1}\tilde{Y}_{n-k_2}\big)V_{t-n}.\end{aligned}$$

Let $EV_t^2 = \sigma^2$. Because of stationarity there exists c_1 [63] such that $\sum_{i=0}^{\infty}\psi_i^2 \leq c_1$. Calculating the variation process $V_T^{(n)} = V\big(\sum_{t=1}^{T}\varphi_t^{(n)}\big)$

$$\begin{aligned}\frac{V_T^{(n)}}{T} &= \frac{1}{T}\sum_{t=1}^{T} E_{t-n-1}(\varphi_t^{(n)})^2 \\ &= 2\psi_{n-k_1}^2\psi_{n-k_2}^2\sigma^4 + \sigma^2\psi_{n-k_2}^2\underbrace{\frac{1}{T}\sum_{t=1}^{T}E_{t-n-1}[\tilde{Y}_{t-k_1}]^2}_{\to\sigma^2\sum_{i=1+n-k_1}^{\infty}\psi_i^2 \text{ a.s.}} \\ &\quad + 2\sigma^2\psi_{n-k_2}\psi_{n-k_1}\underbrace{\frac{1}{T}\sum_{t=1}^{T}E_{t-n-1}[\tilde{Y}_{t-k_1}]E_{t-n-1}[\tilde{Y}_{t-k_2}]}_{\to\sigma^2\sum_{i=1+n-k_2}^{\infty}\psi_{i+k_2-k_1}\psi_i \text{ a.s.}} + \sigma^2\psi_{n-k_1}^2\underbrace{\frac{1}{T}\sum_{t=1}^{T}E_{t-n-1}[\tilde{Y}_{t-k_2}]^2}_{\sum_{i=1+n-k_2}^{\infty}\psi_i^2 \text{ a.s.}}.\end{aligned}$$

[63] see Hamilton (1994)

The a.s. convergences above are obtained by applying the ergodic theorem. Therefore

$$\lim_{T\to\infty}\frac{V_T^{(n)}}{T}=V^{(n)}\le\sigma^4(1+c_1)(|\psi_{n-k_1}|+|\psi_{n-k_2}|)^2.$$

Because $\varphi_t^{(n)}$ also has the product form (B.43), (W2) is satisfied, so we can apply Theorem 2.2 to get

$$\begin{aligned}1&=\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{2V_T^{(n)}\ln\ln 2V_T^{(n)}}}\\&=\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{T2V_T^{(n)}/T\ln\ln(T2V_T^{(n)}/T)}}\\&=\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{T2V^{(n)}\ln\ln T}}\qquad\text{a.s.}\end{aligned}$$

Therefore

$$\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{2T\ln\ln T}}=\sqrt{V^{(n)}}\le\sigma^2\sqrt{1+c_1}(|\psi_{n-k_1}|+|\psi_{n-k_2}|).$$

Because

$$\sum_{t=1}^T(Z_t-E[Z_t])=\sum_{t=1}^T\lim_{N\to\infty}\sum_{n=0}^N\varphi_t^{(n)}=\lim_{N\to\infty}\sum_{n=0}^N\sum_{t=1}^T\varphi_t^{(n)},$$

then

$$\overline{\lim_{T\to\infty}}\lim_{N\to\infty}\sum_{n=0}^N\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{T\ln\ln T}}\le\lim_{N\to\infty}\sum_{n=0}^N\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T\varphi_t^{(n)}}{\sqrt{T\ln\ln T}},$$

therefore

$$\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T(Z_t-E[Z_t])}{\sqrt{T\ln\ln T}}\le\sigma^2\sqrt{1+c_1}\sum_{n=0}^\infty(|\psi_{n-k_1}|+|\psi_{n-k_2}|)=c<\infty.$$

Now for $Z_t=\tilde{Y}_{t-k}\xi_t$ we can easily have $\varphi_t^{(n)}=\xi_t\psi_{n-k}V_{t-n}$. It also has the product form (B.43) and we can have $V_T^{(n)}/T=\sigma^2\psi_{n-k}^2\sum_t\xi_t^2/T$. Then

$$\overline{\lim_{T\to\infty}}\frac{\sum_{t=1}^T(Z_t-E[Z_t])}{\sqrt{T\ln\ln T}}\le\sqrt{\overline{\Xi^2}}\sigma\sum_{n=0}^\infty\psi_{n-k}^2<\infty.$$

□

Remark to the proof of Theorem 2.15

We saw that in Theorem 2.3 the important properties for proving the LR barrier are (i) the P-a.s. convergences of l_T/T and its differentials, (ii) the local uniformity w.r.t parameters of these convergences, and (iii) the control of the convergence speed. After checking the conditions for the structural model we summarize here which properties of the structural model lead to the properties above. It is important that we decompose Y_t into two parts: the exogeneous part and the autoregressive part. The convergences of l_T/T and its differentials are only possible when the autoregressive part is stationary and the exogeneous part satisfies the moment condition (B.36). We used three theorems of strong law of large number: for strict stationary processes, for independent processes, and for mixingales. The local uniformity of the convergences is due to the normal distributions of U_t. Thanks this all series considered for convergences have a product form of parameters and the averages which are P-a.s. convergent. For the control of convergence speed, we use LIL Theorem for martingales. The conditions for this theorem can be satisfied is because of the normal distribution of U_t and the stationarity of (the autoregressive part of) the process.

