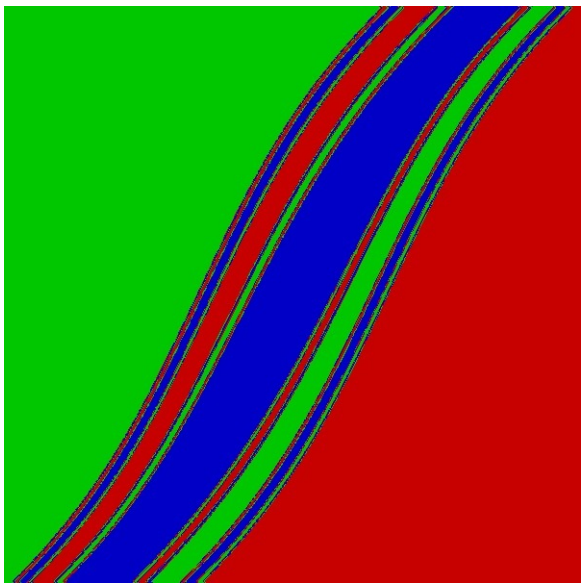
Time-Series Econometrics
A Concise Course



Francis X. Diebold
University of Pennsylvania

Edition 2015
Version 2015.03.22

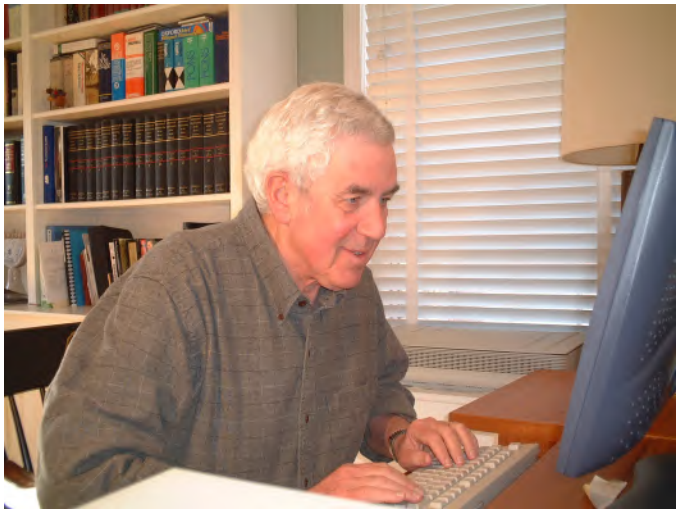
Time Series Econometrics

Time Series Econometrics

A Concise Course

Francis X. Diebold

To Marc Nerlove,
who taught me time series,



and to my wonderful Ph.D. students,
his "grandstudents"

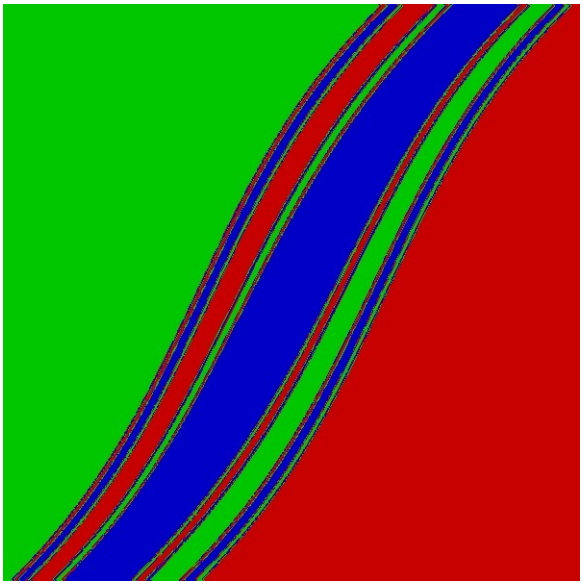
Brief Table of Contents

Detailed Table of Contents

About the Author

Francis X. Diebold is Paul F. and Warren S. Miller Professor of Economics, and Professor of Finance and Statistics, at the University of Pennsylvania and its Wharton School, as well as Faculty Research Associate at the National Bureau of Economic Research in Cambridge, Mass., and past President of the Society for Financial Econometrics. He has published widely in econometrics, forecasting, finance and macroeconomics, and he has served on the editorial boards of numerous scholarly. He is an elected Fellow of the Econometric Society, the American Statistical Association, and the International Institute of Forecasters, and the recipient of Sloan, Guggenheim, and Humboldt fellowships. Diebold lectures actively, worldwide, and has received several prizes for outstanding teaching. He has held visiting appointments in Economics and Finance at Princeton University, Cambridge University, the University of Chicago, the London School of Economics, Johns Hopkins University, and New York University. His research and teaching are firmly rooted in applications; he has served as an economist under Paul Volcker and Alan Greenspan at the Board of Governors of the Federal Reserve System in Washington DC, an Executive Director at Morgan Stanley Investment Management, Co-Director of the Wharton Financial Institutions Center, and Chairman of the Federal Reserve System's Model Validation Council. All his degrees are from the University of Pennsylvania; he received his B.S. from the Wharton School in 1981 and his economics Ph.D. in in 1986. He is married with three children and lives in suburban Philadelphia.

About the Cover



The colorful graphic is by Peter Mills and was obtained from Wikimedia Commons. As noted there, it represents "the basins of attraction of the Gaspard-Rice scattering system projected onto a double impact parameter" (whatever that means). I used it mainly because I like it, but also because it's vaguely reminiscent of a trending time series.

For details see http://commons.wikimedia.org/wiki/File%3AGR_Basins2.tiff. The complete attribution is: By Peter Mills (Own work) [CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons)

Guide to e-Features

- Hyperlinks to internal items (table of contents, index, footnotes, etc.) appear in red.
- Hyperlinks to bibliographic references appear in green.
- Hyperlinks to the web appear in cyan.
- Hyperlinks to external files (e.g., video) appear in blue.
- Many images are clickable to reach related material.
- Additional related materials may appear at http://www.ssc.upenn.edu/~fdiebold. These may include book updates, presentation slides, datasets, and R code.
- Facebook group: Diebold Time Series Econometrics
- Related blog (*No Hesitations*): www.fxdiebold.blogspot.com

Acknowledgments

All media (images, audio, video, ...) were either produced by me (computer graphics using R, original audio/video, etc.) or obtained from the public domain repository at Wikimedia Commons.

List of Figures

List of Tables

Preface

Time Series Econometrics (*TSE*) provides a modern and concise Ph.D.-level course in econometric time series. It can be covered realistically in one semester; indeed I have used the material successfully for many years with first-year Ph.D. students at the University of Pennsylvania.

The elephant in the room is of course Hamilton's *Time Series Analysis*, so let me address it immediately. *TSE* complements it in three key ways. First, *TSE* offers a concise yet precise overview – from the classic early framework of Wold, Wiener, and Kolmogorov, straight through to cutting-edge Bayesian MCMC analysis of non-linear non-Gaussian state space models with the particle filter – and Hamilton's book can be used for more extensive background reading for those topics that overlap.

Second and crucially, however, many of the topics do *not* overlap, as *TSE* treats a variety of more recently-emphasized topics. It stresses Markovian structure throughout, from linear state space, to MCMC, to optimization, to non-linear state space and particle filtering. Bayes features prominently, as do simulation, continuous time, realized volatility, nonparametrics, global optimization, and more.

Finally, *TSE* is in touch with modern computing environments. It uses R throughout, which in this author's opinion is the clear environment of choice for the foreseeable future. Related, *TSE* is generally e-aware, with numerous hyperlinks to internal items, bibliographic references, the internet (web pages, video, etc.), databases, etc.

Francis X. Diebold
Philadelphia

Sunday 22nd March, 2015

Time Series Econometrics

Chapter One

Introduction

1.1 ECONOMIC TIME SERIES AND THEIR ANALYSIS

Any series of observations ordered along a single dimension, such as time, may be thought of as a time series. The emphasis in time series analysis is the study of dependence among the observations at different points in time.[1]

Many economic and financial variables, such as prices, sales, stocks, GDP and its components, stock returns, interest rates and foreign exchange rates, are observed over time; in addition to being interested in the interrelationships among such variables, we are also concerned with relationships among the current and past values of one or more of them, that is, relationships over time.

At its broadest level, time series analysis provides the language for of stochastic dynamics. Hence it's the language of even pure dynamic economic theory, quite apart from empirical analysis. It is, however, a great workhorse of empirical analysis, in "pre-theory" mode (non-structurally "getting the facts straight" before theorizing, always a good idea), in "post-theory" mode (structural estimation and inference), and in forecasting (whether non-structural of structural).

Empirically, the analysis of economic time series is central to a wide range of applications, including business cycle measurement, financial risk management, policy analysis, and forecasting. Special features of interest in economic time series include trends and nonstationarity, seasonality, cycles and persistence, predictability (or lack thereof), structural change, and nonlinearities such as volatility fluctuations and regime switching.

1.2 A PRACTICAL TOOLKIT

1.2.1 Software (and a Tiny bit of Hardware)

Let's proceed from highest level to lowest level.

Eviews is a good high-level environment for economic time-seres analysis. It's a modern object-oriented environment with extensive time series, modeling and forecasting capabilities. It implements almost all of the methods described in this book, and many more.

[1] Indeed what distinguishes time series analysis from general multivariate analysis is precisely the temporal order imposed on the observations.

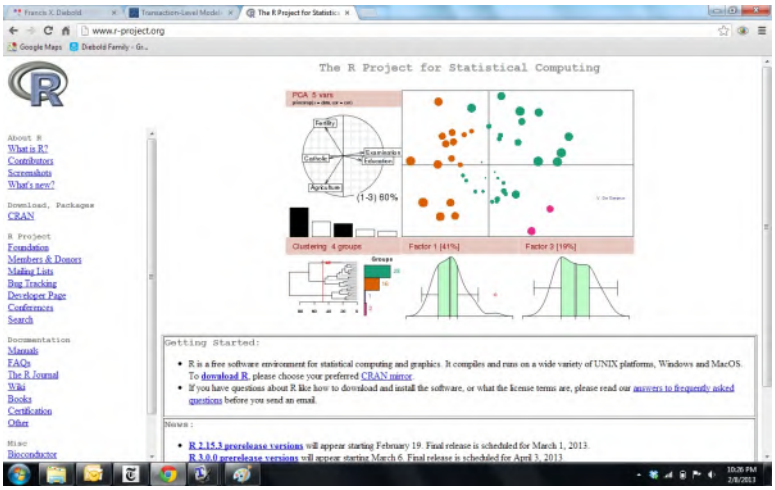


Figure 1.1: The R Homepage

Eviews, however, can sometimes be something of a "black box." Hence you'll also want to have available slightly lower-level ("mid-level") environments in which you can quickly program, evaluate and apply new tools and techniques. R is one very powerful and popular such environment, with special strengths in modern statistical methods and graphical data analysis.[2] R is available for free as part of a massive and highly-successful open-source project. RStudio provides a fine R working environment, and, like R, it's free. A good R tutorial, first given on Coursera and then moved to YouTube, is here. R-bloggers is a massive blog with all sorts of information about all things R.

If you need real speed, such as for large simulations, you will likely need a low-level environment like Fortran or C++. And in the limit (and on the hardware side), if you need blazing-fast parallel computing for massive simulations etc., graphics cards (graphical processing units, or GPU's) provide stunning gains, as documented for example in Aldrich et al. (2011). Actually the real limit is quantum computing, but we're not there yet.

For a compendium of econometric and statistical software, see the software links site, maintained by Marius Ooms at the *Econometrics Journal*.

1.2.2 Data

Here we mention just a few key "must-know" sites. Resources for Economists, maintained by the American Economic Association, is a fine portal to almost anything of interest to economists. It contains hundreds of links to data sources, journals, professional organizations, and so on. FRED (Federal Reserve Economic Data) is a tremendously convenient source for economic data. The National Bureau of Economic Research site has data on U.S. business cycles, and the Real-Time Data Research Center at the Federal Reserve Bank of

[2] Python and Julia are other interesting mid-level environments.

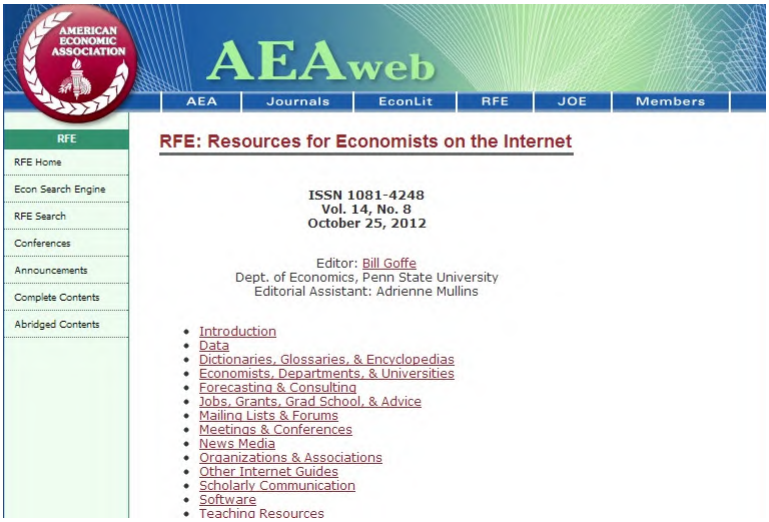


Figure 1.2: Resources for Economists Web Page

Philadelphia has real-time vintage macroeconomic data. Quandl is an interesting newcomer with striking breadth of coverage; it seems to have every time series in existence, and it has a nice R interface.

1.2.3 Markup

Markup languages effectively provide typesetting or "word processing." HTML is the most well-known example. Research papers and books are typically written in LaTeX. MiCTeX is a good and popular flavor of LaTeX, and TeXworks is a good editor designed for LaTeX. knitr is an R package, but it's worth mentioning separately, as it powerfully integrates R and LaTeX./footnoteYou can access everything in RStudio.

Another markup language worth mentioning is Sphinx, which runs under Python. The Stachurchski-Sargent e-book *Quantitative Economics*, which features Python prominently, is written in Sphinx.

1.2.4 Version Control

Git and GitHub are useful for open/collaborative development and version control. For my sorts of small-group projects I find that Dropbox or equivalent keeps me adequately synchronized, but for serious large-scale development, use of git or equivalent appears crucial.

1.3 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Approaches and issues in economic time series analysis.

Consider the following point/counterpoint items. In each case, which do you think would be more useful for analysis of *economic* time series? Why?

- Continuous / discrete
- linear / nonlinear
- deterministic / stochastic
- univariate / multivariate
- time domain / frequency domain
- conditional mean / conditional variance
- trend / seasonal / cycle / noise
- ordered in time / ordered in space
- stock / flow
- stationary / nonstationary
- aggregate / disaggregate
- Gaussian / non-Gaussian

2. Nobel prizes for work involving time series analysis.

   Go to the economics Nobel Prize web site. Read about Economics Nobel Prize winners Frisch, Tinbergen, Kuznets, Tobin, Klein, Modigliani, Friedman, Lucas, Engle, Granger, Prescott, Sargent, Sims, Fama, Shiller, and Hansen. Each made extensive contributions to, or extensive use of, time series analysis. Other econometricians and empirical economists winning the Prize include Leontief, Heckman, McFadden, Koopmans, Stone, Modigliani, and Haavelmo.

1.4 NOTES

- The study of time series of, for example, astronomical observations predates recorded history. Early writers on economic subjects occasionally made explicit reference to astronomy as the source of their ideas. For example, Cournot stressed that, as in astronomy, it is necessary to recognize secular variation that is independent of periodic variation. Similarly, Jevons made clear his approach to the study of short-term fluctuations used the methods of astronomy and meteorology. During the 19th century interest in, and analysis of, social and economic time series evolved into a new field of study independent of developments in astronomy and meteorology. Time-series analysis then flourished. Nerlove et al. (1979) provides a brief history of the field's early development.

- For references old and new, see the "library" of useful books in Appendix A.

Chapter Two

The Wold Representation and its Approximation

2.1 THE ENVIRONMENT

Time series Y_t (doubly infinite)

Realization y_t (again doubly infinite)

Sample path $y_t,\ t\ =\ 1,\ ...,\ T$

Strict Stationarity

Joint cdfs for sets of observations
depend only on displacement, not time.

Weak Stationarity

(Second-order stationarity, wide sense stationarity,
covariance stationarity, ...)

$$Ey_t\ =\ \mu,\ \forall t$$

$$\gamma(t,\ \tau) = E(y_t\ -\ Ey_t)\,(y_{t+\tau}\ -\ Ey_{t+\tau})\ =\ \gamma(\tau),\ \forall t$$

$$0\ <\ \gamma(0)\ <\ \infty$$

Autocovariance Function

(a) symmetric
$\gamma(\tau)\ =\ \gamma(-\tau),\ \forall\tau$

(b) nonnegative definite
$a'\Sigma a\ \geq\ 0,\ \forall a$
where Toeplitz matrix Σ has ij-th element $\gamma(i-j)$

(c) bounded by the variance
$\gamma(0)\ \geq\ |\gamma(\tau)|,\ \forall\tau$

Autocovariance Generating Function

$$g(z) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau)\, z^{\tau}$$

Autocorrelation Function

$$\rho(\tau) \;=\; \frac{\gamma(\tau)}{\gamma(0)}$$

2.2 WHITE NOISE

White noise: $\eta_t \;\sim\; WN(\mu,\; \sigma^2)$ $(serially\ uncorrelated)$

Zero-mean white noise: $\eta_t \;\sim\; WN(0,\; \sigma^2)$

Independent (strong) white noise: $\eta_t \overset{iid}{\sim} (0,\; \sigma^2)$

Gaussian white noise: $\eta_t \overset{iid}{\sim} N(0,\; \sigma^2)$

Unconditional Moment Structure of Strong White Noise

$$E(\eta_t) \;=\; 0$$

$$var(\eta_t) \;=\; \sigma^2$$

Conditional Moment Structure of Strong White Noise

$$E(\eta_t|\Omega_{t-1}) \;=\; 0$$

$$var(\eta_t|\Omega_{t-1}) \;=\; E[(\eta_t \;-\; E(\eta_t|\Omega_{t-1}))^2|\Omega_{t-1}] \;=\; \sigma^2$$

where

$$\Omega_{t-1} \;=\; \eta_{t-1},\; \eta_{t-2},\; ...$$

Autocorrelation Structure of Strong White Noise

$$\gamma(\tau) \;=\; \begin{cases} \sigma^2, \; \tau \;=\; 0 \\ 0, \; \tau \;\geq\; 1 \end{cases}$$

$$\rho(\tau) \;=\; \begin{cases} 1, \; \tau \;=\; 0 \\ 0, \; \tau \;\geq\; 1 \end{cases}$$

An Aside on Treatment of the Mean

In theoretical work we assume a zero mean, $\mu = 0$.

This reduces notational clutter and is without loss of generality.

(Think of y_t as having been centered around its mean, μ,
and note that $y_t - \mu$ has zero mean by construction.)

(In empirical work we allow explicitly for a non-zero mean,
either by centering the data around the sample mean
or by including an intercept.)

2.3 THE WOLD DECOMPOSITION AND THE GENERAL LINEAR PROCESS

Under regularity conditions,
every covariance-stationary process $\{y_t\}$ can be written as:

$$y_t \;=\; \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

where:

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

$$\varepsilon_t \;=\; [y_t \;-\; P(y_t | y_{t-1}, \; y_{t-2}, \; ...)] \;\sim\; WN(0, \; \sigma^2)$$

The General Linear Process

$$y_t \;=\; B(L)\varepsilon_t \;=\; \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \;\sim WN(0,\; \sigma^2)$$

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

Unconditional Moment Structure of the LRCSSP

$$E(y_t) \;=\; E\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) \;=\; \sum_{i=0}^{\infty} b_i E\varepsilon_{t-i} \;=\; \sum_{i=0}^{\infty} b_i \cdot 0 \;=\; 0$$

$$var(y_t) \;=\; var\left(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\right) \;=\; \sum_{i=0}^{\infty} b_i^2 var(\varepsilon_{t-i}) \;=\; \sigma^2 \sum_{i=0}^{\infty} b_i^2$$

Conditional Moment Structure

$$E(y_t|\Omega_{t-1}) \;=\; E(\varepsilon_t|\Omega_{t-1}) + b_1 E(\varepsilon_{t-1}|\Omega_{t-1}) + b_2 E(\varepsilon_{t-2}|\Omega_{t-1}) + ...$$

$$(\Omega_{t-1} \;=\; \varepsilon_{t-1},\; \varepsilon_{t-2},\; ...)$$

$$= \; 0 + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ... \;=\; \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}$$

$$var(y_t|\Omega_{t-1}) \;=\; E[(y_t \;-\; E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}]$$

$$= \; E(\varepsilon_t^2|\Omega_{t-1}) \;=\; E(\varepsilon_t^2) \;=\; \sigma^2$$

(These calculations assume strong WN innovations. Why?)

Autocovariance Structure

$$\gamma(\tau) \;=\; E\left[\left(\sum_{i=-\infty}^{\infty} b_i \varepsilon_{t-i}\right)\;\left(\sum_{h=-\infty}^{\infty} b_h \varepsilon_{t-\tau-h}\right)\right]$$

$$= \sigma^2 \sum_{i=-\infty}^{\infty} b_i b_{i-\tau}$$

(where $b_i \equiv 0$ if $i < 0$)

$$g(z) \;=\; \sigma^2\; B(z)\; B(z^{-1})$$

2.4 APPROXIMATING THE WOLD REPRESENTATION

2.4.1 The $MA(q)$ Process

(Obvious truncation)

Unconditional moment structure, conditional moment structure, autocovariance functions, stationarity and invertibility conditions

2.4.2 The $AR(p)$ Process

(Stochastic difference equation)

Unconditional moment structure, conditional moment structure, autocovariance functions, stationarity and invertibility conditions

2.4.3 The $ARMA(p, q)$ Process

Rational $B(L)$, later rational spectrum, and links to state space.

