CLASS NOTES FOR MACROECONOMETRICS: EC266.

Bent E. Sørensen
Department of Economics
Box B, Brown University
Providence, RI 02912

Spring 1996

Contents

1 References (with comments)

1.1 Methodology

Blaug, M. (1980): "The Methodology of Economics", Cambridge, UK: Cambridge University Press.
A readable introduction to methodology as seen from a methodologist, who is not an econometrician. This is quite a large field; but unfortunately (in spite of lengthy discussions of testability) quite disconnected from econometrics. There is a recent edition.

Caldwell, B. (1982): "Beyond Positivism: Economic Methodology in the 20th Century", London: George Allan and Unwin.
A - maybe slightly more modern - book on general methodology in the same area as Blaug's.

Granger, C.W.J. (Ed.): "Modeling Economic Series", Oxford: Oxford University Press, 1990.
A book af readings in methodology from the perspective of econometricians. Some of the articles are classics that you should know. The focus is too much on the LSE tradition to my taste.

Nickell, S. (1985): "Error Correction, Partial Adjustment and All That: An Expository Note", *Oxford Bulletin of Economics and Statistics"*, 47, no. 2.

Sims,C.A. (1982) : "Policy Analysis with Econometric Models", Brookings Papers on Economic Activity, no. 1.

Summers, L.H. (1991) : "The Scientific Illusion in Empirical Macroeconomics", *Scandinavian Journal of Economics* 93, 129-148.
A somewhat provocative article about macroeconometrics. Summers' basic claim is that only "simple" econometric studies (or even just tabulations of data) has been useful in econometrics. The

articly is quite provocatory and not very precise in it use of terms, but Summers do make a number of clever observations that are worth considering.

Rational expectations econometrics
Hansen, L.P. and T. Sargent (1991): "Rational Expectations Econometrics", Boulder: Westview Press. HB3730.H284.
A collection of older but unpublished notes and articles. Chapter 2 is lecture notes on least squares prediction theory, and the best place that an econometrician can get a rigourous introduction to that area.

Lucas,R. and T. Sargent (1981): "Rational Expectations and Econometric Practice", London: George Allan and Unwin. HD82.
A collection of mostly published articles. The introduction is interesting reading. I think that the book has been reprinted by the University of Minnesota Press in 2 volumes.

Sargent, T. (1987): "Dynamic Economic Theory", Cambridge, Mass.: Harvard University Press.
A graduate textbook. The book uses fairly advanced time series methods, so it will be easier to read after following this course.

1.2 Basic time series analysis

NOTE: Basic time series is an important part of the course. The books below are listed approximately in the order that I think that you may find them useful.

Harvey, A.C. (1993): "Time Series Models. 2nd Edition". Cambridge: MIT Press.
This is an introduction to classical time series analysis written by an econometrician. Very readable - an excellent place to start.

Lütkepohl, H. (1991): "Introduction to Multiple Time Series Analysis", Heidelberg: Springer Verlag. QA280.L87.
A quite comprehensive textbook on multiple time series written for economists. It is mostly about stationary time series; but it has a chapter on co-integrated series. It is solid and well written - a good place to learn about multiple time series.

Harvey, A.C. (1990): "*The Econometric Analysis of Time Series. 2nd Edition*," Cambridge: The MIT Press.
An econometrics book with a time-series perspective. I find it sloppily written here and there, but sometimes this is the most easily accessible place to find a result.

Harvey, A.C. (1989): "Forecasting, Structural Time Series Models and the Kalman Filter", Cambridge, U.K.: Cambridge University Press.
Treats the Kalman filter in detail. This is a very useful tool for time series analysis, and it has the advantage that it is of very general applicability - we will use the Kalman filter to estimate vector ARMA models.

Aoki, M. (1990) : "*State Space Modeling of Time Series. 2nd Edition*, New York: Springer Verlag.
I haven't read the whole book, but it seem to be a quite good book, with more of a "systems engineering" (a huge field that one may have to look into once in a while) feeling to it.

Fuller, W.: "Introduction to Statistical Time Series", New York, Wiley 1976. QA280.F84.
This is an introduction to classical time series written by a statistician. It is also quite readable, and a bit more comprehensive than Harvey (1981). It has a subsection (8.5) on non stationary time series and testing for a unit roots. That chapter is famous among econometricians and helped spark the unit root revolution in econometrics.

Granger, C. and Newbold, "Forecasting Economic Time Series", 2nd Edition, New York, Academic Press, 1986. HB 3730.G67.
This is quite a popular textbook for graduate time series courses. It covers a lot of topics, and has quite a bit of words in it - discussing modelling from the perspective of econometricians. In spite of that it is a quite advanced book. It is a very good book to read; but maybe best after some of the concepts have been learned.

Mills, T.C. (1990) : "Time Series Techniques for Economists", Cambridge University Press. HB135.M54 1990.
A recent introduction to time series - maybe you would find it useful as supplementary reading for the first half of the course. The paperback version is quite inexpensive.

Whittle, P. (1983) : "Prediction and Regulation. By least squares methods", 2nd revised edition. Minnesota: University of Minnesota Press.
This book covers the time series methods that Sargent (in particular) makes heavy use of. Contrary to the other time series books mentioned it does not focus on estimation.

Box, G.E.P. and G.M. Jenkins (1976): "Time Series Analysis. Forecasting and Control", 2nd Edition, Oakland: Holden-Day.
This is a classic book that made the so-called ARIMA modeling (also known as Box-Jenkins modeling) enormously popular.

Anderson, T.W.: "The Statistical Analysis of Time Series, New York, Wiley, 1971.
A classic with a wealth of results. Maybe a little outdated by now (but still worth reading).

Reprinted in the series: Wiley Classics.

Brockwell, P.J. and Davis, R.A.: "Time Series: Theory and Methods", Springer 1987. QA 280.B76.
I like this book, it is clear exposition of classical theory; but also with some chapters on newer methods. It uses more advanced math; but explains it along the way. I think there is a later edition.

Priestley,M.B.: "Spectral Analysis and Time Series", New York: Academic Press, 1981. QA280.
A comprehensive treatment of classical time series methods, with a lot of weight on spectral methods. (Available in two hardcover volumes; but you may prefer the one volume paperback if you have to buy it).

Priestley,M.B. "Non-linear and Non-stationary Time Series Analysis". London: Academic Press, 1988.
This is not the kind of non linearity and non stationarity that we will concentrate on in this course. But take a look if you want to see what a statistician might mean by those terms (and it is potentially useful in economics).

Strang, G. (1980):"Linear Algebra and Its Applications", 2nd Edition, New York: Harcourt Brace Jovanovich.
A very readable math book, that I will use for reference.

A few articles (out of an ocean)

Engle,R.F., D.F. Hendry, and J.-F. Richard (1983): "Exogeneity", Econometrica, 51, 277-305.
This article defines exogenity in term of a precise probability model. A very useful article (even if we may not have time to get to it).

Berk,K.N. (1974): "Consistent Autoregressive Spectral Estimates", *Annals of Statistics*, 2, 489-502.
A classical (though a bit hard for economists) article that shows that one can approximate an infinite AR model by a sequence of AR models of increasing dimension and still obtain consistent estimates.

1.3 General asymptotic theory

Most of the asymptotic results that we will come across is covered in the specific references. Here is just a few general references that give a coverage of the background theory.

Dhrymes, P.J. (1989) : "*Topics in Advanced Econometrics. Probability foundations*", New York: Springer Verlag.HB139.D49.
I have only skimmed it. It may be more accessible than the other two books mentioned here.

Hall, P. and Heyde, C.C. (1980) : "Martingale Limit Theory and its Applications", New York: Academic Press. QA 274.5.H34.
This is an advanced book. It is the source of most of the probability results that are used in the modern theoretical time series papers in econometrics (both GMM and unit-root type results).

White, H (1984) : "Asymptotic Theory for Econometricians", New York, Academic Press. HB 139.W5.
This book treats only linear models. It has some of the important results from Hall and Heyde (1980) in a more accessible form.

1.4 Testing

NOTE: The most general results (outside newer research articles) on testing in non-linear models are in Gallant (1987) that is listed under GMM. We will not have time for talking much about testing in this course, so you may have to consult the literature on your own at some stage. Here are a few scattered references to help you get started.

MacKinnon, J.G. (1992) : "Model Specification Tests and Artificial Regressions", *Journal of Economic Literature*, 30, 102-147.
A recent survey with many references.

Godfrey, L.G. (1988): "Misspecification Tests in Econometrics", Cambridge, UK: Cambridge University Press. HB139.G63.
Treats LM testing in detail. A good place to read about testing.

Engle, R. (1982): "A General Approach to Lagrange Multiplier Model Diagnostics", Journal of Econometrics, 20, 83-104.

Engle, R. (1984): "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics", Handbook of Econometrics, Vol 2, Ch. 13.
A well written (and much read) survey.

Hausman, J. (1978): "Specification Testing in Econometrics", *Econometrica*, 46, 1251-1272.
This introduced the now very famous "Hausman test". You may want to take a look at this article, if you do not know what a "Hausman test" is.

Newey, W. (1985): "GMM Specification Testing", Journal of Econometrics, 29, 229-256.

Newey, W.K. (1985): "Maximum Likelihood Specification Testing and Conditional Moments Test", *Econometrica*, 53, 1047-1071.

Perron, P. (1991) : "Test Consistency with Varying Sampling Frequency", *Econometric Theory* 7, 341-361.
An interesting article. It shows for example that when you test for unit roots in the AR(1) model without drift, then your test is only consistent if the length of the observation period goes to infinity; but not if you obtain closer observations with the same span of the data.

Singleton, K.J. (1985): "Testing Specifications of Economic Agents' Intertemporal Optimum Problems in the Presence of Alternative Models", Journal of Econometrics, 30, 391-413.

Tauchen, G. (1985): "Diagnostic Testing and Evaluation of Maximum Likelihood Models", *Journal of Econometrics*, 30, 415-443.

White, H. (1983) (Ed.): Special issue of *Journal of Econometrics*, 21, no. 1.
This is about non-nested tests, which we will not make much of in the course; but they may come in handy some day.

1.5 Probability theory of importance for unit root and continuous time econometrics

NOTE: Most of this is advanced material that is hard for an economist. We will not go deep into any of these books; but the references may be useful later.

Billingsley, P. (1968) : "*Convergence of Probability Measures*", New York: John Wiley & Sons.
This book gives the background theory for convergence to Wiener processes, that forms the background for most of the modern unit root theory. It is very well written; but written for statisticians.

The following 3 books are about stochastic integration and differential equations. One does not need to master that subject, even to do research in unit root econometrics. The important thing is to understand what a stochastic integral actually is and to understand how to use Ito's lemma.

Chung, K.L. and R.J. Williams (1990) : *"Introduction to Stochastic Integration",* 2nd Edition, Boston: Birkhauser.
This is one of the more accessible introductions to stochastic integrations, although it is rigourous mathematically.

Øksendal, B. (1985) : *"Stochastic Differential Equations",* Berlin: Springer-Verlag. QA274.23.047. There are several later editions.
This is presumably the most accessible introduction to the subject. It is rigourous but without stressing proofs. There is a later edition.

Protter, P. (1990) : *"Stochastic Integration and Differential Equations. A new approach",* New York, Springer Verlag.
This is my favorite reference to the modern theory of stochastic integration. It is a very good book; but it is also hard.

1.6 Unit Roots (and including unit roots-and-cointegration)

You may not believe it, but this is just a small selection of the most important articles.

Surveys
Cambell, J.Y. and P. Perron: "Pittfalls and Opportunities: What Macroeconomists Should Know About Unit Roots", NBER Macroeconometrics Annual 1991, 141-218. (With discussion).
The best of the surveys in my opinion.

Bannerjee, A. and D.F. Hendry (1992) : "An overview", Introduction to special issue on cointegration: *Oxford Bulletin of Economics and Statistics* 54, no. 3.
It is not the most comprehensive survey (nor was it intended to be), but it may be worth taking a look at. There is a few things in it that I cannot quite make sense of.

Diebold, F.X. and Nerlove,M. (1990): "Unit Roots in Economic Time Series: A Selective Survey", in T.B.Fomby and G.F. Rhodes (eds.): Advances in Econometrics, Vol 8, "Cointegration, Spurious Regressions, and Unit Roots". Greenwich: JAI Press.

Dolado, J.J., T. Jenkinson, and S. Sosvilla-Rivero: "Cointegration and Unit Roots", Journal of Economic Surveys, 4, 249-275.

Stock, J.H. (1994) : "Unit Roots and Trend Breaks". Handbook of Econometrics, Vol IV (R.F. Engle and D.L. McFadden Eds.).
A good place to read about recent developments and get ideas for research.

Stock, J.H. and M.W.Watson (1988) : "Variable Trends in Economic Time Series", *Journal of Economic Perspectives* 2, Summer 1988, 147-174.
A good survey for the applied econometrician.

Theory
Andrews, D.W.K. (1993) : "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models", *Econometrica* 61, 139-167.

Beverage, S. and C.R. Nelson (1981): "A New Approach to the Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to the Measurement

of the 'Business Cycle"', *Journal of Monetary Economics*, 7, 151-174.
A quite simple paper (in the Journal of Monetary Economics!), but the decomposition turns out to be *very* useful - in particular as a tool in theoretical work.

Bhargava, A. (1986) : "On the Theory of Testing for Unit Roots in Observed Time Series", *Review of Economic Studies* 53, 369-384.

Chan, N.H. and C.Z. Wei (1988) : "Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes", *Annals of Statistics* 16, 367-402.
A pathbreaking article that solved a lot of hard questions for unstable multivariate processes. Hard reading for an economist, but if you are doing theory it may be worth it.

Cochrane,J.H. (1988) : "How Big is the Random Walk in GNP?", *Journal of Political Economy* 96, 893-920.

Dickey,D.A., W.R. Bell, and R.B.Miller (1986):"Unit Roots in Time Series Models: Tests and Implications", *The American Statistician*, 40, no. 1, 12-26.

Dickey, D.A. and W.A. Fuller (1979) : "Distribution of the Estimators for Autoregressive Time Series With a Unit Root", *Journal of the American Statistical Association* 74, 427-431.

Dickey, D.A. and W.A. Fuller (1981) : "Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root", *Econometrica* 49, 1057-1072.

Elliott, G., T.J. Rothenberg, and J. Stock (1996) : "Efficient Tests for an Autoregressive Unit Root". *Econometrica* 64, 813-836.
Seems like the best unit test, so you may want to among the first to use it.

Evans, G.B.A. and N.E. Savin (1981) : "Testing for Unit Roots: 1", *Econometrica* 49, 753-779.

Evans, G.B.A. and N.E. Savin (1984) : "Testing for Unit Roots: 2", *Econometrica* 52, 1241-1269.
The Evans and Savin papers are classical early papers.

Hall, A. (1992) : "Testing for a Unit Root in Time Series using Instrumental Variable Estimators with Pretest Data Based Model Selection", *Journal of Econometrics* 54, 223-251.
Test for unit roots are usually using an ARMA representation. This paper treats the important problem of how one can choose the order of the ARMA representation from the data, and what effect this has (asymptotically) on the unit root test.

Maddala, G.S. (1992): "Introduction to Econometrics, 2nd Ed.", New York: Macmillan.
Comment: Maddala's book is an advanced undergraduate textbook. It has a quite readable introduction to unit roots and cointegration that supplements these notes - but I don't think many undergraduates will understand a word. (It is not really authoritative though, so you should only take this as an easy introduction).

Nabeya, S. and K. Tanaka (1990) : "A General Approach to the Limiting Distribution for Estimators in Time Series Regression with Nonstable Autoregressive Errors", *Econometrica* 58, 145-163.
A very clever method (going back to T.W. Anderson and maybe earlier) to find the characteristic function of asymptotic distributions. The paper by Nabeya and Sørensen (listed under "Near Unit Roots") shows (among other things) the relation between the expressions for limiting distributions, in terms of integrals of Brownian motions that dominate the literature, and the Nabeya-Tanaka approach.

Phillips, P.C.B. (1987a) : "Time Series Regression with a Unit Root", *Econometrica* 55, 277-301.

Schmidt, P. (1990): "Dickey-Fuller Tests with Drift", in T.B. Fomby and G.F. Rhodes, Jr. (Eds.), *Advances in Econometrics*, Vol 8., p. 161-203.

Schmidt, P. and P.C.B. Phillips (1992) : "LM Tests for a Unit Root in the Presence of Deterministic Trends", *Oxford Bulletin of Economics of Statistics* 54, 257-289.

Schwert, G.W. (1989) : "Tests for Unit Roots: A Monte Carlo Investigation", *Journal of Business & Economic Statistics* 7, 147-161.

West, K.D. (1988) : "Asymptotic Normality when Regressors Have a Unit Root", *Econometrica* 56, 1397-1419.

White, J.S. (1958) : "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case", *Annals of Mathematical Statistics* 29, 1188-1197.

Phillips, P.C.B. and Perron, P. (1988): "Testing for a Unit Root in Time Series Regression", *Biometrika*, 75, 335-346.

Bayesian approaches to unit root theory
There is a lot of other articles in this area, but we will not have time for Bayesian methods unless there is particular student interest. Ask me for further references if you are interested. There is a special issue of Econometric Theory coming out this year (or next).

Berger, J. O. (1985) : "*Statistical Decision Theory and Bayesian Analysis.* 2nd Edition." New York: Springer-Verlag. QA279.4.B46.
A good statistics book to read if you want to really go into Bayesian methods.

Sims, C.A., and H.Uhlig (1991) : "Understanding Unit Rooters: A Helicopter Tour", *Econometrica* 59, 1591-1601.

Phillips, P.C.B. (1991): "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends", *Journal of Applied Econometrics*, 6, no 4.
Special issue with extensive (and animated) discussion of unit roots and Bayesian methods. You can find lots of references here.

Applications

Nelson,C.R. and C.I.Plosser (1982) : "Trends and Random Walks in Macroeconometic Time Series", *Journal of Monetary Economics* 10, 139-162.
A widely read article that sparked a lot of the interest in unit root econometrics.

Campbell,J.Y. and N.G.Mankiw (1987) : "Are Output Fluctuations Transitory?", *Quarterly Journal of Economics*, November, 857-880.

Campbell,J.Y. and N.G.Mankiw (1989) : "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence" *NBER Macroeconomics Annual*, 185-216.

Inference for models with near unit roots

(Relative to its importance this subject has "too many" references; I had them on the machine since I have done work in that area).

Cavanagh, C. (1986) : "Roots Local to Unity", Harvard Institute of Economic Research, Discussion Paper No. 1259.

Chan, N.H. (1988) : "The Parameter Inference for Nearly Nonstationary Time Series", *Journal of the American Statistical Association* 83, 857-862.

Chan, N.H. and C.Z. Wei (1987) : "Asymptotic Inference for Nearly Nonstationary AR(1) Processes", *Annals of Statistics* 15, 1050-1063.

Nabeya, S. and B.E. Sørensen (1994) : "Asymptotic Distributions of the Least Squares Estimators and Test Statistics in the Near Unit Root Model with Non-Zero Initial Value and Local Drift and Trend", *Econometric Theory*, 10, 937-967.

Perron, P. (1989) : "The Calculation of the Limiting Distribution of the Least-Squares Estimator in a Near-Integrated Model", *Econometric Theory* 5, 241-256.

Perron, P. (1991) : "A Continuous Time Approximation to the Unstable First-Order Autoregressive Process: The Case Without an Intercept", *Econometrica*, 59, 211-236 .

Phillips, P.C.B. (1987b) : "Towards a Unified Asymptotic Theory for Autoregression", *Biometrika* 74, 535-547.

Phillips, P.C.B. (1988) : "Regression Theory For Near Integrated Time Series", *Econometrica* 56, 1021-1045.

Sørensen,B.E. (1992):"Continuous Record Estimations in Systems of Stochastic Differential Equations", *Econometric Theory*, 8, 28-51.

Cochrane,J.H. (1991) : "A Critique of the Application of Unit Root Tests", *Journal of Economic Dynamics and Control*, 15, 275-284.

1.7 Cointegration

Survey:
Watson, M.W. (1994) : "Vector Autoregressions and Cointegration". Handbook of Econometrics Vol IV (R.F. Engle and D.L. McFadden, Eds.).
A recent survey by one of the major players in the field.

Books:
Banerjee, A., J. Dolado, J.W. Galbraith, and D.F. Hendry (1993) : "*Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data.*", Oxford: Oxford University Press.
The first book on co-integration. I have used it as a text. The authors are all very good applied people.

Johansen, S. (1995): "*Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*", Oxford University Press.

The "Johansen method" is very powerful and Johansen's coverage is very good and precise. Johansen is a statistician, and you might want to look elsewhere for applications.

Cointegration: Theory

Engle,R.F. and C.W.J. Granger (1987): "Cointegration and Error Correction: Representation, Estimation, and Testing", Econometrica, 55, 251-276.
The classic article. If you want to do empirical work you should be aware that the methods here are already a bit outdated.

Engle,R.F. and S.B. Yoo (1987): "Forecasting and testing in co-integrated systems", *Journal of Econometrics*, 35, 143-159.

Engle,R.F. and S.B. Yoo (1991): "Cointegrated Economic Time Series: An Overview with New Results", p. 237-267 in Engle, R.F. and C.W.J. Granges (Eds.): "*Long Run Economic Relationships*", Oxford University Press, Oxford.

Granger,C.W.J. (1981) : "Some Properties of Time Series Data and Their Use in Econometric Model Specification", *Journal of Econometrics*, 16, 121-130.
The first appearence of the word cointegration. Granger said: "..a very special case...seems to be potentially very important.." You bet!

Granger,C.W.J. and T.-H. Lee (1990) : "Multicointegration", *Advances in Econometrics*, 8, 71-84.

Granger,C.W.J. and P. Newbold (1974) : "Spurious Regressions in Econometrics", *Journal of Econometrics*, 2, 111-120.

Hansen, B. (1992) : "Efficient Estimation and Testing of Cointegrating Vectors in the Presence of Deterministic Trends", *Journal of Econometrics*, 53, 87-123.

Ho, M.S. and B.E. Sørensen (1996) : "Finding Cointegration Rank in High Dimensional Systems Using the Johansen Test", *Review of Economics and Statistics* 78, 726-732.
Show that you have to be careful if the dimension of your VAR gets too large.

Johansen, S. (1988) : "Statistical Analysis of Cointegration Vectors", *Journal of Economic Dynamics and Control*, 12, 231-254.
The first derivation of the maximum likelihood estimator for cointegrated models.

Johansen, S. (1991): "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models", *Econometrica*, 59, 1551-1581.
Extends the 1988 paper, in particular to allow for potential drift (an important extension for the practitioner).

Johansen, S. (1992a): "Cointegration in Partial Systems and the Efficiency of Single-Equation Analysis", *Journal of Econometrics*, 52, 389-403.

Johansen, S. (1992b) : "Determination of Cointegration Rank in the Presence of a Linear Trend", *Oxford Bulletin of Economics of Statistics* 54, 383-399.
This article is useful in particular because it clearly explains how to perform the likelihood ratio tests for number of unit roots in connection with testing for a linear trend in the series.

Johansen, S. (1992c) : "A Representation of Vector Autoregressive Processes Integrated of Order 2", *Econometric Theory*, 8, 188-203.

Johansen, S. (1992d) : "A Statistical Analysis of Cointegation for I(2) variables", *Econometric Theory* 11, 25-59.

Johansen, S. and K. Juselius (1992): "Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and UIP for the UK", *Journal of Econometrics*, 53, 211-245.
It can be hard to understand some of the Johansen-methods. It may be a good idea to read an applied paper before the theory in order to pin down the aims of the method. This is not a bad place to start.

J.J.M. Kremers, N.R. Ericson, and J.J. Dolado (1992) : "The Power of Cointegration Tests", *Oxford Bulletin of Economics and Statistics* 54, 325-349.

MacKinnon, J.G. (1991): "Critical Values for Cointegration Tests", p. 267-276 in Engle, R.F. and C.W.J. Granger (Eds.): "*Long Run Economic Relationships*", Oxford University Press, Oxford.

Park, J. (1992): "Canonical Cointegrating Regressions", *Econometrica*, 60, 119-145.

Phillips, P.C.B. (1986) : "Understanding Spurious Regressions in Econometrics", *Journal of Econometrics* 33, 311-340.

Phillips, P.C.B. (1991a) : "Optimal Inference in Cointegrated Systems", *Econometrica* 59, 283-306.

Phillips, P.C.B. (1991b): "Spectral Regression for Cointegrated Time Series", p. 413-437 in W.A. Barnett, J. Powell and G. Tauchen (Eds.): "*Nonparametric and Semiparametric Methods in Econometrics And Statistics*", Cambridge University Press, Cambridge, UK.

Phillips, P.C.B. and B. Hansen (1990) : "Statistical Inference in Instrumental Variables Regression

with I(1) Processes", *Review of Economic Studies* 57, 99-127.

Phillips, P.C.B. and M. Loretan (1991): "Estimating Long-Run Economic Equilibria", *Review of Economic Studies*, 58, 407-437.

Quah,D. (1990): "Permanent and Transitory Movements in Labor Income: An Explanation for 'Excess Smoothness' in Consumption", *Journal of Political Economy*, 98, 449-476.

Quah,D. (1992): "The Relative Importance of Permanent and Transitory Components: Identification and Some Theoretical Bounds", *Econometrica* 60, 107-119.

Saikkonen, P. (1991): "Asymptotically Efficient Estimation of Cointegration Regressions", *Econometric Theory*, 7, 1-21.

Saikonnen, P. (1992): "Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation", *Econometric Theory*, 8, 1-28.

Sims, C.A., J.H. Stock, and M.W. Watson (1990) : "Inference in Linear Time Series Models with Some Unit Roots", *Econometrica* 58, 113-145.

Stock, J.H. (1987) : "Asymptotic Properties of Least Squares Estimators", *Econometrica* 55, 1035-1056.
A classic article, where Stock showed the super-consistency of cointegration estimators.

Stock, J.H. and M.W. Watson (1988) : "Testing for Common Trends", *Journal of the American Statistical Association*, 83, 1097-1107.

Stock, J.H. and M.W. Watson (1993) : "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems". *Econometrica* 61, 783-821.

Applications

Davidson, J.E., D.F. Hendry, F. Srba, and S. Yeo (1978) : "Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom", *Economic Journal*, 88, 661-692.
An important precursor for cointegration models (before that term was coined). Popularized "error-correction" models in econometrics.

King, R.G., C.I. Plosser, J. Stock, and M. Watson (1991) : "Stochastic Trends and Economic Fluctuations", *American Economic Review*, 81, 819-841.

1.8 Nonlinear-, Instrumental variable-, and GMM-estimation

Books:

Sargan, D. (1988): "Lectures on Advanced Econometric Theory", Oxford: Basil Blackwell. HB139.S26.
This book does not cover non-linear estimation; but it has a nice description of Instrumental Variables estimation that may be a helpful preparation for studying GMM-estimation.

Gallant, R.A.: "Nonlinear Statistical Models", New York, John Wiley, 1987. QA278.2.G35.
This is a comprehensive textbook/ reference in dynamic non-linear modeling. It covers ML and GMM methods. Parts of the books is well written with examples; but the more advanced parts are hard. If you do research involving non linear models, it is a good book to buy.

Theoretical articles

Hall, A. (1992) : "Some Aspects of Generalized Method of Moments Estimation", Mimeo. Forthcoming: *Handbook of Statistics, Vol. 11: Econometrics*, Eds. C.R. Rao and G.S. Maddala.
This is the most accessible introduction to GMM. It overlaps with the econ 266 notes; but is more extensive.

Ogaki, M. (199x) : "Generalized Method of Moments: Econometric Applications", *Handbook of Statistics, Vol. 11: Econometrics*, Eds. C.R. Rao and G.S. Maddala.
A readable survey. One may want to start with Hall's article and then read Ogaki's. Together they give a very good coverage of GMM for a serious applied econometrician.

Hansen,L. (1982) : "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica* 50, 1029-1054.
This is a very famous article. The article covers the theory behind GMM. I found it very hard to read the first 5 times that I tried.

Hansen,L. (1985) : "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators", *Journal of Econometrics* 30, 203-239.

Hayashi,F. and C.Sims (1983) : "Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous Instruments", *Econometrica* 51, 783-798.

Jorgenson,D.W., J.-J.Laffont (1975) : "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances", *Annals of Economic and Social Measurement* 3, 615-640.

Lee, B.-S. and B.F. Ingram (1991) : "Simulation Estimation of Time-Series Models," *Journal of Econometrics* 47, 197-207.
Simulation estimation of GMM. Can be very useful in a situation where the model can not be explicitly "solved" in such a way that usual GMM can be applied.