References

Akaike, H. (1973). Information theory and extension of maximum likelihood priciple. *2nd International Symposium on Information theory*, pages 267–281.

Amemiya, T. (1985). *Advanced Econometrics*. Basil Blackwell, 1st edition.

Amisano, G. and Giannini, C. (1997). *Topics in Structural VAR Econoetrics*. Springer, 2nd edition.

Charemza, W. (1997). *New directions in econometric practice : general to specific modelling, cointegration, and vector autoregression*. Elgar, 2nd edition.

Crowder, M. J. (1976). Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society, Series B*, 38:45–53.

Davidson, R. and Mackinnon, J. (1993a). *Esitmation and Inference in Econometrics*. Oxfort University Press, 1st edition.

— (1993b). *Estimation and Inference in Econometrics*. Oxford University Press, 1st edition.

Dhaene, G. (1997). *Encompassing, Formulation, Properties and Testing*. Springer Verlag, 1st edition.

Dhrymes, P. J. (1993). *Topics in Advanced Econometrics*. Springer-Verlag, 1st edition.

Engle, R. F. and Granger, W. J. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica*, 55:251–276.

Fair, R. C. (1984). *Specification, estimation, and analysis of macroeconomic models*. Harvard Univ. Pr, 1st edition.

Frohn, J. (1995). *Grundausbildung in Okonometrie*. de Gruyter, 2nd edition.

Frydman, R. (1980). A proof of the consistency of maximum likelihood estimators of nonlinear regression models with autocorrelated errors. *Econometrica*, 48:853–860.

Gourieroux, C. and Monfort, A. (1984). Pseudo-maximum likelihood methods: theory. *Ecomonetrica*, 52:681–700.

— (1996). Testing non-nested econometric models. *Handbook of Econometrics*, 4:2583–2637.

Granger, C. W. J. and White, H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, 67:173–187.

Granger, J. (1990). *Modelling Economic Series*. Clarendon, 1st edition.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, 1st edition.

Hargreaves, C. P. (1994). *Nonstationary Time Series Analysis and Cointegration*. Oxfort University Press, 1st edition.

Hendry, D. (1995). *Dynamic Econometrics*. Oxfort University Press, 1st edition.

Hendry, D. F. and Krolzig, H. M. (2001). New development in automatic general to specific modelling. *Econometrics and Philosophy of Economics, edited by B. P. Stigum*.

Hendry, D. F. and Richard (1988). The Real Term Structure and Consumption Growth. *Journal of Financial Economics*, 22:305–333.

Johansen, S. (1991). estimation and hypothesis testing of cointegration vectors in Gausian vector autoregressive models. *Econometrica*, 59:1151–1180.

— (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, 1st edition.

Judge, G. (1985). *The Theory and Practice of Econometirics*. John Willey and Sons, 2nd edition.

Klein, L. R. (1983). *Lectures in Econometrics*. North-Holland, 1st edition.

Koopmans, T. C. and Hood, W. C. (1953). The estimation of simultaneous linear economic relations. *Studies in Econometric Method edited by T. C. Koopmans and W. C. Hood*.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 23:31–43.

Lucas, R. E. (1976). Econometric policy evaluation: a critique. *Carneie-Rochester Conference Series on Public Policy*, pages 19–46.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Econometrics*. Springer Verlag, 1st edition.

Powell, A. A. and Murphy, C. W. (1997). *Inside a modern macroeconometric model : a guide to the Murphy model*. Springer Verlag, 1st edition.

Schlittgen, R. and Streitberg, B. H. J. (1999). *Zeitreihenanalyse*. München,Oldenburg.

Schmidt, P. (1976). *Econometrics*. Marcel Dekker Inc., New York / Basal.

Shibata, R. (1976). Selection of the Order of an Autoregressive Model by Akaik's Information Criterion. *Biometrica*, 63:117–126.

Sims, C. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Spanos, A. (1990). The simultaneous-equation model revisited. *Journal of Econometrics*, 44:87–105.

Theil, H. (1971). *Principles of econometrics*. Wiley, New York.

Tong, H. (1990). *Non-linear Time Series: a Dynamic Approach*. Oxfort University Press, Oxford / New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26.