Unconditional moment structure, conditional moment structure, autocovariance functions, stationarity and invertibility conditions

2.5 WIENER-KOLMOGOROV-WOLD EXTRACTION AND PREDICTION

2.5.1 Extraction

2.5.2 Prediction

$$y_t \;=\; \varepsilon_t \;+\; b_1\, \varepsilon_{t-1} \;+\; ...$$

$$y_{T+h} \;=\; \varepsilon_{T+h} \;+\; b_1\, \varepsilon_{T+h-1} \;+ ... + b_h \varepsilon_T \;+\; b_{h+1} \varepsilon_{T-1} \;+ ...$$

Project on $\Omega_T = \{\varepsilon_T,\; \varepsilon_{T-1},\; ...\}$ to get:

$$y_{T+h,T} \;=\; b_h\,\varepsilon_T \;+\; b_{h+1}\,\varepsilon_{T-1} \;+ ...$$

Note that the projection is on the *infinite* past

Prediction Error

$$e_{T+h,T} = y_{T+h} - y_{T+h,T} = \sum_{i=0}^{h-1} b_i \varepsilon_{T+h-i}$$

(An $MA(h-1)$ process!)

$$E(e_{T+h,T}) = 0$$

$$var(e_{T+h,T}) = \sigma^2 \sum_{i=0}^{h-1} b_i^2$$

Wold's Chain Rule for Autoregressions

Consider an AR(1) process:
$y_t = \phi y_{t-1} + \varepsilon_t$

History:
$\{y_t\}_{t=1}^{T}$

Immediately,

$$y_{T+1,T} = \phi y_T$$
$$y_{T+2,T} = \phi y_{T+1,T} = \phi^2 y_T$$
$$\vdots$$
$$y_{T+h,T} = \phi y_{T+h-1,T} = \phi^h y_T$$

Extension to $AR(p)$ and $AR(\infty)$ is immediate.

2.6 MULTIVARIATE

2.6.1 The Environment

$(y_{1t},\ y_{2t})'$ is covariance stationary if:

$$E(y_{1t}) \;=\; \mu_1 \;\forall\; t$$
$$E(y_{2t}) \;=\; \mu_2 \;\forall\; t$$

$$\Gamma_{y_1 y_2}(t,\ \tau) \;=\; E\left(\begin{array}{c} y_{1t} - \mu_1 \\ y_{2t} - \mu_2 \end{array}\right)(y_{1,t-\tau} - \mu_1,\ y_{2,t-\tau} - \mu_2)$$

$$= \begin{pmatrix} \gamma_{11}(\tau) & \gamma_{12}(\tau) \\ \gamma_{21}(\tau) & \gamma_{22}(\tau) \end{pmatrix}$$
$$\tau = 0,\ 1,\ 2,\ ...$$

Cross Covariances and the Generating Function

$$\gamma_{12}(\tau) \neq \gamma_{12}(-\tau)$$

$$\gamma_{12}(\tau) = \gamma_{21}(-\tau)$$

$$\Gamma_{y_1 y_2}(\tau) = \Gamma'_{y_1 y_2}(-\tau), \quad \tau = 0,\ 1,\ 2,\ ...$$

$$G_{y_1 y_2}(z) = \sum_{\tau=-\infty}^{\infty} \Gamma_{y_1 y_2}(\tau)\ z^{\tau}$$

Cross Correlations

$$R_{y_1 y_2}(\tau) = D^{-1}_{y_1 y_2}\ \Gamma_{y_1 y_2}(\tau)\ D^{-1}_{y_1 y_2},\ \tau = 0,\ 1,\ ,2,\ ...$$

$$D = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

2.6.2 The Multivariate General Linear Process

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

$$y_t = B(L)\varepsilon_t = (I + B_1 L + B_2 L^2 + ...)\varepsilon_t$$

$$E(\varepsilon_t \varepsilon'_s) = \begin{cases} \Sigma & if\ t = s \\ 0 & otherwise \end{cases}$$

$$\sum_{i=0}^{\infty} \parallel B_i \parallel^2 < \infty$$

Autocovariance Structure

$$\Gamma_{y_1 y_2}(\tau) \;=\; \sum_{i=-\infty}^{\infty} B_i \;\Sigma\; B'_{i-\tau}$$

(where $B_i \equiv 0$ if $i < 0$)

$$G_y(z) \;=\; B(z)\;\Sigma\;B'(z^{-1})$$

2.6.3 Vector Autoregressions

N-variable VAR of order p:

$$\underset{(NxN)(Nx1)}{\Phi(L)\;\; y_t} \;=\; \underset{(Nx1)}{\varepsilon_t}$$

$$\varepsilon_t \;\sim\; \underset{(Nx1)(NxN)}{(0,\;\; \Sigma)}$$

– Simple estimation and analysis (OLS)
– Granger-Sims causality
– Getting the facts straight before theorizing; assessing the restrictions implies by economic theory

Impulse Response Functions

How does a shock to y_{it} (alone) dynamically affect y_{jt} ?

Impulse responses are pieces of the MA(∞) representation:

$$y_t \;=\; (I \;+\; \Theta_1 L \;+\; \Theta_2 L^2 \;+\; ...)\,\varepsilon_t$$

$$\varepsilon_t \;\sim\; (0,\; \Sigma)$$

Problem: Σ generally not diagonal

Cholesky Factor Identification

$$(I - \Phi_1 L \;-\; ... \;-\; \Phi_p L^P) y_t \;=\; \varepsilon_t$$

$$(I - \Phi_1 L \ - \ ... \ - \ \Phi_p L^P) y_t \ = \ P \, v_t$$

$$\text{where } v_t \ \sim \ (0, \ I) \text{ and } \Sigma = PP'$$

Impulse Response Function:

$$y_t \ = \ (I \ + \ \Theta_1 L \ + \ \Theta_2 L^2 \ + \ ...) \ P \, v_t$$

$$= \ (P \ + \ \Theta_1 P \ L \ + \ \Theta_2 P \ L^2 \ + \ ...) \ v_t$$

2.7 A SMALL EMPIRICAL TOOLKIT

2.7.1 Nonparametric: Sample Autocovariances

$$\hat{\gamma}(\tau) \ = \ \frac{1}{T} \ \sum_{t=1}^{T-|\tau|} x_t x_{t+|\tau|}, \ \tau \ = \ 0, \ \pm 1, \ ..., \ \pm (T-1)$$

$$\gamma^*(\tau) \ = \frac{1}{T-|\tau|} \ \sum_{t=1}^{T-|\tau|} x_t x_{t+|\tau|}, \ \tau \ = \ 0, \ \pm \ 1 \ , \ ..., \ \pm (T-1)$$

Perhaps surprisingly, $\hat{\gamma}(\tau)$ is better

Asymptotic Distribution of the Sample Autocorrelations

$$\rho \ = \ (\rho(0), \ \rho(1), \ ..., \ \rho(r))'$$

$$\sqrt{T}(\hat{\rho} \ - \ \rho) \ \overset{d}{\rightarrow} \ N(0, \ \Sigma)$$

Important special case (iid):

$$asyvar(\hat{\rho}(\tau)) \ = \frac{1}{T}, \ \forall \ \tau$$

$$asycov(\hat{\rho}(\tau), \ \hat{\rho}(\tau + v)) \ = \ 0$$

"Bartlett standard errors"

2.7.2 Parametric: $ARMA$ Model Selection, Fitting and Diagnostics

2.7.2.1 Fitting and Selection

Fitting: OLS, MLE, GMM.

Model Selection (Relative Model Performance)

What not to do...

$$MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

$$= 1 - \frac{MSE}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Still bad:

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k}$$

$$s^2 = \left(\frac{T}{T-k}\right)\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right)$$

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2 \ / \ T-k}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2 \ / \ T-1}$$

$$= 1 - \frac{s^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2 \ / \ T-1}$$

Good:

$$SIC = T^{(\frac{k}{T})}\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right)$$

More generally,

$$SIC = \frac{-2lnL}{T} + \frac{KlnT}{T}$$

Consistency (oracle property)

2.7.2.2 Diagnostics

Box-Pierce and Related Results:

$$Q_{BP} \;=\; T \sum_{\tau=1}^{m} \hat{\rho}^2(\tau) \;\sim\; \chi^2(m)$$

$$Q_{LB} \;=\; T\,(T+2) \sum_{\tau=1}^{m} \left(\frac{1}{T-\tau}\right) \hat{\rho}^2(\tau)$$

2.8 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Ergodicity.

   We shall say (loosely speaking) that a time series is ergodic if consistent inference regarding its stochastic structure can be made on the basis of one realization. While ergodicity is a deep mathematical property of the distribution function characterizing the time series in question, its meaning for a stationary time series is essentially independence of observations far enough apart in time.

   Ergodicity refers to consistent moment estimability based only on a single realization, as opposed to stationarity, which is concerned with the time−constancy of the probability structure of a stochastic process. It is therefore nonsensical to pose questions regarding the ergodicity of nonstationary processes. We stress that ergodicity cannot be "checked," even with a (doubly) infinitely sample path. The intuition is simple: regardless of whether or not a time−series is ergodic, sample moments converge to a random variable. If the series is ergodic, that random variable is in fact a (degenerate) constant. It is immediately clear, then, that even with an infinitely large sample one cannot tell whether or not sample moments converge to a constant (fixed in repeated realizations) or just one particular realization of a random variable (which will change from realization to realization). To check ergodicity, one must have available an entire ensemble, which is never the case in practice.

   Due to the impossibility of empirically checking ergodicity in observed time series, attention has focused on the study of specific parameterizations for which ergodicity can be theoretically established. For example, the important LRCSSP, discussed below, is always ergodic. More generally, we seek sufficient conditions under which laws of large numbers (LLN) can be shown to hold. For a time series of independent, identically distributed random variables, Kolmogorov's LLN holds. For dependent, identically (unconditionally) distributed time series, sufficient conditions for the LLN are well known. Much recent research examines conditions sufficient for the LLN in

more general situations, such as dependent time series with heterogeneous innovations. The resulting theories of mixing, martingale difference, and near−epoch dependent sequences are discussed in White (1984), Gallant and White (198*), and White (199*), among many others.

2. The autocovariance function of the MA(1) process, revisited.

   In the text we wrote

$$\gamma(\tau) \;=\; E(y_t y_{t-\tau}) \;=\; E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) \;=\; \begin{cases} \theta\sigma^2, \ \tau = 1 \\ \\ 0, otherwise. \end{cases}$$

   Fill in the missing steps by evaluating explicitly the expectation

$$E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})).$$

3. Predicting AR processes.

   Show the following.

   (a) If y_t is a covariance stationary $AR(1)$ process, i.e., $y_t = \alpha y_{t-1} + \epsilon_t$ with $|\alpha| < 1$, then $y_{t+h,t} = \alpha^h y_t$.

   (b) If y_t is $AR(2)$,

$$y_t = (\alpha_1 + \alpha_2) y_{t-1} - \alpha_1\alpha_2 y_{t-2} + \epsilon_t,$$

   with $\mid \alpha_1 \mid, \mid \alpha_2 \mid < 1$, then

$$y_{t+1,t} = (\alpha_1 + \alpha_2) y_t - \alpha_1\alpha_2 y_{t-1}.$$

   (c) In general, the result for an $AR(p)$ process is

$$y_{t+k,t} = \psi_1 y_{t+k-1,t} + ... + \psi_p y_{t+k-p,t},$$

   where $y_{t-j} = y_{t-j}$, for $j = 0, 1, ...$, at time t. Thus for pure autoregressions, the MMSE prediction is a linear combination of only the p most recently observed values.

4. Predicting MA process.

   If y_t is $MA(1)$,

$$y_t = \epsilon_t - \beta\epsilon_{t-1},$$

   where $|\beta < 1$, then

$$y_{t+1} = -\beta \sum_{j=0}^{\infty} \beta^j x_{t-j},$$

and $y_{t+k} = 0$ for all $k > 1$.

For moving-average processes more generally, predictions for a future period greater than the order of the process are zero and those for a period less distant cannot be expressed in terms of a finite number of past observed values.

5. Predicting the $ARMA(1,1)$ process.

   If y_t is $ARMA(1,1)$,

   $$y_t - \alpha y_{t-1} = \epsilon_t - \beta \epsilon_{t-1},$$

   with $\mid \alpha \mid, \mid \beta \mid < 1$, then

   $$y_{t+k,t} = \alpha^{k-1}(\alpha - \beta) \sum_{j=0}^{\infty} \beta^j y_{t-j}.$$

6. Prediction-error dynamics.

   Consider the general linear process with strong white noise innovations. Show that both the conditional (with respect to the information set $\Omega_t = \{\epsilon_t, \epsilon_{t-1}, ...\}$) and unconditional moments of the Wiener-Kolmogorov h-step-ahead prediction error are identical.

7. Truncating the Wiener-Kolmogorov predictor.

   Consider the sample path, $\{y_t\}_{t=1}^{T}$, where the data generating process is $y_t = B(L)\epsilon_t$ and $B(L)$ is of infinite order. How would you modify the Weiner-Kolmogorov linear least squares prediction formula to generate an operational 3-step-ahead forecast? (Hint: truncate.) Is your suggested predictor linear least squares? Least squares within the class of linear predictors using only T past observations?

8. Empirical GDP dynamics.

   (a) Obtain the usual quarterly expenditure-side U.S. GDP_E from FRB St. Louis, 1960.1-present.

   (b) Leaving out the 12 most recent quarters of data, perform a full correlogram analysis for GDP_E logarithmic growth.

   (c) Again leaving out the 12 most recent quarters of data, specify, estimate and defend appropriate $AR(p)$ and $ARMA(p,q)$ models for GDP_E logarithmic growth.

   (d) Using your preferred $AR(p)$ and $ARMA(p,q)$ models for GDP_E logarithmic growth, generate a 12-quarter-ahead linear least-squares path forecast for the "hold-out" sample. How do your $AR(p)$ and $ARMA(p,q)$ forecasts compare to the realized values? Which appears more accurate?

   (e) Obtain ADNSS GDP_{plus} logarithmic growth from FRB Philadelphia, read about it, and repeat everything above.

(f) Contrast the results for GDP_E logarithmic growth and GDP_{plus} logarithmic growth.

9. Time-domain analysis of housing starts and completions.

   (a) Obtain monthly U.S. housing starts and completions data from FRED at FRB St. Louis, seasonally-adjusted, 1960.1-present. Your two series should be of equal length.

   (b) Using only observations $\{1, ..., T-4\}$, perform a full correlogram analysis of starts and completions. Discuss in detail.

   (c) Using only observations $\{1, ..., T-4\}$, specify and estimate appropriate univariate $ARMA(p, q)$ models for starts and completions, as well as an appropriate $VAR(p)$. Discuss in detail.

   (d) Characterize the Granger-causal structure of your estimated $VAR(p)$. Discuss in detail.

   (e) Characterize the impulse-response structure of your estimated $VAR(p)$ using all possible Cholesky orderings. Discuss in detail.

   (f) Using your preferred $ARMA(p, q)$ models and $VAR(p)$ model, specified and estimated using only observations $\{1, ..., T-4\}$, generate linear least-squares path forecasts for the four quarters of "hold out data," $\{T-3, T-2, T-1, T\}$. How do your forecasts compare to the realized values? Discuss in detail.

10. Factor structure.

    Consider the bivariate linearly indeterministic process,

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} B_{11}(L) & B_{12}(L) \\ B_{21}(L) & B_{22}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},$$

    under the usual assumptions. Suppose further that $B_{11}(L) = B_{21}(L) = 0$ and $\varepsilon_{1t} = \varepsilon_{2t} = \varepsilon_t$ (with variance σ^2). Discuss the nature of this system. Why might it be useful in economics?

2.9 NOTES

Characterization of time series by means of autoregressive, moving average, or ARMA models was suggested, more or less simultaneously, by the Russian statistician and economist E. Slutsky and the British statistician G.U. Yule. The Slutsky-Yule framework was modernized, extended, and made part of an innovative and operational modeling and forecasting paradigm in a more recent classic, a 1970 book by Box and Jenkins. In fact, ARMA and related models are often called "Box-Jenkins models."

By 1930 Slutzky and Yule had shown that rich dynamics could be obtained by taking weighted averages of random shocks. Wold's celebrated 1937 decomposition established the converse, decomposing covariance stationary series into weighted averages of random shocks, and paved the way for subsequent path-breaking work by Wiener, Kolmogorov, Kalman and others. The beautiful 1963 treatment by Wold's student Whittle (1963), updated and reprinted as Whittle (1983) with a masterful introduction by Tom Sargent, remains widely-read. Much of macroeconomics is built on the Slutzky-Yule-Wold-Wiener-Kolmogorov foundation. For a fascinating overview of parts of the history in its relation to macroeconomics, see Davies and Mahon (2009), at http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=4348.

Chapter Three

Nonparametric Estimation and Prediction

3.1 DENSITY ESTIMATION

3.1.1 The Basic Problem

$$\{x_i\}_{i=1}^N \overset{iid}{\sim} f(x)$$

f smooth in $[x_0 - h, x_0 + h]$

Goal: Estimate $f(x)$ at arbitrary point $x = x_0$
By the mean-value theorem,

$$f(x_0) \approx \frac{1}{2h} \int_{x_0-h}^{x_0+h} f(u)du = \frac{1}{2h} P(x \in [x_0 - h, x_0 + h])$$

Estimate $P(x \in [x_0 - h,\ x_0 + h])$ by $\frac{\#x_i \in [x_0-h,\ x_0+h]}{N}$

$$\hat{f}_h(x_0) = \frac{1}{2h} \frac{\#x_i \in [x_0 - h, x_0 + h]}{N}$$

$$= \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} I\left(\left|\frac{x_0 - x_i}{h}\right| \leq 1\right)$$

"Rosenblatt estimator"

Kernel density estimator with

kernel: $K(u) = \frac{1}{2} I(|u| \leq 1)$

bandwidth: h

3.1.2 Kernel Density Estimation

Issues with uniform kernels:

1. Why weight distant observations as heavily as nearby ones?

2. Why use a discontinuous kernel if we think that f is smooth?

Obvious solution: Choose *smooth* kernel

Standard conditions:

$\int K(u)du = 1$

$K(u) = K(-u)$
Common Kernel Choices
Standard normal: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u}{2}^2}$

Triangular $K(u) = (1 - |u|)I(|u| \leq 1)$

Epinechnikov: $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$
General Form of the Kernel Density Estimator

$$\hat{f}_h(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right)$$

"Rosenblatt-Parzen estimator"

3.1.3 Bias-Variance Tradeoffs

3.1.3.1 Inescapable Bias-Variance Tradeoff (in Practice, Fixed N)

3.1.3.2 Escapable Bias-Variance Tradeoff (in Theory, $N \to \infty$)

$E(\hat{f}_h(x_0)) \approx f(x_0) + \frac{h^2}{2} \cdot O_p(1)$

$(So\ h \to 0 \implies bias \to 0)$

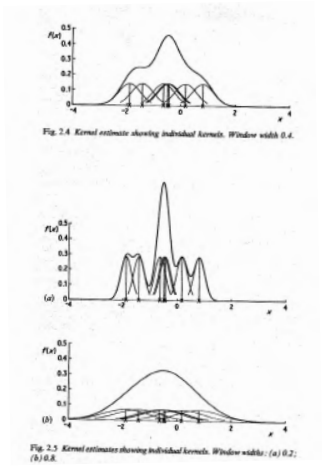
$var\ (\hat{f}_h(x_0)) \approx \frac{1}{Nh} \cdot O_p(1)$

$(So\ Nh \to \infty \implies var \to 0)$

Thus,

$$\left.\begin{array}{c} h \to 0 \\ Nh \to \infty \end{array}\right\} \implies \hat{f}_h(x_0) \overset{p}{\to} f(x_0)$$

Figure 3.1: Bandwidth Choice – from Silverman (1986)



3.1.3.3 Convergence Rate

$$\sqrt{Nh}(\hat{f}_h(x_0) - f(x_0)) \xrightarrow{d} D$$

Effects of K minor; effects of h major.

3.1.4 Optimal Bandwidth Choice

$$MSE\left(\hat{f}_h(x_0)\right) = E\left(\hat{f}_h(x_0) - f(x_o)\right)^2$$

$$IMSE = \int MSE\left(\hat{f}_h(x_0)\right) f(x)\, dx$$

Choose bandwidth to minimize IMSE:

$h^* = \gamma^* N^{-1/5}$
Corresponding Optimal Convergence Rate
Recall:

$$\sqrt{Nh}\left(\hat{f}_h(x_0) - f(x_0)\right) \xrightarrow{d} D$$

$$h^* \propto N^{-1/5}$$

Substituting yields the best obtainable rate:

$$\sqrt{N^{4/5}}\left(\hat{f}_h(x_0) - f(x_0)\right) \overset{d}{\to} D$$

"Stone optimal rate"
Silverman's Rule
For the Gaussian case,

$h^* = 1.06\sigma N^{-1/5}$

So use:

$\hat{h}^* = 1.06\hat{\sigma} N^{-1/5}$

Better to err on the side of too little smoothing:

$\hat{h}^* = \hat{\sigma} N^{-1/5}$

3.2 MULTIVARIATE

Earlier univariate kernel density estimator:

$$\hat{f}_h(x_0) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right)$$

Can be written as:

$$\hat{f}_h(x_0) = \frac{1}{N}\sum_{i=1}^{N} K_h(x_0 - x_i)$$

where $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$

or $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$
Multivariate Version (d-Dimensional)
Precisely follows equation (3.2):

$$\hat{f}_H(x_0) = \frac{1}{N}\sum_{i=1}^{N} K_H(x_0 - x_i),$$

where $K_H(\cdot) = |H|^{-1}K(H^{-1}\cdot)$, and H $(d \times d)$ is psd.