Pötcher and Prucha (1991a): "Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models. I. Consistency and Approximation Concepts." *Econometric Reviews* 10, 125-217.

Pötcher and Prucha (1991b): "Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models. II. Asymptotic Normality." *Econometric Reviews* 10, 253-357. (Followed by discussion)
NOTE: The Pötcher and Prucha articles are advanced surveys containing new results. They are aimed at the theorist rather that the applied researcher, although you may have to consult them if the assumptions that for example Gallant (1987) employs, do not fit your problem.

Singleton,K.J. (1988) : "Econometric Issues in the Analysis of Equilibrium Business Cycle Models", *Journal of Monetary Economics* 21, 361-387.

Tauchen, G. (1986) : "Statistical Properties of GMM estimates of Structural Parameters Using Financial Markets Data", Journal of Business and Economic Statistics, 4, 397-416.

White,H. (1982) : "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50, 483-500.

GMM estimation - selected applications
The applications here are all to macro-economics. There are also lots of applications of GMM in microeconometrics and finance.

Eichenbaum,M. and L.P.Hansen (1990) : "Estimating Models With Intertemporal Substitution Using Aggregate Time Series Data", *Journal of Business & Economic Statistics* 8, 53-71.

Eichenbaum,M., L.P.Hansen, and K.Singleton (1988) : "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty", *Quarterly Journal of Economics* 103, 51-79.

Hansen,L. and K.Singleton (1982) : "Generalized Instrumental Variables Estimators of Nonlinear Rational Expectations Models", *Econometrica* 50, 1269-1286.
A very famous article that received the Frish medal for best empirical paper in Econometrica. Played a big role in making GMM so popular.

Lucas, R.E. Jr. (1978): "Asset Prices in an Exchange Economy", *Econometrica*, 46, 1429-1445.
Note: This reference is not about GMM, but it gives the theory behind the very influential paper of Hansen and Singleton.

Mankiw,N.G., J.Rotemberg, and L.Summers (1985) : "Intertemporal Substitution in Macroeconomics", *Quarterly Journal of Economics*, 10, 225-251.

1.9 Estimation of variance-covariance matrices

Andersen, T.G. and B.E. Sørensen (1996) - "GMM Estimation of Stochastic Volatility Models. A Monte Carlo Study," *Journal of Business & Economic Statistics* 14, 328-352.
This paper shows among other things that the methods of selection lag length when estimating the weighting matrix (the spectral density at frequency 0) may make a large difference in "tricky" models.

Andrews,D.W.K. (1991) : "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica* 59, 817-859.

Andrews,D.W.K. and J.C.Monahan (1992) : "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator", *Econometrica* 60, 953-967.
At the time of this writing the Andrews papers are the state of the art, and this is the estimator that should be used in applied work.

Newey,W.K. and K.D.West (1987) : "A Simple Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55, 703-708.

1.10 Conditional Heteroskedasticity, ARCH

This is only the most important theoretical articles. There is a huge number of applications. If we decide to cover ARCH in more detail, I will give you a revised reading list.

The first 2 papers are survey papers, which you can consult for more references:

Bollerslev, T., R.Y. Chou, and K.F.Kroner (1992): "ARCH Modeling in Finance: A Review of the theory and Empirical Evidence", *Journal of Econometrics*, 52, 5-61.

Bollerslev, T., R.F. Engle, and D.B. Nelson (1994): "Arch Models", Handbook of Econometrics Vol IV (R.F. Engle and D.L.McFadden).

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity", Journal of Econometrics, 31, 307-327.

Engle, R. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation", Econometrica, 50, 307-327. The "original ARCH-paper".

Engle,R.F. and T. Bollerslev (1986): "Modeling the Persistence of Conditional Variances," *Econometric Reviews*, 5, 1-50.

Engle,R.F., D.M. Lillien, and R.P. Robbins (1987): "Estimating Time Varying Risk Premia in the Term Structure: the ARCH-M Model," *Econometrica*, 55, 391-407.

Ho,M.S., W.R.M. Perraudin, and B.E. Sørensen (1996): "A Continuous Time Arbitrage Pricing Model with Conditional Stochastic and Jumps". *Journal of Business & Economic Statistics*, 14, 31-44.
Here you can see how I think that it should be done. This paper also applies GMM estimation (in a somewhat different fashion from the papers mentioned above).

Nelson, D.B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach", Econometrica, 59, 347-371.

2 Method

It is important before discussing various econometrics approaches to have a bit of historical perspective and talk a little about some important philosophies for empirical work.

Books about methodology proper is not of too much help for the applied economist. Two widely read books on methodology are Blaug (1980) and Caldwell (1982). One thing that strikes an econometrician when reading those books is the fact that a major theme for both authors is whether testing of economic theories is possible; but they have absolutely no discussion on how this testing should or might take place. Nevertheless, it is important to read books on methodology, if on has interest in that area, because methodologists develop a precise language by which one can discuss methods.
Blaug seems to be quite positive towards the idea that economic models should be testable, although he realizes that it is not as simple as that, whereas Caldwell has a much more sceptical view towards testability as you might guess from the title ("Beyond Positivism..."). One interesting line of thought is developed by Lakatos. We won't go into details of his theories; but just note that he claims that all theories come with a "protective belt" of supporting hypothesis, so that an empirical rejection can always be attributed to flaws in the protective belt rather than in the theory.
An example from economics. Say you want to test the hypothesis of Rational Expectations. Then you will build a model, usually assuming a form of a utility function etc., and estimate it. If you reject the implications of the model, does that mean that RE is wrong. No! You probably just chose the wrong utility function. Or if you didn't do that, there is an infinity of reasons why the data might not be just right. And so on. You can think of many examples yourself. I find the protective belt analogy very convincing; but I don't see it as devastating for testing. If models generated from a certain theory consistently fail to perform, this theory will sooner of later fall out of fashion. Isn't this a reasonable description of what has happened over the years in economics?

The debate on econometrics, among econometricians, intensified around 1970 and after. There was a widespread dissatisfaction with the general state of empirical work, presumably provoked by the hard times that the big econometrics models had with predicting the developments in the economy in the early 70ies.

A caricature of empirical work in the 60ies could be like this. An article would contain a first half of theory, which could be more or less formal. In the middle of the article the theory would drop out and the researcher would start estimating linear regressions. Presumably the theory had motivated which variables should enter the equations. Now the researcher would run 100, say, regressions and only 5 per cent of those would give significant coefficients. So the other 95 per cent would be discarded and the 5 per cent "good" results presented in the paper. Variants of this method will be termed "Data Mining". If results have been obtained by data mining the test statistics are totally unreliable. (Also: what is a "good" result. Our profession will hopefully soon be mature enough that all results will be published if they represent careful research, whether they

support a given theory or not).

This line of modeling was attacked by several young econometricians and economists. Leamer had a famous article called "Let's Take the Con Out of Econometrics" (see Granger (1990)) and Hendry suggested an approach based on much more explicit recognition of time series structure combined with extensive testing. Sims criticized structural equation modeling and suggested the use of vector time series methods (see Granger (1990) for articles by Hendry and Sims). Hendry was particularly influential in England whereas Sims had a quite large following in the US. In the US, the most influential critique of econometric practice was voiced by Lucas, who claimed that typical econometric equations could not be assumed to be stable to policy interventions and therefore useless for what they were primarily intended for (see Lucas and Sargent (1981)).

Consider the model

$$y_t \ = \ b'x_t \ ,$$

where x_t is a vector of regressors and b a vector of parameters. The early way to estimate this model would consist in adding an innovation term and estimate the model by least squares. One would then look at diagnostics like DW and maybe test for heteroskedasticity and if these diagnostics were satisfactory, so were the model. (This is also a bit of caricature, I know that many econometricians did quite a good job, especially making use of residual plots to check for problems. The problem of this approach is that it is quite subjective as compared to formal tests).

The Hendry critique was primarily centered on the fact that such an approach did not take the actual stochastic structure of the model into account. Hendry suggested that one instead assumed a general *statistical* model for the data, which he described by an unknown density $D(x_t, y_t \ ; \beta)$, where β is a vector of parameters. This unknown density is called the Data Generating Process (DGP) and it is the job of the empirical econometrician to try and uncover this DGP or at least the part of it that is relevant for modeling y_t. In order to do this in practice, one ought to start with the most general model possible and then *test* down to the preferred model. Hendry claims that the three golden rules of econometrics is "test, test and test". Hendry has developed a software package PC-GIVE that makes his method easy to implement. The Hendry school, to which we will return in the Time Series part of the course, is also know as "the LSE tradition" (much of it going back to Dennis Sargan who was professor at LSE) or "British Econometrics". Hendry's method was a precursor for the modern co-integration methods.

It is of course impossible to evaluate broad methods in general; but it is important to notice already the statistical approach. Hendry's method may be superior if you want to obtain the best possible statistical model for the selected series. Notice that this is totally different from what the methodologist were considering, namely whether economic theory is testable.

One development that is less susceptible to data mining is the formulation of economic models

that are non-linear. One of the advances in the 70ies was the widespread use of nonlinear models that were derived from theory. Still the time series structure was not taken into account to any major degree. In their article in the Handbook of Econometrics Granger and Watson (1984) claims that "For many years economists and particularly econometricians behaved as though they did not realize that much of their data was in the form of time series or they did not view this fact as being important". However, the situation has changed a lot since then.

The Rational Expectations (RE) school of econometrics which was pionered by Sargent, evolved as a reaction to the Lucas critique. Let us look at a simple example, as is typical for the RE school this assumes an "agent" maximizing "utility". Here look at a firm maximizing profits over an infinite time horizon. The criterion function is

$$Max_{N\to\infty} E_t \sum_{j=0}^{N} \beta^j [(a_{t+j} - w_{t+j})n_{t+j} - (\gamma/2)n_{t+j}^2 - (\delta/2)(n_{t+j} - n_{t+j-1})^2 \ ]$$

here n_{t-1} is fixed, where n_{t+j} is input of a production factor to time t+j, w_{t+j} is price of the factor at t+j, and a_{t+j} is a stochastic variation in the production process that the producer observes; but that is unobserved to the modeler. γ , β, and δ are positive constants. It is assumed that w_t and a_t follows stochastic processes that are known to the producer.
This is a very typical example of an RE model of early vintage. (The linear quadratic models that can be solved explicitly is treated in some advanced examples by Hansen and Sargent (in Lucas and Sargent (1981)). In the present course we will concentrate on the later non-linear models; primarily because they need more specialized econometric methods (the subject of the course!) but they have also been more popular in recent years).
Hansen and Sargent (see p. 94 ff. in Lucas and Sargent (1991), and also Sargent (1987) App. A.5 for mathematical details) show that the above model has the solution

$$n_t \ = \ \rho n_{t-1} - (\rho/\delta) \sum_{j=0}^{\infty} (\beta\rho)^j E_t[w_{t+j} - a_{t+j}] \ ,$$

where $1/\rho$ is the smallest root of the equation $1 + \frac{\gamma/\delta+1+\beta}{\beta} z + \frac{1}{\beta} z^2 = 0$. Now assume for simplicity that $E_t a_{t+j} \ = \ 0; j > 0$ and that $w_{t+j} \ = \ \alpha^j w_t + u_{t+j}$ for all $j > 0$ where $E_t U_{t+j} = 0$ and positive parameter $\alpha < 1$. Then the solution is

$$n_t \ = \ \rho n_{t-1} - b w_t + a_t \ .$$

for

$$b \ = \ \frac{\rho}{\delta(1 - \beta\rho\alpha)} \ .$$

This model demonstrates all the outstanding features of the RE school.
I. Heavy reliance on economic theory (but often with a representative agent),
II. All the stochastics is *part of the model.*

III. Cross-equation restrictions: To estimate the model you would use the equations

$$n_t \;=\; \rho(\gamma,\delta,\beta)n_{t-1} - \frac{\rho(\gamma,\delta,\beta)}{\delta[1-\beta\rho(\gamma,\delta,\beta)\alpha]}w_t + a_t \ .$$

and

$$w_t \;=\; \alpha w_{t-1} + u_t \ .$$

This is a simple example of cross-equations restriction, where α occurs in both equations. A note on identification. In non-linear equation systems there are no simple conditions for identification, so one has to examine on an ad hoc basis whether there is a unique solution in terms of the structural parameters. Here, the second equation identifies α, but it is not so easy to see if the rest of the parameters are identified - it is pretty obvious that we can not identify both β and δ, and my guess is that you would choose to fix β and the system would then be identified. We won't have time to discuss identification much in this course (and there are not really any general results), but be aware of the problem. I could also mention a 4th typical feature of linear RE models: they are often complicated. Some of the linear RE models that Hansen and Sargent developed in the early 80ies are very complicated (the above were very much simplified), and they may well be too complicated to estimate on most datasets, which may well be the reason why this type of models are not more popular.

The example is also useful in order to illustrate the Lucas critique. The main point of the Lucas critique is that if one estimates the model for n_t, one will get an estimate of the parameter b, but if the model is not derived from a basic optimization problem the model will break down if an intervention (in most examples a policy intervention) changes one of the basic parameters, e.g. α. See Lucas (1976) (printed in Lucas and Sargent (1976)). The Lucas critique was seen as a scathing critique against the large Keynesian macroeconometric models; but even though it was widely influential it is not obvious how important it is in practice (Sims (1982) argues that it may turn out to be more a cautionary footnote than a devastating critique).
For our purpose here it is most important to notice the model building strategy that explicitly takes into account the stochastic nature of the variables *at the model building stage* and typically imposes numerous non-linear restrictions on the parameters of the model.

Can one say that an economic theory approach is better than a statistical approach. I doubt it. There seems to have been quite a bit of discussion of various models advantages (see the articles in Granger (1990)); but in general my point of view is that such a discussion is totally meaningless unless one discusses which method is best for a specific purpose. For short term forecasting a more statistically based model may be preferable; but if one want to test economic theory, the model has to be based tightly on theory. An appreciation of that point would, I think, make for a more enlightened debate.

An interesting example is the consumption function. Here "British" and "American" methods have competed head on. Two very famous articles which are in the British and the American tradition, respectively, are the Error Correction Model (ECM) of Davidson et al. (1978), sometimes

referred to as DHSY (1978) after the authors Davidson, Hendry, Srba and Yeo, and Hall (1978) (reprinted in Lucas and Sargent (1981)), which "showed" that consumption is a martingale. (A time series X_t is a martingale if the conditional expectation of the change in the series is 0, i.e. $E_t(X_{t+1}) = X_t$. A lot of effort has gone into testing Hall's model, and it proved surprisingly hard to reject the simple prediction, although the consensus now is that "consumption" is not a martingale. (Notice that I haven't made very explicit whether I talk about aggregate consumption or individual consumption or how that is defined). Some authors seemed to take that as a rejection of RE. That is of course crazy. In order to derive his conclusion Hall assume quadratic utility and a fixed rate of return that furthermore had to equal the subjective rate of time preference. Talk about protective belt! More understandable some authors might take the fact that the ECM model seems to behave better as support for the British school. I think that there is no doubt the ECM model predict consumption better; but that might be beside the point if you want to criticize Hall, who was not really interested in prediction. Because the Hall model is derived from explicit theoretical assumptions it gives a direct feedback to theory. For example, there seems to have developed a general consensus in the profession that quadratic utility functions do not really do the trick. (An interesting footnote is that the DHSY model can only be derived from a theoretical model by assuming quadratic utility - see Nickell (1985)). The ECM model is still relevant since it gave the impetus to the modern co-integration models (the ECM model is an example of a single equation cointegration model), and much of the applied cointegration modeling is in the tradition of Hendry et al. with a strong focus on statistical fit as opposed to economic theory content.

Bayesian approaches are also gaining terrain. Arnold Zellner from university of Chicago has for many years tried to recruit econometricians to the Bayesian school, but so far has not convinced the mainstream. Also Chris Sims from Yale has been proposing Bayesian methods, as for example in Sims and Uhlig (1991). This paper (and related articles) seem to have provoked strong reactions as witnessed by Phillips (1991) (a special issue of Journal of Empirical Economics). Bayesian theory will, however, not be a focus of the present course. Talk to Tony if that has your interest.

Presently the dominating approach in main stream macro/general journals like the QJE and the AER is the "Cambridge" or "NBER" style, which stresses simplicity and transparancy. This tradition reminds me personally a bit too much of 60ies econometrics. For a provocative methodological paper from an influential Cambridge macro economist see the article by Summers (1991). Summers claims that econometric studies has had very little influence on macroeconomic theory, and only very simple estimations or even tabulations have been convincing. Summers challenges the reader to find one simple instance where an econometric test has changed the way that macroeconomists think about the world. I cannot think of any example of that, but I think that is besides the point. As I argued previously, this is exactly what the "protective belt" analogy would make us expect, but this does not imply that a series of empirical papers (maybe not all performing Tests with capital T) will not change the way that macroeconomists think about the world if they consistently reject (or confirm) a certain economic hypothesis.

3 Time Series

This note gives a short introduction to time series. The aim is to define the most commonly used concepts that you will meet in the literature and treat the concepts that are important for understanding of cointegration theory in more detail. Be aware that time series analysis is a huge area and we will only touch the surface of it.
The book by Harvey (1981) is the easiest introduction for economists, and quite readable. The book by Lütkepohl (1991) treats multivariate processes and it is fairly accessible to economists (it is intended as a graduate textbook for economics students). It has most formulas that you might want to use for standard vector time series models. If you need to go further than that, in the area of vector time series, the book by Hannan and Deistler (1988) is good (but this is a much harder book). If you need more general "probabilistic" results, by which I basically mean more general (or different) assumptions about the probability structure of the model innovations, you may have to go to the statistics or probability texts in the references or to journal articles. Granger and Newbold's book is also a quite popular textbook, it covers a lot of topics, but not quite at the same level of detail as Lütkepohl's book.

A time series is a collection of stochastic variables $x_1, .., x_t, .., x_T$ indexed by an integer value t. The interpretation (which will be the only one applied in this course) is that the series represent a vector of stochastic variables observed at equispaced time intervals. The series is also some times called a stochastic process, but in this course I will try to restrict the term stochastic process to series in continuous time, i.e. series of the form $x(t)$; $t \in [0, T]$. In engineering applications one might have processes that are actually observed in continuous time, and even though stock prices in principle can be observed continuously we will only meet continuous processes in this course in connection with limiting distribution of unit root processes (to be defined). As you might expect, continuous processes are much harder to treat rigorously than are discrete processes. The distinguishing feature of time series is that of temporal dependence: the distribution of x_t *conditioned* on the previous value of the series depends on the outcome of those previous observations, i.e. the outcomes are not independent. For the purpose of analyzing a time series we will usually model the time series over all the non-negative integers: $x_t \, ; t = \{0, 1, .., \infty\}$ or $x_t \, ; t = \{-\infty, .., 0, 1, .., \infty\}$. Time 1 *or* time 0 will be the first period that you observe the series. In a specific model you will have to be explicit about the initial value, as will be clear from the following.

3.1 Stationarity

Definition A time series is called *stationary* (or *strictly stationary* or *completely stationary*) if the distribution of $x_{t1}, x_{t2}, .., x_{tK}$ for any selection of K time indices (for any K) is the same as the distribution of $x_{t1+m}, x_{t2+m}, .., x_{tK+m}$ for any positive integer m. This is the rigorous definition. It implies that the distribution of x_t does not depend on the time index t; but this is not quite enough for a stringent definition, since the marginal distribution conceivably could be constant over time while the covariance between, say, neighboring observations could change. This is ruled out by the

stringent definition.

For the following definition assume for simplicity that the series is univariate.
Definition A time series is called *covariance stationary* (or *weakly stationary*, or *wide sense stationary*, or *2nd order stationary*, or (unfortunately) just *stationary*) if

$$\begin{aligned} &E(x_t) = \mu \ , \\ &E[(x_t - \mu)^2] \ = \ \gamma(0) \ , \\ &E[(x_t - \mu)(x_{t+k} - \mu)] \ = \ \gamma(k) \ ; \ k = 1, 2, ... \ , \end{aligned}$$

where $\gamma(k) \ ; \ k = 0, 1..$ are finite and (this is the stationarity) and independent of t.
Notice that this definition includes the condition that x_t has finite variance. (It used to be possible for econometricians to regard time series with infinite variance as pathological; but stock prices may or may not have finite variance, and the class of ARCH models that are widely used at the present also include models with infinite variance). Notice that there is a quite long tradition in time series to focus on only the first two moments of the process, rather than on the actual distribution of x_t. If the process is normal all information is contained in the first two moments and most of the statistical theory of time series estimators is asymptotic and more often than not only dependent on the first two moments of the process. We will refer to theoretical constructs that only involve the two first moments of the process as **2nd order concepts**.

The first 2nd order concept we will consider is $\gamma(k)$ considered as a function of k. This is called the **autocovariance function**, and is (maybe in particular was) a heavily used tool in applied time series analysis (see f.ex. Harvey's book). It is still an important theoretical tool. You may also come across the term autocorrelation function (often abbreviated ACF), which is the autocovariance function divided through by the variance of x_t, i.e. $[1, \ \rho(1) = \gamma(1)/\gamma(0), ..]$.
Note that strict stationarity and finite variance implies covariance stationarity and that covariance stationarity and normality implies strict stationarity.
For vector processes the definition is exactly the same; but the autocovariance function takes takes matrix values $\Gamma(k)$, where

$$E[(x_t - \mu)(x_{t+k} - \mu)'] \ = \ \Gamma(k) \ ; \ k = .., -2, -1, 0, 1, 2, ... \ .$$

Note that we also define the autocovariance function for negative values. It is easy to see, though, that $\Gamma(k)' \ = \ \Gamma(-k)$.

3.1.1 The lag- (backshift-) operator, the difference operator and their inverses

In this subsection I will try to explain lag-operators in some detail. There seems to be very few explanations at this level in the literature, since the engineering oriented time-series books take the engineers approach of just multiplying and dividing lag-polynomials without worrying the slightest bit about what they are actually doing. There are of course mathematical statistics texts that are

fully precise, but they are often difficult. The explanation I give here is precise, but not to the level of the mathematical statistics texts, since I leave out a precise mathematical discussion of when the limit of infinite sums are meaningful.

Recall that the backshift operator B (or the lag operator L - we will use both symbols for the same thing) is defined by $Bx_t = Bx_{t-1}$. We will also define the symbol B^k as $B^k x_t = x_{t-k}$. You should think of the lag-operator as moving the whole process $\{x_t \; ; t = -\infty, ..., \infty\}$. Notice that it is here practical to assume that the series is defined for all integer t - in applications this may not always make sense, and in these cases you may have to worry about your starting values. We will define **lag polynomials** as polynomials in the lag operator as follows. Let A(L) be the the lag polynomial

$$A(L) = A_0 + A_1 L + ... + A_p L^p$$

which is defined as an operator such that

$$A(L)x_t = A_0 + A_1 x_{t-1} + ... + A_p x_{t-p} \; .$$

This simply means that the equation above defines A(L) by the way if operates on x_t.

You can also invert the lag operator. The inverse lag-operator will be denoted

$$B^{-1} \text{ or } \frac{1}{B} \text{ or } L^{-1} \text{ or } \frac{1}{L} \; .$$

These are just symbols for the operator that is the inverse of the lag-operator, namely the operator that shifts the time-index one period forward in time: $B^{-1}x_t = x_{t+1}$. Obviously $BB^{-1} = I$. Similarly the inverse of the lag-polynomial is defined, we use the notation

$$A(L)^{-1} \text{ or } \frac{1}{A(L)} \; ,$$

where the "fractions" notation is used mostly for scalar processes. Of course $A(L)^{-1}$ is defined as the operator such that $A(L)^{-1}A(L)x_t = A(L)A(L)^{-1}x_t = x_t$.

How do you find the inverse to a lag-polynomial? The key observation is that we can add and multiply lag-polynomials in exactly the same way as we can add and multiply polynomials in complex variables. For a given lag-polynomial $A(L)$ we therefore define the corresponding *z-transform* (also a label from the engineering literature) $A(z)$ where z is a complex number. For the purpose of this course you should always think of z as having length 1 (being on the complex unit circle), even though some clever theorems for z-transforms can be shown using other z's (using complex function theory).

For now, restrict attention to scalar lag-polynomials, which I denote $a(L)$. It is well know that if $a(z) = 1 + a_1 z + ... + a_k z^k$ is a complex polynomial then

$$a(z) = (1 - \alpha_1 z)(1 - \alpha_2 z)...(1 - \alpha_l z)$$

where $\frac{1}{\alpha_1}, ..., \frac{1}{\alpha_k}$ are the roots of the polynomial. Recall that the roots will be complex conjugates if the coefficients of $a(z)$ are real numbers. Of course the lag-polynomial $a(L)$ factors the same way.

This means that all problems concerning inversion of lag polynomials can be reduced to inversion of the first order polynomial $1 - az$. This is really useful and even better: the multidimensional case can also be reduced to considering the scalar first order polynomial. We will go through the gory details below. You all know the formula (valid for $|x| < 1$)

$$\frac{1}{1-x} \;=\; 1 \;+\; x \;+\; x^2 \;+\; x^3 \;+\; ....$$

(If you are not fully familiar with this equation, then you should take a long look at it, prove it for yourself, and play around with it. It is very important for the following).

For $|a| < 1$ we now get

$$\frac{1}{1-az} \;=\; 1 \;+\; az \;+\; (az)^2 \;+\; (az)^3 \;+\; ....$$

for $|z| = 1$ and

$$\frac{1}{I-aL} \;=\; I \;+\; aL \;+\; (aL)^2 \;+\; (aL)^3 \;+\; ....$$

when this expression is well defined. Notice that this may depend on your assumptions about the x_t's being operated upon by the lag operator. We will take the definition just given as the formal definition of $(1 - aL)^{-1}$ - then when you want to use the inverse of the lag operator on a given series you have to make sure the expression

$$x_t \;+\; ax_{t-1} \;+\; a^2x_{t-2} \;+\; ....$$

is well defined.

The inverse of the general scalar lag polynomial is now simply defined by inverting the "component" first order lag polynomials one-by-one just as you invert a complex polynomial.

The difference operator Δ is defined as $\Delta = I - B$ giving $\Delta x_t \;=\; x_t \;-\; x_{t-1}$. The n'th power of the difference operator is obtained by applying the difference operator k times. For example: $\Delta^2 x_t \;=\; (x_t \;-\; x_{t-1}) - (x_{t-1} \;-\; x_{t-2}) \;=\; x_t \;-\; 2x_{t-1} \;+\; x_{t-2}$.

Polynomials in the difference operator are defined parallel to the definition of lag polynomials. The inverse of the difference operator is the summation operator

$$\Delta^{-1} x_t \;=\; \sum_{i=0}^{t} x_i \;.$$

Notice that you have to even more careful with inverse difference operators than with inverse lag operators. This is because the infinite sum $x_t + x_{t-1} + x_{t-2} + ...$ typically is *not* well defined. You therefore have to be explicit about the initial value x_0 in order to obtain

$$\Delta^{-1}\Delta x_t \;=\; \Delta\Delta^{-1} x_t \;=\; x_t$$

3.2 Order of integration

Definition A time series is called *integrated of order k* if Δ^k is stationary. We say that x_t is I(k). Notice that k can be positive or negative, although k=1 and k=2 are most commonly encountered. The definition can also be extended to non-integer k (in which case the series is called fractionally integrated - fractional integration has recently been quite popular in econometric research, and may turn out to become very useful in economics). An I(1) series is also often called difference stationary.

So an I(1) process is an example of a time series that is not stationary. Another example is the process $y_t = \beta t + x_t$, where x_t is stationary, which contain a linear trend. It is usually quite easy to detrend a series containing linear (or polynomial) trends, so the fact that series may contain deterministic trends does not subtract from the importance the theory of stationary processes. Sometimes you may get the impression from a paper, that follows a strategy of testing the series used for stationarity vs. difference stationarity, that the author is not imposing *any* stationarity assumptions a priori. Nothing could be more wrong. The most devastating form of non stationarity is when a time series just not follows the same distribution over time even after detrending, differencing or whatever you could think of. In economics this could for example be caused by changes in policy, cf. the Lucas critique, so doing pure time series analysis does not make one immune to this critique even if all economic assumptions are implicit. As empirical economists you just have to *hope* that the structure of the economy that you are modeling show enough stability to make your models useful. (Testing for structural breaks in time series is, by the way, a very active research area at the present).
Another very common example of a non stationary process is a process where x_0 is considered fixed. Such a process can not be stationary (unless it is degenerate and only takes one value). Sometimes it is simpler to assume that the process is "started" at minus infinity, which is one major reason that processes are often modeled for all integer t. We will discuss the problem of initial values again in connection with AR(1) models and I(1) processes.

Definition: A process x_t is called a martingale with respect to an increasing sequence of sigma-fields (or σ-fields) $\{\mathcal{F}_n, n \in]-\infty, \infty[\}$ if $E\{x_{t+k}|\mathcal{F}_t\} = x_t$ for all $k > 0$. For all practical purposes you should think of $\mathcal{F}_n$ as $y_n, y_{n-1}, ...$ where y_n is a vector of variables that *at least* contains x_n. In economics we will always think of y_n as either the universe of our model or as x_n. We will normally write $E_t x_{t+k}$ as an abbreviation for $E\{x_{t+k}|\mathcal{F}_t\}$. Note that we here are not explicit about the conditioning set. If nothing else is mentioned think of the conditioning set as being the past of the process itself. In modern rational expectations economics, models can have very different implications depending on which σ-field (often called "information set" in that literature) that you condition upon, so you should not be cavalier about that. In this course, we will be often be imprecise about the conditioning sets because we only focus on estimation issues. For a rigorous mathematical introduction to martingale theory you should consult Hall and Heyde (1980) (and if you want to know exactly what σ-fields are, you should consult standard textbooks in measure theory).

Definition: A process e_t is called a *martingale difference sequence* if $E_t(e_{t+1}) = 0$.

Convince yourself that if x_t is a martingale the Δx_t is a martingale difference sequence and vice versa (the partial sum up to time t of a martingale difference sequence is a martingale).

Definition: A stationary process e_t with mean 0 is called *white noise* if $\gamma(k) = 0$ for $k \neq 0$. Notice that a stationary martingale difference sequence is white noise.

Unfortunately not all authors define white noise exactly like this (some define it to be iid), but this is the most common definition. Notice that in these notes, white noise is a 2nd order concept.

3.3 The Wold decomposition:

Wold's theorem

Any stationary process X_t has a decomposition in an infinite MA-part and a deterministic part:

$$X_t \ = \ v_t + \ B(L)u_t \ ,$$

where v_t is "deterministic" and the MA-representation is infinite in general. The series u_t is mean zero and uncorrelated.

A sketch of a proof in the normal case is the following:
Define

$$u_t \ = \ x_t - E[x_t|x_{t-1}, x_{t-2}, ...]$$

Since the conditional expectation is linear in the variables that we condition upon (this is well known in the case where we condition on a finite number of variables), we find

$$u_t \ = \ x_t \ - \ \alpha_1 x_{t-1} \ - \ \alpha_2 x_{t-2} \ + \ ... \ ,$$

or

$$x_t \ = \ u_t \ + \ \alpha_1 x_{t-1} \ + \ \alpha_2 x_{t-2} \ + \ ... \ .$$

Now use this equation shifted one period back to substitute for x_{t-1} and you get

$$x_t \ = \ u_t \ + \ \alpha_1 * [u_{t-1} \ + \ \alpha_1 x_{t-2} \ + \ \alpha_3 x_{t-3} \ + \ ...] \ + \ \alpha_2 x_{t-2} \ + \ \alpha_3 x_{t-3} ... \ ,$$

from which

$$x_t \ = \ u_t \ + \ \alpha_1 u_{t-1} \ + \ [\alpha_1^2 \ + \ \alpha_2] x_{t-2} \ + \ [\alpha_1 \alpha_2 \ + \ \alpha_3] x_{t-3} \ + \ ... \ .$$

Substituting again, we find

$$x_t \ = \ u_t \ + \ \alpha_1 u_{t-1} \ + \ [\alpha_1^2 \ + \ \alpha_2] u_{t-2} \ + \ .....$$

It is now clear that the coefficients of $B(L)$ can be obtained by repeated substitutions. (Of course this quickly becomes very tedious and this derivation is not used much in practice, but it shows that it can be done). The "deterministic" term is then $X_t \ - \ B(L)u_t$. Now you can see that this

term is not necessarily deterministic in a strict since any random term that was decided at $-\infty$ would be part of v_t. In economics, where normally only one realization of a time series is available, I do not see much use for a model where the v_t were not assumed to be deterministic. Since a drift term leaves the process non-stationary the v_t term would have to be a constant. For that reason you can basically think of the Wold representation as saying that any stationary process can be written as an infinite moving average.

3.3.1 2nd-order representations and normality

In the case where x_t is not normally distributed then you will get the same Wold-expansion as if x_t were normal. Why is that? This is because the Wold-expansion is an expansion in terms of uncorrelated terms. In other words a 2nd order concept. The conditional expectation ($E(X|Y)$, say) in a normal model is the function $f(Y)$ of Y that minimizes $E(X - f(Y))^2$. In the normal case this happens to be a linear function. In the general (non-normal) case you find the Wold-representation by finding the linear function that minimizes the variance of the remainder everywhere where the conditional expectation was used in the above sketch of the proof. This is called a *linear projection*. The linear projection gives you the same linear combinations as you get if you assume that the process is normal and use the conditional mean. Think about this for a second or two - it all follows because the formula for the variance of a linear combination is independent of the distribution of the random variables.

For a rigorous proof (which unfortunately needs the introduction of a more advanced mathematical machinery), see chapter 2 of Hansen and Sargent (1991). The rigorous proof is not much different from what is done here, but here we have not shown carefully that all the infinite sums always are well-defined (which they are).

3.4 MA models:

The simplest time series models are the moving average (MA) models:

$$x_t \;=\; \mu \;+\; u_t \;+\; B_1 u_{t-1} \;+\; ... \;+\; B_l u_{t-l} \;=\; \mu \;+\; B(L)u_t,$$

where the innovation u_t is a martingale difference sequence (or white noise) and the lag-polynomial is defined by the equation. The positive integer l is called the **order** of the MA-process. MA processes are quite easy to analyze because they are given as a sum of independent (or uncorrelated) variables. However, they may not always be easy to estimate: since it is only the x_ts that are observed, the u_ts are unobserved, i.e. latent variables.

Consider the simple scalar MA(1)-model (I leave out the mean for simplicity)

$$(*)\; x_t \;=\; u_t \;+\; b u_{t-1}\;.$$

If u_t is an independent series of $N(0, \sigma_u^2)$ variables, then this model really only says that x_t has mean zero and and autocovariance function: $\gamma(0) = (1+b^2)\sigma_u^2;\; \gamma(1) = \gamma(-1) = b\sigma_u^2;\; \gamma(k) = 0;\; k \neq$

$-1, 0, 1$. (Notice that I here figured out what the model says about the distribution of the observed x's. Therefore this is what the model "really" says. The autocorrelation function for the MA(1) model is trivial except for $\rho(-1) = \rho(1) = \frac{b}{1+b^2}$. By elementary calculus, it is easy to show that the MA(1) has a maximum of 0.5 for the first order autocorrelation function.

This same statistical model could also be modeled

$$(**)\ x_t \ = \ v_t \ + \ bv_{t+1} \ .$$

with v_t having the same distribution as u_t. Typically one will rule out the model (**) by assumption, but even the model (*) does not identify the parameter b uniquely. If you only observe the data then you can calculate the empirical autocovariance function and if the data are normally distributed then this is basically all the information that the data is is going to give you (apart from the mean). Assume that the autocovariance function for x_t is $\gamma_x(0) = 1$ (you can always normalize) and $\gamma_x(1) = r$ and the higher order autocorrelations are all 0. Then we find the equation

$$\frac{c}{1+c^2} \ = \ r$$

to determine c. But this is a second order polynomial, which has the solution $c = 0$ if $r = 0$ and $c = \frac{1}{2r} \pm .5\sqrt{r^{-2} - 4}$, otherwise. Notice that this in general gives one solution in the interval $[-1, 1]$ and one solution outside this interval.

Consider equation (*) again. In lag-operator notation it reads

$$x_t \ = \ (1 \ + \ bL)u_t \ ,$$

which can be inverted to

$$u_t \ = \ (1 \ + \ bL)^{-1}x_t \ = \ x_t \ - \ bx_{t-1} \ + \ b^2x_{t-2} \ + \ ...$$

It is quite obvious that this expression is not meaningful if $|b| \geq 1$ since the power term blows up. In the case where $|b| < 1$ the right hand side converges to a well defined random variable. This follows for example from the following theorem for absolutely summable sequences:

Theorem
Let z_t be a sequence of random variables with $Ez_t^2 < c$ for some finite constant c, and let b_i be an *absolutely summable sequence* (i.e. $\sum_{i=1}^{\infty} |b_i| < \infty$) then $\sum_{i=1}^{N} b_i z_i$ converges (for $N \to \infty$) in probability to a well defined random variable, which we will denote by $\sum_{i=1}^{\infty} b_i z_i$

Definition: The scalar MA(q) model is called *invertible* if all the roots of the lag-polynomial b(L) (strictly speaking the corresponding z-transform b(z)) are outside the unit circle.

In the invertible model the innovation term u_t is a function of (the infinite past of) the observable x_ts. This is often the most sensible assumption, and in any event the invertibility assumption is almost always imposed in estimations in order for the maximum likelihood estimator to have a unique maximum (i.e. for the model to be identified). Of course you may have an economic model

where the u_ts have another interpretation - there is nothing inherently wrong with the non-invertible MA model, but then you have to be careful with identification if you are estimating the model.

Let me finish this subsection by discussing the assumption on the u_t terms in a little more detail. In time series books you will often read that the exact assumptions on u_t (white noise, independence etc.) do not matter. It may not matter a lot for your estimations; but an economist/econometrician ought not (except under extreme duress) write down a model without making clear to him- or herself what are the assumptions on the error terms and the initial conditions. This type of modeling goes back to the old days when economists build deterministic models and then at the last moment tagged on an error term before sending the model down in the basement for processing by the research assistants. In modern empirical model building the model is usually formulated explicitly in terms of random variables and the the innovation terms will therefore have a clear interpretation. I think it is almost correct to say that all discussions about what to assume about the innovation term could have been avoided if the researcher had considered the fact that he is working with stochastic variables from the outset. In a stochastic economic model it will very often be the case that the error term has the character of a martingale difference sequence (which stops all discussion of whether the model go backward or forward in time). A statistician may be more prone to assume that the innovation terms are iid (and maybe normal) which is convenient for Maximum Likelihood estimation. Notice that (assuming stationarity)

$$IID \;\Rightarrow\; MDS \;\Rightarrow \text{white noise} ,$$

so the white noise assumption is the weakest assumption.

3.5 AR models:

The most commonly used type of time series models are the auto regressive (AR) models. In vector form it is often denoted a VAR process:

$$x_t \;=\; \mu \;+\; A_1 x_{t-1} \;+\; ... \;+\; A_k x_{t-k} \;+\; u_t ,$$

where the innovation u_t is a martingale difference sequence (or white noise). Here k is a positive integer called the order of the AR-process. Such a process is usually referred to as an AR(k) process. For simplicity I will use the notation AR-processes whether the process is scalar or multi dimensional.

Note that sometimes "VAR-modeling" is used in a much more specific sense as a specific approach (using vector ARs as a tool) to macro econometric modeling. The VAR approach in the narrow sense was suggested by Christopher Sims as an alternative to the big Keynesian macro econometric models. The basic philosophy was that the usual macro models only can be identified under extensive a priori restrictions (in order for individual equations to be identified it is usually assumed that a given endogenous variable only depends on a limited number of other endogenous variables). Sims finds many of these a priori restrictions "incredible" and suggest that one starts

with an unrestricted VAR model instead. Sims' article "Macroeconomics and Reality" is reprinted in Granger (1991). We will not go into any more detail with Sims' methodology; but it is a possible subject for a representation of a paper (the RATS software is particular well suited for VAR modeling).

Most of the intuition for AR processes can be gained from looking at the AR(1) process, which is also by far the commonly applied model. Assume for simplicity that the mean is 0, and let us first consider the scalar case.

$$x_t \;=\; ax_{t-1} \;+\; u_t \;.$$

If $|a| \;<\; 1$ the process is called *stable*. For example assume that x_0 is a fixed number. Then

$$x_t = x_0 a^t \;+\; a^{t-1}u_1 \;+\; ... \;+\; au_{t-1} + u_t \;.$$

The important thing to notice is that asymptotically the influence on x_t of the initial value x_0 becomes negligible. It is also easy to see that if u_t is i.i.d. with variance σ^2 then

$$var(x_t) \;=\; \sigma^2[1 + a^2 + ... + (a^2)^t] \;=\; \frac{1 - a^{2(t+1)}}{1 - a^2} \;\rightarrow\; \frac{1}{1 - a^2} \;,$$

for $t \rightarrow \infty$.
Notice that the stable process is not stationary if x_0 is fixed (as the variance varies). However, for large t it approximately is, and if the process was started at $-\infty$ then x_0 would be stochastic with variance $\frac{1}{1-a^2}$. So the stationary AR(1) is a stable process with the initial condition that x_0 is distributed like u_t with variance $\frac{1}{1-a^2}$.

Notice what happens if $a \;\rightarrow\; 1$: the variance tends to infinity and the stationary initial condition can not be defined - at least in typical case as where $u_t \sim N(0, \sigma^2)$ since a normal distribution with infinite variance is not well defined.
The AR(1) process

$$x_t \;=\; x_{t-1} \;+\; u_t \;,$$

with u_t iid white noise, is called a *random walk*. A random walk is an I(1) process; but in the class of I(1) processes is much more general than the class of random walks (why?). Nevertheless, the random walk can be thought of as the archtypical I(1) process. Notice that the random walk is a martingale, since $E_t x_{t+k} \;=\; x_t$ for all k. In words, it is often said that the random walk has infinite memory. In contrast the stable AR(1) model is called mean reverting. Look at $E_0 x_t = x_0 a^t$. It is obvious that if $|a| < 1$ then the process revert very quickly to its mean (normalized to be 0 here).

In general a scalar AR(k) model is called stable if all the roots $r_1, ..., r_k$ of the lag polynomial a(L) are larger than 1 in absolute value. Let $\alpha_i = \frac{1}{r_i}; k = 1, ...k$ then in the stable case we get

$$x_t \;=\; a(L)^{-1}u_t \;=\; (1 - \alpha_k)^{-1}....(1 - \alpha_1)^{-1}u_t \;;$$

which is well defined since we can invert the stable first order lag polynomials one by one.

Notice what happens if $a \rightarrow 1$: the variance tends to infinity and the stationary initial condition can not be defined - at least in typical case as where $u_t \sim N(0, \sigma^2)$ since a normal distribution with infinite variance is not well defined.
The AR(1) process

$$x_t \ = \ x_{t-1} \ + \ u_t \ ,$$

with u_t iid white noise, is called a *random walk*.

Example: If you are given for example an AR(2) process, like

$$x_t \ = \ 1.5x_{t-1} \ + \ x_{t-2} \ + \ u_t \ ,$$

you should be able to tell if the process is stable. In the example we find the roots of the polynomial $1 - 1.5z - x^2$ to be

$$r_i \ = \ .5 * (1.5 \pm \sqrt{2.25 + 4}) \ ,$$

so the roots are 2 and -.5. Since -.5 is less than one in absolute value the process is not stable.

3.6 Stability of VAR-processes

The basic intuition from the scalar AR(1) model carries over to the VAR(1) case. To see this we need the following theorem from linear algebra:
Theorem (The Jordan Form):
Any quadratic matrix A can be written as

$$A \ = \ C^{-1}JC \ ,$$

where

$$J \ = \ \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{pmatrix}$$

with blocks

$$J_i \ = \ \begin{pmatrix} \lambda_i & 1 & \\ & \ddots & 1 \\ & & \lambda_i \end{pmatrix} \ .$$

The Jordan form is a very important tool, in particular in the analysis of systems of linear differential equations, but as you will see also in the analysis of VARs. In the case where all the eigenvalues are different (which you may sometimes get away with assuming), the J matrix will be diagonal. In the case where A is symmetric, J will be real valued and symmetric.
For a proof of the Jordan theorem, see for example Strang (1980).

For the analysis of the VAR(1) we will now assume that the time series has been transformed (by premultiplying by C^{-1}) into diagonal blocks, which now behaves like univariate AR processes and into Jordan blocks. It is enough to look at the simplest type of Jordan block in order to see what goes on. So consider the time series

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} a & 1 \\ & a \end{pmatrix} \begin{pmatrix} x_{1t-1} \\ x_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} ,$$

where the u-vector is white noise. Now if a is numerically less that 1 then x_{2t} will be stable and otherwise not.
Performing the same substitutions as we did for the scalar AR(1) we can write

$$\begin{aligned} x_{1t} &= u_t + au_{t-1} + a^2u_{t-2} + ... \\ &+ x_{2t-1} + ax_{2t-2} + ... \end{aligned}$$

The first term is asymptotically stationary from our previous reasoning and since x_{2t} is asymptotically stationary, the second term is also asymptotically stationary if it converges to a random variable.

The fact that a matrix has the absolute value of all eigenvalues less than 1, can also be expressed as

$$det(A - zI) \neq 0 ; \quad \text{for } |z| \geq 1 ,$$

which of course is the same as

$$det(I - Az) \neq 0 ; \quad \text{for } |z| \leq 1 .$$

The general VAR(k) case poses no new problems (theoretically) as far as the analysis of stability is concerned, since one can reduce the VAR(k) to a (higher dimensional) VAR(1). We will demonstrate how for the scalar AR(k) case. Consider the AR(k) model

$$x_t = a_1x_{t-1} + ... + a_kx_{t-k} + u_t ,$$

where we assume the mean is zero for simplicity. This model corresponds to the model

$$Y_t = CY_{t-1} + U_t ,$$

where

$$\begin{aligned} Y_t &= [x_t, x_{t-1}, ..., x_{t-k}]' , \\ U_t &= [u_t, 0, ..., 0]' , \end{aligned}$$

and

$$C \;=\; \begin{pmatrix} a_1 & a_2 & \dots & a_{k-1} & a_k \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} .$$

The C-matrix is sometimes known as the companion matrix. (This method is also the standard method of treating systems of linear differential equations, wherefore a lot of results related to this area of linear algebra are readily available in mathematical textbooks).
Now it follows from the previous result that Y_t and therefore x_t is stationary if

$$det(I - Cz) \;\neq\; 0 \;;\quad \text{for } |z| \;\leq\; 1 \;.$$

It is an exercise to show that this is equivalent to

$$1 - a_1 z - \;...\; a_k z^k \;\neq\; 0 \;;\quad \text{for } |z| \;\leq\; 1 \;.$$

This is the standard form of the stability condition for higher order AR processes that you will meet in the literature. If you substitute matrices for the the scalar coefficients (i.e. A_1 for a_1 etc.) the derivation carries over to vector AR(k) processes without change. For the VAR(k) process the polynomial

$$det(I - A_1 z - \;...\; A_k z^k) \;,$$

is know as the reverse characteristic polynomial for the AR(k) process (Lütkepohl (1991), p. 12); but I think that the word "reverse" is often left out in the literature. You can look in Lütkepohl's book for more discussion and references although he does not show the actual derivation. It is usually left out of textbooks but is not really hard (only the proof of Jordan's theorem is hard, but we don't need to check that).

3.7 ARMA models.

The ARMA(k,l) process is defined as

$$x_t \;=\; \mu \;+\; A_1 x_{t-1} \;+\; ... \;+\; A_k x_{t-k} \;+\; u_t \;+\; B_1 u_{t-1} \;+\; ... \;+\; B_l u_{t-l} \;,$$

where the innovation u_t is a martingale difference sequence (or white noise). In the case where x_t is a vector process, the term VARMA(k,l) model is also commonly used. In these notes I will usually use the term ARMA to denote both scalar and vector processes, unless I want to stress the vector nature of a statement.

Defining the lag polynomials $A(L) \;=\; -A_1 L \;-\; ... \;-\; A_k L^k$ and $B(L) \;=\; I \;+\; B_1 L \;+\; ... \;+\; B_l L^l$ the defining equation can be conveniently written as

$$A(L) x_t \;=\; \mu \;+\; B(L) u_t \;.$$

We will think of $A(L)x_t$ as the AR part of the model (strictly speaking the left hand side of the equation is not defined on its own, but this should not cause any ambiguity) and $B(L)u_t$ as the MA part of the model. Then we define an ARMA model to be stable if the AR part of the model is stable.

The stable ARMA(k,l) process therefore also has the representation as an infinite MA-model (assuming mean 0):

$$x_t \;=\; A(L)^{-1}B(L)u_t \ .$$

If the characteristic polynomial $det(I \;-\; Bz)$ corresponding to the MA part has all zeroes outside the unit circle then the process is said to be *invertible* and the process has a representation as an infinite AR process:

$$B(L)^{-1}A(L)x_t \;=\; u_t \ .$$

For the purpose of estimating the ARMA process one would always formulate the model in the ARMA(k,l) form, which only contains a finite number of parameters. For theoretical analysis of processes the inversion of the AR part in particular is a useful tool.

The scalar ARMA model

$$(\#) \quad a(L)x_t \;=\; b(L)u_t$$

is not uniquely defined unless one imposes the restriction of no *common factors*. If $a(L) = (1 - \alpha L)a_1(L)$ (say) and $b(L) = (1 - \alpha L)b_1(L)$ then the equation (#) is identical to the model

$$a_1(L)x_t \;=\; b_1(L)u_t \ ,$$

so in the scalar ARMA case we will always (usually implicitly) impose the assumption that the AR and the MA polynomials do not have common factors. (In the statistical analysis you may have to test for common factors).

In the VARMA case this problem also arise, but here the problem is worse. We will not go into this problem here - there is a quite readable discussion in Lütkepohl (1991) and a more theoretical discussion in Hannan and Deistler (1988).

To find the variance of x_t assume that the innovation u_t has variance Ω_u. We first notice that the variance for a pure MA model, is

$$\Omega_x \;=\; B_1\Omega_u B_1' \;+\; ... \;+\; B_l\Omega_u B_l' \ .$$

To find the covariance matrix for the stationary vector AR(1) model, we note that if

$$x_t \;=\; Ax_{t-1} \;+\; u_t \ ,$$

then (since $var(x_t) = var(x_{t-1})$ by stationarity), we get

$$\Omega_x \;=\; A\Omega_x A' \;+\; \Omega_u \ .$$

This can be written in vector form as

$$vec(\Omega_x) \;=\; (A \otimes A)vec(\Omega_x) \;+ vec(\Omega_u) \;,$$

which has solution

$$vec(\Omega_x) \;=\; (I \;-\; A \otimes A)^{-1} vec(\Omega_u) \;.$$

Note that these operations are simple to perform in a matrix oriented software package like GAUSS. For the general ARMA case you can combine the two derivations for the MA and the AR parts. This can be useful for finding the likelihood function, even though I am going to present another method below.

Another compact representation of the VARMA model is

$$X_t' \;=\; (x_t', ..., x_{t-k+1'}, u_t', ..., u_{t-l+1}') \;,$$

and $U_t' \;=\; (U_t^{1'}, U_t^{1'})$ where U_t^1 is a $kl \times 1$ vector defined by

$$U_t^{1'} \;=\; (u_t', 0, ..., 0) \;,$$

then it is an exercise to show that the original ARMA process can be written as

$$X_t \;=\; \mathbf{A} X_{t-1} \;+\; U_t \;,$$

for a suitable matrix **A**. Right now, as I revise this section, I have forgotten what this representation is particularly useful for, but it is a good exercise anyway, since it is important to see that it is easy to reformulate the model like this, and some times it can make all the difference (for the difficulty of showing some result) whether on uses the right (i.e. convenient for the problem at hand) representation.

The **ARIMA** model is defined as a model which follows and ARMA model after the data has been differenced (one or sometimes more times).

3.8 Box-Jenkins methods and model-identification.

ARIMA modeling is a term that is primarily used to denote the approach to modeling in the widely read book by Box and Jenkins (1976). Box and Jenkins suggested the use of the empirical autocorrelations as a tool for identifying the order of the model, i.e. k and l in the representation of the ARMA model above, and also for identifying the order of integration.
Box and Jenkins concentrated on the univariate case so we will do the same.
The empirical autocorrelations defined as $r_k = \frac{c_k}{c_0}$, where the c_k are the empirical autocovariances

$$r_k \;=\; \frac{1}{n-k} \sum_{i=k}^{n} (x_t - \hat{\mu})(x_{t-k} - \hat{\mu}) \;;\; k = 0, 1, .. \;.$$

(Some time the empirical autocorrelation is called the sample autocorrelation function and abbreviated SACF). A plot of r_k against k is called the covariogram. We will sketch the theoretical

covariogram for a few simple processes. [Look at graphs]. If one has a lot of data then the sample autocorrelation will be close to the true autocorrelation, and since the true autocorrelation function $\rho(k) = 0$ when $k > l$ in the case where the true process is an MA(l) process one may hope to be able to identify an MA(l) model correctly from the data with the help of the sample ACF. For identifying AR model the so-called partial autocorrelation function (PACF) is more suited. The partial autocorrelation function is defined (for k = 1, 2, ..) as the coefficient ϕ_k to x_{t-k} in the regression

$$x_t = \mu \ + \ \alpha_1 x_{t-1} \ + \ .... \ + \ \alpha_{k-1} x_{t-k+1} \ + \ \phi_k x_{t-k} \ \ + \ u_t \ .$$

This autoregression is not necessarily the true model, but rather the k-th order autoregression that minimizes the variance of the error-term. This is obviously a 2nd order concept and you can therefore assume normality (in order to find ϕ_k). In the normal case ϕ_k is the correlation between x_t and x_{t-k} conditional on $x_{t-1}, ..., x_{t-k+1}$. It is obvious that the PACF may be useful for identifying AR models. Maybe one can also identify low order ARMA models from the sample ACF and PACFs, but it is not something that I recommend.
You may ask why one does not instead select a suitably (whatever that is) high order for the AR and MA parts of the model and then estimate from there. Box and Jenkins argue that such a strategy often will lead to models that contain a large number of parameters. In practice one is likely to achieve more robust results with a low number of parameters. This is called the *principle of parsimony.*

The Box-Jenkins method was very popular in th 70ies, and the method was instrumental in sparking the incredible interest in testing for unit roots etc. that is such a hot topic these days. In my opinion (shared with most econometricians) the Box-Jenkins methods are more useful for areas like engineering where large data sets are available. Also in many fields, one can select a model and then replace it if it does not perform well in practice. In economics we are often interested in testing theories and we want more objective procedures than eye balling techniques, since testing on new datasets is usually not an option. Box-Jenkins methods - in spite of their popularity - never became mainstream econometrics, even among model builders (as opposed to theory testers). The standard view among model builders was that "time series methods need much larger data sets than was available to economists". I agree with this statement when "time series methods" are interpreted to mean Box-Jenkins methods; but I think many made the mistake of making that identification and then, having discarded "time series methods", went on to ignore the fact that they were actually working with time series data (for example read the preface to Granger and Newbold (1986)).

One improvement of the Box Jenkins strategy that has been made suggested is the use of criteria that trades of parsimony against residual variance.
The most well-known being the Akaike AIC criterion that minimizes

$$log(\sigma^2) + \frac{2K}{n}$$

where K is the number of freely estimated parameters and σ is the estimated variance of u_t. One can show, however, that the AIC criterion tends to overestimate the order of the models in large samples and recently other criteria has been suggested that are better behaved in large samples (read for example Lütkepohl (1991) or Granger and Newbold (1986)). It seems that this type of order selection methods is gaining popularity (they have the huge advantage that they are automatic, which limits the scope for data mining) and if we had more time we would give them more consideration (this is a possible subject for a student presentation).

3.9 Estimation of ARMA models.

We will first consider estimation of the scalar AR(k) model:

$$x_t \;=\; \mu \;+\; a_1 x_{t-1} \;+\; ... \;+\; a_k x_{t-k} \;+\; u_t \;.$$

The mean μ will always be estimated by empirical mean of the process so let us disregard the mean for simplicity.
The 2 standard methods of estimating the AR model is by ordinary least squares taking the k initial values of the process to be fixed, and by ML methods assuming that the initial values follow the stationary distribution. Asymptotically it does not matter, and usually it is more meaningful in economics to condition on the initial values. In the case of fixed initial values the LS estimator turn out to be identical to the ML estimator. (Note: I will often say ML estimation without mentioning with probability distribution I have in mind - in this case it will always be implicit that I mean the normal likelihood. In the case where the data do actually follow a normal distribution, the term quasi maximum likelihood is often used for the estimator that maximizes the normal likelihood).
In the general VAR case

$$x_t \;=\; \mu \;+\; A_1 x_{t-1} \;+\; ... \;+\; A_k x_{t-k} \;+\; u_t \;,$$

it turns out (for fixed initial values) that again the maximum likelihood corresponds to the least squares estimator. Let

$$B \;=\; (\mu, A_1, ..., A_k) \;,$$

and let

$$Z_t \;=\; \begin{pmatrix} 1 \\ x_t \\ \vdots \\ x_{t-k+1} \end{pmatrix} \;.$$

Now let

$$Z \;=\; (Z_k, \;...\;, Z_{T-1}) \;,$$

and

$$X \;=\; (x_k, \;...\;, x_{T-1}) \;,$$

where T is sample size. Then the least squares estimator of B is

$$\hat{B} \;=\; XZ'(ZZ')^{-1} \;,$$

independently of the variance-covariance of u_t (compare this to the result for "seemingly unrelated regressions").
Estimation of the univariate AR model is covered in all introductory time series texts, whereas the derivation in the multivariate case can be found in Lütkepohl (1991).