Common choice: $K(u) = N(0, I),\ H = hI$

$$\implies K_H(\cdot) = \frac{1}{h^d} K\left(\frac{1}{h}\cdot\right) = \frac{1}{h^d} K\left(\frac{x_0 - x_i}{h}\right)$$

$$\implies \hat{f}_h(x_0) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right)$$

Bias-Variance Tradeoff, Convergence Rate, Optimal Bandwidth, Correpsonding Optimal Convergence Rate

$$\left.\begin{array}{c} h \to 0 \\ Nh^d \to \infty \end{array}\right\} \implies \hat{f}_h(x_0) \overset{p}{\to} f(x_0)$$

$$\sqrt{Nh^d}\left(\hat{f}_h(x_0) - f(x_0)\right) \overset{d}{\to} D$$

$$h^* \propto N^{-\frac{1}{d+4}}$$

$$\sqrt{N^{1-\frac{d}{d+4}}}\left(\hat{f}_h(x_0) - f(x_0)\right) \overset{d}{\to} D$$

Stone-optimal rate drops with d

"Curse of dimensionality"
Silverman's Rule

$$\hat{h}^* = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} \hat{\sigma} N^{-\frac{1}{d+4}}$$

where

$$\hat{\sigma}^2 = \frac{1}{d} \sum_{i=1}^{d} \hat{\sigma}_i^2$$

(average sample variance)

3.3 FUNCTIONAL ESTIMATION

Conditional Mean (Regression)

$$E(y|x) = M(x) = \int y \frac{f(y,x)}{f(x)} dy$$

Regression Slope

$$\beta(x) = \frac{\partial M(x)}{\partial x_j} = \lim_{h \to 0} \frac{(M(x + \frac{h}{2}) - M(x - \frac{h}{2}))}{h}$$

Regression Disturbance Density

$$f(u),\ u = y - M(x)$$

Conditional Variance

$$var(y|x) = V(x) = \int y^2 \frac{f(y,x)}{f(x)} dy - M(x)^2$$

Hazard Function

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

Curvature (Higher-Order Derivative Estimation)

$$C(x) = \frac{\partial}{\partial\ x_j}\ \beta(x) = (\frac{\partial^2}{\partial\ {x_j}^2}) M(x) = \lim_{h \to 0} \frac{\beta(x + \frac{h}{2}) -\ \beta(x - \frac{h}{2})}{h}$$

The curse of dimensionality is much worse for curvature...

d-vector: $r = (r_1, ..., r_d),\ |r| =\ \Sigma_{i=1}^{d}\ r_i$

Define $M^{(r)}(x) \equiv \partial^{\frac{|r|}{\partial^{r_1} x_1, ..., \partial^{r_d} x_d}}\ M(x)$

Then $\sqrt{Nh}^{2|r|+d}\ [\hat{M}^{(r)}(x_0) - M^{(r)}(x_0)] \to_d D$

3.4 LOCAL NONPARAMETRIC REGRESSION

3.4.1 Kernel Regression

$$M(x_0) = \int y f(y|x_0) dy = \int y \frac{f(x_0, y)}{f(x_0)} dy$$

Using multivariate kernel density estimates and manipulating gives the "Nadaraya-Watson" estimator:

$$\hat{M}_h(x_0) = \sum_{i=1}^{N} \left[\frac{K\left(\frac{x_0 - x_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x_0 - x_i}{h}\right)} \right] y_i$$

$$h \to 0,\ Nh \to \infty \implies$$

$$\sqrt{Nh^d}\ (\hat{M}_h(x_0) - M(x_0)) \overset{d}{\to} N(0, V)$$

3.4.2 Nearest-Neighbor Regression

3.4.2.1 Basic Nearest-Neighbor Regression

$\hat{M}_k(x_0) = \frac{1}{k} \sum_{i \in n(x_0)} y_i$ (Locally Constant, uniform weighting)

$k \to \infty,\ \frac{k}{N} \to 0 \ \Rightarrow\ \hat{M}_k(x_0) \overset{P}{\to} M(x_0)$

$\sqrt{k}\ (\hat{M}_k(x_0) - M(x_0)) \overset{d}{\to} D$

Equivalent to Nadaraya-Watson kernel regression with:
$K(u) = \frac{1}{2} I(|u| \leq 1)$ (uniform)
and $h = R(k)$ (distance from x_0 to k^{th} nearest neighbor)
$\Rightarrow$ Variable bandwidth!

3.4.2.2 Locally-Weighted Nearest-Neighbor Regression (Locally Polynomial, Non-Uniform Weighting)

$y_t = g(x_t) + \varepsilon_t$

Computation of $\hat{g}(x^*)$:

$0 < \xi \leq 1$

$k_T = int(\xi \cdot T)$

Find K_T nearest neighbors using norm:

$\lambda(x^*,\ x^*_{k_T}) = [\Sigma_{j=1}^{P}(x^*_{k_T\ j} - x^*_j)^2]^{\frac{1}{2}}$
Neighborhood weight function:

$$v_t(x_t,\ x^*,\ x^*_{k_T}) \ = \ C\left(\frac{\lambda(x_t,\ x^*)}{\lambda(x^*,\ x^*_{k_T})}\right)$$

$$C(u) \ = \ \begin{cases} (1-u^3)^3 & for\ u<1 \\ 0 & otherwise \end{cases}$$

3.5 GLOBAL NONPARAMETRIC REGRESSION

3.5.1 Series (Sieve, Projection, ...)

$M(x_0) \ = \ \Sigma_{j=0}^{\infty}\ \beta_j\ \phi_j\ (x_0)$
(the ϕ_j are orthogonal basis functions)

$\hat{M}_J(x_0) \ = \ \Sigma_{j=0}^{J}\ \hat{\beta}_j\ \phi_j(x_0)$

$J \to \infty,\ \frac{J}{N} \to 0 \ \ \Rightarrow \ \ \hat{M}_J(x_0) \ \underset{\to}{P}\ M(x_0)$

Stone-optimal convergence rate, for suitable choice of J.

3.5.2 Neural Networks

Run linear combinations of inputs through "squashing functions" $i = 1, ..., R$ inputs, $j = 1, ..., S$ neurons

$h_{jt} \ = \ \Psi(\gamma_{jo} + \Sigma_{i=1}^{R}\ \gamma_{ij}\ x_{it}),\ j = 1, ..., S\ (Neuron\ j)$
e.g. $\Psi(\cdot)$ can be logistic (regression), 0-1 (classification)

$O_t \ = \ \Phi(\beta_0 + \Sigma_{j=1}^{S}\ \beta_j\ h_{jt})$
e.g. $\Phi(\cdot)$ can be the identity function

Compactly: $O_t \ = \ \Phi(\beta_0 \ + \ \Sigma_{j=1}^{S}\beta_j\Psi(\gamma_{jo} + \Sigma_{i=1}^{R}\gamma_{ij}x_{it}) \ \equiv \ f(x_t;\theta)$

Universal Approximator: $S \to \infty,\ \frac{S}{N} \to 0 \ \Rightarrow \ \hat{O}(x_0) \to_p O(x_0)$

Same as other nonparametric methods.

3.6 TIME SERIES ASPECTS

1. Many results go through under mixing or Markov conditions.

2. Recursive kernel regression.

   Use recursive kernel estimator:

   $$\hat{f}_N(x_0) \ = \ (\tfrac{N-1}{N})f_{N-1}(x_0) \ + \ \tfrac{1}{Nh^d}K(\tfrac{x_0-x_N}{h})$$

to get:

$$\hat{M}_N(x_0) \;=\; \frac{(N-1)h^d \hat{f}_{N-1}(x_0)\hat{M}_{N-1}(x_0) \;+\; Y_N K(\frac{x_0 - x_N}{h})}{(N-1)h^d \hat{f}_{N-1}(x_0) \;+\; K(\frac{x_0 - x_N}{h})}$$

3. Bandwidth selection via recursive prediction.

4. Nonparametric nonlinear autoregression.

   $y_t \;=\; g(y_{t-1}, ..., y_{t-p}) \;+\; \varepsilon_t$
   $E(y_{t+1} \mid y_t, ..., y_{t-p+1}) = \;\int y_{t+1}\; f(y_{t+1} \mid y_t, ..., y_{t-p+1})\; dy$
   $= \;\int y_{t+1} \frac{f(y_{t+1}, ..., y_{t-p+1})}{f(y_t, ..., y_{t-p+1})}\; dy$

   Implementation: Kernel, Series, NN, LWR

5. Recurrent neural nets.

   $h_{jt} \;=\; \Psi(\gamma_{jo} \;+\; \Sigma_{i=1}^{R} \gamma_{ij} x_{it} \;+\; \Sigma_{l=1}^{S} \delta_{jl} h_{l,\; t-1}),\; j \;=\; 1,\; ...,\; S$

   $O_t \;=\; \Phi(\beta_0 \;+\; \Sigma_{j=1}^{S} \beta_j h_{jt})$

   Compactly: $O_t \;=\; \Phi(\beta_0 \;+\; \Sigma_{j=1}^{S} \beta_j \Psi\,(\gamma_{j0} \;+\; \Sigma_{i=1}^{R} \gamma_{ij} x_i t \;+\; \Sigma_{l=1}^{S} \delta_{jl} h_{l,\; t-1})$

   Back substitution:
   $O_t \;=\; g(x_t, x_{t-1}, ..., x_1; \theta)$

3.7 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Tightly parametric models are often best for time-series prediction.

   Generality isn't so great; restrictions often help!

2. Semiparametric and related approaches.

   $\sqrt{N}$ consistent estimation. Adaptive estimation.

3.8 NOTES

Chapter Four

Spectral Analysis

4.1 THE MANY USES OF SPECTRAL ANALYSIS

Spectral Analysis

- As with the acov function, "getting the facts straight"
- Trend and persistence (power near zero)
- Integration, co-integration and long memory
- Cycles (power at cyclical frequencies)
- Seasonality (power spikes at fundamental and harmonics)
- Filter analysis and design
- Maximum-likelihood estimation (including band spectral)
- Assessing agreement between models and data
- Robust (HAC) variance estimation

4.2 THE SPECTRUM AND ITS PROPERTIES

Recall the General Linear Process

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

Autocovariance generating function:

$$g(z) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau)\ z^{\tau}$$

$$= \sigma^2 B(z)B(z^{-1})$$

$\gamma(\tau)$ and $g(z)$ are a z-transform pair

Spectrum
Evaluate g(z) on the unit circle, $z = e^{-i\omega}$:

$$g(e^{-i\omega}) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau)\, e^{-i\omega\tau}, \;\; -\pi < \omega < \pi$$

$$= \sigma^2 B(e^{i\omega})\; B(e^{-i\omega})$$

$$= \sigma^2 \mid B(e^{i\omega}) \mid^2$$

Spectrum
Trigonometric form:

$$g(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-i\omega\tau}$$

$$= \gamma(0) + \sum_{\tau=1}^{\infty} \gamma(\tau) \left(e^{i\omega\tau} + e^{-i\omega\tau}\right)$$

$$= \gamma(0) + 2\sum_{\tau=1}^{\infty} \gamma(\tau)\;\cos(\omega\tau)$$

Spectral Density Function

$$f(\omega) = \frac{1}{2\pi} g(\omega)$$

$$f(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-i\omega\tau} \quad (-\pi < \omega < \pi)$$

$$= \frac{1}{2\pi}\gamma(0) + \frac{1}{\pi}\sum_{\tau=1}^{\infty} \gamma(\tau)\cos(\omega\tau)$$

$$= \frac{\sigma^2}{2\pi} B\left(e^{i\omega}\right) B\left(e^{-i\omega}\right)$$

$$= \frac{\sigma^2}{2\pi} \mid B\left(e^{i\omega}\right) \mid^2$$

Properties of Spectrum and Spectral Density

1. symmetric around $\omega = 0$
2. real-valued
3. 2π-periodic
4. nonnegative

A Fourier Transform Pair

$$g(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-i\omega\tau}$$

$$\gamma(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\omega) e^{i\omega\tau} d\omega$$

A Variance Decomposition by Frequency

$$\gamma(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\omega) e^{i\omega\tau} d\omega$$

$$= \int_{-\pi}^{\pi} f(\omega) e^{i\omega\tau} d\omega$$

Hence

$$\gamma(0) = \int_{-\pi}^{\pi} f(\omega) d\omega$$

Robust Variance Estimation
$\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$
$var(\bar{x}) = \frac{1}{T^2} \sum_{s=1}^{T} \sum_{t=1}^{T} \gamma(t-s)$
("Add row sums")
$= \frac{1}{T} \sum_{\tau=-(T-1)}^{T-1} \left(1 - \frac{|\tau|}{T}\right) \gamma(\tau)$
("Add diagonal sums," using change of variable $\tau = t - s$)
Hence:

$$\sqrt{T}(\bar{x} - \mu) \sim \left(0, \sum_{\tau=-(T-1)}^{T-1} \left(1 - \frac{|\tau|}{T}\right) \gamma(\tau)\right)$$

$$\sqrt{T}(\bar{x} - \mu) \xrightarrow{d} N\left(0, g_x(0)\right)$$

4.3 RATIONAL SPECTRA

White Noise Spectral Density

$$y_t = \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2)$$

$$f(\omega) = \frac{\sigma^2}{2\pi} B\left(e^{i\omega}\right) B\left(e^{-i\omega}\right)$$

$$f(\omega) = \frac{\sigma^2}{2\pi}$$

AR(1) Spectral Density

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2)$$

$$f(\omega) = \frac{\sigma^2}{2\pi} B(e^{i\omega}) B(e^{-i\omega})$$

$$= \frac{\sigma^2}{2\pi} \frac{1}{(1 - \phi e^{i\omega})(1 - \phi e^{-i\omega})}$$

$$= \frac{\sigma^2}{2\pi} \frac{1}{1 - 2\phi \cos(\omega) + \phi^2}$$

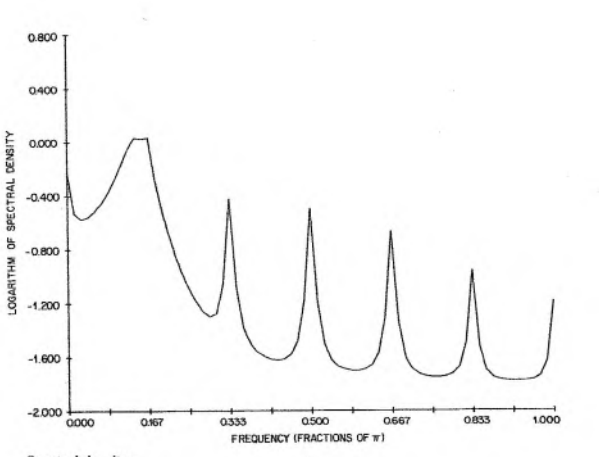
How does shape depend on ϕ? Where are the peaks?
ARMA(1, 1) Spectral Density

$$(1 - \phi L) y_t = (1 - \theta L) \varepsilon_t$$

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{1 - 2\theta \cos(\omega) + \theta^2}{1 - 2\phi \cos(\omega) + \phi^2}$$

"Rational spectral density"
Internal peaks? What will it take?

Figure 4.1: Granger's Typical Spectral Shape of an Economic Variable



4.4 MULTIVARIATE

Multivariate Frequency Domain

Covariance-generating function;

$$G_{yx}(z) = \sum_{\tau=-\infty}^{\infty} \Gamma_{yx}(\tau) z^{\tau}$$

Spectral density function:

$$F_{yx}(\omega) = \frac{1}{2\pi} G_{yx}(e^{-i\omega})$$

$$= \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \Gamma_{yx}(\tau)\ e^{-i\omega\tau}, \ \ -\pi < \omega < \pi$$

(Complex-valued)

Co-Spectrum and Quadrature Spectrum

$$F_{yx}(\omega) = C_{yx}(\omega) + iQ_{yx}(\omega)$$

$$C_{yx}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \Gamma_{yx}(\tau) \cos(\omega\tau)$$

$$Q_{yx}(\omega) = \frac{-1}{2\pi} \sum_{\tau=-\infty}^{\infty} \Gamma_{yx}(\tau) \sin(\omega\tau)$$

Cross Spectrum

$f_{yx}(\omega) = ga_{yx}(\omega) exp(i\ ph_{yx}(\omega))$ (generic cross spectrum)

$ga_{yx}(\omega) = [C_{yx}^2(\omega) + Q_{yx}^2(\omega)]^{\frac{1}{2}}$ (gain)
$ph_{yx}(\omega) = arctan\left(\frac{Q_{yx}(\omega)}{C_{yx}(\omega)}\right)$ (phase)
(Phase shift in time units is $\frac{ph(\omega)}{\omega}$)

$$coh_{yx}(\omega) = \frac{|f_{yx}(\omega)|^2}{f_{xx}(\omega)f_{yy}(\omega)} \quad (coherence)$$

Squared correlation decomposed by frequency
Useful Spectral Results for Filter Design and Analysis

- (Effects of a linear filter)

  If $y_t = B(L)x_t$, then:

  - $f_{yy}(\omega) = |B(e^{-i\omega})|^2 f_{xx}(\omega)$
  - $f_{yx}(\omega) = B(e^{-i\omega})f_{xx}(\omega)$.

  $B(e^{-i\omega})$ is the filter's *frequency response function.*

- (Effects of a series of linear filters (follows trivially))

  If $y_t = A(L)B(L)x_t$, then

  - $f_{yy}(\omega) = |A(e^{-i\omega})|^2 |B(e^{-i\omega})|^2 f_{xx}(\omega)$
  - $f_{yx}(\omega) = A(e^{-i\omega})B(e^{-i\omega})f_{xx}(\omega)$.

- (Spectrum of an independent sum)

  If $y = \sum_{i=1}^N x_i$, and the x_i are independent, then

$$f_y(\omega) = \sum_{i=1}^N f_{x_i}(\omega).$$

Nuances... Note that

$$B(e^{-i\omega}) = \frac{f_{yx}(\omega)}{f_{xx}(\omega)}$$

$$\implies B(e^{-i\omega}) = \frac{ga_{yx}(\omega)e^{i\ ph_{yx}(\omega)}}{f_{xx}(\omega)}$$

Phases of $f_{yx}(\omega)$ and $B(e^{-i\omega})$ are the same.
Gains are closely related.
Example

$$y_t = .5x_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0,1)$$

$$x_t = .9x_{t-1} + \eta_t$$

$$\eta_t \sim WN(0,1)$$

Correlation Structure

Autocorrelation and cross-correlation functions are straightforward:

$$\rho_y(\tau) = .9^{|\tau|}$$

$$\rho_x(\tau) \propto .9^{|\tau|}$$

$$\rho_{yx}(\tau) \propto .9^{|\tau-1|}$$

(What is the qualitative shape of $\rho_{yx}(\tau)$?)

Spectral Density of x

$$x_t = \frac{1}{1-.9L}\eta_t$$

$$\implies f_{xx}(\omega) = \frac{1}{2\pi}\ \frac{1}{1-.9e^{-i\omega}}\frac{1}{1-.9e^{i\omega}}$$

$$= \frac{1}{2\pi}\ \frac{1}{1-2(.9)\cos(\omega)+(.9)^2}$$

$$= \frac{1}{11.37-11.30\cos(\omega)}$$

Shape?

Spectral Density of y

$$y_t = 0.5Lx_t + \varepsilon_t$$

$$\implies f_{yy}(\omega) = \mid 0.5e^{-i\omega} \mid^2 f_{xx}(\omega) + \frac{1}{2\pi}$$

$$= 0.25f_{xx}(\omega) + \frac{1}{2\pi}$$

$$= \frac{0.25}{11.37 - 11.30\cos(\omega)} + \frac{1}{2\pi}$$

Shape?

Cross Spectrum

$B(L) = .5L$

$B(e^{-i\omega}) = 0.5e^{-i\omega}$

$f_{yx}(\omega) = B(e^{-i\omega}) f_{xx}(\omega)$

$= 0.5e^{-i\omega} f_{xx}(\omega)$

$= (0.5 f_{xx}(\omega))\, e^{-i\omega}$

$g_{yx}(\omega) = 0.5 f_{xx}(\omega) = \frac{0.5}{11.37 - 11.30\cos(\omega)}$

$Ph_{yx}(\omega) = -\omega$

(In time units, $Ph_{yx}(\omega) = -1$, so y leads x by -1)

Coherence

$$Coh_{yx}(\omega) = \frac{|\ f_{yx}(\omega)\ |^2}{f_{xx}(\omega) f_{yy}(\omega)} = \frac{.25 f_{xx}^2(\omega)}{f_{xx}(\omega) f_{yy}(\omega)} = \frac{.25 f_{xx}(\omega)}{f_{yy}(\omega)}$$

$$= \frac{.25\ \frac{1}{2\pi}\ \frac{1}{1 - 2(.9)\cos(\omega) + .9^2}}{.25\ \frac{1}{2\pi}\ \frac{1}{1 - 2(.9)\cos(\omega) + .9^2} + \frac{1}{2\pi}}$$

$$= \frac{1}{8.24 + 7.20\cos(\omega)}$$

Shape?

4.5 FILTER ANALYSIS AND DESIGN

Filter Analysis: A Trivial (but Important) High-Pass Filter

$$y_t = x_t - x_{t-1}$$

$$\Longrightarrow\ B(e^{-i\omega}) = 1 - e^{-i\omega}$$

Hence the filter gain is:

$$\left|B(e^{-i\omega})\right| = \left|1 - e^{-i\omega}\right| = 2(1 - cos(\omega))$$

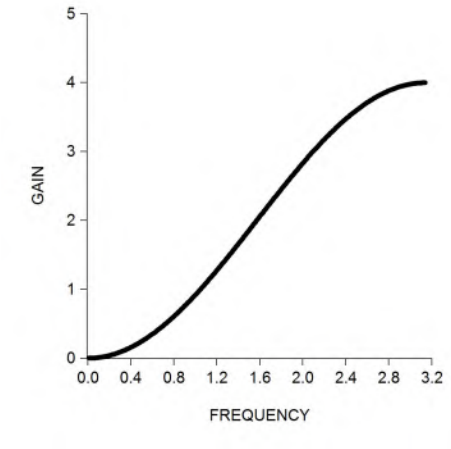
How would the gain look for $B(L) = 1 + L$?

Filter Analysis: Kuznets' Infamous Filters

Low-frequency fluctuations in aggregate real output growth.