Let us now consider the scalar MA process.

$$x_t \;=\; \mu \;+\; u_t \;+\; b_1 u_{t-1} \;+\; ... \;+\; b_l u_{t-l} \;,$$

If you assume that the initial values $u_0, u_{-1}, ..., u_{-l}$ are all zero then we have

$$u_1 \;=\; x_1 \;-\; \mu$$

$$u_2 \;=\; x_2 \;-\; \mu \;-\; b_1 u_1$$

and in general

$$u_t \;=\; x_t \;-\; \mu \;-\; b_1 u_{t-1} \;...\; - b_l u_{t-l} \;.$$

In order to use the above equations for estimation one has to calculate u_1 first and then u_2 etc. recursively.
Now the u_t terms has been found as functions of the parameters and the observed variables x_t. These equations are very convenient to use for estimation since the u_ts are identically independently distributed (under stationarity), so that the likelihood function $\mathcal{L}_u$ in terms of the u_t has the simple form

$$\mathcal{L}_u(u_1, ..., u_T; \psi) \;=\; \Pi_{t=1}^{T} \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-u_t^2}{2\sigma^2}} \;,$$

where ψ is the vector of parameters of the model. Now, unfortunately it is not the u_t's that we observe; but rather the x_t vector. The equations above however gives u_t as a function of the x_ts so the likelihood function $\mathcal{L}_x(x_1, ..., x_t; \psi)$ (where $\mathbf{b}$ is the vector of parameters of the MA-model) is just

$$\mathcal{L}_x(x_1, ..., x_T \,; \psi) \;=\; \mathcal{L}_u(u_1(x_1), ..., u_T(x_1, ..., x_T); \psi) \Pi_{t=1}^{T} \tfrac{1}{\sqrt{2\pi\sigma^2}} \; e^{\frac{-u_t(x_1,...,x_t)^2}{2\sigma^2}} \;.$$

Be aware that most of the parameters of the likelihood function in this notation are implicit in the mapping from x_ts to u_ts. Note that in general, you have to be careful when making this kind of substitutions in likelihood functions. The rule for changing the variable of the likelihood function through a transformation is that if

$$\mathbf{y} \;=\; f(\mathbf{x}) \;,$$

where $\mathbf{x}$ and $\mathbf{y}$ are both T-dimensional vectors, and f is a one-to-one mapping, that often will depend on parameters, of R^T onto R^T (or relevant subsets), then

$$\mathcal{L}_y(y_1, ..., y_T) \;=\; \mathcal{L}_x(f^{-1}(y_1, ..., y_T))|Df^{-1}(y)| \;=\; \mathcal{L}_x(f^{-1}(y_1, ..., y_T)) \frac{1}{|Df(f^{-1}y)|} \;.$$

The last two forms are equivalent; but the last mentioned is often the most convenient form. The matrix Df with i,jth element $Df_{ij} = \frac{\partial f_i}{\partial x_j}$ is known as the Jacobian matrix of the mapping (or transformation). In the application to the MA-process you can check that **u** as a function of **x** has unit Jacobian (so that the Jacobi-determinant is unity). You should also be aware that if the Jacobi-determinant is a function of the observations but not of the parameters, then it can be ignored for the purpose of maximizing the likelihood function, and this is often done without comment in the literature.

The strategy of assuming the initial values of the innovation to be zero will not have any influence in large samples; but it may not be advisable in small datasets. It is possible to find the exact likelihood function, see Granger and Newbold (1986) p 91 ff., but this is quite messy in general. The reason that it is complicated to estimate the MA model (and therefore also the ARMA models) is the the parameterizations involves terms that are unobserved (namely the u_ts). It turns out that one can estimate the model by a very general algorithm, called the Kalman Filter, that is incredibly useful - in particular for estimating models with unobserved components. This covers many more models than the ARMA models, for example the Kalman filter has long been used to estimate models with missing observations, but there are lots of other applications. One area where the Kalman filter has recently been successfully introduced is to estimate continuous time stochastic differential equations (here it is the differential that is unobserved), see Harvey (1989) for an introduction. We will show how to use the Kalman filter to estimate ARMA model; but keep it in mind for other potential applications.

3.10 The Kalman filter.

The Kalman filter is treated many places. There is an easy introduction in Harvey (1981), but there is also another book by Harvey (1989) that gives a systematic coverage of the Kalman filter and many of its applications.

We assume that we have a model that concerns a series of (vectors of) variables α_t, which are called "state vectors". These variables are supposed to describe the current state of the system in question. These state variables will typically not be observed and the other main ingredient is therefore the observed variables y_t. The first step is to write the model in **state space form** which is in a form of a linear system which consists of 2 sets of linear equations. The first set of equations describes the evolution of the system and is called the "Transition Equation":

$$\alpha_t = K\alpha_{t-1} + R\eta_t \ ,$$

where K and R are matrices of constants and η is $N(0,Q)$. The second set of equations describes the relation between the state of the system and the observations and is called the "Measurement Equation":

$$y_t = Z\alpha_t + \xi_t \ ,$$

where ξ_t is $N(0, H)$ and $E(\xi_t \eta_{t-j}) = 0$ for all t and j.
It turns out that a lot of models can be put in the state space form with a little imagination. The main restriction is of course on the linearity of the model. The state-space model as it is defined here is not the most general possible - it is in principle easy to allow for non-stationary coefficient matrices, see for example Harvey(1989). There is also extension of Kalman filter methods to non-linear models. These are known as extended Kalman filters and they are also treated in Harvey (1989) - be aware, however, that extended Kalman filters usually can not be used to evaluate likelihood functions exactly; but only gives an approximation.
As an example, let us look at a model where the economy consists of two sectors producing a homogeneous product, where we only observe the aggregate output subject to measurement error. Assume that the output of the individual sectors follow a scalar VAR(1) model. Then the state-space system becomes as follows. Transition equation:

$$\begin{pmatrix} \alpha_t^1 \\ \alpha_t^2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \alpha_{t-1}^1 \\ \alpha_{t-1}^2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_t^1 \\ \eta_t^2 \end{pmatrix} ,$$

and measurement equation:

$$y_t = (1 \;\; 1) \begin{pmatrix} \alpha_t^1 \\ \alpha_t^2 \end{pmatrix} + \xi_t ,$$

In the case of a general ARMA process, one can use several representations; but the most compact (and useful) one is the following. Assume that we have given the scalar ARMA process (where I leave out the mean for simplicity):

$$x_t = a_1 X_{t-1} + ... + a_k X_{t-m} + u_t + b_1 u_{t-1} + ... + b_l u_{t-l} ,$$

where $m = max\{k, l+1\}$. This process can be represented in the following state space form: Transition equation

$$\alpha_t = \begin{pmatrix} a_1 & 1 & 0 & \dots & 0 \\ a_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m-1} & 0 & 0 & \dots & 1 \\ a_m & 0 & 0 & \dots & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_{m-2} \\ b_{m-1} \end{pmatrix} u_t ,$$

and measurement equation

$$x_t = (1, 0, \dots, 0)\, \alpha_t .$$

Example. The MA(1) model has the state-space representation

$$\alpha_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 \\ b_1 \end{pmatrix} u_t .$$

If $\alpha_t = (\alpha_{1t}, \alpha_{2t})'$, then $\alpha_{2t} = b_1 u_t$ and $\alpha_{1t} = \alpha_{2,t-1} + u_t = u_t + b_1 u_{t-1}$.
It is an exercise to show that the general ARMA-case does indeed have the state-space representation given above.

The Kalman filter is very useful when you want to calculate the likelihood function. You will typically have a general maximization algorithm at your disposal (e.g. the MAXMUM algorithm in GAUSS). Such an algorithm takes as input a subroutine that evaluates the value of the likelihood function for a given set of parameters. This is where you will use the Kalman filter algorithm.
For a start look at the general likelihood function:

$$f(y_T, ..., y_1; \theta) \ ,$$

where θ is a vector of parameters. One can always factor such a likelihood function (independent of which model or distribution the that generates the variables), as

$$f(y_T, ..., y_1; \theta) \ = \ f(y_T | y_{T-1}, ..., y_1; \theta) f(y_{T-1}, ..., y_1; \theta) \ .$$

Iterating this formula we find

$$f(y_T, ..., y_1; \theta) \ = \ \Pi_{t=p+1}^{T} f(y_t | y_{t-1}, ..., y_1; \theta) f(y_p, ..., y_1; \theta) \ ,$$

From which we find the log-likelihood function in *recursive form* as

$$L \ = \ lnf \ = \ \sum_{t=p+1}^{T} lnf(y_t | y_{t-1}, ..., y_1; \theta) \ + \ lnf(y_p, ..., y_1; \theta) \ .$$

In the case of the normal likelihood function this becomes

$$L \ = \ \sum_{t=p+1}^{T} -\frac{1}{2} ln|F_t| - \frac{1}{2} \nu_t' F_t^{-1} \nu_t \ + \ constant \ + \ lnf(y_p, ..., y_1; \theta) \ ,$$

where

$$\nu_t \ = \ y_t - E(y_t | y_{t-1}, ..., y_1) \text{ and } F_t \ = \ E[\nu_t \nu_t' | y_{t-1}, ..., y_1] \ .$$

Now if we knew F_t and ν_t we would have a quite convenient way of evaluating the likelihood function. This is what the Kalman filter equations below are designed to do.
At this stage note the following aside. The likelihood equations in recursive form allows you to evaluate the "impact" of a new observation arriving, in the sense that it immediately shows the conditional likelihood. In engineering it is often important to be able to *update* a parameter estimate instantly when a new observation occurs - and hopefully without having to reestimate using all the data. The Kalman Filter does exactly that and it is therefore used extensively by engineers. More surprising is the fact that it at the same time is so convenient to use that it is also a good choice to use for the purpose of a single estimation on a given data set.

The ingredients of the Kalman filter (besides the state-space representation) consist of *predicting equations* and *updating equations*.

For any vector x_t define $x_{t|t-1} = E(x_t|y_{t-1}, ..., y_1)$, where y_j are the observed variables. This definition gives the best guess of x_t based on all the information available at time $t-1$, $x_{t|t-1}$ is the prediction of x_t at $t-1$. As you may guess, the Kalman filter evolves around predicting and updating the prediction of the state vector. Also define $P_{t|t-1} = E\{(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'\}$ - $P_{t|t-1}$ is the conditional variance of the prediction error.

The prediction equations are then

$$\begin{aligned} \alpha_{t|t-1} &= K\alpha_{t-1|t-1} \\ y_{t|t-1} &= Z\alpha_{t|t-1} \\ P_{t|t-1} &= KP_{t-1|t-1}K' + RQR' . \end{aligned}$$

Define

$$\nu_t = y_t - y_{t|t-1} .$$

We will need the variance matrix

$$F_t = E\{\nu_t\nu_t'\} = E\{(y_t - y_{t|t-1})(y_t - y_{t|t-1})'\} .$$

To finish the Kalman filter we finally need the updating equations:

$$\begin{aligned} \alpha_{t|t} &= \alpha_{t|t-1} + P_{t|t-1}Z'F_t^{-1}\nu_t \\ P_{t|t} &= P_{t|t-1} - P_{t|t-1}Z'F_t^{-1}ZP_{t|t-1} , \end{aligned}$$

where

$$F_t = ZP_{t|t-1}Z' + H .$$

The term

$$P_{t|t-1}Z'F_t^{-1}\nu_t ,$$

is called the *Kalman gain*. Any new information enters the system through the Kalman gain.

The Kalman filter can be derived from the rules of the Normal distribution.
We can write

$$\begin{aligned} \nu_t &= Z(\alpha_t - \alpha_{t|t-1}) + \xi_t \\ \alpha_t &= \alpha_{t|t-1} + (\alpha_t - \alpha_{t|t-1}) , \end{aligned}$$

One can show that

$$\left(\begin{array}{c} \nu_t \\ \alpha_t \end{array} \right) \Bigg|_{y_{t-1},...,y_1} = \mathrm{N}\left(\left(\begin{array}{c} 0 \\ \alpha_{t|t-1} \end{array} \right), \left[\begin{array}{cc} F_t & ZP_{t|t-1} \\ P_{t|t-1}Z' & P_{t|t-1} \end{array} \right] \right)$$

Recall the following rule for the conditional Normal distribution (or see e.g. Lütkepohl (1991), pp. 480-81). If

$$\left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) \sim \mathrm{N}\left(\left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right), \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right] \right) ,$$

then the conditional distribution of x_1 given x_2 is

$$N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]) \ .$$

Using this rule on the conditional distribution of (ν_t, α_t) we find

$$\mathcal{L}(\alpha_t|y_t, ..., y_1) \ = \ \mathcal{L}(\alpha_t|\nu_t, y_{t-1}, ..., y_1) \ = \ N(\alpha_{t|t-1} + Z'P_{t|t-1}F_t^{-1}\nu_t, \ P_{t|t-1} - P'_{t|t-1}ZF_t^{-1}Z'P_{t|t-1}) \ ,$$

which is the updating equation.
One problem is how to initialize the filter. It is natural to choose $\alpha_{t|t-1} = 0$ since this is the unconditional mean of α_t. It is often also natural to choose the stationary value of the variance as the initial value for the variance, even though this is only an option in the stable case). Other choices are possible, for example you may want to condition on initial values as discussed earlier; but in that case the special cases has to be considered one by one. One can, however, find the stationary variance by the following method.
Combining the updating and the prediction equations we find

$$P_{t+1|t} = KP_{t|t-1}K' \ - \ KP_{t|t-1}Z'(ZP_{t|t-1}Z' \ + \ H \ )^{-1}ZP_{t|t-1}K' \ + \ RQR' \ ,$$

which is known as the Riccatti equation. If the model is stable $P_{t|t-1}$ will converge to the solution $\bar{P}$ of the *algebraic Ricatti equation*

$$\bar{P} = K\bar{P}K' \ - \ K\bar{P}Z'(Z\bar{P}Z' \ + \ H \ )^{-1}Z\bar{P}K' \ + \ RQR' \ .$$

In order to apply the Kalman filter one has to choose a set of starting values. The most natural choice for a stable system is the unconditional mean and variance. Since

$$\alpha_t \ = \ K\alpha_{t-1} \ + \ R\eta_t \ ,$$

has the form of an AR(1) model, we will then choose $\alpha_{1|0} = 0$ and P_0 such that

$$vec(P_0) \ = \ (I \ - \ K \otimes K)^{-1}vec(RQR') \ .$$

If you want, you can choose other initial conditions, for example chosen from a Bayesian prior, or if you want to condition on initial values. In the non-stationary case it is obviously not possible to choose the initial distribution from the stationary distribution.
Example: The state-space representation for an AR(2) model is

$$\begin{pmatrix} x_t \\ a_2x_{t-1} \end{pmatrix} \ = \ \begin{pmatrix} a_1 & 1 \\ a_2 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ a_2x_{t-2} \end{pmatrix} \ + \ \begin{pmatrix} 1 \\ 0 \end{pmatrix} u_t \ ,$$

so here conditioning on inital observations just corresponds to an initial variance of zero; but in the ARMA case one has to be a bit more sophisticated unless one want to condition on initial values

of the innovation terms to be zero. This is not always advisable as discussed previously.

Now in order to use the Kalman filter you will need a general optimization routine as OPTMUM in GAUSS. Such a routine need a subroutine that returns the criterion function and the optimization algorithm will then use some specified algorithm to minimize (or maximize) the value of the criterion function. Let me sketch how a subroutine using the Kalman filter for evaluating a Normal likelihood function would look. Assume the subroutine evaluating the likelihood function is called CRITFUNC. If we evaluate the value of the likelihood function for a scalar ARMA model if will have the parameters $a_1, ...a_k, b_1, ..., b_l, \sigma$ of the ARMA model as arguments. The structure of a GAUSS subroutine would be something like this:

$CRITFUNC(a_1, ...a_k, b_1, ..., b_l, \sigma)$;
@ First create the matrices that is used in state-space form @
$K[1,1] = a_1$;
$K[1,2] = a_2$;
etc.
@ Initialize @
$L = 0$ @ L is the value of the likelihood function @
$t = 0$
@Initialize the Kalman Filter.@
P_0 = ..
etc.
@ Loop @
DO UNTIL t == (T-1)

Prediction Equations for time t

Evaluate the condition likelihood = $L(y_t|y_{t-1}, ..., y_1)$
$L = L + L(y_t|y_{t-1}, ..., y_1)$

Updating equations for t ;
ENDO ;
@return the value of the likelihood function for all points@
RETURN(L);
END OF CRITFUN

4 The Power Spectrum

Definition The *power spectrum* (or spectral density) of a stationary vector process x_t is defined as the function

$$F(\lambda) \;=\; \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\Gamma(k)e^{i\lambda k}\;,$$

where i is the complex number i. Note that the term spectral density is sometimes used for the power spectrum divided by the variance of the process (Harvey's book is one place). It may be more natural to call this a density since the spectral density then integrates to 1. Of course it does not matter; but you should be aware of the variation in usage - which definition an author uses will be clear from the context - for example Phillips (1991) uses the term spectral density in the sense that I defined it.

In the case of a scalar process, where the covariance function is an even function this reduces to

$$f(\lambda) \;=\; \frac{1}{2\pi}[\gamma(0) \;+\; 2\sum_{k=1}^{\infty}\gamma(k)cos(\lambda k)]\;.$$

One can show that

$$\int_{-\pi}^{\pi} f(\lambda)d\lambda \;=\; \gamma(0)\;,$$

so the power spectrum can be interpreted as a decomposition of the variance into the variance at different frequencies. In economics it is more natural to think of time series in the *time domain* than in term of frequencies (*the frequency domain*), so to an economist it may sound slightly surprising that estimation in the frequency domain was well developed before estimation in the time domain (see e.g. Priestley (1981)). It is still a quite efficient strategy to estimate time series models in the frequency domain; but we will not go further into the subject at this stage.
The most important reason for bringing up the subject here is that you will meet references to spectra spread over the time series literature. Here is one connection where you will often meet it: If x_t is a covariance stationary process with power spectrum $F(\lambda)$, then

$$Var\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_t\} \;\rightarrow 2\pi f(0)\;,$$

for $T \rightarrow \infty$. Check this by using the definition of the spectral density. This limit shows up in the theory of I(1) processes for the following reason. If y_t is a an I(1) process with

$$\Delta y_t \;=\; x_t\;;y_0 \;=\; c\;,$$

where c is a constant, then $lim_{T\rightarrow\infty}var(\frac{1}{\sqrt{T}}y_T)$ converges to 2π times the spectral density of x_t at frequency zero. You will meet that exact use of words in the literature.

4.1 Estimation of the spectrum.

We will need the estimate of the spectrum at frequency zero.
If we estimate the auto covariances, given a sample of size T, by

$$\frac{\sum_{t=k}^{T}[(x_t-\mu)(x_{t-k}-\mu)']}{T-k} \;=\; C(k)\;;\; k=..,0,1,2,...\;,$$

(and for negative k by $C(k)=C(-k)'$), we can estimate the power spectrum by

$$\hat{F}(\lambda) \;=\; \frac{1}{2\pi}\sum_{k=-T}^{T} C(k)e^{i\lambda k}\;.$$

Take the mean of that expression and it is easy to see that this is an unbiased estimator of the theoretical spectrum. However, since we are summing over T terms, we are likely to have a lot of variance even if each element in the sum is consistent, and indeed it turns out that the spectral estimator is not even consistent.
Instead one usually apply an estimator of the spectrum that has the form

$$\hat{F}(\lambda) \;=\; \frac{1}{2\pi}\sum_{k=-T}^{T} w(\frac{k}{M(T)})\; C(k)e^{i\lambda k}\;,$$

where the function $w(.)$ is called a weighting function, and $M(T)$ is called a bandwidth parameter. $M(T)$ is assumed to converge to infinity but at a slower rate than T. A typical form for the weights are

$$w_T(k) \;=\; 1-\frac{k}{M(T)}\;,$$

for $k<M(T)$ and 0 otherwise. This is a kernel suggested by Newey and West (1987); but that has long been used in the time series literature under the name of a Bartlett kernel (and also under other names). (One can show that this weighting corresponds to smoothing the graph of the spectrum in the way that is taught in the Nonparametrics course. This is also the origin of the terminology "kernel"). In connection with GMM estimation we will return to this subject.

4.2 z-transforms and spectra

It turns out that the z-transforms are useful for finding spectra. Let us define the spectrum as

$$F(z) \;=\; \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \Gamma(k)z^k\;,$$

for $z=e^{i\lambda}$.
Now one can show that if u_t has spectrum $F_u(z)$ (if u_t is white noise it will just be constant), then if $A(L)x_t=B(L)u_t$ then x will have the spectrum

$$F_x(z) \;=\; A(z)^{-1}B(z)F_u(z)B'(\bar{z})A'(\bar{z})^{-1}$$

for $z = e^{i\lambda}$. In the scalar case this reduces to

$$f_x(z) \;=\; \frac{|b(z)|^2}{|a(z)|^2} f_u(z) \;.$$

This is sometimes taken as the definition of the spectrum of an ARMA process. Now if you consider the random walk this will give a value for the spectrum at frequency 0 of infinity. You will sometimes meet that statement in seminars or in articles. From the "basic" definition of the spectrum this is of course meaningless since the spectrum is not defined for processes like the random walk that does not have finite variance. (Be aware, however, that the so-called fractionally integrated processes have finite variance but a value of infinity of the spectrum at frequency zero).

4.3 Periodicities and spectra

Given what we have already covered it is easy to demonstrate one of the main applications of spectral theory - namely searching for periodicities. If we start with an extreme example you will get the idea.
Assume that x_t is periodic with period 4 (this is also a unit root model), so that

$$x_t \;=\; x_{t-4} \;+\; u_t \;,$$

where u_t is white noise. Then the power spectrum for u_t is $\frac{\sigma_u^2}{2\pi}$, and it then follows from the rule above that

$$f_x(\lambda) \;=\; \frac{\sigma_u^2}{2\pi|1 - e^{i4\lambda}|^2} \;,$$

which is easily found to be

$$f_x(\lambda) \;=\; \frac{\sigma_u^2}{2\pi(2 - 2cos(4\lambda))} \;.$$

Now it is easy to see that the spectrum is infinite at $\lambda = \frac{\pi}{2}, \pi, ...$ This is an extreme case but one can convince oneself that if

$$x_t \;=\; ax_{t-4} \;+\; u_t \;.$$

where a is close to one, then the spectrum will have a spike a the "quarterly frequencies", and that this will still be the case even if u_t itself is not white noise.

Look at some spectral densities e.g. in Granger and Newbold's book. Granger showed long time ago in the sixties that almost all economic time series have the same shape of the empirical spectrum, namely one that tends to infinity (or at least is very large) for $\lambda \to 0$ and then decreases rapidly (sort of looking like an irregular version of the graph of e^{-x}). Granger called this *the typical spectral shape* of economic time series.

5 Least Squares Estimation of VAR Models

Consider the VAR(k) model

$$x_t \;=\; \mu \;+\; A_1 x_{t-1} \;+\; ... \;+\; A_k x_{t-k} \;+\; u_t \;,$$

where x_t is an L dimensional vector and u_t are Nid with variance Σ. Let

$$B \;=\; (\mu, A_1, ..., A_k) \;,$$

and let

$$Z_t \;=\; \begin{pmatrix} 1 \\ x_t \\ \vdots \\ x_{t-k+1} \end{pmatrix} .$$

Now let

$$Z \;=\; (Z_k, \; ... \; , Z_{T-1}) \;,$$

and

$$X \;=\; (x_{k+1}, \; ... \; , x_T) \;,$$

and

$$U \;=\; (u_{k+1}, \; ... \; , u_T) \;,$$

where T is sample size. Then we can write the VAR compactly as

$$X = BZ + U \;. \tag{1}$$

The GLS estimator of B is

$$\hat{B} \;=\; XZ'(ZZ')^{-1} \;.$$

independently of the variance-covariance, Σ_u, of u_t. Note that this implies that the coefficients for x_{it}, i.e. the ith row, B_i, of B – is estimated by

$$B_i \;=\; (ZZ')^{-1} Z X_i \;,$$

where X_i is the ith row of X, which is the values $x_{it}, x_{it-1}, ..., x_{ik+1}$. This means that the coefficients on the ith equation in the VAR, when we consider each row of the VAR system as an equation, can be estimated individually by OLS and one will get the same result as using system GLS. This is a special case of a very well known theorem for "seemingly unrelated regressions (SURE)".

I believe most of you have not seen the SURE result derived and I will therefore do it here. Rewrite (**??**) as

$$vec(X) \;=\; vec(BZ) + vec(U) \;,$$

or

$$vec(X) \;=\; (Z' \otimes I)vec(B) + vec(U) \;.$$

The equations are now written as a standard linear equation system. The variance-covariance matrix for $vec(U)$ is $I \otimes \Sigma$. The ML estimator of $vec(B)$ – conditional on the initial observations – is the GLS estimator

$$\begin{aligned}
vec(\hat{B}) &= [(Z \otimes I)(I \otimes \Sigma)^{-1}(Z' \otimes I)]^{-1}(Z \otimes I)(I \otimes \Sigma)^{-1}vec(X) \\
&= [(Z \otimes I)(I \otimes \Sigma^{-1})(Z' \otimes I)]^{-1}(Z \otimes I)(I \otimes \Sigma^{-1})vec(X) \\
&= [(ZZ' \otimes \Sigma^{-1})]^{-1}(Z \otimes \Sigma^{-1})vec(X) \\
&= (ZZ')^{-1} \otimes \Sigma)(Z \otimes \Sigma^{-1})vec(X) \\
&= (ZZ')^{-1}Z \otimes I)vec(X) \\
&= vec(X(ZZ')^{-1}Z) \;.
\end{aligned}$$

The estimated variance matrix $\hat{\Sigma}$ for U is $\frac{1}{T}\Sigma_{t=K+1}^{T} e_t e_t'$ where $e_t = X - \hat{B}Z$ (i.e. e_t is just the residuals defined in the usual fashion), and the estimated variance of $\hat{B}$ is then $(ZZ')^{-1} \otimes \hat{\Sigma}$.

6 Granger Causality.

6.1 Linear Prediction.

Assume that you want to predict the value of y_{t+k} based on the information set $\mathcal{F}_t$. How do you do that in the best possible way? This depends on your cost of making a wrong prediction, so if you have a formal model for your cost of making an error of a given size, then you should minimize that function (in statistics this is usually called a *loss function*). In econometrics it is usual to choose to minimize the mean square error (MSE) of the forecast, i.e.

$$\min E\{(y_{t+k} - \hat{y}_{t+k})^2\}$$

where $\hat{y}_{t+k}$ is the predictor of y_{t+k}. One can show that the conditional mean $E\{y_{t+k}|\mathcal{F}_t\}$ is the best mean square predictor. If the information set $\mathcal{F}_t$ consists of a vector of observations z_t (which would usually include $y_t, y_{t-1}, ..., y_1$), then the conditional mean in the case of normally distributed variables is linear (as we know). In the case where the observations are not normally distributed the conditional mean is not a linear function of the conditioning varibles, so if you can find the true conditional mean you may want to do that, however, timeseries analysis is, as mentioned, mostly in the 2nd order tradition, so often people use the *best linear predictor* rather than the conditional mean. You find the best linear predictor as that linear function of the conditioning variables that would give you the conditional mean if the data had been normally distributed.

Assume that your data are described by a VAR(2) model:

$$y_t \;=\; \mu \;+\; A_1 y_{t-1} \;+\; A_2 y_{t-2} \;+\; u_t \;.$$

What would be the best (linear) forecast of y_{t+1} based on $y_1, ..., y_t$? Obviously,

$$\hat{y}_{t+1} \;=\; \mu \;+\; A_1 y_t \;+\; A_2 y_{t-1}.$$

It turns out that we can iterate this formula to find

$$\hat{y}_{t+k} \;=\; \mu \;+\; A_1 \hat{y}_{t+k-1} \;+\; A_2 \hat{y}_{t+k-2}.$$

for any k. Another approach would be to reformulate the model as a higher dimensional VAR(1) system, since it is easy to see that

$$\hat{y}_{t+k} \;=\; (I + A + ... + A^{k-1})\mu \;+\; A^k y_t \;,$$

in this case. (Note that the best linear predictor in the stable case converges (for $k \to \infty$) to the unconditional mean of the process).