"Kuznets cycle" – 20-year period

Figure 4.2: Gain of Differencing Filter $1 - L$



Filter 1 (moving average):

$$y_t = \frac{1}{5}\sum_{j=-2}^{2} x_{t-j}$$

$$\implies B_1(e^{-i\omega}) = \frac{1}{5}\sum_{j=-2}^{2} e^{-i\omega j} = \frac{sin(5\omega/2)}{5sin(\omega/2)}$$

Hence the filter gain is:

$$\left|B_1(e^{-i\omega})\right| = \left|\frac{sin(5\omega/2)}{5sin(\omega/2)}\right|$$

Kuznets' Filters, Continued

Kuznets' Filters, Continued

Filter 2 (fancy difference):

$$z_t = y_{t+5} - y_{t-5}$$

$$\implies B_2(e^{-i\omega}) = e^{i5\omega} - e^{-i5\omega} = 2sin(5\omega)$$

Hence the filter gain is:

$$\left|B_2(e^{-i\omega})\right| = |2sin(5\omega)|$$

Kuznets' Filters, Continued

Kuznets' Filters, Continued

Composite gain:

Figure 4.3: Gain of Kuznets' Filter 1

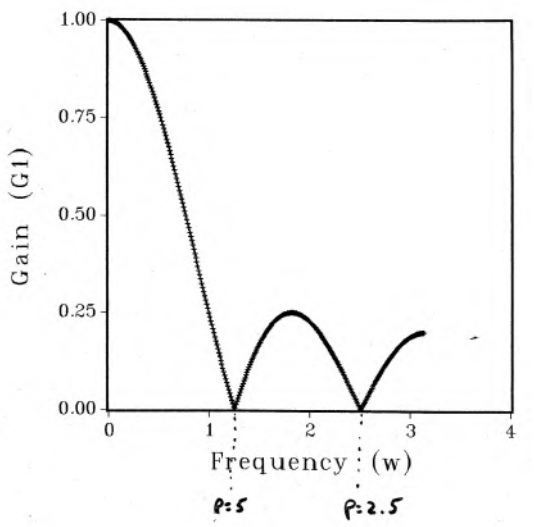


Figure 4.4: Gain of Kuznets' Filter 2

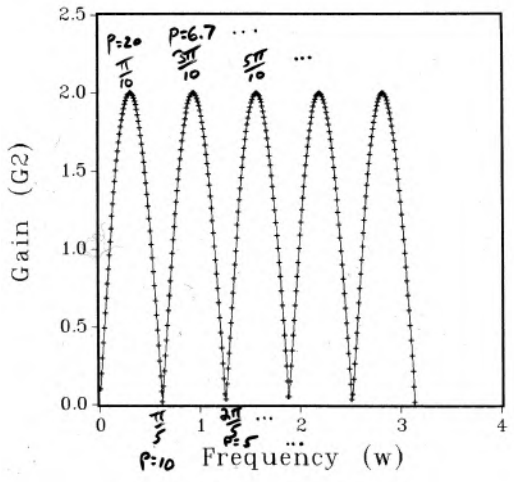
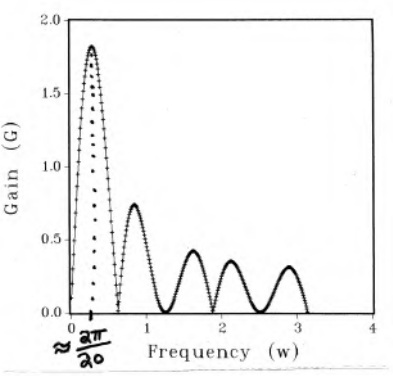


Figure 4.5: Composite Gain of Kuznets' two Filters



$$\left|B_1(e^{-i\omega})B_2(e^{-i\omega})\right| = \left|\frac{sin(5\omega/2)}{5sin(\omega/2)}\right| |2sin(5\omega)|$$

Kuznets' Filters, Continued
Filter Design: A Bandpass Filter
Canonical problem:
Find $B(L)$ s.t.

$$f_y(\omega) = \begin{cases} f_x(\omega) \text{ on } [a,b] \cup [-b,-a] \\ 0 \text{ otherwise}, \end{cases}$$

where

$$y_t = B(L)x_t = \sum_{j=-\infty}^{\infty} b_j \varepsilon_{t-j}$$

Bandpass Filter, Continued
Recall
$f_y(\omega) = |B(e^{-i\omega})|^2 f_x(\omega)$.
Hence we need:

$$B(e^{-i\omega}) = \begin{cases} 1 \text{ on } [a,b], \cup [-b,-a], \ 0 < a < b < \pi \\ 0 \text{ otherwise} \end{cases}$$

By Fourier series expansion ("inverse Fourier transform"):

$$b_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} B(e^{-i\omega}) e^{i\omega j}$$

$$= \frac{1}{\pi} \left(\frac{sin(jb) - sin(ja)}{j} \right), \ \forall j \in Z$$

Bandpass Filter, Continued
Many interesting issues:

- What is the weighting pattern? Two-sided? Weights symmetric around 0?
- How "best" to make this filter feasible in practice? What does that mean? Simple truncation?
- One sided version?
- Phase shift?

4.6 ESTIMATING SPECTRA

4.6.1 Univariate

Estimation of the Spectral Density Function
Periodogram ordinate at frequency ω:

$$I(\omega) = \frac{2}{T}\left|\sum_{t=1}^{T} y_t e^{-i\omega t}\right|^2 = \left(\sqrt{\frac{2}{T}}\sum_{t=1}^{T} y_t e^{-i\omega t}\right)\left(\sqrt{\frac{2}{T}}\sum_{t=1}^{T} y_t e^{i\omega t}\right)$$

$-\pi \leq \omega \leq \pi$
Usually examine frequencies $\omega_j = \frac{2\pi j}{T}$, $j = 0, 1, 2, ..., \frac{T}{2}$
Sample Spectral Density

$$\hat{f}(\omega) = \frac{1}{2\pi}\sum_{\tau=-(T-1)}^{T-1} \hat{\gamma}(\tau) e^{-i\omega\tau}$$

$$\hat{f}(\omega) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} y_t e^{-i\omega t}\right|^2$$

$$= \left(\frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} y_t e^{-i\omega t}\right)\left(\frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} y_t e^{i\omega t}\right)$$

$$= \frac{1}{4\pi} I(\omega)$$

Properties of the Sample Spectral Density
(Throughout we use $\omega_j,\ j = \frac{2\pi j}{T},\ j = 0, 1, ..., \frac{T}{2}$)

- Ordinates asymptotically unbiased

- Ordinates asymptotically uncorrelated
- But variance does not converge to 0
  (degrees of freedom don't accumulate)
- Hence inconsistent

For Gaussian series we have:

$$\frac{2\hat{f}(\omega_j)}{f(\omega_j)} \overset{d}{\rightarrow} \chi^2_2,$$

where the χ^2_2 random variables are
independent across frequencies

Consistent (Lag Window) Spectral Estimation

$$\hat{f}(\omega) = \frac{1}{2\pi} \sum_{\tau=-(T-1)}^{T-1} \hat{\gamma}(\tau) e^{-i\omega\tau} = \frac{1}{2\pi}\ \hat{\gamma}(0) + \frac{2}{2\pi} \sum_{\tau=1}^{T-1} \hat{\gamma}(\tau) \cos(\omega\tau)$$

$$f^*(\omega) = \frac{1}{2\pi} \sum_{\tau=-(T-1)}^{T-1} \lambda(\tau)\hat{\gamma}(\tau) e^{-i\omega\tau}$$

Common lag windows with truncation lag M_T:

$\lambda(\tau) = 1,\ |\tau| \leq M_T$ and 0 otherwise (rectangular, or boxcar)

$\lambda(\tau) = 1 - \frac{|\tau|}{M_T}, \tau \leq M_T$ and 0 otherwise
(triangular, or Bartlett, or Newey-West)

Consistency: $M_T \rightarrow \infty$ and $\frac{M_T}{T} \rightarrow 0$

Truncation lag must increase "appropriately" with T

Other (Closely-Related) Routes to
Consistent Spectral Estimation

Fit parametric approximating models (e.g., autoregressive)

("Model-based estimation")

– Let order increase appropriately with sample size

Smooth the sample spectral density
("spectral window estimation")

– Let window width decrease appropriately with sample size

4.6.2 Multivariate

Spectral density matrix:

$$F_{yx}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \Gamma_{yx}(\tau) e^{-i\omega\tau}, \quad -\pi < \omega < \pi$$

Consistent (lag window) estimator:

$$F^*_{yx}(\omega) = \frac{1}{2\pi} \sum_{\tau=-(T-1)}^{(T-1)} \lambda(\tau) \hat{\Gamma}_{yx}(\tau)\ e^{-i\omega\tau}, \quad -\pi < \omega < \pi$$

Different lag windows may be used for different elements of $F_{yx}(\omega)$
Or do model-based...

4.7 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Seasonality and Seasonal Adjustment

2. HAC Estimation

3. Applied spectrum estimation.

   Pick an interesting and appropriate real time series. Compute and graph (when appropriate) the sample mean, sample autocovariance, sample autocorrelation, sample partial autocorrelation, and sample spectral density functions. Also compute and graph the sample coherence and phase lead. Discuss in detail the methods you used. For the sample spectra, try to discuss a variety of smoothing schemes. Try smoothing the periodogram as well as smoothing the autocovariances, and also try autoregressive spectral estimators.

4. Sample spectrum.

   Generate samples of Gaussian white noise of sizes 32, 64, 128, 256, 512, 1024 and 2056, and for each compute and graph the sample spectral density function at the usual frequencies. What do your graphs illustrate?

5. Lag windows and spectral windows.

   Provide graphs of the rectangular, Bartlett, Tukey-Hamming and Parzen lag windows. Derive and graph the corresponding spectral windows.

6. Bootstrapping sample autocorrelations.

   Assuming normality, propose a "parametric bootstrap" method of assessing the finite-sample distribution of the sample autocorrelations. How would you generalize this to assess the sampling uncertainty associated with the entire autocorrelation function? How might you dispense with the normality assumption?

Solution: Assume normality, and then take draws from the process by using a noram random number generator in conjunction with the Cholesky factorization of the data covariance matrix. This procedure can be used to estimate the sampling distribution of the autocorrelations, taken one at a time. One will surely want to downweight the long-lag autocorrelations before doing the Cholesky factorization, and let this downweighting adapt to sample size. Assessing sampling uncertainty for the entire autocorrelation function (e.g., finding a 95% confidence "tunnel") appears harder, due to the correlation between sample autocorrelations, but can perhaps be done numerically. It appears very difficult to dispense with the normality assumption.

7. Bootstrapping sample spectra.

   Assuming normality, propose a "parametric bootstrap" method of assessing the finite-sample distribution of a consistent estimator of the spectral density function at various selected frequencies. How would you generalize this to assess the sampling uncertainty associated with the entire spectral density function?

   Solution: At each bootstrap replication of the autocovariance bootstrap discussed above, Fourier transform to get the corresponding spectral density function.

8. Bootstrapping spectra without normality.

   Drop the normality assumption, and propose a "parametric bootstrap" method of assessing the finite-sample distribution of (1) a consistent estimator of the spectral density function at various selected frequencies, and (2) the sample autocorrelations.

   Solution: Make use of the asymptotic distribution of the periodogram ordinates.

9. Sample coherence.

   If a sample coherence is completed directly from the sample spectral density matrix (without smoothing), it will be 1, by definition. Thus, it is important that the sample spectrum and cross-spectrum be smoothed prior to construction of a coherence estimator.

   Solution:

$$coh(\omega) \;=\; \frac{|f_{yx}(\omega)|^2}{f_x(\omega)\; f_y(\omega)}$$

   In unsmoothed sample spectral density analogs,

$$\begin{aligned} \hat{coh}(\omega) &= \frac{[\Sigma y_t e^{-i\omega t}\Sigma x_t e^{+i\omega t}][\Sigma y_t e^{+i\omega t}\Sigma x_t e^{-i\omega t}]}{[\Sigma x_t e^{-i\omega t}\Sigma x_t e^{+i\omega t}][\Sigma y_t e^{-i\omega t}\Sigma y_t e^{+i\omega t}]} \\ &\equiv 1. \end{aligned}$$

10. De-meaning.

    Consider two forms of a covariance stationary time series: "raw" and de-meaned. Contrast their sample spectral density functions at ordinates $2\pi j/T$, $j = 0, 1, ...,$

T/2. What do you conclude? Now contrast their sample spectral density functions at ordinates that are *not* multiples of 2πj/T. Discuss.

Solution: Within the set 2πj/T, j = 0, 1, ..., T/2, only the sample spectral density at frequency 0 is affected by de-meaning. However, de-meaning *does* affect the sample spectral density function at all frequencies in [0, π] outside the set 2πj/T, j = 0, 1, ..., T/2. See Priestley (1980, p. 417). This result is important for the properties of time- versus frequency-domain estimators of fractionally-integrated models. Note in particular that

$$I(\omega_j) \ \propto \ \frac{1}{T}|\sum y_t e^{i\omega_j t}|^2$$

so that

$$I(0) \ \propto \ \frac{1}{T}|\sum y_t|^2 \ \propto \ T\bar{y}^2,$$

which approaches infinity with sample size so long as the mean is nonzero. Thus it makes little sense to use I(0) in estimation, regardless of whether the data have been demeaned.

11. Schuster's periodogram.

    The periodogram ordinates can be written as

$$I(\omega_j) \ = \ \frac{2}{T}([\sum_{t=1}^{T} y_t \cos\omega_j t]^2 \ + \ [\sum_{t=1}^{T} y_t \sin\omega_j t]^2).$$

    Interpret this result.

12. Applied estimation.

    Pick an interesting and appropriate real time series. Compute and graph (when appropriate) the sample mean, sample autocovariance, sample autocorrelation, sample partial autocorrelation, and sample spectral density functions. Also compute and graph the sample coherence and phase lead. Discuss in detail the methods you used. For the sample spectra, try and discuss a variety of smoothing schemes. If you can, try smooting the periodogram as well as smoothing the autocovariances, and also try autoregressive spectral estimators.

13. Periodogram and sample spectrum.

    Prove that $I(\omega) \ = \ 4\pi\hat{f}(\omega)$.

14. Estimating the variance of the sample mean.

    Recall the dependence of the variance of the sample mean of a serially correlated time series (for which the serial correlation is of unknown form) on the spectral density of the series evaluated at $\omega = 0$. Building upon this result, propose an estimator of

the variance of the sample mean of such a time series. If you are very ambitious, you might want to explore in a Monte Carlo experiment the sampling properties of your estimator of the standard error vs. the standard estimator of the standard error, for various population models (e.g., AR(1) for various values of ρ) and sample sizes. If you are not feeling so ambitious, at least conjecture upon the outcome of such an experiment.

15. Coherence.

a. Write out the formula for the coherence between two time series x and y.
b. What is the coherence between the filtered series, (1 - b_1L) x_t and (1 - b_2L) y_t? (Assume that $b_1 \neq b_2$.)
c. What happens if $b_1 = b_2$? Discuss.

16. Multivariate spectra.

Show that for the multivariate LRCSSP,

$$F_y(\omega) \;=\; B(e^{-i\omega})\;\Sigma\;B^*(e^{-i\omega})$$

where "*" denotes conjugate transpose.

17. Filter gains.

Compute, plot and discuss the squared gain functions associated with each of the following filters.
(a) B(L) = (1 - L)
(b) B(L) = (1 + L)
(c) B(L) = (1 - .5 L12)-1
(d) B(L) = (1 - .5 L12).

Solution:

(a) G2 = 1 - e-iω2 is monotonically increasing on [0, π]. This is an example of a "high pass" filter.
(b) G2 = 1 + e-iω2 is monotonically decreasing on [0, π]. This is an example of a "low pass" filter.
(c) G2 = (1 - .5 e-12iω)2 has peaks at the fundamental seasonal frequency and its harmonics, as expected. Note that it corresponds to a seasonal autoregression.
(d) G2 = (1 - .5 e-12iω)2 has troughs at the fundamental seasonal frequency and its harmonics, as expected, because it is the inverse of the seasonal filter in (c) above.

Thus, the seasonal process associated with the filter in (c) above would be appropriately "seasonally adjusted" by the present filter, which is its inverse.

18. Filtering

(a) Consider the linear filter B(L) = 1 + θ L. Suppose that yt = B(L) xt,
where xt $\sim$ WN(0, σ2). Compute fy(ω).

(b) Given that the spectral density of white noise is $\sigma 2/2\pi$, discuss how the filtering theorem may be used to determine the spectrum of any LRCSSP by viewing it as a linear filter of white noise.

Solution:
(a) fy(ω) = 1 + θe-iω2 fx(ω)
= $\sigma 2/2\pi$ (1 + θe-iω)(1 + θeiω)
= $\sigma 2/2\pi$ (1 + $\theta 2$ + 2θ cos ω),
which is immediately recognized as the sdf of an MA(1) process.
(b) All of the LRCSSP's that we have studied are obtained by applying linear filters to white noise. Thus, the filtering theorem gives their sdf's as
f(ω) = $\sigma 2/2\pi$ B(e-iω)2
= $\sigma 2/2\pi$ B(e-iω) B(eiω)
= $\sigma 2/2\pi$ B(z) B(z-1),
evaluated on |z| = 1, which matches our earlier result.

19. Zero spectra.

Suppose that a time series has a spectrum that is zero on an interval of positive measure. What do you infer?

Solution: The series must be deterministic, because one could design a filter such that the filtered series has zero spectrum everywhere.

20. Period.

Period is $2\pi/\omega$ and is expressed in time/cycle. 1/P, cycles/time. In engineering, time is often measured in seconds, and 1/P is Hz.

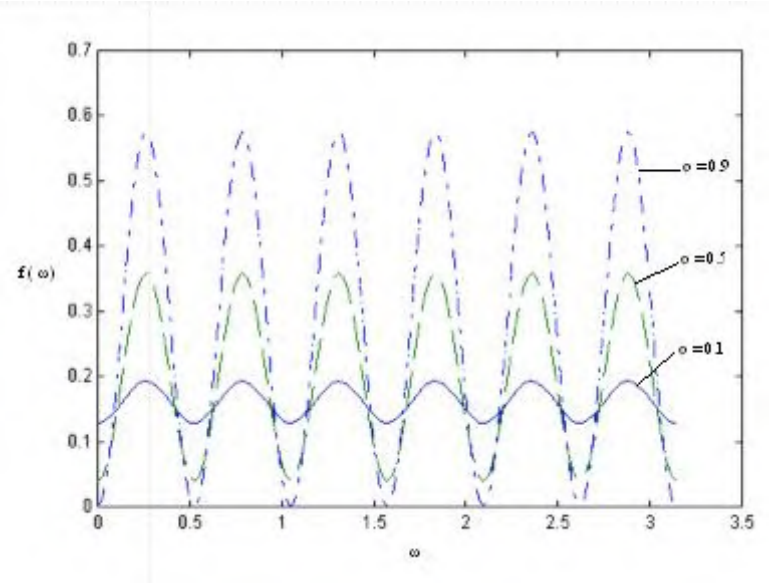
21. Seasonal autoregression.

Consider the "seasonal" autoregression (1 - ϕ L12) yt = ϵt.
(a) Would such a structure be characteristic of monthly seasonal data or quarterly seasonal data?

(b) Compute and plot the spectral density f(ω), for various values of ϕ. Does it have any internal peaks on (0, π)? Discuss.
(c) The lowest-frequency internal peak occurs at the so-called fundamental seasonal frequency. What is it? What is the corresponding period?
(d) The higher-frequency spectral peaks occur at the harmonics of the fundamental seasonal frequency. What are they? What are the corresponding periods?
Solution:
(a) Monthly, because of the 12-period lag.
(b)

$$f(\omega) \;=\; \frac{\sigma^2}{2\pi}\,(1 \;+\; \phi^2 \;-\; 2\phi\,\cos(12\omega))$$

The sdf has peaks at $\omega = 0$, $\pi/6$, $2\pi/6$, ..., $5\pi/6$, and π.
(c) The fundamental frequency is $\pi/6$, which corresponds to a period of 12 months.
(d) The harmonic frequencies are $2\pi/6$, ..., $5\pi/6$, and π, corresponding to periods of 6 months, 4 months, 3 months, 12/5 months and 2 months, respectively.



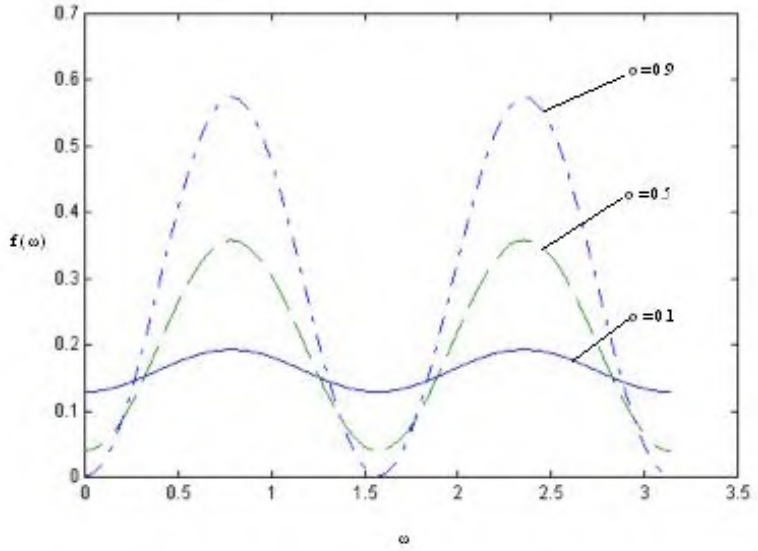
22. More seasonal autoregression.

Consider the "seasonal" autoregression
(1 - ϕ L4) yt = ϵt.
(a) Would such a structure be characteristic of monthly seasonal data or quarterly seasonal data?
(b) Compute and plot the spectral density f(ω), for various values of ϕ. Does it have any internal peaks on (0, π)? Discuss.
(c) The lowest-frequency internal peak occurs at the so-called fundamental seasonal frequency. What is it? What is the corresponding period?

(d) The higher-frequency spectral peaks occur at the harmonics of the fundamental seasonal frequency. What are they? What are the corresponding periods?
Solution: (a) Quarterly, because of the 4-period lag.
(b)

$$f(\omega) = \frac{\sigma^2}{2\pi}(1 + \phi^2 - 2\phi \cos(4\omega))$$

The sdf has peaks at $\omega = 0$, $\pi/2$ and π.
(c) The fundamental frequency is $\pi/2$, which corresponds to a period of 4 quarters.
(d) The only harmonic is π, corresponding to a period of 2 quarters.



23. The long run.

Discuss and contrast the economic concept of "long run" and the statistical concept of "low frequency". Give examples showing when, if ever, the two may be validly and fruitfully equated. Also give examples showing when, if ever, it would be inappropriate to equate the two concepts.
Solution:
A potential divergence of the concepts occurs when economists think of the "long run" as "in steady state," which implies the *absence* of dynamics.