For models with MA components things are harder. Recall that one can write the ARMA model as a high order VAR(1) (the state-space representation), so one can use the formula above, but the complication is that even at time t one does not know u_t. The Kalman filter does however,

as a byproduct, give you the best guess of $u_t, u_{t-1}, ...,$ (namely as part of $\alpha_{t|t}$), so you can use the Kalman filter to generate $\alpha_{t|t}$ and then you can use the formula above. For more elaborations, see Harvey (1989).

6.2 Granger Causality.

Assume that the information set $\mathcal{F}_t$ has the form $(x_t, z_t, x_{t-1}, z_{t-1}, ..., x_1, z_1)$, where x_t and z_t are vectors (that includes scalars of course) and z_t usually will include y_t and z_t may or may not include other variables than y_t.

Definition: We say that x_t is Granger causal for y_t wrt. $\mathcal{F}_t$ if the variance of the optimal linear predictor of y_{t+h} based on $\mathcal{F}_t$ has smaller variance than the optimal linear predictor of y_{t+h} based on $z_t, z_{t-1}, ...$ - for any h. In other word x_t is Granger causal for y_t if x_t helps predict y_t at some stage in the future.

Often you will have that x_t Granger causes y_t and y_t Granger causes x_t. In this case we talk about a *feedback system*. Most economists will interpret a feedback system as simply showing that the variables are related (or rather they do not interpret the feedback system).

Sometimes econometrians use the shorter terms "causes" as shorthand for "Granger causes". You should notice, however, that Granger causality is not causality in a deep sense of the word. It just talk about linear prediction, and it only has "teeth" if one thing happens before another. (In other words if we only find Granger causality in one direction). In economics you may often have that all variables in the economy reacts to some unmodeled factor (the Gulf war) and if the response of x_t and y_t is staggered in time you will see Granger causality even though the real causality is different. There is nothing we can do about that (unless you can experiment with the economy) - Granger causality measures whether one thing happens before another thing and helps predict it - and nothing else. Of course we all secretly hope that it partly catches some "real" causality in the process. In any event, you should try and use the full term Granger causality if if is not obvious what you are refering to

The definition of Granger causality did not mention anything about possible instantaneous correlation between x_t and y_t. If the innovation to y_t and the innovation to x_t are correlated we say there is *instantaneous causality*. You will usually (or at least often) find instantaneous correlation between two time series, but since the causality (in the "real" sense) can go either way, one usually does not test for instantaneous correlation. However, if you do find Granger causality in only one direction you may feel that the case for "real" causality is stronger if there is no instantaneous causality, because then the innovations to each series can be thought of as actually being generated from this particular series rather than part of some vector innovations to the vector system. Of

course, if your data is ampled with a long sampling period, for example annually, then you would have to explain why one variable would only cause the other after such a long lag (you may have a story for that or you may not, depending on your application).

Granger causality is particularly easy to deal with in VAR models. Assume that our data can be described by the model

$$\begin{bmatrix} y_t \\ z_t \\ x_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} A_{11}^1 & A_{12}^1 & A_{13}^1 \\ A_{21}^1 & A_{22}^1 & A_{23}^1 \\ A_{31}^1 & A_{32}^1 & A_{33}^1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \\ x_{t-1} \end{bmatrix} + ... + \begin{bmatrix} A_{11}^k & A_{12}^k & A_{13}^k \\ A_{21}^k & A_{22}^k & A_{23}^k \\ A_{31}^k & A_{32}^k & A_{33}^k \end{bmatrix} \begin{bmatrix} y_{t-k} \\ z_{t-k} \\ x_{t-k} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}$$

Also assume that

$$\Sigma_u = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12}' & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13}' & \Sigma_{23}' & \Sigma_{33} \end{bmatrix}.$$

This model is a totally general VAR-model - only the data vectors has been partitioned in 3 sub-vectors - the y_t and the x_t vectors between which we will test for causality and the z_t vector (which may be empty) which we condition on.

In this model it is clear (convince yourself!) that x_t does *not* Granger cause y_t with respect to the information set generated by z_t if either $A_{13}^i = 0$ and $A_{23}^i = 0$; $i = 1, ..., k$ or $A_{13}^i = 0$ and $A_{12}^i = 0$; $i = 1, ..., k$. Note that this is the way you will test for Granger causality. Usually you will use the VAR approach if you have an econometric hypothesis of interest that states that x_t Granger causes y_t but y_t does not Granger cause x_t. Sims (1972) is a paper that became very famous because it showed that money Granger causes output, but output does not Granger cause money. (This was in the old old days when people still took monetarism seriously, and here was a test that could tell whether the Keynesians or the monetarists were right!!). Later Sims showed that this conclusion did not hold if interest rates were included in the system. This also shows the major drawback of the Granger causality test - namely the dependence on the right choice of the conditioning set. In reality one can never be sure that the conditioning set has been choosen large enough (and in short macro-economic series one is forced to choose a low dimension for the VAR model), but the test is still a useful (although not perfect) test.

I think that the Granger causality tests are most useful in situations where one is willing to consider 2-dimensional systems. If the data are reasonably well described by a 2-dimensional system ("no z_t variables") the Granger causality concept is most straightforward to think about and also to test. By the way, be aware that there are special problems with testing for Granger causality in co-integrated relations (see Toda and Phillips (1991)).

In summary, Granger causality tests are a useful tool to have in your toolbox, but they should be used with care. It will very often be hard to find any clear conclusions unless the data can be described by a simple "2-dimensional" system (since the test may be between 2 vectors the system

may not be 2-dimensional is the usual sense), and another potentially serious problem may be the choice of sampling period: a long sampling period may hide the causality whereas for example VAR-systems for monthly data may give you serious measurement errors (e.g. due to seasonal adjustment procedures).

Extra reference:

Toda, H.Y. and P.C.B. Phillips (1994) : "Vector Autoregressions and Causality: A Theoretical Overview and Simulation Study", *Econometric Reviews* 13, 259-285.

7 Asymptotic theory and testing.

Maximum likelihood theory was developed for models with independent identical observations, but it turns out that most of the standard results and asymptotic formulae emerge in exactly the same way for time series models as they do for iid models. This statement is only true for stable models, so until further notice it is implicitly assumed that all the models are stable. The results below are asymptotic so initial values will likewise be ignored.

A general result in maximum likelihood theory is that if ψ is a vector of parameters and $L(\psi)$ is the likelihood function for a single observation then the maximum likelihood estimator $\tilde{\psi}$ has an asymptotically normal distribution with the true parameter as the mean and the inverse information matrix as the variance.
The information matrix is defined as

$$I(\psi) \;=\; -E(\frac{\partial^2 logL}{\partial\psi\partial\psi'}) \;,$$

where L is the likelihood function corresponding to a single observation, and in the iid case one can show that

$$\sqrt{T}(\psi - \tilde{\psi}) \;\Rightarrow\; N(0, I(\psi)^{-1}) \;,$$

subject to regularity conditions. The most important regularity conditions are sufficient differentiability of the likelihood function (typically 2 times differentiable with continuous second derivative) and that the likelihood function has a unique solution for ψ, i.e. that ψ is identified.

In the time series case we define the *asymptotic information matrix*

$$IA(\psi) \;=\; -plimT^{-1}(\frac{\partial^2 logL_T}{\partial\psi\partial\psi'}) \;,$$

where $L_T(\psi)$ is the likelihood function corresponding to the full set of observations. Under regularity conditions (see below) one can then show that

$$\sqrt{T}(\psi - \tilde{\psi}) \;\Rightarrow\; N(0, IA(\psi)^{-1}) \;,$$

where $\Rightarrow$ indicates convergence in distribution for the sample size (T) going to infinity.
Following Harvey (1980,1989) I will express this as

$$\tilde{\psi} \;\sim\; AN(\psi, T^{-1}IA(\psi)^{-1}) \;.$$

Example For an AR(1) model with N(0,σ^2)errors, it is straightforward to find the asymptotic information matrix as

$$IA(a,\sigma^2) \;=\; \begin{pmatrix} \frac{1}{1-a^2} & \\ & \frac{1}{2\sigma^4} \end{pmatrix} \;.$$

I will show it in detail for this example. The contribution from a single observation x_t is:

$$logL(x_t|x_{t-1}) \;=\; -\frac{1}{2}log\sigma^2 - \frac{(x_t - ax_{t-1})^2}{2\sigma^2} \;.$$

We find that

$$\begin{aligned}
\frac{\partial logL}{\partial a} &= \frac{(x_t - ax_{t-1})x_{t-1}}{\sigma^2} \\
\frac{\partial logL}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{(x_t - ax_{t-1})^2}{2\sigma^4} \\
\frac{\partial^2 logL}{\partial a \partial a} &= -\frac{x_{t-1}^2}{\sigma^2} \\
\frac{\partial^2 logL}{\partial \sigma^2 \partial a} &= -\frac{(x_t - ax_{t-1})x_{t-1}}{\sigma^4} \\
\frac{\partial^2 logL}{\partial \sigma^2 \partial \sigma^2} &= \frac{1}{2\sigma^4} - \frac{(x_t - ax_{t-1})^2}{\sigma^6} \;.
\end{aligned}$$

Now, by the law of large numbers, the asymptotic information matrix can be found by taking the mean of the contribution to the likelihood function from a single observation.

In the general ARMA case it will still be the case that the estimator of the error variance will be distributed independently of the parameters in the lag polynomials. If ψ is the vector of parameters $a_1, ..., a_k, b_1, ..., b_l$ of an ARMA(k,l) model with normally distributed error terms then (as in the example above) we find

$$\frac{\partial^2 logL}{\partial \psi \partial \psi'} \;=\; -\frac{z_t z_t'}{\sigma^2} \;,$$

where $z_t \;=\; \frac{-\partial u_t}{\partial \psi}$.

Example For the MA(1) we have

$$u_t \;=\; x_t - bu_{t-1} \;,$$

from which we find the recursion equation

$$\frac{\partial u_t}{\partial b} \;=\; -b\frac{\partial u_{t-1}}{\partial b} \;-\; u_{t-1} \;.$$

Now use the following very elegant argument. The recursion implies that z_t follows an AR(1) process

$$z_t \;= bz_{t-1} \;+\; u_{t-1} \;,$$

where u_{t-1} is uncorrelated with z_{t-1} (why?). Now from our results fom AR(1) models we find that

$$E(z_t^2) \;=\; var(z_t) \;=\; \frac{\sigma^2}{1-b^2} \;,$$

from which we conclude that $\hat{b}$ is asymptotically normally distributed with mean b and variance $1 - b^2$.

For the ARMA(1,1) the same type of reasoning can be applied, see Harvey (1980) p. 131 where it is shown that the asymptotic variance Avar(a,b) of $\sqrt{T}(a,b)$ is

$$Avar(a,b) \;=\; \frac{1+ab}{(a+b)^2}\begin{pmatrix} (1-a^2)(1+ab) & -(1-a^2)(1-b^2) \\ -(1-a^2)(1-b^2) & (1-b^2)(1+ab) \end{pmatrix} .$$

The formulae become more complicated for higher order processes, and one way around this is to evaluate the asymptotic covariance matrix numerically, rather than analytically. Harvey (1989) p. 140-143, shows how one can extend the Kalman filter to include the derivatives of the likelihood function in the updating and predicting loops so as to arrive at an estimate of the asymptotic variance.

For the VAR(k) model I have already mentioned that the least squares estimator is identical to the ML estimator. Consider the VAR(k) model for a p-dimensional x_t-process.

$$x_t \;=\; \mu \;+\; A_1 x_{t-1} \;+\; ... \;+\; A_k x_{t-k} \;+\; u_t \;,$$

where $var(u_t) = \Omega_u$. This model corresponds to the model

$$x_t \;=\; BZ_t \;+\; u_t \;,$$

where

$$B \;=\; (\nu, A_1, ..., A_k)$$

and

$$Z_t \;=\; \begin{bmatrix} 1 \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-k} \end{bmatrix} .$$

If we define

$$X \;=\; (x_{k+1}, ..., x_T)$$

and

$$Z \;=\; (Z_{k+1}, ..., Z_T)$$

then the model can be written compactly (!) as

$$X \;=\; BZ \;+\; U \;,$$

where

$$U \;=\; (u_{k+1}, ..., u_T) \;.$$

The least squares estimator of B is

$$\hat{B} \;=\; XZ'(ZZ')^{-1} \;,$$

which I have not proved (although I am very willing to do it if there is a demand).
Note: The indices are shifted a little bit compared to the presentation earlier.
I will repeat the definitions of X and Z here with more detail, so that you can compare with the GAUSS-program.

$$X = \begin{pmatrix} x_{1,k+1} & ... & x_{1t} & ... & x_{1T} \\ \vdots & & \vdots & & \vdots \\ x_{p,k+1} & ... & x_{pt} & ... & x_{pT} \end{pmatrix}$$

and

$$Z = \begin{pmatrix} 1 & ... & 1 & ... & 1 \\ x_{1k} & ... & x_{1,t-1} & ... & x_{1,T-1} \\ \vdots & & \vdots & & \vdots \\ x_{p,k} & ... & x_{p,t-1} & ... & x_{p,T-1} \\ \vdots & & \vdots & & \vdots \\ x_{11} & ... & x_{1,t-k} & ... & x_{1,T-k} \\ \vdots & & \vdots & & \vdots \\ x_{p,1} & ... & x_{p,t-k} & ... & x_{p,T-k} \end{pmatrix} .$$

Theorem If

$$a) \ \Gamma \stackrel{def}{=} plim ZZ'/T \ \text{ exist and is nonsingular}$$

and if

$$b) \ \frac{1}{\sqrt{T}} vec(UZ') = \frac{1}{\sqrt{T}} (Z \otimes I) vec(U) \Rightarrow N(0, \Gamma \otimes \Omega_u)$$

then

$$\sqrt{T} vec(\hat{B} - B) \Rightarrow N(0, \Gamma^{-1} \otimes \Omega_u) .$$

Comment: Notice that the conditions are true for the stable VAR model with normally distributed error terms.
The conditions are also true if x_t is stable and u_t satisfies the conditions: $Eu_t = 0$, $E(u_t u_t') = \Omega_u$ is constant and non-singular, u_s and u_t are independent (but not necessarily identically distributed) and

$$E|u_{it} u_{jt} u_{kt} u_{lt}| < c \ \text{ for } i, j, k, l = 1, ..., p \text{ and all } t$$

for some finite constant c.

This is the kind of assumptions that you will always need: i) some assumption of independence of successive observations and ii) some condition on the size of the moments and of course iii) stability of the model.

In order to use this theorem in practice one needs to estimate the variance parameters. You will use

$$\hat{\Gamma} = ZZ'/T ,$$

and

$$\hat{\Omega}_u \;=\; \frac{\hat{U}\hat{U}'}{T-kp-1}\;,$$

where

$$\hat{U} \;=\; X \;-\; \hat{B}Z\;.$$

Lütkepohl (1991) shows that $\hat{\Gamma}^{-1}\otimes\hat{\Omega}_u$ is a consistent estimator of $\Gamma^{-1}\otimes\Omega_u$.

In the scalar case the theorem says that

$$\sqrt{T}\begin{pmatrix}\hat{\nu}-\nu\\ \hat{a}_1-a_1\\ \hat{}\\ \vdots\\ \hat{a}_k-a_k\end{pmatrix} \;\Rightarrow\; N(0,\Gamma^{-1}\otimes\sigma_u^2)\;.$$

Which for the scalar AR(1) reduces to (disregrading the mean)

$$a_1-\hat{a}_1 \;\Rightarrow\; N(0,\frac{\sigma_u^2}{1-a^2})\;,$$

which we already knew. In the scalar AR(1) case this is widely known as the Mann-Wald theorem.

Testing

I will here give a brief introduction to the 3 common test of restrictions in econometrics. These tests are the LR (likelihood ratio) test, The W (Wald) test (named after the brilliant statistician Abraham Wald (?-1950)), and the LM (Lagrange Multiplier) test (the LM test is also know as the *score* test).

The LR test.

Assume that you are estimating a vector of parameters ψ, and denote the logarithm of the likelihood function evaluated at ψ by $logL(\psi)$. Denote the value of ψ that maximizes the likelihood function by $\hat{\psi}$. Assume that you want to test a hypothesis (often this will be the model suggested by theory) that involves constraints on your parameters. We will restrict ourselves to linear constraints which can be written as

$$R\psi \;=\; 0\;,$$

where R is an $m\times n$ matrix of full rank. In other words, there are m restrictions. Note that the most common test, namely that of an individual coefficient (the first say) being zero is covered by setting $R=(1,0,...,0)$. Note that R is always a matrix that the econometrician chooses a priori. The value of ψ that maximizes the likelihood function under the constraints will be denote $\hat{\psi}_0$. The LR test is simply

$$LR \;=\; 2logL(\hat{\psi}) \;-\; 2\log L(\hat{\psi}_0)$$

Under the null hypothesis that the restrictions are true the LR statistic is distributed as $\chi^2(m)$ - the chi-square distribution with m degrees of freedom. A sketch of the proof is the following. A Taylor series expansion of $L(\psi)$ around ψ_0 gives

$$logL(\hat{\psi}_0) \;=\; logL(\hat{\psi}) \;+\; (\hat{\psi}-\hat{\psi}_0)'DlogL(\hat{\psi}) \;+\; \frac{1}{2}(\hat{\psi}-\hat{\psi}_0)'D^2logL(\hat{\psi}-\hat{\psi}_0) \;,$$

where $DlogL = \frac{\partial logL}{\partial \psi}$ and $D^2logL = \frac{\partial^2 logL}{\partial \psi \partial \psi'}$. Now the first terms in this expansion is equal to 0 because $DlogL(\hat{\psi}) = 0$ (why?), so that

$$LR \;=\; (\hat{\psi}-\hat{\psi}_0)'D^2logL(\hat{\psi}-\hat{\psi}_0) \;,$$

It is obvious that $\hat{\psi}-\hat{\psi}_0$ has mean 0 under the null hypothesis. Also we know that if a k-dimensional random variable $x \sim N(0,\Sigma)$ then $x'\Sigma^{-1}x$ is χ^2-distributed with k degrees of freedom. Now compare this to the asymptotic formulae on the first page of the present chapter and you see where the asymptotic distribution comes from. (Note: to make this into a proper proof we have to be more careful with the Taylor series expansion and we also have to figure out where the m-degrees of freedom for the asymptotic χ^2 distribution comes from).

The Wald test.
Recall that asymptotically

$$\sqrt{T}\hat{\psi} \Rightarrow N(\psi, IA(\psi)^{-1}) \;,$$

where $\hat{\psi}$ is the unrestricted ML estimator. If $R\psi = r$ (where ψ is the true value of the parameter, then

$$\sqrt{T}(R\hat{\psi}-r) \Rightarrow N(0, R\,IA(\psi)^{-1}\,R') \;,$$

wherefore

$$T(R\hat{\psi}-r)'[R\,IA(\psi)^{-1}\,R']^{-1}(R\hat{\psi}-r)$$

has a limiting χ^2 distribution. This is the basis for the Wald test which is obtained by choosing a matrix $I(\hat{\psi})$ with the property that $plim T^{-1}I(\hat{\psi}) = IA(\psi)$, such that the limiting χ^2 distribution still obtains even if the asymptotic variance matrix is also estimated. So the Wald statistic is

$$W = (R\hat{\psi}-r)'[R\,I(\hat{\psi})^{-1}\,R']^{-1}(R\hat{\psi}-r) \;.$$

It is fairly obvious from the heuristics given here that the Wald test is asymptotically equivalent to the LR test. The Wald test is simpler to use than the LR test since only one estimation has to be performed. In the simple case where the test involves only one parameter ψ, say, the Wald test is

$$W \;=\; (\hat{\psi}-\psi_0)^2/avar(\hat{\psi}) \;,$$

where $avar(\hat{\psi})$ is a consistent estimate of the asymptotic variance of $\hat{\psi}$. Equivalently you may want to perform a 2-sided test based on the statistic

$$t \;=\; (\hat{\psi}-\psi_0)/\sqrt{avar(\hat{\psi})} \;,$$

which is asymptotically normally distributed. This is very often referred to as an asymptotic t-test. The reason for this is of course that in the standard normal regression model (without time-series structure) this statistic is just the standard t-distributed t-statistic. Most applied researchers do not actually test their models for normality, which really is an implicit statement that expresses that they rely on the *asymptotic* t-test. I find this practice quite ok.

All the tests (LR, Wald and ML) are equally valid for tests of non-linear restrictions. This is most easily seen for the Wald test. Assume that we are interested in the non-linear hypothesis $r(\psi) = 0$. Now if $\hat{\psi}$ is distributed as $N(\psi, \Sigma)$ then as well known theorem states that

$$r(\hat{\psi}) \ \sim \ N(r(\psi), R\Sigma R') \ ,$$

when r is differentiable with Jacobian R. This can be derived from the Taylor series expansion

$$r(\hat{\psi}) \ = \ r(\psi) \ + \ (\hat{\psi} - \psi)R \ .$$

For that reason this theorem is often referred to as the "Delta-method" (a bit odd to my taste; but you will probably meet that expression).
Anyway, it is now fairly obvious that the Wald test in the case of possibly non-linear restrictions of the form $r(\psi) = 0$, takes the form

$$W \ = \ r(\hat{\psi})' \, [R \, I(\hat{\psi})^{-1} \, R' \,]^{-1} r(\hat{\psi}) \ .$$

where R now is the matrix of derivatives of r evaluated at $\hat{\psi}$.
Take some time to go over this non-linear extension. When we get to the non-linear GMM estimators this will be useful. I always guide my intuition in non-linear models by thinking of functions as linear, and this has never - to my knowledge - mislead me (because of the "delta-method").

The LM-test.
Loook at the Taylor series expansion of $DlogL(\psi)$ around $\hat{\psi}$:

$$DlogL(\hat{\psi}_0) \ = \ Dlog(\hat{\psi}) \ + \ (\hat{\psi} - \hat{\psi}_0)' D^2 LogL \ ,$$

where $D^2 logL$ is evaluated at $\hat{\psi}$. But of course $Dlog(\hat{\psi}) = 0$ so the Taylor series expansion gives

$$DlogL(\hat{\psi}_0) \ = (\hat{\psi} - \hat{\psi}_0)' D^2 LogL \ .$$

Now define the test statistic

$$LM \ = \ DlogL(\hat{\psi}_0)' \, IA^{-1}(\hat{\psi}_0) \, DlogL(\hat{\psi}_0) \ .$$

Using the Taylor series expansion gives

$$(\hat{\psi} - \hat{\psi}_0)' \, D^2 logL \ \ IA^{-1} \, D^2 logL \, (\hat{\psi} - \hat{\psi}_0) \ .$$

Now $plim\ D^2logL = -IA$, so this expression is in large samples equivalent to

$$-(\hat{\psi} - \hat{\psi}_0)' D^2 logL\, (\hat{\psi} - \hat{\psi}_0) \ ,$$

which is exactly the same expression as the one given for the ML test above. Therefore the LM test will also have an asymptotic $\chi^2(m)$ distribution. The main advantage of the LM-test is that one only has to evaluate the criterion function under the null hypothesis. In linear models that does rarely make a big difference; but in non-linear models, where convergence of the optimization alogorithms may be hard achieve, it can be a major advantage. Also note, that is often the case in non-linear models, that if the null hypothesis is actually true, then the parameters of the more general model will be badly determined from the data - something that often hampers convergence. In the literature about LM-testing a gooddeal is often made out of the fact that the LM test often can ve written in a convenient form that allows for easy computation using linear regression packages. To derive theorems of this sort you will have to find the derivative a particular likelihood function explicitly. With a modern computer language like GAUSS, you can, however, just ask GAUSS to evaluate the derivative of the likelihood function numerically and the just apply the general form of the test. This is just as goo except in special cases, as for example where the estimated parameters are near the boundary of the allowed region, in which case the derivatives sometimes are badly determined (this all depends on your specific model).

The Portmanteau test.
In one has estimated an ARMA model (or another time series model) a "quick-and-dirty" test for whether the model is underparameterrized is the "portmanteau" test. It was originally suggested by Box and Pierse (1970) and modified to the form given below by Ljung and Box (1978).
Assume that $\hat{\epsilon}_t$ are the residuals from an estimated ARMA(k,l) model. If $\hat{a}_i,\ \hat{b}_j$ are the estimated (for example by Maximum Likelihood) parameters of the ARMA-model, then

$$\hat{\epsilon}_t \ = \ x_t - \hat{a}_1 x_{t-1} - ... - \hat{a}_k x_{t-k} \ - \ \hat{b}_1 \hat{\epsilon}_{t-1} - ... - \hat{b}_l \hat{\epsilon}_{t-l} \ ,$$

where one will usually start the recursion by setting the inital $\hat{\epsilon}_t$ to zero. The portmanteau test is

$$Q \ = \ T(T+2) \sum_{k=1}^{M} \frac{r_k^2}{T-k} \ ,$$

where r_k is the k-th autocorrelation of ϵ_t estimated in the usual fashion (as desribed earlier, where we used the notation $c(k)$ for the estimated autocorrelation); but from the residual process. The Q statistic asymptotically distributed as $\chi^2(M-k-l)$ under the hypothesis that the ARMA model is correctly specified. Intuitively it is clear that if the residuals of the true model are white noise, then the residuals from the estimated process will also tend to be white noise in which case the estimated auto-correlations will all be close to 0.

There are quite a few other common tests. We will not go into them unless there is special interest. One group of tests is known as misspecification tests. (These are also often called specification

tests. I am not quite certain by I think that some authors may define "test for specification" as different from "test for misspecification"; so don't get confused). Misspecification tests are tests that intned to evaluate the "fit" of the model; but not necessarily with any particular alternative in mind. The most well known misspecification test is the Hausman test (Haussman (1978)), and the White information matrix test. See Godfrey (1978) for references.

Another group of tests that one may need is tests for non-nested models. Sometimes it is hard to come up with a general model nesting two competing models, and even if you can come up with such an *encompassing* model it may be too unwieldy to estimate. In this case you may want ot perform a test that allows for non-nested models. The most famous class of non-nested tests are known as Cox-tests. There is a special issue of the *Journal of Econometrics* (White (1983)), that can serve as an introduction to this area.

8 Unit Roots.

A very good place to look after you have read this section is the survey in *Handbook of Econometrics* Vol. IV by Jim Stock.
In the statistical literature it has long been known that unit root processes behave differently from stable processes.
For example in the scalar AR(1) model, consider the distribution of the OLS estimator of the parameter a in the simple first order process,

$$(1)\ y_t \ = \ a\ y_{t-1} \ + \ e_t\ .$$

If e_t are independently identically normally distributed (niid) variables and a_N denotes the least squares estimator of b based on $y_0, y_1, .., y_N$, then Mann and Wald (1943) showed that $N^{1/2}(a_N - a)$ has a limiting normal distribution if $a < 1$. White (1958) showed that $|a|\ N(a^2 - 1)\ (a_N - a)$ has a limiting Cauchy distribution if $a > 1$, whereas $N(a_N - 1)$ has a limiting distribution that can be written in terms of a ratio of two functionals of a Wiener process, when $a = 1$. In the later years a lot of theoretical work has been done on the distribution of least squares estimators in the presence of unit roots. Some notable early contributions are Fuller (1976), Dickey and Fuller (1979,1981), and Evans and Savin (1981,1984). The authoritative paper by Phillips (1987) sums up most of the theory.

It was the paper by Nelson and Plosser (1982) that sparked the huge surge in interest for unit root models among economists. They examined time series for some of the most important U.S. aggregate economic variables and concluded that almost all them were better described as being integrated of order one rather than stable. (They further went on to suggest that this favored real-business-cycle type of classical models in favor of monetarist and Keynesian models. In my opinion it is plain crazy to try and derive such sweeping conclusions from atheoretical time series modeling, and it is even crazier to do so on aggregate data; but maybe the huge interest that this paper generated is partly due to this provocative statement).