24. Variance of the sample mean.

The variance of the sample mean of a serially correlated time series is proportional to the spectral density function at frequency zero.
Solution:

Let

$$\bar{x} \ = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

Then

$$\begin{aligned} var(\bar{x}) \ &= \tfrac{1}{T^2}\textstyle\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma(t-s) \\ &= \tfrac{1}{T}\textstyle\sum_{\tau=-(T-1)}^{T-1}(1-\frac{|\tau|}{T})\gamma(\tau), \end{aligned}$$

where $\gamma(\tau)$ is the autocovariance function of x. So for large T, we have

$$var(\bar{x}) \ \approx \ \frac{2\pi f_x(0)}{T}.$$

25. ARCH process spectrum.

Consider the ARCH(1) process x_t, where

$$x_t \mid x_{t-1} \ \sim \ N(0, \ .2 \ + \ .8 \ x_{t-1}^2)$$

a. Compute and graph its spectral density function.
b. Compute and graph the spectral density function of x_T^2. Discuss.

Solution:
a. By the law of iterated expectations, we have

$$E(x_t) \ = \ E[E(x_t|x_{t-1})] \ = \ E(0) \ = \ 0$$

$$\gamma(0) \ = \ E(x_t^2) \ = \ E[E(x_t^2|x_{t-1})] \ = \ E[E(0.2+0.8x_{t-1}^2)] \ = \ 0.2+0.8\gamma(0)$$

$$\gamma(0) \ = \ \frac{0.2}{1-0.8} \ = \ 1$$

$$\gamma(\tau) \ = \ E(x_t x_{t-\tau}) \ = \ E[E(x_t x_{t-\tau}|x_{t-1}, x_{t-2}, \cdots)] \ = \ E[x_{t-\tau}E(x_t|x_{t-1}, x_{t-2}, \cdots)] \ = \ E(x_{t-\tau}0) \ = \ 0$$

for τ=1,2,....
Therefore

$$f(\omega) \ = \ \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma(\tau)e^{-i\omega\tau} \ = \frac{1}{2\pi}\gamma(0) \ = \frac{1}{2\pi}$$

Because it is white noise, the spectrum will be flat.
b. Because the kurtosis of normal random variables is 3, we have

$$E(x_t^4) \ = \ E[E(x_t^4|x_{t-1}, x_{t-2}\cdots)] \ = \ E[3(0.2+0.8x_{t-1}^2)^2] \ = \ 3[0.04+0.32E(x_{t-1}^2)+0.64E(x_{t-1}^4)].$$

$$\therefore \ \gamma_{x^2}(0) \ = \ E(x_t^4) - (E(x_t^2))^2 \ = \ 3-1 \ = 2$$

Because

$$x_t^2$$

follows an AR(1) process, it follows that

$$\gamma_{x^2}(\tau) \;=\; 0.8\gamma_{x^2}(\tau-1)$$

for τ=1,2,.... We can write the s.d.f . as

$$f(\omega) \;=\frac{1}{2\pi}\gamma_{x^2}(0)+2\sum_{\tau=1}^{\infty}\gamma_{x^2}(\tau)\cos(\omega\tau) \;=\frac{1}{2\pi}(1+2\sum_{\tau=1}^{\infty}0.8^{\tau}\cos(\omega\tau))$$

which will look like an AR(1) process' s.d.f .

26. Computing spectra.

Compute, graph and discuss the spectral density functions of

a. $y_t = .8y_{t-12}+\epsilon_t$

b. $y_t = .8\epsilon_{t-12}+\epsilon_t$.

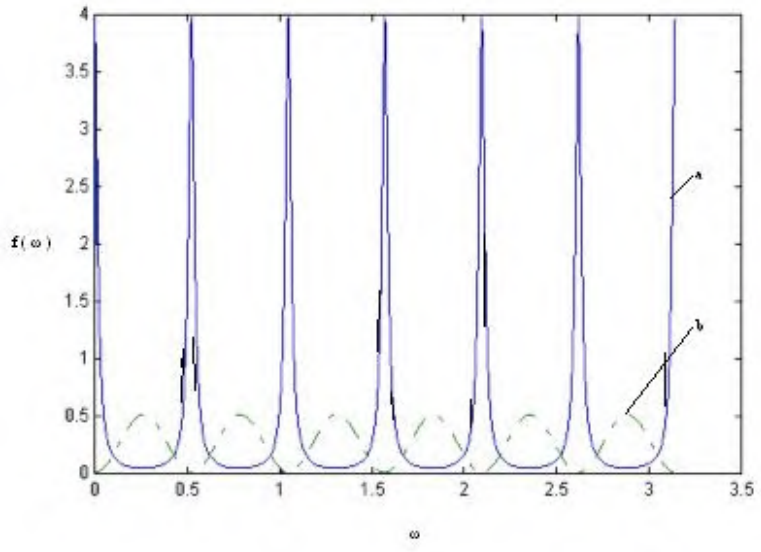
Solution:

a.

$$f(\omega) \;=\; \frac{\sigma^2}{2\pi}[(1-0.8e^{12i\omega})(1-0.8e^{-12i\omega})]^{-1} \;=\; \frac{\sigma^2}{2\pi}(1-1.6cos12\omega+0.64)^{-1}$$

b.

$$f(\omega) \;=\; \frac{\sigma^2}{2\pi}[(1+0.8e^{12i\omega})(1+0.8e^{-12i\omega})] \;=\; \frac{\sigma^2}{2\pi}(1+1.6cos12\omega+0.64)$$



27. Regression in the frequency domain.

    The asymptotic diagonalization theorem provides the key not only to approximate (i.e., asymptotic) MLE of time-series models in the frequency domain, but also to many other important techniques, such as Hannan efficient regression:

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

    But asymptotically,

$$\Sigma = P'DP$$

    so

$$\Sigma^{-1} = P'D^{-1}P$$

    Thus asymptotically

$$\hat{\beta}_{GLS} = (X'P'D^{-1}PX)^{-1}X'P'D^{-1}PY$$

    which is just WLS on Fourier transformed data.

28. Band spectral regression

4.8 NOTES

Harmonic analysis is one of the earliest methods of analyzing time series thought to exhibit some form of periodicity. In this type of analysis, the time series, or some simple transformation of it, is assumed to be the result of the superposition of sine and cosine waves of different frequencies. However, since summing a finite number of such strictly periodic functions always results in a perfectly periodic series, which is seldom observed in practice, one usually allows for an additive stochastic component, sometimes called "noise." Thus, an observer must confront the problem of searching for "hidden periodicities" in the data, that is, the unknown frequencies and amplitudes of sinusoidal fluctuations hidden amidst noise. An early method for this purpose is periodogram analysis, initially used to analyse sunspot data, and later to analyse economic time series.

Spectral analysis is a modernized version of periodogram analysis modified to take account of the stochastic nature of the entire time series, not just the noise component. If it is assumed that economic time series are fully stochastic, it follows that the older periodogram technique is inappropriate and that considerable difficulties in the interpretation of the periodograms of economic series may be encountered.

These notes draw in part on Diebold, Kilian and Nerlove, New Palgrave, ***.

Chapter Five

Markovian Structure, Linear Gaussian State Space, and Optimal (Kalman) Filtering

5.1 MARKOVIAN STRUCTURE

5.1.1 The Homogeneous Discrete-State Discrete-Time Markov Process

$$\{X_t\}, t = 0, 1, 2, \ldots$$

Possible values ("states") of X_t: $1, 2, 3, \ldots$

First-order homogeneous Markov process:

$$Prob(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0)$$

$= Prob(X_{t+1} = j | X_t = i) = p_{ij}$
1-step transition probabilities:

$$\begin{array}{rc} & [time\,(t+1)] \\ [time\ t] & \\ P \equiv & \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdots \\ & & \cdot \end{pmatrix} \end{array}$$

$p_{ij} \geq 0, \quad \sum_{j=1}^{\infty} p_{ij} = 1$

5.1.2 Multi-Step Transitions: Chapman-Kolmogorov

m-step transition probabilities:

$$p_{ij}^{(m)} = Prob(X_{t+m} = j \mid \ X_t = i)$$

Let $P^{(m)} \equiv \left(p_{ij}^{(m)}\right)$.

Chapman-Kolmogorov theorem:

$$P^{(m+n)} = P^{(m)}P^{(n)}$$

Corollary: $P^{(m)} = P^m$

5.1.3 Lots of Definitions (and a Key Theorem)

State j is accessible from state i if $p_{ij}^{(n)} > 0$, for some n.

Two states i and j communicate (or are in the same class) if each is accessible from the other. We write $i \leftrightarrow j$.

A Markov process is irreducible if there exists only one class (i.e., all states communicate).

State i has period d if $p_{ii}^{(n)} = 0 \ \forall n$ such that $n/d \notin Z$, and d is the greatest integer with that property. (That is, a return to state i can only occur in multiples of d steps.) A state with period 1 is called an aperiodic state.

A Markov process all of whose states are aperiodic is called an aperiodic Markov process.

Still more definitions....

The first-transition probability is the probability that, starting in i, the first transition to j occurs after n transitions:

$$f_{ij}^{(n)} = Prob(X_n = j, X_k \neq j, k = 1, ..., (n-1) | X_0 = i)$$

Denote the eventual transition probability from i to j by f_{ij} $(= \sum_{n=1}^{\infty} f_{ij}^{(n)})$.

State j is recurrent if $f_{jj} = 1$ and transient otherwise.

Denote the expected number of transitions needed to return to recurrent state j by $\mu_{jj} (= \sum_{n=1}^{\infty} n f_{jj}^{(n)})$.

A recurrent state j is positive recurrent if $\mu_{jj} < \infty$ null recurrent if $\mu_{jj} = \infty$.

One final definition...

The row vector π is called the stationary distribution for P if:

$\pi P = \pi$.

The stationary distribution is also called the steady-state distribution.

Theorem: Consider an irreducible, aperiodic Markov process.

Then either:

(1) All states are transient or all states are null recurrent

$p_{ij}^{(n)} \to 0$ as $n \to \infty$ $\forall i, j$. No stationary distribution.

or

(2) All states are positive recurrent.

$p_{ij}^{(n)} \to \pi_j$ as $n \to \infty$ $\forall i, j$. $\{\pi_j, j = 1, 2, 3, ...\}$ is the unique stationary distribution.
π is any row of $\lim_{n\to\infty} P^n$.

5.1.4 A Simple Two-State Example

Consider a Markov process with transition probability matrix:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Call the states 1 and 2.

We will verify many of our claims, and we will calculate the steady-state distribution.

5.1.4.1 Valid Transition Probability Matrix

$$p_{ij} \geq 0 \quad \forall i, j$$
$$\sum_{j=1}^{2} p_{1j} = 1, \quad \sum_{j=1}^{2} p_{2j} = 1$$

5.1.4.2 Chapman-Kolmogorov Theorem (for $P^{(2)}$)

$$P^{(2)} = P \cdot P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

5.1.4.3 Communication and Reducibility

Clearly, $1 \leftrightarrow 2$, so P is irreducible.

5.1.4.4 Periodicity

$$\begin{aligned} \text{State 1:} \quad & d(1) = 2 \\ \text{State 2:} \quad & d(2) = 2 \end{aligned}$$

5.1.4.5 First and Eventual Transition Probabilities

$$\begin{aligned} f_{12}^{(1)} = 1,\ f_{12}^{(n)} = 0\ \forall n > 1 \quad &\Rightarrow\ f_{12} = 1 \\ f_{21}^{(1)} = 1,\ f_{21}^{(n)} = 0\ \forall n > 1 \quad &\Rightarrow\ f_{21} = 1 \end{aligned}$$

5.1.4.6 Recurrence

Because $f_{21} = f_{12} = 1$, both states 1 and 2 are recurrent.

Moreover,

$$\mu_{11} = \sum_{n=1}^{\infty} n f_{11}^{(n)} = 2 < \infty \quad \text{(and similarly } \mu_{22} = 2 < \infty)$$

Hence states 1 and 2 are positive recurrent.

5.1.4.7 Stationary Probabilities

We will guess and verify.

Let $\pi_1 = .5$, $\pi_2 = .5$ and check $\pi P = \pi$:

$$(.5, .5) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (.5, .5).$$

Hence the stationary probabilities are 0.5 and 0.5.

Note that in this example we can *not* get the stationary probabilities by taking $\lim_{n\to\infty} P^n$. Why?

5.1.5 Constructing Markov Processes with Useful Steady-State Distributions

In section 5.1.4 we considered an example of the form, "for a given Markov process, characterize its properties." Interestingly, many important tools arise from the reverse consideration, "For a given set of properties, find a Markov process with those properties."

5.1.5.1 Markov Chain Monte Carlo

(e.g., Gibbs sampling)
We want to sample from $f(z) = f(z_1, z_2)$
Initialize ($j = 0$) using $z_2^{(0)}$
Gibbs iteration $j = 1$: Draw $z_{(1)}^1$ from $f(z_1|z_2^{(0)})$, draw $z_2^{(1)}$ from $f(z_2|z_1^{(1)})$
Repeat $j = 2,\ 3,\ ....$

5.1.5.2 Global Optimization

(e.g., simulated annealing)
If $\theta^c \notin N(\theta^{(m)})$ then $P(\theta^{(m+1)} = \theta^c|\theta^{(m)}) = 0$
If $\theta^c \in N(\theta^{(m)})$ then $P(\theta^{(m+1)} = \theta^c \,|\, \theta^{(m)}) \ = \ \ exp\ (\min[0,\ \Delta/T(m)])$

5.1.6 Variations and Extensions: Regime-Switching and More

5.1.6.1 Markovian Regime Switching

$$P = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}$$

$$s_t \sim P$$

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

$$\varepsilon_t \sim iid\, N(0, \sigma_{s_t}^2)$$

"Markov switching," or "hidden Markov," model

Popular model for macroeconomic fundamentals

5.1.6.2 Heterogeneous Markov Processes

$$P_t = \begin{pmatrix} p_{11,t} & p_{12,t} & \cdots \\ p_{21,t} & p_{22,t} & \cdots \\ . & . & \cdots \\ . & . & \cdots \end{pmatrix}.$$

e.g., Regime switching with time-varying transition probabilities:

$$s_t \sim P_t$$

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

$$\varepsilon_t \sim iid\, N(0, \sigma^2_{s_t})$$

Business cycle duration dependence: $p_{ij,t} = g_{ij}(t)$
Credit migration over the cycle: $p_{ij,t} = g_{ij}(cycle_t)$
General covariates: $p_{ij,t} = g_{ij}(x_t)$

5.1.6.3 Semi-Markov Processes

We call semi-Markov a process with transitions governed by P, such that the state durations (times between transitions) are themselves random variables. The process is not Markov, because conditioning not only the current state but also time-to-date in state may be useful for predicting the future, but there is an embedded Markov process.

Key result: The stationary distribution depends only on P and the expected state durations. Other aspects of the duration distribution are irrelevant.

Links to Diebold-Rudebush work on duration dependence: If welfare is affected only by limiting probabilities, then the mean of the duration distribution is the only relevant aspect. Other, aspects, such as spread, existence of duration dependence, etc. are irrelevant.

5.1.6.4 Time-Reversible Processes

Theorem: If $\{X_t\}$ is a stationary Markov process with transition probabilities p_{ij} and stationary probabilities π_i, then the reversed process is also Markov with transition probabilities

$$p^*_{ij} = \frac{\pi_j}{\pi_i} p_{ji}.$$

In general, $p^*_{ij} \neq p_{ij}$. In the special situation $p^*_{ij} = p_{ij}$ (so that $\pi_i p_{ij} = \pi_j p_{ji}$), we say that the process is time-reversible.

5.1.7 Continuous-State Markov Processes

5.1.7.1 Linear Gaussian State Space Systems

$$\alpha_t \;=\; T\alpha_{t-1} \;+\; R\eta_t$$

$$y_t \;=\; Z\alpha_t \;+\; \varepsilon_t$$

$$\eta_t \sim N,\; \varepsilon_t \sim N$$

5.1.7.2 Non-Linear, Non-Gaussian State Space Systems

$$\alpha_t = Q(\alpha_{t-1}, \eta_t)$$

$$y_t = G(\alpha_t, \varepsilon_t)$$

$$\eta_t \sim D^{\eta},\ \varepsilon_t \sim D^{\varepsilon}$$

Still Markovian!

5.2 STATE SPACE REPRESENTATIONS

5.2.1 The Basic Framework

Transition Equation

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$t = 1,\ 2,\ ...,\ T$$

Measurement Equation

$$\underset{1x1}{y_t} = \underset{1xm}{Z}\ \underset{mx1}{\alpha_t} + \underset{1xL}{\Gamma}\ \underset{Lx1}{w_t} + \underset{1x1}{\varepsilon_t}$$

(This is for univariate y. We'll do multivariate shortly.)

$$t = 1,\ 2,\ ...,\ T$$

(Important) Details

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(0, \text{diag}(\underbrace{Q}_{g\times g}, \underbrace{h}_{1\times 1})\right)$$

$$E(\alpha_0\ \eta_t\ ') = 0_{mxg}$$

$$E(\alpha_0\ \varepsilon_t) = 0_{mx1}$$

All Together Now

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{1x1}{y_t} = \underset{1xm}{Z}\ \underset{mx1}{\alpha_t} + \underset{1xL}{\Gamma}\ \underset{Lx1}{w_t} + \underset{1x1}{\varepsilon_t}$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(\mathbf{0}, \text{diag}(\underbrace{Q}_{g\times g}, \underbrace{h}_{1\times 1})\right)$$

$$E(\alpha_0\ \varepsilon_t) = 0_{mx1} \qquad E(\alpha_0\ \eta_t\ ') = 0_{mxg}$$

State Space Representations Are Not Unique
Transform by the nonsingular matrix B.
The original system is:

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{1x1}{y_t} = \underset{1xm}{Z}\ \underset{mx1}{\alpha_t} + \underset{1xL}{\Gamma}\ \underset{Lx1}{w_t} + \underset{1x1}{\varepsilon_t}$$

Rewrite the system in two steps
First, write it as:

$$\underset{mx1}{\alpha_t} = \underset{mxm}{T}\ \underset{mxm}{B^{-1}}\ \underset{mxm}{B}\ \underset{mx1}{\alpha_{t-1}} + \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{1x1}{y_t} = \underset{1xm}{Z}\ \underset{mxm}{B^{-1}}\ \underset{mxm}{B}\ \underset{mx1}{\alpha_t} + \underset{mxL}{\Gamma}\ \underset{Lx1}{w_t} + \underset{1x1}{\varepsilon_t}$$

Second, premultiply the transition equation by B to yield:

$$\underset{mx1}{(B\ \alpha_t)} = \underset{mxm}{(B\ T\ B^{-1})}\ \underset{mx1}{(B\ \alpha_{t-1})} + \underset{mxg}{(B\ R)}\ \underset{gx1}{\eta_t}$$

$$\underset{1x1}{y_t} = \underset{1xm}{(Z\ B^{-1})}\ \underset{mx1}{(B\ \alpha_t\ )} + \underset{mxL}{\Gamma}\ \underset{Lx1}{w_t} + \underset{1x1}{\varepsilon_t}$$

(Equivalent State Space Representation)

5.2.2 ARMA Models

State Space Representation of an AR(1)

$$y_t = \phi\ y_{t-1} + \eta_t$$

$$\eta_t \sim WN(0, \sigma_\eta^2)$$

Already in state space form!

$$\alpha_t = \phi\ \alpha_{t-1} + \eta_t$$

$$y_t = \alpha_t$$

$(T = \phi,\ R = 1,\ Z = 1,\ \Gamma = 0,\ Q = \sigma_\eta^2,\ h = 0)$
MA(1)

$$y_t = \Theta(L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0,\ \sigma^2)$$

where

$$\Theta(L) = 1 + \theta_1 L$$

MA(1) in State Space Form

$$y_t = \eta_t + \theta\ \eta_{t-1}$$

$$\eta_t \sim WN(0,\ \sigma_\eta^2)$$

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ \theta \end{pmatrix} \eta_t$$

$$y_t = (1,\ 0)\ \alpha_t = \alpha_{1t}$$

MA(1) in State Space Form
Why? Recursive substitution from the bottom up yields:

$$\alpha_t = \begin{pmatrix} y_t \\ \theta\eta_t \end{pmatrix}$$

MA(q)

$$y_t = \Theta(L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0,\ \sigma^2)$$

where

$$\Theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$$

MA(q) in State Space Form

$$y_t = \eta_t + \theta_1\ \eta_{t-1} + ... + \theta_q\ \eta_{t-q}$$

$$\eta_t \sim WN\ N(0, \sigma_\eta^2)$$

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{q+1,t} \end{pmatrix} = \begin{pmatrix} 0 & \\ 0 & I_q \\ \vdots & \\ 0 & 0' \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{q+1,t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_q \end{pmatrix} \eta_t$$

$$y_t = (1,\ 0,\ ...,\ 0)\ \alpha_t = \alpha_{1t}$$

MA(q) in State Space Form
Recursive substitution from the bottom up yields:

$$\alpha_t \equiv \begin{pmatrix} \theta_q\eta_{t-q} + \ldots + \theta_1\eta_{t-1} + \eta_t \\ \vdots \\ \theta_q\eta_{t-1} + \theta_{q-1}\eta_t \\ \theta_q\eta_t \end{pmatrix} = \begin{pmatrix} y_t \\ \vdots \\ \theta_q\eta_{t-1} + \theta_{q-1}\eta_t \\ \theta_q\eta_t \end{pmatrix}$$

AR(p)

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

$$\varepsilon_t \sim WN(0, \sigma^2)$$

where

$$\Phi(L) = (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$$

AR(p) in State Space Form

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \eta_t$$

$$\eta_t \sim WN(0, \sigma_\eta^2)$$

$$\alpha_t = \begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{pt} \end{pmatrix} = \begin{pmatrix} \phi_1 & \\ \phi_2 & I_{p-1} \\ \vdots & \\ \phi_p & 0' \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{p,t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \eta_t$$

$$y_t = (1, 0, ..., 0)\, \alpha_t = \alpha_{1t}$$

AR(p) in State Space Form
Recursive substitution from the bottom up yields:

$$\alpha_t = \begin{pmatrix} \alpha_{1t} \\ \vdots \\ \alpha_{p-1,t} \\ \alpha_{pt} \end{pmatrix} = \begin{pmatrix} \phi_1\alpha_{1,t-1} + \ldots + \phi_p\alpha_{1,t-p} + \eta_t \\ \vdots \\ \phi_{p-1}\alpha_{1,t-1} + \phi_p\alpha_{1,t-2} \\ \phi_p\alpha_{1,t-1} \end{pmatrix}$$

$$= \begin{pmatrix} y_t \\ \vdots \\ \phi_{p-1}y_{t-1} + \phi_p y_{t-2} \\ \phi_p y_{t-1} \end{pmatrix}$$

ARMA(p,q)

$$\Phi(L)y_t \;=\; \Theta(L)\varepsilon_t$$

$$\varepsilon_t \;\sim\; WN(0,\;\sigma^2)$$

where

$$\Phi(L) \;=\; (1-\phi_1 L \;-\; \phi_2 L^2 \;-\; ...\; -\; \phi_p L^p)$$

$$\Theta(L) \;=\; 1 \;+\; \theta_1 L \;+\; ...\; +\; \theta_q L^q$$

ARMA(p,q) in State Space Form

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \eta_t + \theta_1 \eta_{t-1} + ... + \theta_q \eta_{t-q}$$

$$\eta_t \sim WN(0, \sigma_\eta^2)$$

Let $m = max(p, q+1)$ and write as ARMA$(m, m-1)$:

$$(\phi_1,\; \phi_2,\; ...,\; \phi_m) \;=\; (\phi_1,\; ...,\; \phi_p,\; 0,\; ...,\; 0)$$

$$(\theta_1,\; \theta_2,\; ...,\; \theta_{m-1}) \;=\; (\theta_1,\; ...,\; \theta_q,\; 0,\; ...,\; 0)$$

ARMA(p,q) in State Space Form

$$\alpha_t \;=\; \begin{pmatrix} \phi_1 & \\ \phi_2 & I_{m-1} \\ \vdots & \\ \phi_m & 0' \end{pmatrix} \alpha_{t-1} \;+\; \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{pmatrix} \eta_t$$

$$y_t \;=\; (1, 0, ..., 0)\;\alpha_t$$

ARMA(p,q) in State Space Form Recursive substitution from the bottom up yields:

$$\begin{pmatrix} \alpha_{1t} \\ \vdots \\ \alpha_{m-1,t} \\ \alpha_{mt} \end{pmatrix} \;=\; \begin{pmatrix} \phi_1\alpha_{1,t-1} \;+\; \phi_p\alpha_{1,t-p} \;+\; \eta_t \;+\; \theta_1\eta_{t-1} \;+\; \ldots \;+\; \theta_q\eta_{t-q} \\ \vdots \\ \phi_{m-1}\alpha_{1,t-1} \;+\; \alpha_{m,t-1} \;+\; \theta_{m-2}\eta_t \\ \phi_m\alpha_{1,t-1} \;+\; \theta_{m-1}\eta_t \end{pmatrix}$$

$$= \begin{pmatrix} y_t \\ \vdots \\ \phi_{m-1}y_{t-1} \;+\; \phi_m y_{t-2} \;+\; \theta_{m-1}\eta_{t-1} \;+\; \theta_{m-2}\eta_t \\ \phi_m y_{t-1} \;+\; \theta_{m-1}\eta_t \end{pmatrix}$$

Multivariate State Space

(Same framework, $N > 1$ observables)

$$\underset{mx1}{\alpha_t} \quad = \quad \underset{mxm}{T} \quad \underset{mx1}{\alpha_{t-1}} \quad + \quad \underset{mxg}{R} \quad \underset{gx1}{\eta_t}$$

$$\underset{Nx1}{y_t} = \underset{Nxm}{Z}\ \underset{mx1}{\alpha_t} + \underset{NxL}{\Gamma}\ \underset{Lx1}{W_t} + \underset{Nx1}{\varepsilon_t}$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(0, \text{diag}(\underbrace{Q}_{g\times g}, \underbrace{H}_{N\times N})\right)$$

$$E(\alpha_0\ \eta_t\ ')\ =\ 0_{mxg} \qquad E(\alpha_0\ \varepsilon_t\ ')\ =\ 0_{mxN}$$

N-Variable $VAR(p)$

$$\underset{Nx1}{y_t} = \underset{NxN}{\Phi_1}\ y_{t-1}\ +\ ...\ +\ \Phi_p y_{t-p}\ +\ \underset{Nx1}{\eta_t}$$

$$\eta_t\ \sim\ WN(0,\ \Sigma)$$

State Space Representation

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{pt} \end{pmatrix}_{Npx1} = \begin{pmatrix} \Phi_1 & \\ \Phi_2 & I_{N(p-1)} \\ \vdots & \\ \Phi_p & 0' \end{pmatrix}_{NpxNp} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{p,t-1} \end{pmatrix}_{Npx1} + \begin{pmatrix} I_N \\ 0_{NxN} \\ \vdots \\ 0_{NxN} \end{pmatrix}_{NPxN} \eta_t$$

$$\underset{Nx1}{y_t} = \underset{NxNp}{(I_N,\ 0_N\ ,\ ...,\ 0_N)}\ \underset{Npx1}{\alpha_t}$$

Multivariate ARMA(p,q)

$$\underset{Nx1}{y_t} = \underset{NxN}{\Phi_1}\ y_{t-1}\ +\ ...\ +\ \underset{NxN}{\Phi_p}\ y_{t-p}$$

$$+\ \eta_t\ +\ \underset{NxN}{\Theta_1}\ \eta_{t-1}\ +\ ...\ +\ \underset{NxN}{\Theta_q}\ \eta_{t-q}$$

$$\eta_t\ \sim\ WN(0,\ \Sigma)$$

Multivariate ARMA(p,q)

$$\underset{Nmx1}{\alpha_t} = \begin{pmatrix} \Phi_1 & \\ \Phi_2 & I_{N(m-1)} \\ \vdots & \\ \Phi_m & 0_{NxN(m-1)} \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} I \\ \Theta_1 \\ \vdots \\ \Theta_{m-1} \end{pmatrix} \eta_t$$

$$y_t = (I,\ 0,\ ...,\ 0)\ \alpha_t = \alpha_{1t}$$

$$\text{where } m = \max(p,\ q+1)$$

5.2.3 Linear Regression with Time-Varying Parameters and More

Linear Regression Model, I

Transition: Irrelevant

Measurement:

$$y_t = \beta' x_t + \varepsilon_t$$

(Just a measurement equation with exogenous variables)

$(T = 0,\ R = 0,\ Z = 0,\ \gamma = \beta',\ W_t = x_t,\ H = \sigma_\varepsilon^2)$

Linear Regression Model, II

Transition:

$$\alpha_t = \alpha_{t-1}$$

Measurement:

$$y_t = x_t' \alpha_t + \varepsilon_t$$

$(T = I,\ R = 0,\ Z_t = x_t',\ \gamma = 0,\ H = \sigma_\varepsilon^2\ )$

Note the time-varying system matrix.

Linear Regression with ARMA(p,q) Disturbances

$$y_t = \beta x_t + u_t$$

$$u_t = \phi_1\ u_{t-1} + ... + \phi_{p\ u_{t-p}} + \eta_t + \phi_1\ \eta_{t-1} + ... + \theta_q\ \eta_{t-q}$$

$$\alpha_t = \begin{pmatrix} \phi_1 & \\ \phi_2 & I_{m-1} \\ \vdots & \\ \phi_m & 0' \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{pmatrix} \eta_t$$

$$y_t \;=\; (1,\; 0,\; ...,\; 0)\alpha_t \;+\; \beta x_t$$

$$\text{where } m \;=\; \max(p,\; q+1)$$

Linear Regression with Time-Varying Coefficients
Transition:

$$\alpha_t \;=\; \phi\;\alpha_{t-1} \;+\; \eta_t$$

Measurement:

$$y_t \;=\; x_t'\alpha_t \;+\; \varepsilon_t$$

$(T = \phi,\; R = I,\; Q = cov(\eta_t),\; Z_t = x_t',\; \gamma = 0,\; H = \sigma_\varepsilon^2)$
– Gradual evolution of tastes, technologies and institutions
– Lucas critique
– Stationary or non-stationary

5.2.3.1 Simultaneous Equations

N-Variable Dynamic SEM
+
Structure:

$$\phi_0 y_t \;=\; \phi_1 y_{t-1} \;+\; ... \;+\; \phi_p y_{t-p} \;+\; P\,\eta_t$$

$$\eta_t \;\sim\; WN(0,\; I)$$

Reduced form:

$$y_t \;=\; \phi_0^{-1}\phi_1 y_{t-1} \;+\; ... \;+\; \phi_0^{-1}\phi_p y_{t-p} \;+\; \phi_0^{-1} P\,\eta_t$$

Assume that the system is identified.
SEM State Space Representation

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \vdots \\ \alpha_{pt} \end{pmatrix} = \begin{pmatrix} \phi_0^{-1}\phi_1 & \\ \phi_0^{-1}\phi_2 & I \\ \vdots & \\ \phi_0^{-1}\phi_p & 0' \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ \vdots \\ \alpha_{p,t-1} \end{pmatrix} + \begin{pmatrix} \phi_0^{-1}P \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} \eta_{1t} \\ \vdots \\ \eta_{Nt} \end{pmatrix}$$

$$\underset{Nx1}{y_t} \;=\; \underset{NxNp}{(I_N,\quad 0_N\quad ,\; ...,\; 0_N)} \quad \underset{Npx1}{\alpha_t}$$

5.2.4 Dynamic Factor Models and Cointegration

Dynamic Factor Model – Single AR(1) factor

(White noise idiosyncratic factors uncorrelated with each other and uncorrelated with the factor at all leads and lags...)

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix} F_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$F_t = \phi F_{t-1} + \eta_t$$

Already in state-space form!

Dynamic Factor Model – Single ARMA(p,q) Factor

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix} F_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$\Phi(L)\, F_t = \Theta(L)\, \eta_t$$

Dynamic Factor Model – Single ARMA(p,q) Factor State vector for F is state vector for system:

$$\alpha_t = \begin{pmatrix} \phi_1 & \\ \phi_2 & I_{m-1} \\ \vdots & \\ \phi_m & 0' \end{pmatrix} \alpha_{t-1} + \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{pmatrix} \eta_t$$

Dynamic Factor Model – Single ARMA(p,q) factor System measurement equation is then:

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix} (1,\ 0,\ ...,\ 0)\ \alpha_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$= \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ \vdots & & & \\ \lambda_N & 0 & ... & 0 \end{pmatrix} \alpha_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

Cointegration (A Special Dynamic Factor Model)

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \alpha_t + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

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

"Common trend" α_t

Note that $\frac{y_{1t}}{\lambda_1} - \frac{y_{2t}}{\lambda_2} = \frac{\varepsilon_{1t}}{\lambda_1} - \frac{\varepsilon_{2t}}{\lambda_2}$

That is, $I(1) - I(1) = I(0)$

"CI(1,0)"

5.2.5 Unobserved-Components Models

– Separate components for separate features
  – Signal extraction
  – Trends and detrending
  – Seasonality and seasonal adjustment
  – Permanent-transitory decompositions

Cycle ("Signal") + Noise Model

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

$$y_t \;=\; x_t \;+\; \varepsilon_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \;\sim\; WN\left(0,\; \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}\right)$$

$(\alpha_t = x_t,\ T = \phi,\ R = 1,\ Z = 1,\ \gamma = 0,\ Q = \sigma_\eta^2,\ H = \sigma_\varepsilon^2)$

Cycle + Seasonal + Noise

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

$$c_t \;=\; \phi\; c_{t-1} \;+\; \eta_{ct}$$

$$s_t \;=\; \gamma\; s_{t-4} \;+\; \eta_{st}$$

Cycle + Seasonal + Noise

Transition equations for the cycle and seasonal:

$$\alpha_{ct} \;=\; \phi\; \alpha_{c,t-1} \;+\; \eta_{ct}$$

$$\alpha_{st} \;=\; \begin{pmatrix} 0 & \\ 0 & I_3 \\ 0 & \\ \gamma & 0' \end{pmatrix} \alpha_{s,t-1} \;+\; \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \eta_{st}$$

Cycle + Seasonal + Noise Stacking transition equations gives the grand transition equation:

$$\begin{pmatrix} \alpha_{st} \\ \alpha_{ct} \end{pmatrix} \;=\; \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi \end{pmatrix} \begin{pmatrix} \alpha_{s,t-1} \\ \alpha_{c,t-1} \end{pmatrix} \;+\; \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{st} \\ \eta_{ct} \end{pmatrix}$$

Finally, the measurement equation is:

$$y_t \;=\; (1,\ 0,\ 0,\ 0,\ 1) \begin{pmatrix} \alpha_{st} \\ \alpha_{ct} \end{pmatrix} \;+\; \varepsilon_t$$

5.3 THE KALMAN FILTER AND SMOOTHER

State Space Representation

$$\underset{mx1}{\alpha_t} \;=\; \underset{mxm}{T}\ \underset{mx1}{\alpha_{t-1}} \;+\; \underset{mxg}{R}\ \underset{gx1}{\eta_t}$$

$$\underset{Nx1}{y_t} \;=\; \underset{Nxm}{Z}\ \underset{mx1}{\alpha_t} \;+\; \underset{NxL}{\gamma}\ \underset{Lx1}{W_t} \;+\; \underset{Nx1}{\varepsilon_t}$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \sim WN\left(0, \operatorname{diag}(\underbrace{Q}_{g\times g}, \underbrace{H}_{N\times N})\right)$$

$$E(\alpha_0\ \eta_t') \;=\; 0_{mxg}$$

$$E(\alpha_0\ \varepsilon_t') \;=\; 0_{mxN}$$

5.3.1 Statement(s) of the Kalman Filter

I. Initial state estimate and MSE

$$a_0 \;=\; E(\alpha_0)$$

$$P_0 \;=\; E(\alpha_0 \;-\; a_0)\,(\alpha_0 \;-\; a_0)'$$

Statement of the Kalman Filter

II. Prediction Recursions

$$a_{t/t-1} \;=\; T\,a_{t-1}$$

$$P_{t/t-1} \;=\; T\;P_{t-1}\;T' \;+\; R\;Q\;R'$$

III. Updating Recursions

$$a_t \;=\; a_{t/t-1} \;+\; P_{t/t-1}\;Z'\;F_t^{-1}\;(y_t \;-\; Za_{t/t-1} \;-\; \gamma W_t)$$

$$(\text{where } F_t \;=\; Z\,P_{t/t-1}\;Z' \;+\; H)$$

$$P_t \;=\; P_{t/t-1} \;-\; P_{t/t-1}\;Z'\,F_t^{-1}\;Z\;P_{t/t-1}$$

t = 1, ..., T

State-Space in Density Form (Assuming Normality)

$$\alpha_t|\alpha_{t-1} \sim N(T\alpha_{t-1},\; RQR')$$

$$y_t|\alpha_t \sim N(Z\alpha_t,\; H)$$

Kalman Filter in Density Form (Assuming Normality)

Initialize at a_0, P_0

State prediction:

$\alpha_t|\tilde{y}_{t-1} \sim N(a_{t/t-1},\; P_{t/t-1})$

$a_{t/t-1} = Ta_{t-1}$

$P_{t/t-1} = TP_{t-1}T' + RQR'$

Data prediction:

$y_t|\tilde{y}_{t-1} \sim N(Za_{t/t-1},\; F_t)$

Update:

$\alpha_t|\tilde{y}_t \sim N(a_t,\; P_t)$

$a_t = a_{t/t-1} + K_t(y_t - Za_{t/t-1})$
$P_t = P_{t/t-1} - K_t Z P_{t/t-1}$
where $\tilde{y}_t = \{y_1, ..., y_t\}$

5.3.2 Derivation of the Kalman Filter

Useful Result 1: Conditional Expectation is MMSE Extraction Under Normality Suppose that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\mu, \Sigma\right)$$

where x is unobserved and y is observed.

Then

$$E(x|y) = argmin_{\hat{x}(y)} \int\int (x - \hat{x}(y))^2 \; f(x,y) \; dx \; dy$$

Useful Result 2: Properties of the Multivariate Normal

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\mu, \Sigma\right) \quad \mu = (\mu_x, \mu_y)' \qquad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

$$\Longrightarrow x|y \sim \quad N\left(\mu_{x|y} \; , \; \Sigma_{x|y}\right)$$

$$\mu_{x|y} \;=\; \mu_x \;+\; \Sigma_{xy} \; \Sigma_{yy}^{-1} \; (y \;-\; \mu_y)$$

$$\Sigma_{x|y} \;=\; \Sigma_{xx} \;-\; \Sigma_{xy} \; \Sigma_{yy}^{-1} \; \Sigma_{yx}$$

Constructive Derivation of the Kalman Filter
Under Normality

Let $E_t(\cdot) \;\equiv\; E(\cdot\,|\Omega_t)$, where $\Omega_t \equiv \{y_1, ..., y_t\}$.

Time 0 "update":

$$a_0 \;=\; E_0\left(\alpha_0\right) \;=\; E\left(\alpha_0\right)$$

$$P_0 \;=\; var_0(\,\alpha_0) \;=\; E\left[(\alpha_0 - a_0)\;(\alpha_0 - a_0)'\right]$$

Derivation of the Kalman Filter, Continued...

Time 0 prediction

$$\alpha_1 \;=\; T\alpha_0 \;+\; R\eta_1$$

$$\Longrightarrow\ a_{1/0}\ =\ E_0(\alpha_1)\ =\ T\,E_0(\alpha_0) + RE_0(\eta_1)$$

$$=\ Ta_0$$

Derivation of the Kalman Filter, Continued...

$$P_{1/0}\ =\ E_0\Big((\alpha_1 - a_{1/0})\ (\alpha_1 - a_{1/0})'\Big)$$

$$(\text{subst. } a_{1/0})\ =\ E_0\Big((\alpha_1 - Ta_0)\ (\alpha_1 - Ta_0)'\Big)$$

$$(\text{subst. } \alpha_1)\ =\ E_0\Big((T(\alpha_0 - a_0)\ +\ R\eta_1)\ (T(\alpha_0 - a_0)\ +\ R\eta_1)'\Big)$$

$$=\ TP_0\,T'\ +\ RQR'$$

$$(\text{using }\ E(\alpha_0\,\eta_t')\ =\ 0\ \forall t)$$

Derivation of the Kalman Filter, Continued...

Time 1 updating

We will derive the distribution of:

$$\left(\begin{array}{c}\alpha_1\\ y_1\end{array}\right)\Big|\Omega_0$$

and then convert to

$$\alpha_1|(\Omega_0 \cup y_1)$$

or

$$\alpha_1|\Omega_1$$

Derivation of the Kalman Filter, Continued...

Means:

$$E_0(\alpha_1)\ =\ a_{1/0}$$

$$E_0(y_1)\ =\ Za_{1/0}\ +\ \gamma W_1$$

Derivation of the Kalman Filter, Continued...

Variance-Covariance Matrix:

$$var_0(\alpha_1) \;=\; E_0\Big((\alpha_1 - a_{1/0})\;(\alpha_1 - a_{1/0})\Big) \;=\; P_{1/0}$$

$$var_0(y_1) \;=\; E_0\Big((y_1 - Za_{1/0} - \gamma W_1)\;(y_1 - Za_{1/0} - \gamma W_1)'\Big)$$

$$=\; E_0\Big((Z(\alpha_1 - a_{1/0}) \;+\; \varepsilon_1)\;(Z(\alpha_1 - a_{1/0}) \;+\; \varepsilon_1)'\Big)$$

$$=\; Z\;P_{1/0}Z' \;+\; H \text{ (using } \varepsilon \perp \eta)$$

$$cov_0\,(\alpha_1, y_1) \;=\; E_0(\alpha_1 - a_{1/0})\;(Z\,(\alpha_1 \;- a_{1/0}) + \varepsilon_1)'$$

$$=\; P_{1/0}Z' \text{ (using } \varepsilon \perp \eta)$$

Derivation of the Kalman Filter, Continued...

Hence:

$$\begin{pmatrix} \alpha_1 \\ y_1 \end{pmatrix}\Big|\Omega_0 \;\sim\; N\left(\begin{pmatrix} a_{1/0} \\ Za_{1/0} + \gamma W_1 \end{pmatrix},\; \begin{pmatrix} P_{1/0} & P_{1/0}\;Z' \\ ZP_{1/0} & ZP_{1/0}\;Z' + H \end{pmatrix}\right)$$

$$\text{Now by Result 2, } \alpha_1\,|\,\Omega_0 \cup y_1 \;\sim\; N(a_1,\; P_1)$$

$$a_1 \;=\; a_{1/0} \;+\; P_{1/0}\;Z'\;F_1^{-1}\;(y_1 \;-\; Za_{1/0} \;-\; \gamma W_1)$$

$$P_1 \;=\; P_{1/0} \;-\; P_{1/0}\;Z'\;F_1^{-1}\;Z\;P_{1/0}$$

$$(F_1 \;=\; Z\;P_{1/0}\;Z' \;+\; H)$$

Repeating yields the Kalman filter.

What Have We Done?

Under normality,

we proved that the Kalman filter delivers

MVU predictions and extractions.

Dropping normality,

one can also prove that the Kalman filter delivers

BLU predictions and extractions.

5.3.3 Calculating P_0

Treatment of Initial Covariance Matrix: $P_0 = \Gamma(0)$ (Covariance stationary case: All eigenvalues of T inside $|z| = 1$)

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

$$\implies \Gamma_\alpha(0) = E(T\alpha_{t-1} + R\eta_t)(T\alpha_{t-1} + R\eta_t)' = T\Gamma_\alpha(0)T' + RQR'$$

$$\implies P_0 = TP_0T' + RQR'$$

$$\implies vec(P_0) = vec(TP_0T') + vec(RQR')$$

$$= (T \otimes T)vec(P_0) + vec(RQR')$$

$$\implies vec(P_0) = (I - (T \otimes T))^{-1}vec(RQR')$$

5.3.4 Predicting y_t

Point prediction:

$$y_{t/t-1} \;=\; Za_{t/t-1} \;+\; \gamma W_t$$

Prediction error:

$$v_t \;=\; y_t \;-\; (Za_{t/t-1} \;+\; \gamma W_t)$$

Density Prediction of y_t

$$y_t|\Omega_{t-1} \;\sim\; N(y_{t/t-1},\; F_t)$$

or equivalently

$$v_t \,|\, \Omega_{t-1} \;\sim\; N\,(0,\; F_t)$$

$$E_{t-1}v_t \;=\; E_{t-1}[y_t \;-\; (Za_{t/t-1} + \;\gamma W_t)]$$

$$=\; E_{t-1}[Z\,(\alpha_t \;- a_{t/t-1}) \;+\; \varepsilon_t] \;=\; 0$$

$$E_{t-1}v_t v_{t'} \;=\; E_{t-1}[Z\,(\alpha_t \,-\, a_{t/t-1}) \;+\; \varepsilon_t]\,[Z\,(\alpha_t \,-\, a_{t/t-1}) \;+\; \varepsilon_t]'$$

$$=\; ZP_{t/t-1}Z' \;+\; H \;\equiv\; F_t$$

Normality follows from linearity of all transformations.