The reason why unit roots are so important is that the limiting distributions of estimates and test statistics are very different from the stationary case. Most importantly, one can not (in general) obtain limiting χ^2 (or t- or F-) distributions. Rather one obtains limiting distributions that can be expressed as functionals of Brownian motions. Also notice that in the case where you have a stable model that is "close" to an integrated process, then the distributions of estimators will look more like the distributions from unit root models than it will look like the asymptotic (normal type) distribution in small samples. This phenomenon is treated in the literature under the heading of near-integrated (or near unit-root, or nearly non-stationary) models. We may not have time to go into this; but the reading list contains a quite detailed bibliography (since I happened to have written a paper in that area - I had the references typed up already). I personally doubt whether many series in economics are best thought of as genuinely non-stationary, and I don't think that one really can decide that on the basis of the statistical evidence (there has been written lots of

papers on that question since the influential paper of Nelson and Plosser (1982)). My point of view is that it does not really matter. The models will have very different predictions for the very long run whether they truly have unit roots or not; but to cite Lord Keynes: "In the long run we are all dead"; or to say it less dramatically - I do not think that simple time series models are useful for forecasting 20 years ahead under any circumstances. What matters is that the small sample distributions look like the asymptotic unit root distributions, so if you do not use those you will make wrong statistical inferences.

8.1 Brownian Motions and Stochastic Integrals.

The easiest way to think of the Brownian motions is in the following way (which corresponds exactly to the way that you will simulate Brownian motions on the computer): Let

$$B_N(t) \ = \frac{1}{\sqrt{N}} \ (\ e_1 \ + \ e_2 \ + \ ... \ + e_{[Nt]}) \ ; \ t \in [0,T] \ ,$$

where $e_1, ..., e_{[NT]}$ are iid $N(0,1)$. The notation $[Nt]$ means the integer part of Nt, i.e. the largest integer less than of equal to Nt. Note that $B_N(t)$ is a stochastic function from the closed interval $[0,T]$ to the real numbers. If N is large $B_N(.)$ is a good approximation to the Brownian motion $B(t); t \in [0,T]$ which is defined as

$$B(t) \ = \ \lim_{N \to \infty} B_N(t) \ .$$

For a fixed value of t it is obvious that $B_N(t)$ converges to a normally distributed random variable with mean zero and variance t. To show that $B_N(t)$ converges as a function to a continuous function $B(t)$ takes a large mathematical apparatus. You can find that in the classical text of Billingsley (1968); but be warned that this is a book written for mathematicians (but given that it is very well written).
For the purpose of the present course this is all you need to know about the Brownian motion. One can show that any stationary continuous stochastic process $B(t)$ for which

$$B(t_4) - B(t_3) \ \text{ and } \ B(t_2) - B(t_1)$$

are independent for all $t_4 \geq t_3 \geq t_2 \geq t_1$ and with

$$E\{B(t_2) - B(t_1)\} \ = \ 0 \ , \text{ and } \ Var\{B(t_2) - B(t_1)\} \ = \ t_2 - t_1 \ ;$$

has to be a Brownian Motion. The Brownian motion is an example of a *process with identical independent increments*. You can convince yourself from the definition I gave of Brownian motion, that this formula for the variance is true. Notice that if you start the process at time 0 then

$$Var(B(t)) \ = \ t \ .$$

So it is obvious that the unconditional variance of $B(t)$ tends to infinity as t tends to infinity. This corresponds to the behavior of the discrete time random walk, which again is just an AR(1) with

an autoregressive coefficient of 1. So it is not surprising that Brownian motions show up in the (properly normalized) asymptotic distribution of estimators of AR models. Brownian motions are quite complicated if you look into some of the finer details of their sample paths. One can show that Brownian motions with probability 1 are only differentiable on a set of measure zero. You can also show that a Brownian motion that you start at zero at time zero will cross the x-axis infinitely many times in any finite interval that includes 0. Properties like those are very important in continuous time finance, so if you want to specialize in that field you should go deeper into the theory of Brownian motions. I have supplied a list of references in the reading list, that will be useful for that purpose; but for unit root theory this is not absolutely necessary.
You will also need to be able to integrate with respect to Brownian motions. We want to give meaning to the symbol $\int_0^1 f(s)dB(s)$, where f is a function that will often be stochastic itself. In many asymptotic formulae $f(s)$ is actually the same as $B(s)$. We will define the so-called Ito-integral (named after the Japanese mathematician Kiyosi Ito):

$$\int_0^1 f(s)dB(s) = \lim_{K\to\infty} \sum_{k=0}^{K} f(\frac{k-1}{K})\Delta B(\frac{k}{K}) \ ,$$

where $\Delta B(\frac{k}{K}) = B(\frac{k}{K} - B(\frac{k-1}{K})$, and where the limit is *in probability*. You can *not* obtain convergence sample path by sample path almost surely (with probability one). If we temporarily call the stochastic integral on the left hand side for I and the approximating sum on the right hand side for I_K then the convergence in probability means that there exists a stochastic variable I such that

$$\lim_{K\to\infty} P\{|I - I_K| > \epsilon\} = 0$$

for any given $\epsilon > 0$. This is however a probability statement and it does not preclude that I_K for a given sample path can be found arbitrarily far from I for arbitrarily large K. For our purpose that does not really matter, since convergence in probability is all we need. If you want to go through the mathematics (which is detailed in the probability books in the reading list, with the book by Øksendal as the most accessible), then the hard part is to show the existence of the limit I to which I_K converges, and if you don't want to go through the mathematics you should just take it for a fact that the limit exists in a well defined sense.

Notice that the sum in the approximating sum is over values of the function multiplied by a "forward looking" increment in the integrator B. This is essential and you will not get the right answer if you do not do it like that. This is in contrast to the standard Riemann-Stieltjes integral where this does not matter. The definition of the stochastic integral is given in the way that you can actually simulate the distribution, and this is the way it is done in many Monte Carlo studies in the literature.

Ito's Lemma
The main tool of modern stochastic calculus is Ito's lemma. It is usually formulated as

$$df(X,t) \ = \ \frac{\partial f}{\partial t}dt \ + \ \frac{\partial f}{\partial X}dX \ + \ \frac{1}{2}\frac{\partial^2 f}{\partial X^2}(dX)^2 \ \text{ where } \ dt^2 \ = \ 0 \ , \text{and } \ dB^2 \ = \ dt \ .$$

In this course Ito's lemma is not so central; but you may meet the following equality

$$(*) \quad \int_0^1 B(s)dB(s) \;=\; \frac{1}{2}(B(1)^2 - 1) \;,$$

which is a simple consequence of Ito's lemma. Ito's lemma is essential in continuous time finance and in more advanced examinations of unit root theory.
I leave the proof of (*) for the homework.
Example:
Find $d\log(\log B)$. We do this is 2 stages. First we set $X = \log(B)$. Then (since $(dB)^2 = dt$)

$$dX \;=\; \frac{1}{B}dB \;-\; \frac{1}{2}\frac{1}{B^2}dt \;.$$

Then

$$\begin{aligned} d\log(\log B) &= \frac{1}{B}dX \;-\; \frac{1}{2}\frac{1}{B^2}(dX)^2 \\ &= \frac{1}{B}(\frac{1}{B}dB \;-\; \frac{1}{2B^2}dt) \;-\; \frac{1}{2B^2}\frac{1}{B^2}dt \\ &= \frac{1}{B^2}dB \;-\; (\frac{1}{2B^3} \;+\; \frac{1}{2B^4})dt \end{aligned}$$

(I cannot think of an application of this particular result but it illustrates the method clearly.)

Now consider the process (1) again. Notice that if $a = 1$ then

$$y_t \;=\; y_{t-1} \;+\; e_t \;=\; \sum_{k=1}^{t} e_k \;.$$

We will consider the least squares estimator

$$\hat{a} \;=\; \frac{\sum_{t=1}^T y_t y_{t-1}}{\sum_{t=1}^T y_{t-1}^2} \;=\; \frac{\sum_{t=1}^T (y_{t-1} + \Delta y_t) y_{t-1}}{\sum_{t=1}^T y_{t-1}^2} \;=\; 1 \;+\; \frac{\sum_{t=1}^T y_{t-1}\Delta y_t}{\sum_{t=1}^T y_{t-1}^2}$$

Now notice that this implies that

$$T(\hat{a} \;-\; 1) \;=\; T\frac{\sum_{t=1}^T y_{t-1}\Delta y_t}{\sum_{t=1}^T y_{t-1}^2} \;=\; \frac{\sum_{t=1}^T (y_{t-1}/\sqrt{T})(e_t/\sqrt{T})}{\sum_{t=1}^T (y_{t-1}/\sqrt{T})^2 \frac{1}{T}}$$

One can now see from the way that we defined the Brownian motion that $y_{t-1}/\sqrt{T}$ converges to a Brownian motion, and from the way that we defined the stochastic integral, one can see (at least one would guess) that

$$T(\hat{a} \;-\; 1) \;\Rightarrow\; \frac{\int_0^1 B(s)dB(s)}{\int_0^1 B(s)^2 ds} \;.$$

Intuitively you should always think of e_t as $dB(t)$ and y_t which under the null of a unit root is equal to $\sum_{k=0}^t e_t$ corresponds then to $\int_0^{t/T} dB(s) = B(s/T)$ (for $B(0) = 0$). From our application of Ito's lemma one can see that another expression would be

$$T(\hat{a} \;-\; 1) \;\Rightarrow\; \frac{1}{2}\frac{B(1)^2 - 1}{\int_0^1 B(s)^2 ds} \;.$$

Notice that $B(1)$ is just a standard normal distribution, so that $B(1)^2$ is just a standard $\chi^2(1)$ distributed random variable. Contrary to the stable model the denominator of the expression for the least squares estimator does not converge to a constant almost surely; but rather to a stochastic variable that is strongly correlated with the numerator. For these reasons the asymptotic distribution does not look like a normal, and it turns our that the limiting distribution of the least squares estimator is highly skewed, with a long tail to the left - look at the graph of the distribution in e.g. Evans and Savin (1981).

OK, if you couldn't quite follow the "derivation" of the limiting distribution, don't despair. One needs quite a bit more machinery and notation to make sure that all the limit operations are legal; but all you need is the basic intuition, so try and get that. Most of the part of the unit root literature that is concerned with asymptotic theory contains the limiting distribution given here. We will refer to the distribution of

$$\frac{\int_0^1 B(s)dB(s)}{\int_0^1 B(s)^2 ds}$$

as *the unit root distribution*. It is not possible to find any simple expression for the density of this distribution but one can find the characteristic function, which can be inverted in order to tabulate the distribution function - (see Evans and Savin (1981)). You can also evaluate the distribution by Monte Carlo simulation, which is performed by choosing a large value of T, and then drawing the innovation terms from a pseudo random number generator (this is very easy in GAUSS) and the generating the series y_t from the defining equation (1). For large T the distribution of the LS estimator is close to the limiting distribution, which can be graphed by repeating this exercise like 10- or 20,000 times and plotting the result.

TS and DS models

If you look at a plot of a typical macro economic time series, like real GNP, it is obvious that it displays are very pronounced *trend*. What is a trend? Well, for many years that question was not considered for many seconds - a trend was simply assumed to be linear function of time, and econometricians would routinely "detrend" their series by using the residuals from a regression on time (and a constant) rather than the original series. This practice was challenged by the Box-Jenkins methodology, which became somewhat popular in economics in the seventies, although it originated from engineering. The Box-Jenkins methodology had as one of its major steps the "detrending" of variables by the taking of differences. In the 80ies a major battle between thes two approaches raged, with the difference-detrenders seemingly having the upper hand in the late 80ies, although challenged from many sides - the Bayesians being the most aggressive.
During that period the following terminology took hold. The model

$$(DS)\ y_t \ = \ \mu \ + \ \ y_{t-1} \ + \ e_t$$

is called Difference Stationary (DS) since it is stationary after the application of the differencing operation, and the model and

$$(TS)\ y_t \ = \ \mu \ + \ \beta t \ + \ a \ y_{t-1} \ + \ e_t \ ; \ \ a < 1 \ ,$$

is called Trend Stationary (TS) since it is stationary after good old-fashioned detrending by regressing on a time-trend. Most tests for unit roots are formulated as testing TS versus DS.

8.2 Unit Root tests

8.2.1 Dickey-Fuller tests

The most famous of the unit root tests are the ones derived by Dickey and Fuller and described in Fuller (1976).
Dickey and Fuller considered the estimation of the parameter a from the models

$$(1) \ \ y_t \ = \ \rho \ y_{t-1} \ + \ e_t \ ,$$

$$(2) \ \ y_t \ = \ \mu \ + \ \rho \ y_{t-1} \ + \ e_t \ .$$

and

$$(3) \ \ y_t \ = \ \mu \ + \ \beta t \ + \ \rho \ y_{t-1} \ + \ e_t \ .$$

The parameter ρ is just the AR-parameter that we have denoted by a so far; but here I use the notation of Fuller (1976). It is assumed that $y_0 = 0$.

The simplest Dickey-Fuller test is simply to estimate (1) by least squares and compare $T(\hat{\rho} - 1)$ to a table of the distribution derived from a Monte Carlo study (or, as shown in Evans and Savin (1981), one can find the characteristic function and invert it). This test is sometimes known as the Dickey-Fuller ρ-test. The critical values for this and the following Dickey-Fuller (DF) tests can be found in Fuller (1976), p. 373. Simplified versions of the tables from Fuller can be found many places, f. ex. in Maddala (1992). (The Monte Carlo simulations were actually done as part of David Dickey's Ph.D. thesis). In practice the model (1) is often too simple and one would like to allow for a mean term in the estimation. Dickey and Fuller suggested that one estimates (2) by first calculating the average $\bar{y}$ of the observations $y_2, ..., y_T$ and the average $\bar{y}_0$ of $y_1, ..., y_{T-1}$ and then calculates the least squares estimator $\hat{\rho}_\mu$ by regressing $y_t - \bar{y}$ on $y_{t-1} - \bar{y}_0$. Comparing $T(\hat{\rho}_\mu - 1)$ to the critical values in Fuller is known as the Dickey-Fuller ρ_μ-test. Dickey and Fuller also suggested estimating $\hat{\rho}_\tau$ from model (3) by a standard regression and they tabulated the critical values of $T(\hat{\rho}_\tau - 1)$ under the composite null hypothesis $\rho = 1$ and $\beta = 0$.
Note that this last test is a test for the DS model against the TS model, and it is known as Dickey-Fuller ρ_τ-test.

It is often not realistic that a data series should follow as simple a model as (1), (2), or (3) with iid

error terms. In most economic data series there will also be substantial short term autocorrelations, so it may be more reasonable to assume the model

$$(4) \quad y_t \;=\; a_1 y_{t-1} \;+ ... + \; a_p y_{t-p} \;+\; e_t \;,$$

where e_t is iid normal, rather than model (1) (and equivalent for models (2) and (3)). An equivalent way of writing model (4) is

$$(4') \quad y_t \;=\; \theta_1 y_{t-1} \;+\; \theta_2 \Delta y_{t-1} + ... + \; \theta_p \Delta y_{t-p+1} \;+\; e_t \;.$$

It is simple to show the equivalence of (4) and (4') and it is left as an exercise. A unit root in (4) will show up as $\theta_1 = 1$ in (4'). So to test for a unit root in (4) you might want to use $T(\hat{\theta}_1 - 1)$. One would expect the distribution of this statistic to depend on nuisance parameters as e.g. the minimum order p of the autoregression, and it does. One can show (see Fuller (1976) that $c\,T(\hat{\theta}_1 - 1)$ has the same limiting distribution as $T(\hat{\rho} - 1)$ has, for some constant c, where c is the sum of the terms in the MA representation for e_t. It is somewhat complicated to find c which makes this test less appealing; but it turns out that if you instead of using the distribution of the coefficient you use the distribution of the t-statistic, then the test will *not* depend on the form of the autoregression. The t-statistic has the usual form, that is calculated by any regression package, namely

$$\hat{\tau} \;=\; \frac{\hat{\rho} - 1}{\sqrt{s^2 (\sum_{t=2}^{T} y_{t-1}^2)^{-1}}} \;,$$

in the simplest model with zero mean and no trend.
For model (1) Dickey and Fuller simulated the critical values under the null of a unit root and the critical values of the test (called the $\hat{\tau}$-test) is tabulated in Fuller (1976) p. 373. Dickey and Fuller also simulated the corresponding t-test for model (2) and (3), which they denoted $\hat{\tau}_\mu$ and $\hat{\tau}_\tau$.
Testing for unit roots using one of the models (1)-(3) is known as Dickey-Fuller tests, whether the $\hat{\rho}$ or the $\hat{\tau}$-tests are used. Estimating a higher order autoregressive process (maybe with non-zero mean and trend) and using the critical $\hat{\tau}$ tables are known as Augmented Dickey-Fuller (ADF) tests. ADF-tests are the most widely applied tests for unit roots. In the influential paper by Engle and Granger (1987) on co-integration, they recommended ADF tests after examining ADF and several of its competitors. (In the case of co-integration test the critical values are different, see the section on co-integration).

Since the publication of Engle and Granger's article scores of new tests have been developed and we will look at the most popular contenders below. But let us consider how to choose between the different DF tests. Experience from Monte Carlo studies, see for example Dickey, Bell, and Miller (1986), shows clearly that the $\hat{\rho}$-tests have the highest power, as was to be expected. Which $\hat{\rho}$-test should you choose then? That depends, but in any raw series you typically must assume a non-zero mean, which means that you should "at least" use the $\hat{\rho}_\mu$- test. If you think that it is likely that there may be a linear trend, then you should use the $\hat{\rho}_\tau$-test; but as a general rule this is not "for free". In small samples (as we quite commonlu use in economics) there is a loss in power that

increases with the number of parameters that you estimate. On the other hand it you allow for too few parameters, then your model is misspecified and your tests will be wrong. In the present case this means that you will prefer not to include a time-trend if you are reasonably sure that it can safely be left out. Note, however, that the model 2 under the null hypothesis has

$$(2') \quad y_t \;=\; \mu t \;+\; \sum e_t \;,$$

which is *not* invariant to the μ- term after subtracting the mean y_t. In the stable process on the contrary, the de-meaned value of y_t will be independent of the mean term μ (since μ here shows up as a mean term $\mu/[1-\rho]$). Therefore you have to use the $\hat{\tau}_\tau$ test if you think that model (2) might be true with a unit root and a non-zero μ.

What do phrases like "reasonably sure" mean? That is hard to tell - it depends strongly on the number of observations that you have, if you have lots of data: take the most general model. If you have few data points, you will have to rely partly on the knowledge that you have about the particular area of economics, that you are researching. You may even feel that you need higher order deterministic terms than Fuller's tables allow for, in which case you may have to simulate the distribution under the null yourself (although you should consult a time series specialist first to see if they are already available). Be aware that it is not very hard to simulate the null distribution in GAUSS (or most other modern computer languages).

Now then should you use DF or the ADF test? For example for U.S. macro economic time series it is not safe to assume that there is no autocorrelation apart from the potential unit root, which means that you should use the ADF test. How many lags should you include? Again, it depends on the number of data points that you have and on your a priori expectations - the considerations are totally parallel to those about mean and trend terms. What if the process really is an ARIMA(k,1,1)? The MA term corresponds to an infinite AR process, so you may hope that including enough AR-terms and letting the number of included terms go to infinity with T, then you will obtain a consistent test. Said and Dickey (1984) show that this is exactly the case, given that on chooses $k = k(T) = k_0 T^{1/3}$. This is nice to know; but since k_0 is unknown it is pretty much useless as a practical guideline. You may sometimes see the ADF test referred to as the Said-Dickey test - that is a way of indicating that the author does not consider the order of the AR model as more than an approximation.

8.2.2 Phillips-Perron tests

The tests that has been most popular next to the DF-tests are the Phillips-Perron (PP) tests suggested in Phillips and Perron (1988).
In the case where you are convinced that there is no autocorrelation apart from the unit root, the PP test is identical to the DF test. The difference between the PP and the DF tests lies in the treatment of the autocorrelation. If there is autocorrelation that is not accounted for you will get

bias. The hard part is to rid of that bias. The ADF does that by choosing an AR(p) process, but the weak point in that procedure is exactly the choice of p. Not only does a researcher usually only have vague ideas about the appropriate chice of p; but for many applied researches it is just too tempting to try out a bunch of different values for p until the results corresponds to what the researcher wanted a priori. If results are derived by data-mining of that sort the asymptotic test statistics reported will be meaningless.

The idea of the Phillips-Perron test is to run a non-augmented Dickey Fuller regression, and then to adjust for the bias that might occur due to correlation in the innovation term so that the Dickey-Fuller tables can be used anyway.
Phillips-Perron suggest estimating $\hat{\rho}$ from model (2), then estimate the innovation error variance $s^2 = \sum_{t=2}^{T} \hat{e}_t^2$ and 2π times the spectral density at frequency zero by, $2\pi \hat{f}_e(0)$, (we will cover techniques for doing this later on). The Phillips-Perron $Z(\hat{\alpha})$ test is then

$$Z(\hat{\alpha}) \;=\; T(\hat{\rho}_\mu - 1) \;-\; \frac{T^2(2\pi \hat{f}(0) - s^2)}{2\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{-1})^2} \;,$$

where $\bar{y}_{-1}$ is the mean of y_{t-1}. (It has its name because Phillips-Perron uses α rather that ρ for the AR-parameter; but here we will not change notation again). The corresponding t-statistic is

$$Z(t_{\hat{\alpha}}) \;=\; \tau_\mu \sqrt{\frac{s^2}{2\pi \hat{f}(0)}} \;-\; \frac{T(2\pi \hat{f}(0) - s^2)}{2\sqrt{2\pi \hat{f}(0)\;\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{-1})^2}} \;.$$

Correspondingly for the model that allows for a time trend under the alternative Phillips and Perron suggest the adjusted DF tests

$$Z(\tilde{\alpha}) \;=\; T(\hat{\rho}_\tau - 1) \;-\; \frac{T^2(2\pi \hat{f}(0) - s^2)}{2\sum_{t=2}^{T}(y_{t-1} - \hat{y})^2} \;,$$

and

$$Z(t_{\tilde{\alpha}}) \;=\; \tau_\tau \sqrt{\frac{s^2}{2\pi \hat{f}(0)}} \;-\; \frac{T(2\pi \hat{f}(0) - s^2)}{2\sqrt{2\pi \hat{f}(0)\sum_{t=2}^{T}(y_{t-1} - \hat{y})^2}} \;,$$

where $\hat{y}$ is the projection of y_{t-1} on $1, t$ (i.e. the fitted value from a regression of lagged y on a constant and a time trend) and s^2 is defined as before but now from the residuals from the regression of y_t on $y_{t-1}, 1, t$. The PP-tests now consist of comparing those adjusted DF-values with the corresponding tables used for the basic Dickey-Fuller test. You will hear those tests referred to as the Phillips- Perron tests or as the $Z\alpha$ and Zt tests.

An interesting (and extensive) Monte Carlo study is Schwert (1987). Schwert examines what happens to the size of the above mentioned tests if the true process is an ARIMA(1,1,1) and the Phillips-Perron or the ADF tests are used. A very clear conclusion from Schwert's study is that the ADF test, used with "many" lags, have the best size-properties in the sense that the size of

the small sample distribution is close to the true 5% size if the Dickey-Fuller tables are used. The other tests can be *very* far of in terms of size, especially if the ARIMA-process has a large negative MA-coefficient, in which case the PP tests (and also the ADF test when the MA-coefficient is very close to -1), almost always reject the unit root model (even if it is true). Note that the happens because the model has the form

$$(1-L)y_t \;=\; (1+bL)e_t \;.$$

One can see that the two lag polynomials "nearly cancel" when b is close to -1. Some authors call processes like that "nearly white noise processes".

Does this mean that one should always use the ADF-test with a high number of AR-terms? Not necessarily, since there is a cost in terms of power. It seems that this cost is not nearly as high as the cost incurred by falsely allowing for a trend-term (see Dickey, Bell and Miller (1986)), but there will presumably still be some loss in power. In an interesting Monte Carlo study, Campbell and Perron (1991) show that both the Said-Dickey test (with 6 lags) and the Phillips-Perron test falsely reject the unit root almost every time in the case of nearly white noise processes. They also show, however, that for simple short term forecasting, this will not necessarily result in a decline in the ability to make simple forecasts, since the trend stationary model forecasts just as well in the cases where the unit root model tends to be falsely rejected.

8.2.3 Approximate POI-tests

In a recent interesting paper Elliott, Rothenberg and Stock (1992) (ERS) examine "point optimal invariant tests" (POI) for unit roots. An invariant test is a test – like the augmented Dickey-Fuller test – that is invariant to nuisance parameters. In the unit root case that will be tests that are invariant to the parameters that capture the stationary movements around the unit roots (i.e., the parameters to $\Delta X_{t-1}, \Delta X_{t-2}, \cdots, \Delta X_{t-k}$ in the ADF-case).

The most commonly used test are the "τ-tests" that allows for a deterministic time trend (in the case where you do not allow for a time trend most tests are pretty well behaved), so I will only consider this type of test in this subsection.

When

$$y_t = \mu + \sigma t + u_t \;;\quad u_t = \rho u_{t-1} + \epsilon_t \;,$$

ERS show that the POI test for a unit root against $\rho = \overline{\rho}$ has the form

$$M_T = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \;,$$

where

$$\tilde{\sigma}^2 \;=\; T^{-1}\sum_{t=1}^{T}\tilde{e}_t^2 \;; \qquad \hat{\sigma}^2 \;=\; T^{-1}\sum_{t=1}^{T}\hat{e}_t^2 \;,$$

and $\hat{e}_t$ and $\tilde{e}_b$ are the residuals from the GLS estimation of y_t on $z_t = (1, t)$ under $\rho = 1$ and $\rho = \overline{\rho}$ respectively:

$$\hat{e} \;=\; \hat{y}_t - \hat{\beta}' \hat{z}_t \;; \qquad (\hat{\beta} \;=\; (\sum \hat{z}_t \hat{z}_t')^{-1} (\sum \hat{z}_t y_t) \;,$$

where

$$(\hat{y}_1, \cdots, \hat{y}_T) \;=\; (y_1, \Delta y_2, \cdots, \Delta y_T) \;,$$

and

$$(\hat{z}_1, \cdots, \hat{z}_T) \;=\; (z_1, \Delta z_2, \cdots, \Delta z_T) \;,$$

and correspondingly for $\tilde{e}$ except

$$(\tilde{y}_1, \cdots, \tilde{y}_T) \;=\; (y_1, (1 - \overline{\rho} L) y_2, \cdots, (1 - \overline{\rho} L) y_T) \;,$$

and

$$(\tilde{z}_1, \cdots, \tilde{z}_T) \;=\; (z_1, (1 - \overline{\rho} L) z_2, \cdots, (1 - \overline{\rho} L) z_T) \;.$$

This last transformation is the GLS-detrending (if you are not quite sure what I mean by that, check your first year econometrics text (look under AR(1) disturbances) - most often the first observation will also be rescaled, but asymptotically that does not make a difference).

The critical value for the test will depend on $\overline{c}$ where

$$\overline{\rho} \;=\; 1 - \frac{\overline{c}}{T} \;.$$

When

$$u_t = \rho u_{t-1} + \beta_2 \Delta u_{t-2} + \cdots + \beta_k \Delta u_{t-k} + \epsilon_t \;,$$

the critical value of the test statistic has to be adjusted in the same way as for the Phillips-Perron (PP) test (in practice you adjust the test-statistic rather than the critical value, (and you don't have to assume a finite AR-representation), exactly as in the PP case).

In practice, $\overline{\rho}$ is not known, so ERS suggest choosing an approximate value of $\overline{c}$. They suggest using the value of $\overline{c}$ that asymptotically gives a power of 50%. These values are $\overline{c} = -7$ in case of a mean and $\overline{c} = -13.5$ in the case of a trend). ERS show that this way we get approximately the POI test asymptotically.

Also, it turns out that if we instead do the GLS-adjustment and then perform the ADF-test (now without allowing for a mean or trend) we get approximately the POI-test. Elliott et al. call this test the DF-GLS$^\tau$ test. They argue that the theoretical examinations above point to the GLS-adjustment as the critical factor, and they then show by Monte Carlo simulation that the DF-GLS$^\tau$ seems to behave as well as the approximate POI test (using the values for $\overline{c}$ that I listed above).

The critical values depend on T. The critical values are

T	1%	5%
50	-3.77	-3.19
100	-3.58	-3.03
200	-3.46	-2.93
500	-3.47	-2.89
∞	-3.48	-2.89

If you need critical values for the variant of the test with mean but not trend, see the article by ERS. (They call that test the DF-GLS$^{\mu}$ test.)