5.3.4.1 Combining State Vector Prediction and Updating

(For notational convenience we now drop W_t)

(1) Prediction: $a_{t+1/t} \;=\; Ta_t$

(2) Update: $a_t \;=\; a_{t/t-1} \;+\; P_{t/t-1}\; Z'\; F_t^{-1}\; (y_t \;-\; Za_{t/t-1})$

$=\; a_{t/t-1} \;+\; K_t\; v_t$

where

$$K_t \;=\; P_{t/t-1}\, Z'\, F_t^{-1}$$

Substituting (2) into (1):

$$a_{t+1/t} \;=\; Ta_{t/t-1} \;+\; TK_t v_t$$

5.3.5 Steady State and the Innovations Representation

Recall the "Two-Shock" State Space Representation

$$\alpha_t \;=\; T\alpha_{t-1} \;+\; R\eta_t$$

$$y_t \;=\; Z\alpha_t \;+\; \varepsilon_t$$

$$E(\eta_t\eta_t') \;=\; Q$$

$$E(\varepsilon_t\varepsilon_t') \;=\; H$$

(Nothing new)

5.3.5.1 Combining Covariance Matrix Prediction and Updating

(1) Prediction: $P_{t+1/t} \;=\; T\,P_t\,T' \;+\; RQR'$

(2) Update: $P_t \;=\; P_{t/t-1} \;-\; K_t\,Z\,P_{t/t-1}$

Substitute (2) into (1):

$$P_{t+1/t} \;=\; T\,P_{t/t-1}\,T' \;-\; T\,K_t\,Z\,P_{t/t-1}\,T' \;+\; RQR'$$

(Matrix Ricatti equation)

Why Care About Combining Prediction and Updating?

It leads us to the notion of steady state of the Kalman filter...

...which is the bridge from the Wold representation
to the state space representation

5.3.5.2 "One-Shock" ("Prediction Error") Representation

We have seen that

$$a_{t+1|t} \;=\; Ta_{t|t-1} \;+\; TK_t\;v_t \;\;\text{(transition)}$$

Moreover, it is tautologically true that

$$y_t \;=\; Z\;a_{t|t-1} \;+\; (y_t - Za_{t|t-1})$$

$$=\; Z\;a_{t|t-1} \;+\; v_t \;\;\text{(measurement)}$$

Note that one-shock state space representation
has time-varying system matrices:

- "R matrix" in transition equation is TK_t

- Covariance matrix of v_t is F_t

5.3.5.3 "Innovations" (Steady-State) Representation

$$a_{t+1|t} \;=\; Ta_{t|t-1} \;+\; T\bar{K}v_t$$

$$y_t \;=\; Z\;a_{t|t-1} \;+\; v_t$$

where

$$\bar{K} \;=\; \bar{P}\,Z'\,\bar{F}^{-1}$$

$$E(v_t v_t') \;=\bar{F} \;=\; Z\bar{P}Z' \;+\; H$$

$\bar{P}$ solves the matrix Ricatti equation
– Effectively Wold-Wiener-Kolmogorov prediction and extraction

– Prediction $y_{t+1/t}$ is now the projection of y_{t+1} on *infinite* past, and one-step prediction errors v_t are now the Wold-Wiener-Kolmogorov innovations

Remarks on the Steady State

1. Steady state will be approached if:
   - underlying two-shock system is time invariant
   - all eigenvalues of T are less than one
   - $P_{1|0}$ is positive semidefinite
2. Because the recursions for $P_{t|t-1}$ and K_t don't depend on the data, but only on P_0, we can calculate arbitrarily close approximations to $\bar{P}$ and $\bar{K}$ by letting the filter run

5.3.6 Kalman Smoothing

1. (Kalman) filter forward through the sample, $t = 1,\ ...,\ T$
2. Smooth backward, $t = T,\ (T-1),\ (T-2),\ ...,\ 1$

$$\text{Initialize: } a_{T,T} = a_T,\ P_{T,T} = P_T$$

Then:

$$a_{t,T} = a_t + J_t(a_{t+1,T} - a_{t+1,t})$$

$$P_{t,T} = P_t + J_t(P_{t+1,T} - P_{t+1,t})J_t'$$

where

$$J_t = P_t T' P_{t+1,t}^{-1}$$

5.4 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Markov process theory.

   Prove the following for Markov processes.

   (a) Communication is an equivalence relation; i.e., $i \leftrightarrow i$ (reflexive), $i \leftrightarrow j \Longleftrightarrow j \leftrightarrow i$ (symmetric), and $i \leftrightarrow j$, $j \leftrightarrow k \Longrightarrow i \leftrightarrow k$ (transitive).

   (b) Any two classes are either disjoint or identical.

   (c) Periodicity is a class property.

   (d) Let f_{ij} denote the probability of an eventual transition from i to j. Then $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^n$.

(e) The expected number of returns to a recurrent state is infinite, and the expected number of returns to a transient state is finite. That is,

$$\text{State } j \text{ is recurrent } \iff \sum_{n=1}^{\infty} P_{jj}^n = \infty,$$
$$\text{State } j \text{ is transient } \iff \sum_{n=1}^{\infty} P_{jj}^n < \infty.$$

(f) Recurrence is a class property. That is, if i is recurrent and $i \leftrightarrow j$, then j is recurrent.

(g) The expected number of transitions into a transient state is finite. That is,

$$j \text{ transient } \implies \sum_{n=1}^{\infty} P_{ij}^n < \infty, \ \forall \ i.$$

(h) Positive and null recurrence are class properties.

(i) If the probability distribution of X_0 (call it $\pi_j = P(X_0 = j), j \geq 1$) is a stationary distribution, then $P(X_t = j) = \pi_j$, $\forall t$. Moreover, the process is stationary.

(j) A stationary Markov process is time-reversible iff each path from i to i has the same probability as the reversed path, $\forall i$.

2. A Simple Markov Process.

Consider the Markov process:

$$P = \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix}$$

Verify and/or calculate:

(a) Validity of the transition probability matrix

(b) Chapman-Kolmogorov theorem

(c) Accessibility/communication

(d) Periodicity

(e) Transition times

(f) Recurrence/transience

(g) Stationarity

(h) Time reversibility

Solution:

(a) See Ross*** or any other good text.

(b) Valid transition probability matrix:

$$P_{ij} \geq 0 \qquad \forall i, j$$
$$\sum_{j=1}^{2} P_{1j} = 1, \quad \sum_{j=1}^{2} P_{2j} = 1$$

(c) Illustrating the Chapman-Kolmogorov Theorem:

$$P^3 = \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} \begin{pmatrix} .9 & .1 \\ .3 & .7 \end{pmatrix} = \begin{pmatrix} .804 & .196 \\ .588 & .412 \end{pmatrix}$$

(d) Communication and reducibility:

$1 \leftrightarrow 2$, so the Markov process represented by P is irreducible.

(e) Periodicity:

$$\text{State 1: } \mathrm{d}(1) = 1$$
$$\text{State 2: } \mathrm{d}(2) = 1$$

Hence both states are aperiodic.

(f) First and eventual transition probabilities:

First:

$$f_{12}^{(1)} = .1$$
$$f_{12}^{(2)} = .9 * .1 = .09$$
$$f_{12}^{(3)} = .9^2 * .1 = .081$$
$$f_{12}^{(4)} = .9^3 * .1 = .0729$$
$$\dots$$

$$f_{12}^{(1)} = .3$$
$$f_{21}^{(2)} = .7 * .3 = .21$$
$$f_{21}^{(3)} = .7^2 * .3 = .147$$
$$f_{21}^{(4)} = .7^3 * .3 = .1029$$
$$\dots$$

Eventual:

$$f_{12} = \sum_{n=1}^{\infty} f_{12}^{(n)} = \frac{.1}{1 - .9} = 1$$

$$f_{21} = \sum_{n=1}^{\infty} f_{21}^{(n)} = \frac{.3}{1-.7} = 1$$

(g) Recurrence:

Because $f_{12} = f_{21} = 1$, both states 1 and 2 are recurrent. In addition,

$$\begin{aligned} \mu_{11} &= \textstyle\sum_{n=1}^{\infty} n\, f_{11}^{(n)} < \infty \\ \mu_{22} &= \textstyle\sum_{n=1}^{\infty} n\, f_{22}^{(n)} < \infty \end{aligned}$$

States 1 and 2 are therefore positive recurrent and (given their aperiodicity established earlier) ergodic.

(h) Stationary distribution

We can iterate on the P matrix to see that:

$$\lim_{n\to\infty} P^n = \begin{pmatrix} .75 & .25 \\ .75 & .25 \end{pmatrix}$$

Hence $\pi_1 = 0.75$ and $\pi_2 = 0.25$.

Alternatively, in the two-state case, we can solve analytically for the stationary probabilities as follows.

$$\begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix}$$

$$\Longrightarrow \begin{aligned} \pi_1 P_{11} + \pi_2 P_{21} &= \pi_1 \quad (1) \\ \pi_1 P_{12} + \pi_2 P_{22} &= \pi_2 \quad (2) \end{aligned}$$

Using $\pi_2 = 1 - \pi_1$, we get from (1) that

$$\pi_1 P_{11} + (1-\pi_1) P_{21} = \pi_1$$

$$\begin{aligned} \pi_1 &= \frac{P_{21}}{1 - P_{11} + P_{21}} \\ \pi_2 &= \frac{1 - P_{11}}{1 - P_{11} + P_{21}} \end{aligned}$$

Thus,

$$\lim_{n\to\infty} P^n = \frac{1}{(1 - P_{11} + P_{21})} \begin{pmatrix} P_{21} & 1 - P_{11} \\ P_{21} & 1 - P_{11} \end{pmatrix}$$

(i) Time reversibility:

$$\begin{aligned} P_{12}^* &= \tfrac{\pi_2}{\pi_1} P_{21} = .1 \\ P_{21}^* &= \tfrac{\pi_2}{\pi_1} P_{12} = .3 \end{aligned}$$

We have a time-reversible process.

3. $AR(p)$ in state-space form.

   Find a state-space representation different from the one used in the text, in which the state vector is $(y_1, ..., y_p)'$. This is precisely what one expects given the intuitive idea of state: in an $AR(p)$ all history relevant for future evolution is $(y_1, ..., y_p)'$.

4. ARMA(1,1) in state space form.

$$y_t \;=\; \phi\; y_{t-1} \;+\; \eta_t \;+\; \theta\; \eta_{t-1}$$

$$\eta_t \;\sim\; WN(0, \sigma_\eta^2)$$

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix} \;=\; \begin{pmatrix} \phi & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{pmatrix} \;+\; \begin{pmatrix} 1 \\ \theta \end{pmatrix} \eta_t$$

$$y_t \;=\; (1, 0)\; \alpha_t \;=\; \alpha_{1t}$$

   Recursive substitution from the bottom up yields:

$$\alpha_t \;=\; \begin{pmatrix} \phi\; y_{t-1} \;+\; \theta\; \eta_{t-1} \;+\; \eta_t \\ \theta\; \eta_t \end{pmatrix} \;=\; \begin{pmatrix} y_t \\ \theta \eta_t \end{pmatrix}$$

5. Rational spectra and state space forms.

   Can all linear discrete-time systems be written in state-space form? In particular, what is the relationship, if any, between rational spectra and existence of a state-space form?

6. Impulse-responses and variance decompositions.

   If given a model in state space form, how would you calculate the impulse-responses and variance decompositions?

7. Identification in UCM's.

   Discuss the identifying assumption that UC innovations are orthogonal at all leads and lags. What convenient mathematical properties does it entail for the observed sum of the unobserved components? In what ways is it restrictive?

   Solution: Orthogonality of component innovations implies that the spectrum of the observed time series is simply the sum of the component spectra. Moreover, the orthogonality facilitates identification. The assumption is rather restrictive, however, in that it entails no interaction between cyclical and secular economic fluctuations.

8. ARMA(1,1) "Reduced Form" of the Signal Plus Noise Model.

   Consider the "structural" model:

$$x_t = y_t + u_t$$

$$y_t = \alpha y_{t-1} + v_t.$$

   Show that the reduced form is $ARMA(1,1)$:

$$y_t = \alpha y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}$$

   and provide expressions for σ_ε^2 and β in terms of the underlying parameters α, σ_v^2 and σ_u^2.

   Solution:

   Box and Jenkins (1976) and Nerlove et al. (1979) show the ARMA result and give the formula for β. That leaves σ_ε^2. We will compute $var(x)$ first from the UCM and then from the ARMA(1,1) reduced form, and equate them.

   From the UCM:

$$var(x) \;=\; \frac{\sigma_v^2}{1 \;-\; \alpha^2} \;+\; \sigma_u^2$$

   From the reduced form:

$$var(x) \;=\; \sigma_\epsilon^2 \;\frac{(1 \;+\; \beta^2 \;-\; 2\alpha\beta)}{1 \;-\; \alpha^2}$$

   Equating yields

$$\sigma_\epsilon^2 \;=\; \frac{\sigma_v^2 \;+\; \sigma_u^2(1 \;-\; \alpha^2)}{1 \;+\; \beta^2 \;-\; 2\alpha\beta}$$

9. Properties of optimal extractions.

   In the case where the seasonal is modeled as independent of the nonseasonal and the observed data is just the sum of the two, the following assertions are (perhaps surprisingly) both true:

   i) optimal extraction of the nonseasonal is identically equal to the observed data minus optimal extraction of the seasonal (i.e., $\mathrm{y} \equiv \hat{\mathrm{y}}_s + \hat{\mathrm{y}}_n \equiv \mathrm{y}_n$), so it doesn't matter whether you estimate the nonseasonal by optimally extracting it directly or instead optimally

extract the seasonal and substract–both methods yield the same answer;
ii) $\hat{y}_s$, the estimated seasonal, is less variable than y_s, the true seasonal, and $\hat{y}_n$, the estimated nonseasonal, is less variable than y_n, the true nonseasonal. It is paradoxical that, by (ii), both estimates are less variable than their true counterparts, yet, by (i), they still add up to the same observed series as their true counterparts. The paradox is explained by the fact that, unlike their true counterparts, the estimates $\hat{y}_s$ and $\hat{y}_n$ are correlated (so the variance of their sum can be more than the sum of their variances).

5.5 NOTES

Chapter Six

Frequentist Time-Series Likelihood Evaluation, Optimization, and Inference

6.1 LIKELIHOOD EVALUATION: PREDICTION-ERROR DECOMPOSITION AND THE KALMAN FILTER

Brute-Force Direct Evaluation:

$$y_t \ \sim \ N(\mu, \ \Sigma(\theta))$$

Example: AR(1)

$$(y_t - \mu) \ = \ \phi(y_{t-1} - \mu) \ + \ \varepsilon_t$$

$$\Sigma_{ij}\,(\phi) \ = \ \frac{\sigma^2}{1 \ - \ \phi^2} \ \phi^{|i-j|}$$

$$L(y;\theta) = (2\pi)^{T/2}|\Sigma(\theta)|^{-1/2} exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(\theta)(y-\mu)\right)$$

$$lnL(y;\theta) = const - \frac{1}{2}ln|\Sigma(\theta)| - \frac{1}{2}(y-\mu)'\Sigma^{-1}(\theta) \ (y-\mu)$$

TxT matrix $\Sigma(\theta)$ can be *very* hard to calculate (we need analytic formulas for the auto-covariances) and invert (numerical instabilities and inaccuracies; slow even if possible)

Prediction-error decomposition and the Kalman filter:

Schweppe's prediction-error likelihood decomposition is:

$$L(y_1, \ldots, y_T; \theta) = \prod_{t=1}^{T} L_t(y_t | y_{t-1}, \ldots, y_1; \theta)$$

or:

$$\ln L(y_1, \ldots, y_T; \theta) = \sum_{t=1}^{T} \ln L_t(y_t | y_{t-1}, \ldots, y_1; \theta)$$

"Prediction-error decomposition"

In the univariate Gaussian case, the Schweppe decomposition is

$$\ln L = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln \sigma_t^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(y_t-\mu_t)^2}{\sigma_t^2}$$

$$= -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln F_t - \frac{1}{2}\sum_{t=1}^{T}\frac{v_t^2}{F_t}$$

Kalman filter delivers v_t and F_t!

No autocovariance calculation or matrix inversion!

In the N-variate Gaussian case, the Schweppe decomposition is

$$\ln L = -\frac{NT}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln |\Sigma_t| - \frac{1}{2}\sum_{t=1}^{T}(y_t-\mu_t)'\Sigma_t^{-1}(y_t-\mu_t)$$

$$= -\frac{NT}{2}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln |F_t| - \frac{1}{2}\sum_{t=1}^{T}v_t'F_t^{-1}v_t$$

Kalman filter again delivers v_t and F_t.

Only the small matrix F_t $(N \times N)$ need be inverted.

6.2 GRADIENT-BASED LIKELIHOOD MAXIMIZATION: NEWTON AND QUASI-NEWTON METHODS

The key is to be able to evaluate lnL for a given parameter configuration

Then we can climb uphill to maximize lnL to get the MLE

Crude Search

Function $lnL(\theta)$ to be optimized w.r.t. θ,

$\theta \in \Theta$, a compact subset of R^k

- Deterministic search: Search k dimensions at r locations in each dimension.
- Randomized Search: Repeatedly sample from Θ, repeatedly evaluating $lnL(\theta)$

– Absurdly slow (curse of dimensionality)

6.2.1 The Generic Gradient-Based Algorithm

So use the gradient for guidance.

"Gradient-Based" Iterative Algorithms ("Line-Search")

Parameter vector at iteration m: $\theta^{(m)}$.

$\theta^{(m+1)} = \theta^{(m)} + C^{(m)}$, where $C^{(m)}$ is the **step**.
Gradient algorithms: $C^{(m)} = -t^{(m)} D^{(m)} s^{(m)}$
$t^{(m)}$ is the step length, or step size (a positive number)
$D^{(m)}$ is a positive definite direction matrix
$s^{(m)}$ is the score (gradient) vector evaluated at $\theta^{(m)}$
General Algorithm

1. Specify $\theta^{(0)}$
2. Compute $D^{(m)}$ and $s^{(m)}$
3. Determine step length $t^{(m)}$

   (Often, at each step, choose $t^{(m)}$ to optimize the objective function ("variable step length"))
4. Compute $\theta^{(m+1)}$
5. If convergence criterion not met, go to 2.

Convergence
Convergence Criteria
$\| s^{(m)} \|$ "small"
$\| \theta^{(m)} - \theta^{(m-1)} \|$ "small"
Convergence Rates
p such that $\lim_{m\to\infty} \frac{\|\theta^{(m+1)} - \hat{\theta}\|}{\|\theta^{(m)} - \hat{\theta}\|^p} = O(1)$
Method of Steepest Decent
Use $D^{(m)} = I, t^{(m)} = 1, \ \forall \ m.$
Properties:

1. May converge to a critical point other than a minimum
   (of course)
2. Requires only first derivative of the objective function
3. Very slow to converge ($p = 1$)

6.2.2 Newton Algorithm

Take $D^{(m)}$ as the inverse Hessian of $lnL(\theta)$ at $\theta^{(m)}$

$$D^{(m)} = H^{-1(m)} = \begin{pmatrix} \frac{\partial^2 lnL}{\partial \theta_1^2}|_{\theta^{(m)}} & \cdot & \cdot & \cdot & \frac{\partial^2 lnL}{\partial \theta_1 \partial \theta_k}|_{\theta^{(m)}} \\ & \cdot & & & \\ & \cdot & & & \\ & \cdot & & & \\ \frac{\partial^2 lnL}{\partial \theta_k \partial \theta_1}|_{\theta^{(m)}} & \cdot & \cdot & \cdot & \frac{\partial^2 lnL}{\partial \theta_k^2}|_{\theta^{(m)}} \end{pmatrix}^{-1}$$

Also take $t^{(m)} = 1$

Then $\theta^{(m+1)} = \theta^{(m)} - H^{-1(m)} s^{(m)}$

Derivation From Second-Order Taylor Expansion

Initial guess: $\theta^{(0)}$

$lnL(\theta) \approx lnL(\theta^{(0)}) + s^{(0)}(\theta - \theta^{(0)}) + \frac{1}{2}(\theta - \theta^{(0)})' H^{(0)} (\theta - \theta^{(0)})$

F.O.C.:

$s^{(0)} + H^{(0)}(\theta^* - \theta^{(0)}) = 0$

or

$\theta^* = \theta^{(0)} - H^{-1(0)} s^{(0)}$

Properties of Newton

$lnL(\theta)$ quadratic $\Rightarrow$ full convergence in a single iteration

More generally, iterate to convergence:

$\theta^{(m+1)} = \theta^{(m)} - H^{-1(m)} s^{(m)}$

Faster than steepest descent ($p = 2$ vs. $p = 1$)

But there is a price:

Requires first *and* second derivatives of the objective function

Requires inverse Hessian at each iteration

6.2.3 Quasi-Newton Algirithms

e.g., Davidon-Fletcher-Powell (DFP):

$D^{(0)} = I$

$$D^{(m)} = D^{(m-1)} + f\left(D^{(m-1)}, \delta^{(m)}, \phi^{(m)}\right), \quad m = 1, 2, 3, ...$$

$\delta^{(m)} = \theta^{(m)} - \theta^{(m-1)}$

$\phi^{(m)} = s^{(m)} - s^{(m-1)}$

- Second derivatives aren't needed, but first derivatives are
- Approximation to Hessian is built up during iteration
- If $lnL(\theta)$ is quadratic, then $D^{(k)} = H$, where $k = dim(\Theta)$
- Intermediate convergence speed ($p \approx 1.6$)

6.2.4 "Line-Search" vs. "Trust Region" Methods: Levenberg-Marquardt

An interesting duality...