8.3 The importance of unit roots

The discussion above indicates that there is not necessarily a sharp distinction between unit root processes and stable processes, which has led some researchers (Cochrane (1991), Christiano and Eichenbaum (1989) - the later paper titled "Unit roots in real GNP: Do we know and do we care?") to question the importance of unit root tests.
I do not think that arguments that stresses that some stable processes look a lot like unit root processes in finite time, makes it less important to test for unit roots. But it is true that there are models, like the ARIMA models where the MA-process nearly cancels with the unit root, in which it really isn't interesting whether the model "truly" is a unit root model. In my opinion no time series model is true in econometrics anyway. If one accepts that these models are just approximations then there isn't really a problem - if the stable process does a good enough job at modeling the data, then let it!

It turns out that one can always decompose a unit root process into the sum of a random walk and a stable process. This is known as the Beveridge-Nelson (1981) composition. The Beveridge-Nelson (BN) decomposition states that any I(1) time series can be decomposed into the sum of a random walk an I(0) time series. In the vector case we can state the results as

Theorem: Beveridge-Nelson decomposition
Any I(1) process y_t can be written as the sum of a random walk s_t and a stable process c_t (s_t and c_t will *not* be independently distributed):

$$y_t \;=\; s_t \;+\; c_t$$

Proof: Since $(1-L)y_t$ is a stable process it has a Wold-decomposition

$$(1-L)y_t \;=\; \psi(L)e_t \;,$$

now write

$$(1-L)y_t \;=\; \psi(L)e_t \;=\; \psi(1)e_t \;+\; (\psi(L)-\psi(1))e_t \;,$$

which is equivalent to

$$(1-L)y_t \;=\; \psi(L)e_t \;=\; \psi(1)e_t \;+\; \psi^{**}(L)e_t \;,$$

where $\psi^{**}(1)=0$. This implies that

$$y_t \;=\; \psi(1)(1-L)^{-1}e_t \;+\; (1-L)^{-1}\psi^{**}(L)e_t \;,$$

which has the form

$$y_t \;=\; s_t \;+\; \psi^{*}(L)e_t \;,$$

where $\psi^*(L)e_t$ is a stable process and s_t is the random walk $\psi(1)(1-L)^{-1}e_t = \psi(1)\sum_s^t e_s$.
QED.

Note that $\psi(1)e_t$ is the long run effect of e_t. For example if $y_0=0$ then $y_T \to \psi(1)e_1 + h_T(e_2,e_3,..)$ for some function h. Also note that the derivation of the Campbell and Mankiw (1987) suggested the use of $\psi(1)$ as a measure of the persistence in y_t; whereas Cochrane (1988) suggests the use of

$$\psi(1)^2\sigma_e^2/\sigma_{\Delta y}^2 \;,$$

where $\sigma_{\Delta y}^2$ and σ_e^2 is the variance of Δy_t and e_t, respectively. (Note the importance of the spectral density for Δy at frequency zero). Cochrane refers to his measure as the *size of the unit root*. This may be a slightly unfortunate term, and some researchers prefer not to use it; but you will meet it in the literature. Cochrane's measure is best if you want to measure the importance of the random walk component of a series relative to the stationary part of the series (if Cochrane's measure is very small an ADF test would most likely conclude falsely that the process was stable) whereas Campbell and Mankiw's measure is better if you are a macroeconomist who wants to know the impact in the infinite future of a shock happening today.

9 Cointegration.

The survey by Campbell and Perron (1991) is a very good supplement to this chapter - for further study read Watson's survey for the handbook of econometrics Vol. IV, and for multivariate models use Johansen's (1995) book.

Cointegration theory is definitely the innovation in theoretical econometrics that has created the most interest among economists in the last decade. The definition in the simple case of 2 time series x_t and y_t, that are both integrated of order one (this is abbreviated I(1), and means that the process contains a unit root), is the following:

Definition:

x_t and y_t are said to be cointegrated if there exists a parameter α such that

$$u_t \;=\; y_t - \alpha x_t$$

is a stationary process.

This turns out to be a pathbreaking way of looking at time series. Why? because it seems that lots of lots of economic series behaves that way and because this is often predicted by theory. The first thing to notice is of course that economic series behave like I(1) processes, i.e. they seem to "drift all over the place"; but the second thing to notice is that they seem to drift in such a way that the they do not drift away from each other. If you formulate this statistically you come up with the cointegration model.

The famous paper by Davidson, Hendry, Srba and Yeo (1978), argued heuristically for models that imposed the "long run" condition that the series modeled should not be allowed to drift arbitrarily far from each other.

The reason unit roots and cointegration is so important is the following. Consider the regression

$$y_t = \alpha_0 + \alpha_1 x_t + u_t \ . \tag{2}$$

A: Assume that x_t is a random walk and that y_t is an *independent* random walk (so that x_t is independent of y_s for all s). Then the true value of α_1 is of course 0, but the limiting distribution of $\hat{\alpha_1}$ is such that $\hat{\alpha_1}$ converges to a function of Brownian motions. This is called a **spurious regression**, and was first noted by Monte Carlo studies by Granger and Newbold (1974) and Phillips (1986) analyzed the model rigorously and found expressions for the limiting distribution. One can furthermore show that the t-statistic *diverges* at rate $\sqrt{T}$, so you can not detect the problem from the t-statistics. The problem does reveal itself in typical OLS regression output though - if you find a very high R^2 and a very low Durbin-Watson value, you usually face a regression where unit roots should be taken into account. I usually consider any reported time series regression with R^2 coefficients above (say) .95 with extreme suspicion. Of course, if you test and find unit roots, then you can get standard consistent estimators by running the regression

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta x_t + e_t \ , \tag{3}$$

since you now regress a stationary variable on a stationary variable, and the classical statistical theory applies.

B: Assume that x_t is a random walk and the y_t is another random walk, such that (1) holds for non-zero α_1, but with the error term u_t following a unit root distribution. In this case you still get inconsistent estimates and you need to estimate the relation (2). This may also be called a spurious regression, even though there actually is a relation between x_t and y_t – I don't think there is quite an established way of referring to this situation (which is often "forgotten" in the discussion of case A and Case C).

C: Now assume that (1) holds with a stationary error term. This is exactly the case where x and y are cointegrated. In this case, $\hat{\alpha}_1$ is not only consistent, but it converges to the true value at rate T. We say that the OLS estimater is **superconsistent**. In the case where x_t is a simple random walk and u_t is serially uncorrelated you will find that $T * (\hat{\alpha}_1 - \alpha_1)$ is asymptotically distributed as a $\int_0^1 B_2 dB_1 / (\int_0^1 B_2^2 dt)$, where B_1 and B_2 are independent Brownian motions. This limiting distribution has mean zero, but more importantly the standard t-test is asymptotically normally distributed. In the situation where there is serial correlation or x_t may be correlated with u_s for some s you do not get a symmetric asymptotic distribution of $T * (\hat{\alpha}_1 - \alpha_1)$, and even if the estimate converges at the very fast rate T, this may translate into non-negligible bias in typical macro data samples. In this situation the t-statistic is no longer asymptotically normal. The best way to test is therefore to use the multivariate Johansen estimator (see below), although you can also use the so-called "Fully Modified" estimator of Phillips (the idea is to estimate a correction term, similarly to what is done in the Phillips-Perron unit root tests), or you can allow for more dynamics in the relation (1). A good place to start reading about this issue is the book by Banerjee, Dolado, Galbraith, and Hendry (1993).
Note that in the situation where x_t and y_t are cointegrated the regression (2) is still consistent (although you would introduce a unit root in the MA representation for the error term). So the differenced regression is always consistent and one could argue that it would be "safe" to always estimate this relation. The loss of efficiency is, however, very big and most macro time series are so short that the gain in efficiency from running the cointegrating regression can be critical for getting decent results.

A stochastic process is said to be integrated of order p, abbreviated as $I(p)$, if it need to be differenced p times in order to achieve stationarity. More generally x_t and y_t are said to be *co-integrated of order CI(d,p)* if x_t and y_t are both integrated of order d; but there exist an α such that $y_t - \alpha x_t$ is integrated of order d-p. For the rest of this chapter I will only treat the CI(1,1) case, which will be referred to simple as cointegration. Most applications of cointegration methods treats that case, and it will allow for a much simpler presentation to limit the development to the CI(1,1) case.

For many purposes the above definition of cointegration is too narrow. It is true that economic series tend to move together but in order to obtain a linear combination of the series, that is stationary one may have to include more variables. The general definition of co-integration (for the I(1) case) is therefore the following:

Definition *A vector of I(1) variables y_t is said to be cointegrated if there exist at vector β_i such that $\beta_i' y_t$ is trend stationary. If there exist r such linearly independent vectors β_i, i = 1,...,r, then y_t is said to be cointegrated with cointegrating rank r. The matrix $\beta = (\beta_1, ...\beta_r)$ is called the cointegrating matrix.*

Note that $\beta' y_t$ is an r-dimensional vector of trend-stationary variables.
Also note that this definition is symmetric in the variables, i.e. there is no designated left-hand side variable. This is usually an advantage for the statistical testing but of course it makes it harder for the economic intuition.
Note the β_i vectors are individually identified only up to scale since $\beta_i y_t$ stationary implies that $c\beta_i y_t$ is stationary. This of course implies that one can normalize one of the coefficients to one - but only in the case where one is willing to impose the a priori restriction that this coefficient is not zero. As far as the identification of the matrix β is concerned it is also clear that if β is a cointegrating matrix then $\beta F'$ is also a cointegrating matrix for any non-singular matrix F.
Finally note that the treatment of constants and drift terms are suppressed here. One has to consider those for any practical applications of cointegration methods.

9.1 Cointegration in the autoregressive representation

The general VAR(k) model can be written as

$$\Delta y_t \;=\; \Pi y_{t-1} \;+\; \sum_{j=1}^{k-1} \Gamma_j \Delta y_{t-j} \;+\; e_t \;,$$

as considered earlier.
If Π is equal to zero this means that there is no cointegration. This is the model that is implicit in the Box-Jenkins method. The variables may be I(1); but that can easily be "cured" by taking differences (in order to achieve the usual asymptotic distribution theory).
If Π has full rank then all y_t must be stationary since the left hand side and the other right hand side variables are stationary (since we limit ourselves to variables that are either I(0) or I(1)).
The most interesting case is when Π has less than full rank but is not equal to zero. This is the case of cointegration. In this case Π can be written as $\Pi = \alpha\beta'$ (yes, this β corresponds to the cointegration matrix introduced above), where α and β are $n \times r$ matrices. Note that α and β are only identified up to non-singular transformations since $\Pi = \alpha\beta' = \alpha F^{-1}(\beta F')'$ for any non-singular F. This lack of identification can sometimes render results from multivariate cointegration analysis impossible to interpret and finding a proper way of normalizing β (and thereby α) is often the hardest part of the work. α can be interpreted as a "speed of adjustment towards equilibrium".

9.2 Cointegration in the moving average representation

The multivariate Wold-representation states that the stationary series Δy_t can be written as

$$(1-L)y_t \;=\; \Psi(L)e_t \;,$$

which, by the Beveridge-Nelson decomposition, can be written as

$$(1)\quad y_t \;=\; \Psi(1)S_t \;+\; \Psi^*(L)e_t \;,$$

where S_t is the n-dimensional random walk $S_t = \sum_{s=1}^{t} e_s$ and $\Psi^*(L) = (1-L)^{-1}(\Psi(L) - \Psi(1))$. Now βy_t is stationary in the case of cointegration, so that

$$\beta y_t \;=\; \beta\Psi(1)S_t \;+\; \beta\Psi^*(L)e_t \;,$$

is stationary, which implies that $\beta\Psi(1)S_t$ is equal to 0. This gives another characterization of cointegration that may be useful for testing.

One can show that the representation (1) can be reformulated in the case of cointegration as

$$y_t \;=\; \tilde{\Psi}S_t^* \;+\; \Psi^*(L)e_t \;,$$

where S_t^* is the (n-r)-dimensional random walk. This is called a common trend representation in Stock and Watson 1988, and this representation can also be used as the basis for cointegration tests (some of which are included in coint package for GAUSS).

9.3 Testing for cointegration

9.4 The Engle-Granger test

The most well known test, suggested by Engle and Granger (1987) (sometimes known as the EG test) is to run a static regression (after first having verified that y_t and x_t both are I(1))

$$y_t \;=\; \theta' x_t \;+\; e_t \;,$$

where x_t is one- or higher-dimensional. The asymptotic distribution of θ is not standard, but the test suggested by Engle and Granger was to estimate $\hat{\theta}$ by OLS and the test for unit roots in

$$\hat{e}_t \;=\; y_t \;-\; \hat{\theta}' x_t \;.$$

Note, that since the unit root tests test the null-hypothesis of a unit root, most cointegration tests test the **Null of no cointegration**. Unfortunately the limiting distribution of for example the

t-test, does not have the limiting distribution tabulated by Dickey and Fuller. The limiting distribution does, however, resemble the Dickey-Fuller distribution even though you need a separate table for each dimension of the regressor. Typically, you will allow for dynamics in the residual and perform the equivalent of the ADF test (using the slightly different critical values in this case). Such a procedure is usually called a Cointegration ADF test, abbreviated CADF-test. Engle and Granger (1987) compared different tests and recommended the CADF test. They supplied critical values based on Monte Carlo simulations for the case of just one regressor. Engle and Yoo (1987) extend those tables to the case of more than one regressor, and MacKinnon (1991) has the most complete tables available so far. You can find the critical values for this residual based test in Hamilton Table B.9.

New tests for unit roots in residuals from a potentially cointegrating relation (like the Phillips-Perron tests) have been suggested since the publication of Engle and Granger (1987) and critical values have been simulated for some of those (see Phillips and Ouliaris (1990) for critical values for the PP test - these values are built into the COINT package), but it seems that the CADF test stands up pretty well. Again, you have to be careful if the series contains trends. If the x_t series contain a trend (or may contain a trend) then you should be careful to include a trend in the cointegrating regression, otherwise the asymptotic critical values will be different. In the case of a one-dimensional x_t, that include a deterministic trend, a regression of y_t on x_t that does not include the trend will give you an asymptotically normal coefficient (this is not too surprising since a deterministic trend always will dominate a unit root trend). Bruce Hansen's article in Journal of Econometrics (Hansen (1992)) treats this topic in more detail. Also note that Campbell and Perron (1991) refer to the case where there is a deterministic trend in y_t and x_t, but not in $y_t - \theta' x_t$ as "deterministic cointegration". If you include a trend in the CADF test, as I suggested, you use (e.g.) Table IIc in Phillips and Ouliaris (1990). Hamilton p 596-7 suggests (based on Hansen (1992)) a slightly more efficient test where you do not include a time trend in the regression – I find the biggest drawback of this strategy that this test is not invariant to whether there is a trend in the data or not.

9.4.1 Estimation of the parameters in case of cointegration

NOTE: Be aware that the issue of efficient *estimation* of parameters in cointegrating relationships is quite a different issue from the issue of *testing* for cointegration.

Engle and Granger (1987) suggested the following simple two-step estimator (which is not efficient and therefore *not* recommended). First estimate the static cointegrating relationship $y_t = \theta' x_t + e_t$, then define $z_t = y_t - \hat{\theta}' x_t$, and finally estimate the **error correction model**

$$\Delta y_t = A_1(L)\Delta y_{t-1} + A_2(L)\Delta x_{t-1} + \gamma z_t + e_t \ .$$

The fact that $\hat{\theta}$ is super-consistent implies that the parameters of the lag polynomials have the same (asymptotically normal) distribution as they would have if θ had been known.

9.5 The Johansen ML estimator

The best way of testing for unit roots is by using the system ML estimator of Johansen (1988,1991) is a test for cointegration restrictions in a VAR representation. "Johansen" estimation is treated in much detail in the book by Johansen (1995). This estimator also gives you asymptotically efficient estimates of the cointegrating vectors (the β's) and of the adjustment parameters (the α's).

"Johansen's method" is the maximum likelihood estimator of the so-called reduced rank model. We start with the AR(k) model

$$\Delta y_t \;=\; \mu \;+\; \Gamma_1 \Delta y_{t-1} \;+\; ... \;+\; \Gamma_{k-1} \Delta y_{t-k+1} \;+\; \Pi y_{t-k} \;+\; e_t \;,$$

which under the assumption of cointegration of order k can be written as

$$\Delta y_t \;=\; \mu \;+\; \Gamma_1 \Delta y_{t-1} \;+\; ... \;+\; \Gamma_{k-1} \Delta y_{t-k+1} \;+\; \alpha\beta' y_{t-k} \;+\; e_t \;, \tag{4}$$

where α and β both have dimension $p \times k$. The number of parameters in the unrestricted model is $p + kp^2 + p(p+1)/2$. Let $Z_{0t} = \Delta y_t,\ Z_{1t} = (\Delta y'_{t-1}, ..., \Delta y'_{t-k+1}, 1)'$ and $Z_{kt} = y_{t-k}$. Define the moment matrices as

$$M_{ij} \;=\; T^{-1} \sum_{t=1}^{T} Z_{it} Z'_{jt} \qquad (i, j = 0, 1, k) \;,$$

We first regress $Z_{it}\ ,\ i = 0, k$ on Z_{1t} and get residuals $R_{it}\ , i = 0, k$. You should think of this as "purging" Δy_t and y_{t-k} of the short-run parameters which can be considered "noise" in the cointegrating relation. We are also purging those variables of their mean, and if you want to include exogenous variables (e.g. dummy variables) they should also be included in the Z_{1t} vector. Denote the residual sum of squares from regressing Z_0 and Z_k on Z_1 as $S_{ij}\ ;\ i, j = 0, k$, in other words

$$S_{ij} = \frac{1}{T} \sum_{t=1}^{T} R_{it} R'_{jt} \;.$$

The maximum likelihood estimator of α and β is a function of these residuals. Johansen (1988,1991) shows that $\hat{\beta}$ can be found from choosing the eigenvectors $(\hat{v_1}, ..., \hat{v_r})$, where $\hat{V} = (\hat{v}_1, ..., \hat{v}_p)$ are the eigenvectors of the equation

$$(*) \quad |\lambda S_{kk} - S_{k0} S_{00}^{-1} S_{0k}| \;=\; 0 \;,$$

normalized such that $\hat{V}' S_{kk} \hat{V} = I$, and ordered in the order of the corresponding eigenvalues such that $\hat{\lambda}_1 > ... > \hat{\lambda}_p > 0$. Make sure you get the intuition here: Cointegration of order r implies that $\lambda_1 \neq 0, ..., \lambda_r \neq 0$, while $\lambda_{r+1} = ... = \lambda_p = 0$. (Since the estimated eigenvalues are continuous random variables they are different from zero (and from each other) with probability 1.) And it is intuitively clear now that you want the eigenvectors corresponding to the non-zero eigenvalues to be the estimators of the cointegrating vectors.

In order to find those eigenvalues pre- and post-multiply the equation above by $S_{kk}^{-1/2}$ (you can use the Cholesky factorization in e.g. GAUSS to get $S_{kk}^{-1/2}$, but the inverse of any matrix X that satisfies $XX' = S_{kk}$ will do) and get the equivalent problem

$$(**) \quad |\lambda - S_{kk}^{-1/2} S_{k0} S_{00}^{-1} S_{0k} S_{kk}^{'-1/2}| \;=\; 0 \;.$$

Note that this is a standard eigenvalue problem that programs like GAUSS can solve directly. The eigenvalues will be the ones that you are looking for. The eigenvectors (u_i, say) that GAUSS gives you will be normalized such that $u_i'u_i = 1$ so you will use $(\hat{v_1}, ..., \hat{v_r}) = S_{kk}^{-1/2}u_1, ..., S_{kk}^{-1/2}u_r$.

In order to give some interpretation of this equation remember that the least squares $\hat{\Pi}$ can be obtained by regressing R_{0t} on R_{kt}, by the Frisch-Waugh theorem. So the least squares estimate of Π is

$$\hat{\Pi} \;=\; S_{kk}^{-1}S_{k0} \; .$$

Now note that

$$S_{kk}^{-1/2}S_{k0}S_{00}^{-1}S_{0k}S_{kk}^{-1/2} \;=\; S_{kk}^{1/2}S_{kk}^{-1}S_{k0}S_{00}^{-1/2}S_{00}^{-1/2}S_{0k}S_{kk}^{-1}S_{kk}^{1/2} \;=\; (S_{kk}^{1/2}\hat{\Pi}S_{00}^{-1/2})(S_{kk}^{1/2}\hat{\Pi}S_{00}^{-1/2})' \; .$$

The intuitively natural approach would be to consider the eigenvalues of $\hat{\Pi}\hat{\Pi}'$ and you can see that this is actually what the Maximum Likelihood algorithm does apart from the fact that $\hat{\Pi}$ has been normalized by pre-multiplying by $S_{kk}^{-1/2}$ and post-multiplying by $S_{00}^{-1/2}$.

The maximized likelihood function is

$$L_{\max}^{-2/T}(r) \;=\; |S_{00}| \prod_{i=1}^{r}(1 - \hat{\lambda_i}) \; .$$

Notice that this is a function of the estimated eigenvalues where all the eigenvalues except the largest r eigenvectors are set equal to zero. So for example the test for one cointegrating vector against no cointegrating vectors consist of testing whether the largest eigenvalue is significantly different from zero. Johansen further finds

$$\hat{\alpha} \;=\; S_{0k}\hat{\beta} \; ,$$

$$\{\hat{\Gamma}_1, ..., \hat{\Gamma}_{k-1}, \hat{\mu}\} \;=\; (M_{01} \;-\; \hat{\alpha}\hat{\beta}'M_{k1})M_{11}^{-1} \; ,$$

and

$$\hat{\Lambda} \;=\; S_{00} - \hat{\alpha}\hat{\alpha}' \; .$$

The likelihood ratio test statistic H for the hypothesis that $\Pi = \alpha\beta'$ is of rank r against the unrestricted model where Π has full rank p is

$$H \;=\; -2\ln(Q) \;=\; -T\sum_{i=r+1}^{p}\ln(1 - \hat{\lambda}_i) \; .$$

Note that the Null hypothesis here is that there are $(p - r)$ unit roots. This corresponds to the simple residual based test previously, where we have $p = 2$ (if the X variable is one dimensional), and we test for 1 cointegrating relation, the null is then that there are 2 unit roots. This test statistic is often referred to as the "trace"-statistic, see e.g. Johansen and Juselius (1992). Note that this statistic is expected to be close to zero if there is at most r (linearly independent) cointegrating vectors. Another test that is often used it the "$\lambda - max$" test which looks at $-T\ln(1 - \hat{\lambda}_{r+1})$ - the idea being that if the $(r + 1)$th eigenvalue can be accepted to be zero, then all the smaller eigenvalues can also. This test is a test of $r+1$ cointegrating vectors against r cointegrating vectors.

The asymptotic distribution of the likelihood ratio test is a functional of multivariate Brownian motion (Johansen (1991)), and is tabulated for values of p up to 11 in Osterwald-Lenum (1992) and reproduced in Hamilton – Table B. 10. The case that allows for a a deterministic trend in the variables is the one that you will "normally" use – this is denoted "Case 3" in Table B. 10.

Often you do not really want to test whether there is (say) 3 cointegrating vectors against no cointegrating vectors, rather you want to make a *decision* on to what is the *number of cointegrating vectors*. In the situation where you directly want to test $r+1$ cointegrating vectors against r cointegrating vectors you should of course use the "$\lambda - max$" test, but this test will not give you a consistent way of deciding the cointegration rank. A consistent (in the sense that you with probability 1 will not underestimate the number of cointegrating vectors) way to do this, using the trace test, is to *start by testing for zero* cointegrating vectors. (I.e. if your system is 4 dimensional, you compare the test statistic $-T\sum_{i=1}^{4}\ln(1-\hat{\lambda}_i)$ to the row labelled 4 in Hamilton Table B.10). If you reject zero cointegrating vectors, you then test for (at most) 1 cointegrating vectors. (In the 4-dimensional case, you compare the test statistic $-T\sum_{i=2}^{4}\ln(1-\hat{\lambda}_i)$ to the row labelled 3 in Hamilton Table B.10). If this is not rejected you stop and decide that $r=1$ - if you reject this you move on until you can not longer reject and stop there. See Johansen (1992b) for further details.

Even though there is a constant in the error correction representation (eqn. (**??**)), this may not translate into a deterministic trend in y_t. Note that this is not the same as what Campbell and Perron (1992) refer to as "deterministic cointegration", namely the case where there is trend in y_t but no trend in $\beta' y_t$. Johansen (1991) derives the likelihood ratio test (which we will denote H^*) for reduced rank in the case where there is a constant in the ECM but no trend in y_t, see Johansen (1991) or Johansen (1995) for the full explanation. Johansen (1992b) discusses how to obtain a consistent test for the number of stochastic trends and for trend in y_t at the same time. See Johansen (1991) for the derivation of the maximum likelihood estimator when there may or may not be trend. It turns out to be very convenient to program the Maximum Likelihood estimator in this case: all you have to do is to move the vector of ones in to Z_{kt} and delete if from Z_{1t}. (The Johansen (1991) article also has the most readable proof of the Granger representation theorem in my opinion).

There are two drawbacks of the Johansen method. One is that it takes a little getting used to interpreting the results and formulating hypotheses in this setting. In the VAR system all variables are treated symmetrically, as opposed to the standard univariate models that usually have a clear interpretation in terms of exogenous and endogenous variables. The other drawback of the VAR system is one has to model all the variables at the same time, which will be a problem if the relation for some variable is flawed. This may give bias in the whole system and one may have been better of conditioning on that variable. Further, the multidimensional VAR model uses many degrees of freedom

9.6 Hypothesis testing in the Johansen model

This is just a very short introduction - Bannerjee et al. p. 276-277 have a slightly longer (but also much too short for the purpose of really learning it) introduction and Hamilton p. 648-650 has

a discussion that is also very brief. The best place to look for more results is the Johansen and Juselius (1992) paper, which explains the method and uses them on an actual economic model, or you may look in Johansen (1995), chapter 7. Here is just an example. Assume that your system consists of 3 variables (y_{1t}, y_{2t}, y_{3t}) and that you have found that there are 2 cointegrating vectors. In other words you have an estimate of β where

$$\begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{pmatrix} .$$

Assume then that you have a long-run hypothesis of the form $y_{1t} - y_{3t} = 0$ (e.g. the PPP-model has this form). This would be the case if for example $\beta_{11} = -\beta_{13}$ and $\beta_{12} = 0$. Note that the model only identifies the parameters up to scale. But even worse the β matrix is only identified up to rotations so it is not obvious how one can do a Wald test. However a Likelihood Ratio test will give you a standard χ^2-test (if you condition on cointegration of order 2). Under the alternative you have to parameterize β as (H, ψ), where ψ is a 3×1 vector of parameters, and

$$H = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} .$$

Johansen and Juselius explain how one can estimate models with restrictions of this and similar forms (and also how to impose linear restrictions on α). (It is actually surprisingly easy to modify the program that does the "standard" Johansen ML-estimator to impose these restrictions). I will not go further into this here, but notice that the hypoteses have to formulated in the form of linear relationships rather than as explicit null-restrictions. I need to explain how to find the degrees of freedom (df) for such a hypothesis. Theoretically we can normalize the β vector to

$$\begin{pmatrix} 1 & 0 \\ \beta_{21} & 1 \\ \beta_{31} & \beta_{32} \end{pmatrix} ,$$

so there are 3 df in the β matrix. The reason I use the word "theoretically" is that you may have to re-order the rows since we have not imposed the restriction that any particular β_{ij} is equal to zero. (You may have a model that implies such a restriction, but the general ML estimator does not do that and Johansen himself is quite critical against imposing such a restriction *a priori*). In the restricted model there are 2 df (since you can normalize one of the coefficients (but again you may not know which one *a priori*) to be 1. So the "PPP" hypothesis that one of the cointegration vectors is $(1, 0, -1)$ can then be tested against the $\chi^2(1)$ distribution. The null hypothesis is estimated by modifying the ML-algorithm for the unrestricted model - I will not go into the details, see Johansen and Juselius (1992), but it is not exceedingly hard to do.

9.6.1 "Common Trends"

Following Johansen (1988,1991) one can choose a set of vectors $\beta_\perp$ such that the matrix $\{\beta, \beta_\perp\}$ has full rank and $\beta'\beta_\perp = 0$. In other words the $p \times (p-r)$ matrix $\beta_\perp$ is orthogonal to the matrix β and the columns of $\beta_\perp$ are orthogonal to the columns of β. The vectors $\beta'_\perp y_t$ constitute the noncointegrated part of the time series y_t. We call $\beta_\perp$ the *common trends loading matrix* and will refer to the space spanned by $\beta'_\perp y_t$ as the *unit root space* of the series y_t. See also Stock and Watson (1988).

9.7 Multicointegration

"Multicointegration" denotes the case where there is cointegration between processes of different order of integration, f.ex. where z_t is I(2), x_t is I(1), and there exist γ such that $z_t - \gamma x_t$ is I(0). We will not have time to go into the theory of multicointegration, but you should now it exists and where to look in the literature. Some models imply a relation between I(2) and I(1) variables, and by extending the ideas of cointegration for I(1) variables above, these models can be analyzed. See Granger and Lee (1990) for a simple introduction, and Johansen (1992c) and Johansen (1992d) for a comprehensive treatment. Also see Stock and Watson (1992.

9.8 Asymptotically efficient single equation methods

The simple two-step estimator of Engle and Granger is not asymptotically efficient; but recently several asymptotically efficient single equation methods have been proposed. Phillips (1991b) suggests a regression in the spectral domain, Phillips and Loretan (1991) suggests a non-linear error-correction estimation, Phillips and Hansen (1990) suggests an instrumental regression with a correction a la Phillips-Perron, Saikkonen (1991) suggests a simple trick of including leads as well as lags in the lag-polynomials of the error correction model in order to achieve asymptotic efficiency, Saikkonen (1992) suggests a simple GLS type estimator, whereas Park's (1991) CCR estimator transforms the data so that OLS afterwards gives asymptotically efficient estimators, and finally Engle and Yoo (1991) suggest a 3 step estimator starting from the static Engle-Granger estimation. From all of those estimators it is possible to obtain simple t-values for the short term adjustment parameters.
The routines for implementing these single equation procedures are all available in the COINT package for GAUSS, except Engle and Yoo (1991). NOTE, however, that it is not obvious what you are estimating if the system contains more than one cointegrating relation. This makes a very strong case for using the Johansen test in the case of a higher dimensional system, where you rarely can rule out the possibility of more than one conintegrating vector *a priori*.

10 GMM.

Generalized Method of Moment (GMM) estimation is the only other development in econometrics in the 80ies that might threaten the position of cointegration for the number one spot on the hit parade. The path breaking articles being Hansen (1982) and Hansen and Singleton (1982).

Until quite recently GMM theory was quite inaccessible, but the surveys by Hall (1992) and Ogaki (1992) has made the theory much easier to get into. Davidson and MacKinnon (1993) also has a quite comprehensive chapter on GMM. For the harder theory you should still consult Hansen (1982) or Gallant (1987).

I think that one can claim that there wasn't that much material in Hansen (1982) that was not known already (although the article definitely was not redundant); but the demonstration in Hansen and Singleton (1982), that this allowed for the estimation of non-linear rational expectations models, that could not be estimated by other methods, really catapulted Hansen and Singleton to major fame. We will start by reviewing linear instrumental variables estimation, since that will contain all the ideas and intuition for the general GMM estimation.

10.1 Linear IV estimation

Consider the following simple model

$$(1) \quad y_t = x_t\theta \ + \ u_t \ , \ t \ = 1, ..., T$$

where y_t and u_t scalar and x_t is $1 \times K$. NOTE from the beginning that even though I use the index "t" - indicating time, that GMM methods are applicable, and indeed much used, in cross sectional studies.
In vector form the equation (1) can be written

$$(2) \quad Y = X\theta \ + \ U \ ,$$

in the usual fashion. If x_t and u_t are potentially correlated, one will obtain a **consistent** estimator by instrumental variables (IV) estimation. The idea is to find a $1 \times L$ vector z_t that is as highly correlated with x_t as possible and at the same time is independent of u_t - so if x_t is actually uncorrelated with u_t you will use x_t itself as instruments - in this way all the simple estimators that you know, like OLS, are special cases of GMM-estimation. If Z denotes the $T \times L$ $(L \leq K)$ vector of the z-observations then we get by premultiplying (2) by Z that

$$Z'Y = Z'X\theta + Z'U \ .$$

If we now denote $Z'Y$ by $\tilde{Y}$, $Z'X$ by $\tilde{X}$, and $Z'U$ by $\tilde{U}$ then the system has the form

$$\tilde{Y} \ = \ \tilde{X} \ + \ \tilde{U} \ ,$$

which corresponds to a standard OLS formulation with L observations. Now the standard OLS estimator of θ is

$$\hat{\theta} \;=\; (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \;,$$

which is consistent and unbiased with variance

$$Var(\tilde{U}) \;=\; Z'Var(U)Z \;.$$

For simplicity let us now consider drop the tilde's, and just remember that the system (of the form (2)) often will have been obtained via the use of instrumental variables. (Most of the GMM-literature uses very sparse notation, which is nice when you are familiar with it, but makes it hard to get started on).
If U does not have a variance matrix that is proportional to the identity matrix the OLS estimator is not efficient. Remember that the OLS estimator is chosen to minimize the criterion function

$$U'U \;=\; (Y - X\theta)'(Y - X\theta) \;.$$

To obtain a more **efficient** estimator than the OLS estimator we have to give different weights to the different equations. Assume that we have given a **weighting matrix** W (the choice of weighting matrices is an important subject that we will return to) and instead choose $\hat{\theta}$ to minimize

$$U'WU \;=\; (Y - X\theta)'W(Y - X\theta) \;,$$

or (in the typical compact notation)

$$\hat{\theta} \;=\; argmin_{\theta}U'WU \;.$$

In this linear case one can then easily show that $\hat{\theta}$ is the GLS-estimator

$$\hat{\theta} \;=\; (X'WX)^{-1}X'WY \;.$$

Let the variance of U be denoted Ω and we find that $\hat{\theta}$ have variance

$$var((X'WX)^{-1}X'WU \;=\; (X'WX)^{-1}X'W\Omega WX(X'WX)^{-1} \;.$$

We want to choose the weighting matrix optimally, so as to achieve the lowest variance of the estimator. It is fairly obvious that one will get the most efficient estimator by weighing each equation by the inverse of its standard deviation which suggests choosing the weighting matrix Ω^{-1}. In this case we find by substituting Ω^{-1} for W in the previous equation that

$$var((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}U \;=\; (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \;=\; (X'\Omega^{-1}X)^{-1} \;.$$

We recognize this as the variance of the GLS estimator. Since we know that the GLS estimator is the most efficient estimator we have indeed shown Ω^{-1} is the optimal weighting matrix.
For practical purposes one would have to do a 2-step estimation. First perform a preliminary estimation by OLS (for example), then estimate Ω (from the residuals), and perform a second

step using this estimate of Ω to perform "feasible GLS". This is asymptotically fully efficient. It sometimes can improve finite sample performance to iterate one step more in order to get a better estimate of the weighting matrix.

The derivations above illustrate all the concepts of GMM. Personally I always guide my intuition by the GLS model. For the general GMM estimators the formulas look just the same (in particular the formulas for the variance) except that if we consider the nonlinear estimation

$$(1) \quad Y = h(X, \theta) \; + \; U \; , \; t \; = 1, ..., T \; ,$$

then "X" in the GLS-formulas should be changed to $\frac{\partial h}{\partial \theta}$.

In GMM jargon the model would usually be formulated as

$$U \; = \; Y - h(X, \theta) \; ,$$

or more often as

$$(**) \quad U \; = \; f(\tilde{X}, \theta) \; ,$$

(where $\tilde{X} = Y, X$ and $f(\tilde{X}, \theta) = Y - X\theta$. The later - very compact - notation is the one that is commonly used in the GMM literature and we will follow it here. We again drop the tilde and denote all the variables by X. It is typical for the newer methods (typically inspired from statistics) that the variables are treated symmetrically.
In the language of GMM the whole model is summarized by the **orthogonality conditions**:

$$EU \; = \; 0 \; ,$$

or (when you want to be really explicit!):

$$EU(X, \theta) \; = \; 0 \; .$$

Here you should thing of U as being a theoretical model. It is not quite explicit here whether we think of U as equations that have been premultiplied by instrument vectors or not. In rational expectations models, the theory often implies which variables will be valid instruments; but this is not always so. For the statistical development the terse notation is good; but in applications you will of course have to be more explicit.

Before developing the general theory for non-linear models let us look at a famous example.

10.2 Hansen and Singleton's 1982 model

The model in Hansen and Singleton (1982) is a simple non-linear rational expectations representative agent model for the demand for financial assets. The model is a simple version of the model of Lucas (1978), and here the model is simplified even more in order to highlight the structure. Note that the considerations below are very typical for implementations of non linear rational expectations models.

We consider an agent that maximize a time-separable von Neumann-Morgenstern utility function over an infinite time horizon. In each period the consumer has to choose between consuming or investing. It is assumed that the consumers utility index is of the constant relative risk aversion (CRRA) type. There is only one consumption good (as in Hansen and Singleton) and one asset (a simplification here).
The consumers problem is

$$\text{Max } E_t [\sum_{j=0}^{\infty} \beta^j \frac{1}{\gamma} C_{t+j}^{\gamma}]$$

$$\text{s.t. } C_{t+j} + I_{t+j} \leq r_{t+j} I_{t+j-1} + W_{t+j} ; \quad j = 0, 1, .., \infty$$

where E_t is the comsumer's expectations at time t and

C_t	:	Consumption
I_t	:	Investment in (one-period) asset
W_t	:	Other Income
r_t	:	Rate of Return
β	:	Discount Factor
γ	:	Parameter of Utility Function

If you knew how C_t and I_t was determined this model could be used to find r_t (which is why it called an asset pricing model), but here we will consider this optimization problem as if it was part of a larger unknown system. Hansen and Singleton's purpose was to estimate the unknown parameters (β and γ), and to test the model.

The first order conditions (called the "Euler equation") for maximum in the model is that

$$C_t^{\gamma-1} = \beta E_t [C_{t+1}^{\gamma-1} r_{t+1}] .$$

The model can not be solved for the optimal consumption path and the major insight of Hansen and Singleton (1982) was that knowledge of the Euler equations are sufficient for estimating the model.
The assumption of rational expectations is critical here - if we assume that the agents expectations at time t (as expressed through E_t corresponds to the true expectations as derived from the probability measure that describes that actual evolution of the variables then the Euler equation can be used to form the "orthogonality condition"

$$U(C_t, \theta) = \beta C_{t+1}^{\gamma-1} r_{t+1} - C_t^{\gamma-1} ,$$

where $E_t U = 0$ (why?), where we now interpret E_t as the "objective" or "true" conditional expectation. Note that $E_t U = 0$ implies that $EU = 0$ by the "law of iterated expectations", which is all that is needed in order to estimate the parameters by GMM. The fact that the *conditional* expectation of U is equal to zero can be quite useful for the purpose of selecting instruments. In

the Hansen-Singleton model we have one orthogonality condition and that is not enough in order to estimate two parameters (more about that shortly), but if we can find two or more independent instrumental variables to use as instruments then we effectively have more than 2 orthogonality conditions.
We denote the agents information set at time t by Ω_t. Ω_t will typically be a set of previous observations of economic variables $\{z_{1t}, z_{1t-1}, ...; z_{2t}, z_{2t-1}, ...; z_{Kt}, z_{Kt-1}, ...\}$. (Including C_t, and I_t among the z's. Then any variable in Ω_t will be a valid instrument in the sense that

$$E[z_t U(C_t, \theta)] \ = \ 0$$

for any z_t in Ω_t. Notice that z_t here denotes any valid instrument at time t, for example z_t could be z_{1t-3} - this convention indexing the instruments will prove quite convenient. The $E[.,.]$ operation can be considered an inner product, so this equation is really the origin of the term orthogonality conditions. For those of you who want to see how this can be developed rigorously, see the book by Hansen and Sargent (1991).

10.3 The GMM estimator

Here we will develop the general theory for GMM-estimation - it is probably a good idea to understand the previous sections fully before reading this.
We have given an economic model

$$u_t(x_t, \theta) \ ,$$

where the vector u_t satisfies

$$Eu_t(x_t, \theta) \ = \ 0 \ .$$

Also assume that we have given a vector z_t of instruments, such that

$$Eu_t \otimes z_t = 0 \ .$$

We will use the notation

$$f_t \ = \ u_t \otimes z_t \ ,$$

where f_t now has dimension $K \times 1$, say. Here we have suppressed the explicit dependence on the underlying series, but when we will use $f_t(\theta)$ rather than f_t when it is important to make the dependence on the parameter explicit.
Now comes the most important feature of GMM. We will work not with g_t itself; but with the *time average* of g_t. Define

$$g_T \ = \ \frac{1}{T} \sum_{t=1}^{T} f_t \ .$$

We will use the notation g_T or $g_T(\theta)$, but from now on the dependence of g_T on the underlying series will be implicit. The GMM estimator will be the estimator that makes $g_T(\theta)$ as close to zero as possible. Notice that g_T is the empirical first moment of the series g_t which is why the estimator is called a moment estimator. Also note that the standard idea of moment estimation, which

consists of equating as well as possible a series of moments. This would be achieved by choosing $g_T' = [\overline{x}_T - Ex_t, \overline{x^2}_T - E\{x_t^2\}, ..., \overline{x^K}_T - E\{x_t^K\}]$. (The moments used for the GMM-estimator in Ho, Perraudin, and Sørensen (1992) are of that form).
We now define the **GMM-estimator** as

$$\hat{\theta} \ = \ \text{argmin}_\theta \ g_T' W_T g_T \ ,$$

where W_T is a weighting matrix that (typically) depends on T such that there exist a positive definite matrix W_0, such that $W_T \to W_0$ $(a.s.)$. The latter condition allows us to let the weighting matrix be dependent on an initial consistent estimator, which is very important since the optimal GMM estimator will be a two step estimator, just as in the GLS-case above.

10.4 Asymptotic theory

We will assume that the series $(x_t', z_t')'$ is **ergodic** which means that

$$\frac{1}{T}\sum_{t=1}^{T} h(x_t) \ \to Eh(x_t)$$

for all functions h(.) (for which the mean is well defined). Notice that the right hand side of the above equation is assumed to not be a function of t.

We will also assume that the series $f_t(\theta)$ satisfies a central limit theorem, i.e. that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\theta) \ \Rightarrow \ \text{N}(0, \Omega) \ ,$$

where $\Omega = E[f_t^2]$ if f_t is not autocorrelated, but in general

$$\Omega = \lim_{t\to\infty} \sum_{j=-J}^{J} E[f_t f_{t-j}'] \ .$$

So intuitively, where we in the GLS model had T (or L in the IV case) normally distributed error terms, we here have K asymptotically normally distributed moment (or orthogonality) conditions.

Let $Df = E\frac{\partial f_t}{\partial \theta}$. One can then show that for any convergent sequence of weighting matrices the GMM-estimator is consistent and asymptotically normal with

$$\sqrt{T}(\hat{\theta} - \theta \ ) \Rightarrow \ \text{N}(0, \Sigma) \ ,$$

where

$$\Sigma \ = \ (Df' W_0 Df)^{-1} Df' W_0 \Omega W_0 Df (Df' W_0 Df)^{-1} \ .$$

Notice that this formula corresponds exactly to the one obtained in the linear case if you substitute X for Df. Of course the reasoning behind the GLS estimator also carries over and the *optimal*

GMM-estimator is the one where $W_T \rightarrow \Omega^{-1}$ in which case the asymptotic covariance of the GMM-estimator is

$$\Sigma^o \ = \ (Df'\Omega^{-1}Df)^{-1} \ .$$

Notice that this is the optimal estimator for a **given set of instruments**. The problem of finding the best instruments is much harder and no satisfactory solution exists to that problem in general (although often for special cases, like the OLS model).

10.5 Estimation of the asymptotically optimal weighting matrix

To find an estimate of the optimal weighting matrix one has to start with a consistent estimate

$$\hat{\theta_1} \ = \ \text{argmin}_\theta \ g_T' I g_T \ ,$$

where we have used the identity matrix as the weighting matrix for the first stage estimator. One can choose any initial weighting matrix but it is very common to use the identity matrix in the first step, and if you use any other matrix as the initial weighting matrix you will be expected to supply some argument for your choice - even if one can use any positive definite matrix and still get a consistent estimate.
Now estimate the auto covariances, given a sample of size T, the same way as we did earlier in the course by

$$C(k) \ = \ \frac{\sum_{t=k}^{T}[(f_t(\hat{\theta_1})(f_{t-k}(\hat{\theta_1})']}{T-k} \ ; \ k = -T, -T+1, ..., T-1, T \ .$$

We can estimate the asymptotic covariance matrix for $\sqrt{T}g_T$ (which corresponds to the spectrum at frequency zero for f_t) by

$$\hat{\Omega} \ = \ \sum_{k=-T}^{T} w(\frac{k}{M(T)}) \ C(k) \ ,$$

where the function $w(.)$ is the weighting function, and $M(T)$ is the bandwidth parameter. (Note that it is not necessary to divide by $\sqrt{2\pi}$ since the GMM estimator is unaffected by a scalar multiplication of the weighting matrix). The Newey-West/Bartlett kernel is very often used, although one can improve a bit on this asymptotically (see Andrews (1991)). More important is the choice of $M(T)$, and there is an approximately asymptotically optimal method available for choosing $M(T)$, see Andrews (1991) and Andrews and Monahan (1991). The basic idea of the choice of $M(T)$ is that there is a trade-off between bias and variance as previously explained, and the bias will be worse the lower $M(T)$ is chosen. This is however most serious if there is high autocorrelation (if there is no autocorrelation $M(T)$ should be chosen as low as possible), so the Andrews selection of $M(T)$ is based on an estimation of the degree of autocorrelation. If that estimate is high then $M(T)$ is chosen to be large and vice versa. We will not go into detail with those methods here, but you should probably apply them in an actual application. I think that whatever their other merits, it is important to have automatic methods for selecting aspects of the estimation, since it reduces fiddling. The more fiddling the more useless the asymptotic t-values etc. are.

When an estimator $\hat{\Omega}_T$ of the asymptotic variance-covariance matrix has been formed, then the asymptotically optimal 2nd stage GMM estimator can be found as

$$\hat{\theta}_2 \;=\; \text{argmin}_\theta \; g_T' \hat{\Omega}_T^{-1} g_T \; .$$

This is the GMM-estimator as it is typically found.

10.6 Testing in a GMM-framework

Hansen (1982) suggested the following test for misspecification: Consider

$$J_T \;=\; T g_T(\hat{\theta}_2)' \hat{\Omega}_T^{-1} g_T(\hat{\theta}_2) \; .$$

If the model is correctly specified this statistic is asymptotically χ^2 distributed with degrees of freedom equal to $K - p$, where p is the number of parameters estimated. So a value that is far out in the tail indicates that the whole model is mis-specified. By the whole model I do not mean that *all* parts of the model are mis-specified; but rather that *some* part of the model is mis-specified - it could be that it was just the instruments that were not pre-determined. This test is known as the **test for overidentifying restrictions** or sometimes as the "Hansen J-test". In Hansen and Singleton (1982) the model was rejected by the J-test, and my subjective impression is that from then on it really became just as acceptable to present an econometric estimation that rejected the model, as one that accepted the model. Exaggerating a bit one could claim that in the earlier period models were *never* rejected; but at least I am strongly convinced that they were not rejected the 5% of the times that they should have been even in the case that they were true.

I often find the J-test useless. Models are never exactly true so the result of the J-test will usually be that it accepts the model (due to lack of power) if the number of observations is low, and rejects the model if the number of observations is high.

More useful (to my taste) are the standard tests for specific restrictions. There exists equivalents of the standard Wald-, LM-, and ML-test in the case of GMM estimation. Note: This is only true in the case where the optimal weighting matrix has been applied. In a case where you apply a non-optimal weighting matrix then there is no equivalent of the ML-test available. (Ho, Perraudin, and Sørensen (1992) is an example of a paper that applies a non-optimal weighting matrix). For the development here I will assume that the optimal GMM-estimator has been estimated - for the more general case, see Gallant (1987).

Consider a test for s nonlinear restrictions

$$R(\theta) \;=\; 0 \; ,$$

where R is a $s \times 1$ vector of functions.
Let DR be $\frac{dR}{d\theta}$ (and we assume that DR is evaluated at the optimal GMM-estimator in the unrestricted model, then the Wald test is

$$T R' [DR(\hat{D} g_T' \hat{\Omega}_T^{-1} \hat{D} g_T)^{-1} DR']^{-1} R \; ,$$

where Dg (and DR if this is dependent on the parameters) are evaluated at the restricted estimator of θ. Sometimes you can find the analytical derivatives, but otherwise you will have to use numerical derivatives, which are easy to evaluate in GAUSS. In this formula $(\hat{Dg}_T'\hat{\Omega}_T^{-1}\hat{Dg}_T)^{-1}$ is our estimator of the variance of $\hat{\theta}$ and when we pre- and post-multiply this by DR we get an estimate of the asymptotic variance of $R(\hat{\theta})$. Let us define

$$\hat{\Sigma} \;=\; (\hat{Dg}_T'\hat{\Omega}_T^{-1}\hat{Dg}_T)^{-1} \;,$$

The LM-test is then

$$LM \;=\; T g_T\,\hat{\Omega}_T^{-1}\hat{Dg}_T\,\hat{\Sigma}\,DR'\,[DR\,\hat{\Sigma}\,DR']^{-1}\,DR\,\hat{\Sigma}\,\hat{Dg}_T'\,\hat{\Omega}_T^{-1}g_T \;,$$

which is pretty ugly looking, so you may want to check against Gallant (1987) on your own to see if you get the same (there is a formula in Ogaki (1992) that I cannot quite get to agree with Gallant's formula and a much simpler looking formula in Davidson and MacKinnon, that I cannot see how they get, so I recommend using the terrible Gallant formula if you do need the LM-test).

Finally the LR-test (of course it should strictly speaking be "LR-type test") is

$$LR \;=\; T[J_T(\theta_2^r) - J_T(\theta_2^u)] \;,$$

where J_T is the criterion function (NB) evaluated at the *same* estimator or Ω and where the superscripts u and r of course indicates that the estimators were found in the unrestricted and the restricted models respectively.

The Wald-, LM-, and LR-test can all be shown to converge in distribution to a χ^2-distribution with s degrees of freedom in the case where the restrictions are true.

11 ARCH and generalizations.

For further or alternative readings, the very up-to-date survey by Bollerslev, Chou, and Kroner (1992) is highly recommended.

Many financial time series exhibit *volatility clustering*, which means that the series have periods where volatility is low and other periods where volatility is high. As econometricians we will understand "volatility" to mean "conditional variance". The conditioning set will always be the behavior of relevant variables up to "time t" - the time when we observe the series. There have recently been developed many different models for data that shows volatility clustering, and it is still a *very* active research area. The main differences between competing models will often be the choice of which variables to condition on, and (as usual) the choice of functional forms. The problems of modeling series with varying volatility is of course well known in econometrics under the heading of heteroskedasticity; but most of the interest in time-varying volatility models comes from finance.
One of the main ideas of asset-pricing is that the variability of an asset should be reflected in its price. One would expect an asset with high variance ("risk") to give a higher return (for investors to want to hold it). With some polishing and generalizations this is the main point of the CAPM-, and APT-models that are common in finance. But standard economic reasoning also says that risk should not be measured as the unconditional variance but rather as the conditional variance, and therefore finance characters are interested in modeling conditional heteroskedasticity in its own right (e.g. for the purpose of pricing options).

11.1 Engle's ARCH model

Consider the following simple scalar model

$$
\begin{aligned}
& y_t = \mu + e_t \ , \ t = 1, ..., T \\
(1) \qquad & e_t = z_t \sigma_t \\
& \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 = \omega + a(L)\epsilon_t^2
\end{aligned}
$$

where $\alpha_1, .., \alpha_q$, μ and ω are scalar parameters to be estimated. z_t is supposed to have mean zero and variance one, and will often (but not always) be assumed to be normally distributed. One has to assume that ω and α_i are all positive in order to obtain positive values for the estimate of the condition variance. In practice you have to assume this by either penalize the likelihood by setting it to a large negative number when negative values are met or by parameterizing it for example as the square of the parameter. I tend to prefer the later method since the former method potentially will create problems because the likelihood then will not be differentiable.

11.2 The GARCH model

Bollerslev (1986) suggested the following natural generalization of the ARCH model. Let the $e_t = \sigma_t$ as before, but now let

$$\sigma_t^2 \;=\; \omega \;+\; \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \;+\; \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$$

which is a natural generalization corresponding to an ARMA model for the variance. This model is called a GARCH(p,q) model. Also in the GARCH model one will restrict the parameters to be positive, which will ensure a positive estimate of the conditional variance (even though I am not sure whether this is also a necessary condition in higher order models). More compactly we write

$$\sigma_t^2 \;=\; \omega \;+\; b(L)\sigma_t^2 \;+\; a(L)\epsilon_t^2 \;.$$

Bollerslev and Engle (1986) looks at the case where the variance process follows the equivalent of an ARIMA model, allowing for unit roots in the lag polynomials. In the case where

$$\alpha_1 \;+\; ... \;+\; \alpha_q \;+\; \beta_1 \;+\; ... \;+\; \beta_p \;=\; 1$$

they refer to the model as an IGARCH(p,q) model.

11.3 The E-GARCH model

One limitation of the GARCH models is that it a priori restricts the shocks to the model to have the same effect on the conditional variance whether the shocks are negative or positive. This may or may not be a reasonable assumption but one would like to be able to test this. The positivity constraints on the parameters can also be viewed as restrictive, since it rules out cyclical behavior in the conditional variance. For those reasons (among others) Nelson (1991) suggests the EGARCH(p,q) model:

$$log(\sigma_t^2) \;=\; \omega \;+\; \sum_{i=1}^{p} \beta_i log(\sigma_{t-i}^2) \sum_{i=1}^{q} \alpha_i (\phi z_{t-i} + \gamma[|z_{t-i}| - E|z_{t-i}|]) \;.$$

In the EGARCH the parameters are not restricted to be positive. Note that the term $|z_{t-i}| - E|z_{t-i}|$ is positive if the error term is larger than its expected value and negative otherwise. One can use other models for $log(\sigma_t^2)$, but for the particular model suggested by Nelson, he shows that the model seems to behave well asymptotically. We will look a little bit at the issue of stationarity of *ARCH models.

11.4 Stationarity of ARCH models

The GARCH model is covariance stationary if $A(1)+B(1) < 1$. It turns out that if $A(1)+B(1) = 1$ then the process is still stationary; but not covariance stationary since the variance is infinite. Notice that this is very different from the ARMA models where strict stationarity and covariance stationarity coincides (if the intial conditions are chosen properly). One can also show for the

standard GARCH model that if ω is equal to zero then the conditional variance of the process will converge to zero almost surely. I will not go into details of the stochastic properties of *ARCH processes, but you should be aware that they can be quite tricky. It is obvious that the form of the ARCH models is chosen to give convenient estimations, and not to give convenient theoretical properties.

11.5 ARCH-M models

As mentioned in the introduction, one of the major motivations for looking at conditional variances is that financial theory says that the expected return on an asset should be correlated with its conditional variance. (The expected return will actually depend on the covariance with other assets as well as on the variance, but we will not go into finance theory here). Therefore one may want to combine the models for σ_t^2, whether one prefers ARCH, GARCH, or whatever, with a model for the mean (f.eks. a mean return). So now one would model

$$y_t \ = \ g(\sigma_t, b) \ + \ e_t \ , \ t \ = 1, ..., T \ ,$$

where b is a parameter. (Of course one will usually also want to include regressors but that is supressed here). The ARCH-M model was first suggested by Engle, Lillien and Roberts (1987).
It is usually more complicated to estimate ARCH-M models, because of the fact that the model for the conditional mean now depends on the conditional variance, making the model a lot more non-linear.

11.6 Estimation of ARCH models

The most commonly used estimation strategy is Maximum Likelihood, with an assumption of normality of the error terms. (One may also use the normal likelihood function without wanting to claim that the error terms are normally distributed, in which case one speaks of Quasi Maximum Likelihood estimation). For financial data this is often not a reasonable assumption and there has been articles in the literature that performs Maximum Likelihood using distributions like the t-distribution, that has heavier tails than the normal distribution.
There are also articles in the literature that estimates ARCH models using GMM.

11.7 Other ARCH models

A lot of research is still being devoted to ARCH models. Some other of the newer research concerns factor-ARCH models, non-parametric ARCH models (you can have a nonparametric representation of the conditional variance of the probability density), multivariate ARCH, and all possible combinations. There is also STARCH (structural ARCH), and threshold ARCH and probably a lot of others. How about a non-parametric multivariate threshold factor-GARCH-M model? (I don't know if anybody has done that one yet). In the paper Ho, Perraudin and Sørensen (1992) we suggest an alternative to ARCH, modeling conditional heteroskedasticity in continuous time, which has some major advantages, but this becomes a bit technical.

%newpage