Line search: First determine direction, then step

Trust region: First determine step, then direction

– Approximate the function locally in a trust region containing all admissible steps, and then determine direction

Classic example: Levenberg-Marquardt

Related R packages:
trust (trust region optimization)
minpack.lm (R interface to Levenberg-Marquardt in MINPACK)

6.3 GRADIENT-FREE LIKELIHOOD MAXIMIZATION: EM

Recall Kalman smoothing
1. Kalman filter the state forward through the sample, $t = 1,\ ...,\ T$
2. Kalman smooth the state backward, $t = T,\ (T-1),\ (T-2),\ ...,\ 1$
Initialize:

$$a_{T,T} = a_T$$

$$P_{T,T} = P_T$$

Smooth:

$$a_{t,T} = a_t + J_t(a_{t+1,T} - a_{t+1,t})$$

$$P_{t,T} = P_t + J_t(P_{t+1,T} - P_{t+1,t})J_t'$$

$$where\ J_t = P_t T' P_{t+1,t}^{-1}$$

Another Related Smoothing Piece, Needed for EM
3. Get smoothed predictive covariance matrices as well:

$$P_{(t,t-1),T} = E[(\alpha_t - a_{t,T})(\alpha_{t-1} - a_{t-1,T})']$$

Initialize:

$$P_{(T,T-1),T} = (I - K_T Z)TP_{T-1}$$

Then:

$$P_{(t-1,t-2),T} = P_{t-1}J_{t-2}' + J_{t-1}(P_{(t,t-1),T} - TP_{t-1})J_{t-2}'$$

where K_t is the Kalman gain.

The EM "Data-Augmentation Algorithm" Think of $\{\alpha_t\}_{t=1}^T$ as data that are unfortunately missing in

$$\alpha_t = T\alpha_{t-1} + \eta_t$$

$$y_t = Z\alpha_t + \varepsilon_t$$

Incomplete Data Likelihood:

$$lnL(y;\theta)$$

Complete Data Likelihood: (If only we had complete data!)

$$lnL\left(y,\{\alpha_t\}_{t=0}^{T};\theta\right)$$

Expected Complete Data Likelihood:

$$lnL^{(m)}\left(y;\theta\right)\approx\mathbf{E}_{\alpha}\left[lnL\left(y,\{\alpha_t\}_{t=0}^{T};\theta\right)\right]$$

EM iteratively constructs and maximizes the expected complete-data likelihood, which (amazingly) has same
maximizer as the (relevant) incomplete-data likelihood.

The EM (Expectation/Maximization) Algorithm

1. E Step:

Construct $lnL^{(m)}\left(y;\theta\right)\approx\mathbf{E}_{\alpha}\left[lnL\left(y,\{\alpha_t\}_{t=0}^{T};\ \theta\right)\right]$

2. M Step:

$\theta^{(m+1)}=argmax_{\theta}\left\{lnL^{(m)}(y;\ \theta)\right\}$

3. If convergence criterion not met, go to 1

But how to do the E and M steps?

6.3.1 "Not-Quite-Right EM"
(But it Captures and Conveys the Intuition)

1. E Step:

Approximate a "complete data" situation by replacing
$\{\alpha_t\}_{t=0}^{T}$ with $a_{t,T}$ from the Kalman smoother

2. M Step:

Estimate parameters by running regressions:

$a_{t,T}\rightarrow a_{t-1,T}$

$y_t\rightarrow a_{t,T}$

3. If convergence criterion not met, go to 1

6.3.2 Precisely Right EM

6.3.2.1 Complete Data Likelihood

The complete data are $\left\{\alpha_0,\ \{\alpha_t,y_t\}_{t=1}^{T}\right\}$.

Complete-Data Likelihood:

$$f_\theta(y, \alpha_0, \{\alpha_t\}_{t=1}^T) = f_{a_0,P_0}(\alpha_0) \prod_{t=1}^T f_{T,Q}(\alpha_t|\alpha_{t-1}) \prod_{t=1}^T f_{Z,H}(y_t|\alpha_t)$$

Gaussian Complete-Data Log-Likelihood:

$$\begin{aligned}\ln L(y, \{\alpha_t\}_{t=1}^T; \theta) = const &- \frac{1}{2}\ln|P_0| - \frac{1}{2}(\alpha_0 - a_0)' P_0^{-1}(\alpha_0 - a_0) \\ &- \frac{T}{2}\ln|Q| - \frac{1}{2}\sum_{t=1}^T (\alpha_t - T\alpha_{t-1})' Q^{-1}(\alpha_t - T\alpha_{t-1}) \\ &- \frac{T}{2}\ln|H| - \frac{1}{2}\sum_{t=1}^T (y_t - Z\alpha_t)' H^{-1}(y_t - Z\alpha_t)\end{aligned}$$

6.3.2.2 E Step

Construct: $lnL^{(m)}(y;\theta) \approx \mathbf{E}_\alpha \left[lnL\left(y, \{\alpha_t\}_{t=0}^T;\ \theta\right)\right]$

$$\begin{aligned}\mathbf{E}_\alpha\left[\ln L(y, \{\alpha_t\}_{t=0}^T; \theta)\right] = const &- \frac{1}{2}\ln|P_0| - \frac{1}{2}\mathbf{E}_\alpha\left[(\alpha_0 - a_0)' P_0^{-1}(\alpha_0 - a_0)\right] \\ &- \frac{T}{2}\ln|Q| - \frac{1}{2}\sum_{t=1}^T \mathbf{E}_\alpha\left[(\alpha_t - T\alpha_{t-1})' Q^{-1}(\alpha_t - T\alpha_{t-1})\right] \\ &- \frac{T}{2}\ln|H| - \frac{1}{2}\sum_{t=1}^T \mathbf{E}_\alpha\left[(y_t - Z\alpha_t)' H^{-1}(y_t - Z\alpha_t)\right]\end{aligned}$$

– Function of $\{a_{t,T}(\theta^{(m)})$, $P_{t,T}(\theta^{(m)})$, and $P_{(t,t-1),T}(\theta^{(m)})\}_{t=1}^T$
– So the E step is really just running the three smoothers

6.3.2.3 M Step Formulas

Find: $\theta^{(m+1)} = argmax_\theta \left\{ lnL^{(m)}(y;\theta)\right\}$

$$\hat{T}' = \left(\sum_{t=1}^T \mathbf{E}_\alpha\left[\alpha_t \alpha_{t-1}'\right]\right)\left(\sum_{t=1}^T \mathbf{E}_\alpha\left[\alpha_{t-1}\alpha_{t-1}'\right]\right)^{-1}$$

$$\hat{\eta}_t = (\alpha_t - \hat{T}\alpha_{t-1})$$

$$\hat{Q} = \frac{1}{T}\sum_{t=1}^T \mathbf{E}_\alpha\left[\hat{\eta}_t \hat{\eta}_t'\right]$$

$$\hat{Z}' = \left(\sum_{t=1}^T y_t \mathbf{E}_\alpha\left[\alpha_t\right]'\right)\left(\sum_{t=1}^T \mathbf{E}_\alpha\left[\alpha_t \alpha_t'\right]\right)^{-1}$$

$$\hat{\epsilon}_t = (y_t - \hat{Z}\alpha_t)$$

$$\hat{H} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_\alpha \left[\hat{\epsilon}_t \hat{\epsilon}_t'\right]$$

where:

$$\mathbf{E}_\alpha \left[\alpha_t\right] = a_{t|T}$$

$$\mathbf{E}_\alpha \left[\alpha_t \alpha_t'\right] = a_{t|T} a_{t|T}' + P_{t|T}$$

$$\mathbf{E}_\alpha \left[\alpha_t \alpha_{t-1}'\right] = a_{t|T} a_{t-1|T}' + P_{(t,t-1)|T}$$

Simply replacing α_t with $a_{t,T}$ won't work because $E\left[\alpha_t \alpha_t' | \Omega_T\right] \neq a_{t,T} a_{t,T}'$
Instead we have $E\left[\alpha_t \alpha_t' | \Omega_T\right] = E\left[\alpha_t |\ \Omega_T\right] E\left[\alpha_t | \Omega_T\right]' + Var(\alpha_t | \Omega_T) = a_{t,T} a_{t,T}' + P_{t,T}$

6.4 LIKELIHOOD INFERENCE

6.4.1 Under Correct Specification

$$\ln L(\theta) \ = \ \sum_{t=1}^{T} \ln L_t(\theta)$$

- Always true in iid environments
- Also holds in time series, via prediction-error decomposition of $\ln L$

Expected Fisher Information
Score at true parameter θ_0:

$$s(\theta_0) \ = \ \left.\frac{\partial \ln L(\theta)}{\partial \theta}\right|_{\theta_0} \ = \ \sum_{t=1}^{T} \left.\frac{\partial \ln L_t(\theta)}{\partial \theta}\right|_{\theta_0} \ = \ \sum_{t=1}^{T} s_t(\theta_0)$$

Expected Fisher Information at true parameter θ_0:

$$I_{EX,H}(\theta_0) \ = \ -E\left(\left.\frac{\partial^2 \ln L(\theta)}{\partial \theta\ \partial \theta'}\right)\right|_{\theta_0}$$

$$= \ -EH(\theta_0) \ = \ \sum_{t=1}^{T} -E\left(\left.\frac{\partial^2 \ln L_t(\theta)}{\partial \theta\ \partial \theta'}\right)\right|_{\theta_0} \ = \ \sum_{t=1}^{T} -EH_t(\theta_0)$$

More on Expected Information Matrices
(1) We had already: $I_{EX,H}(\theta_0) \ = \ \sum_{t=1}^{T} -E\, H_t(\theta_0)$
"Expected information based on the Hessian"
(2) Can also form: $I_{EX,s}(\theta_0) \ = \ \sum_{t=1}^{T} E\ s_t(\theta_0)\ s_t'(\theta_0)$
"Expected information based on the score"
(In a moment we'll see why this is of interest)
Distribution of the MLE Under Correct Specification

Under correct specification and regularity conditions,

$$\sqrt{T}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, V_{EX}(\theta_0)) \tag{6.1}$$

where

$$V_{EX}(\theta_0) = V_{EX,H}(\theta_0) = plim_{T\to\infty}\left(\frac{I_{EX,H}(\theta_0)}{T}\right)^{-1}$$

$$= V_{EX,s}(\theta_0) = plim_{T\to\infty}\left(\frac{I_{EX,s}(\theta_0)}{T}\right)^{-1}$$

$\hat{\theta}_{ML}$ consistent, asymptotically normal, asymptotically efficient (Cramer-Rao lower bound met)

Observed Information Matrices

(1) $I_{OB,H}(\theta_0) = \sum_{t=1}^{T} - H_t(\theta_0)$

("Observed information based on the Hessian")

(2) $I_{OB,s}(\theta_0) = \sum_{t=1}^{T} s_t(\theta_0)\, s_t'(\theta_0)$

("Observed information based on the score")

In a moment we'll see why these are of interest.

Consistent Estimators of $V_{EX}(\theta_0)$

$$\hat{V}_{EX,H}(\theta_0) = \left(\frac{I_{EX,H}(\hat{\theta}_{ML})}{T}\right)^{-1} \quad \hat{V}_{EX,s}(\theta_0) = \left(\frac{I_{EX,s}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

$$\hat{V}_{OB,H}(\theta_0) = \left(\frac{I_{OB,H}(\hat{\theta}_{ML})}{T}\right)^{-1} \quad \hat{V}_{OB,s}(\theta_0) = \left(\frac{I_{OB,s}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

Under correct specification, $plim_{T\to\infty}\, \hat{V}_{EX,H}(\theta_0) = plim_{T\to\infty}\, \hat{V}_{EX,s}(\theta_0) = V_{EX}(\theta_0)$

$plim_{T\to\infty}\, \hat{V}_{OB,H}(\theta_0) = plim_{T\to\infty}\, \hat{V}_{OB,s}(\theta_0) = V_{EX}(\theta_0)$

6.4.2 Under Possible Mispecification

6.4.2.1 Distributional Misspecification

Under possible distributional misspecification (but still assuming correct conditional mean and variance function specifications),

$$\sqrt{T}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N\left(0, V_{EX}^{m}(\theta_0)\right) \tag{6.2}$$

where:

$$V_{EX}^{m}(\theta_0) = V_{EX,H}(\theta_0)^{-1}\ V_{EX,s}(\theta_0)V_{EX,H}(\theta_0)^{-1}$$

"Quasi MLE" or "Pseudo-MLE"

Under correct specification: (6.2) collapses to standard result (6.1)

Under distributional misspecification:

$\hat{\theta}_{ML}$ consistent, asymptotically normal, asymptotically inefficient (CRLB not met), but we can nevertheless do credible

(if not fully efficient) inference

Consistent estimators of $V_{EX}^{m}(\theta_0)$:

$$\hat{V}_{EX}^{m}(\theta_0) = \left(\frac{I_{EX,H}(\hat{\theta}_{ML})}{T}\right)^{-1}\left(\frac{I_{EX,s}(\hat{\theta}_{ML})}{T}\right)\left(\frac{I_{EX,H}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

$$\hat{V}_{OB}^{m}(\theta_0) = \left(\frac{I_{OB,H}(\hat{\theta}_{ML})}{T}\right)^{-1}\left(\frac{I_{OB,s}(\hat{\theta}_{ML})}{T}\right)\left(\frac{I_{OB,H}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

"Sandwich Estimator"

6.4.2.2 General Misspecification

Under possible general mispecification,

$$\sqrt{T}(\hat{\theta}_{ML} - \theta_*) \xrightarrow{d} N\left(0, V_{EX}^{m}(\theta_*)\right) \tag{6.3}$$

where:

$$V_{EX}^{m}(\theta_*) = V_{EX,H}(\theta_*)^{-1}\ V_{EX,s}(\theta_*)V_{EX,H}(\theta_*)^{-1}$$

Under correct specification: Collapses to standard result (6.1)

Under purely distributional specification: Collapses to result (6.2)

Under general misspecification:

$\hat{\theta}_{ML}$ consistent for KLIC-optimal pseudo-true value θ_*,

asymptotically normal, and we can do credible inference

Consistent estimators of $V_{EX}^{m}(\theta_*)$:

$$\hat{V}_{EX}^{m}(\theta_*) = \left(\frac{I_{EX,H}(\hat{\theta}_{ML})}{T}\right)^{-1}\left(\frac{I_{EX,s}(\hat{\theta}_{ML})}{T}\right)\left(\frac{I_{EX,H}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

$$\hat{V}_{OB}^{m}(\theta_*) \;=\; \left(\frac{I_{OB,H}(\hat{\theta}_{ML})}{T}\right)^{-1}\left(\frac{I_{OB,s}(\hat{\theta}_{ML})}{T}\right)\left(\frac{I_{OB,H}(\hat{\theta}_{ML})}{T}\right)^{-1}$$

6.5 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Approximate (asymptotic) frequency domain Gaussian likelihood

   Recall that for Gaussian series we have as $T \to \infty$:

$$x_j = \frac{2\hat{f}(\omega_j)}{f(\omega_j;\theta)} \;\overset{d}{\to}\; \chi_2^2$$

   where $f(\omega_j;\theta)$ is the spectral density and the χ_2^2 random variables are independent across frequencies

$$\omega_j = \frac{2\pi j}{T},\; j = 0, 1, ..., \frac{T}{2}$$

   $\Rightarrow$ MGF of any one of the x_j's is

$$M_x(t) \;= \frac{1}{1-2t}$$

   Let

$$y_j \;= \hat{f}(\omega_j) \;=\; \frac{f(\omega_j;\theta)\; x_j}{2}$$

$$\Rightarrow M_y\,(t) \;=\; M_x\left(\frac{f(\omega_j;\theta)}{2}\; t\right) = \frac{1}{1-f(\omega_j;\theta)\; t}$$

   This is the MGF of exponential rv with parameter $1/f(\omega_j;\theta)$.

$$\Rightarrow\; g(\hat{f}(\omega_j);\theta) \;= \frac{1}{f(\omega_j;\theta)}\; e^{\frac{-\hat{f}(\omega_j)}{f(\omega_j;\theta)}}$$

   Univariate asymptotic Gaussian log likelihood:

$$\ln L(\hat{f};\theta) = -\sum_{j=0}^{T/2} \ln f(\omega_j;\theta) - \sum_{j=0}^{T/2} \frac{\hat{f}(\omega_j)}{f(\omega_j;\theta)}$$

   Multivariate asymptotic Gaussian log likelihood:

$$\ln L(\hat{f};\theta) = -\sum_{j=0}^{T/2} \ln |F(\omega_j;\theta)| - trace\left(\sum_{j=0}^{T/2} F^{-1}(\omega_j;\theta)\,\hat{F}(\omega_j)\right)$$

2. State space model fitting by exact Gaussian pseudo-MLE using a prediction-error decomposition and the Kalman filter.

   Read Aruoba et al. (2013), and fit the block-diagonal dynamic-factor model (3)-(4) by exact Gaussian pseudo-MLE using a predecition-error decomposition and the Kalman filter. Obtain both filtered and smoothed estimates of the series of states. Compare them to each other, to the "raw" GDP_E and GDP_I series, and to the current vintage of GDPplus from FRB Philadelphia.

3. Method of scoring

   Slight variation on Newton:

   Use $\left(E(H^{(m)})\right)^{-1}$ rather than $H^{-1(m)}$

   (Expected rather than observed Hessian.)

4. Constrained optimization.

   (a) Substitute the constraint directly and use Slutsky's theorem.
   - To keep a symmetric matrix nonnegative definite, write it as PP'
   - To keep a parameter in [0, 1], write it as $p^2/(1+p^2)$.

   (b) Barrier and penalty functions

6.6 NOTES

Chapter Seven

Simulation for Economic Theory, Econometric Theory, Estimation, Inference, and Optimization

7.1 GENERATING U(0,1) DEVIATES

Criteria for (Pseudo-) Random Number Generation

1. Statistically independent
2. Reproducible
3. Non-repeating
4. Quickly-generated
5. Minimal in memory requirements

The Canonical Problem: Uniform (0,1) Deviates
Congruential methods
$a = b(mod\ m)$
"a is congruent to b modulo m"
"$(a - b)$ is an integer multiple of m"
Examples:
$255 = 5(mod 50)$
$255 = 5(mod 10)$
$123 = 23(mod 10)$

Key recursion: $x_t = ax_{t-1}(mod\ m)$
To get x_t, just divide ax_{t-1} by m and keep the remainder
Multiplicative Congruential Method

$$x_t = ax_{t-1}(mod\ m), \quad x_t,\ a,\ m\ \in\ Z_+$$

Example:

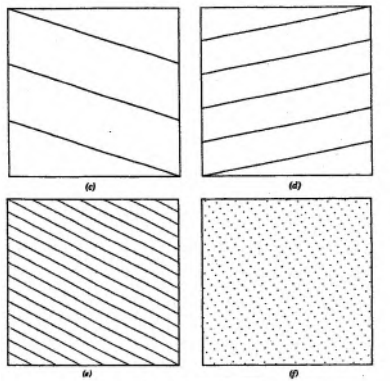


Figure 7.1: *Ripley's "Horror" Plots of pairs of* (U_{i+1}, U_i) *for Various Congruential Generators Modulo 2048 (from Ripley, 1987)*

$$x_t = 3x_{t-1}(mod\ 16), \qquad x_0 = 1$$

$$x_0 = 1, x_1 = 3, x_2 = 9, x_3 = 11, x_4 = 1, x_5 = 3, ...$$

Perfectly periodic, with a period of 4.

Generalize:

$$x_t(ax_{t-1} + c)(mod\ m) \quad (x_t,\ a,\ c,\ m\ \in\ Z_+)$$

Remarks

1. $x_t \in [0, m-1],\ \forall t$. So take $x_t^* = \frac{x_t}{m},\ \ \forall t$
2. Pseudo-random numbers are perfectly periodic.
3. The maximum period, m, can be attained using the mixed congruential generator if:
   - c and m have no common divisor
   - $a = 1(mod\ p)\ \forall$ prime factors p of m
   - $a = 1(mod\ 4)$ if m is a multiple of 4
4. m is usually determined by machine wordlength; e.g., 2^{64}
5. Given $U(0,\ 1)$, $U(\alpha,\ \beta)$ is immediate

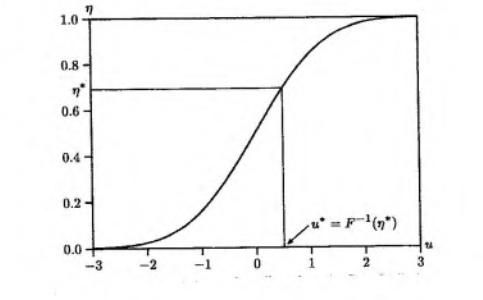


Figure 7.2: *Transforming from U(0,1) to f (from Davidson and MacKinnon, 1993)*

7.2 THE BASICS: C.D.F. INVERSION, BOX-MUELLER, SIMPLE ACCEPT-REJECT

7.2.1 Inverse c.d.f.

Inverse cdf Method ("Inversion Methods")

Desired density: $f(x)$

1. Find the analytical c.d.f., $F(x)$, corresponding to $f(x)$
2. Generate T $U(0,1)$ deviates $\{\mathrm{r}_1,...\mathrm{r}_T\}$
3. Calculate $\{\mathrm{F}^{-1}(\mathrm{r}_1),..., \mathrm{F}^{-1}(\mathrm{r}_T)\}$

Graphical Representation of Inverse cdf Method

Example: Inverse cdf Method for exp(β) Deviates

$f(x) = \beta e^{-\beta x}$ where $\beta > 0,\ x \geq 0$

$\Rightarrow F(x) = \int_0^x \beta e^{-\beta t}\, dt$

$= \frac{\beta e^{-\beta t}}{-\beta} \Big|_o^x = -e^{-\beta x} + 1 = 1 - e^{-\beta x}$

Hence $e^{-\beta x} = 1 - F(x)$ so $x = \frac{\ln(1 - F(x))}{-\beta}$

Then insert a $U(0,1)$ deviate for $F(x)$

Complications Analytic inverse cdf not always available (e.g., $N(0,1)$ distribution).

- Approach 1: Evaluate the cdf numerically
- Approach 2: Use a different method

  e.g., CLT approximation:

  Take $\left(\sum_{i=1}^{12} U_i(0,1) - 6\right)$ for $N(0,1)$

7.2.2 Box-Muller

An Efficient Gaussian Approach: Box-Muller

Let x_1 and x_2 be i.i.d. $U(0,1)$, and consider

$y_1 = \sqrt{-2\ln x_1}\ \cos(2\pi x_2)$

$y_2 = \sqrt{-2\ln x_1}\ \sin(2\pi x_2)$

Find the distribution of y_1 and y_2. We know that

$$f(y_1, y_2) = f(x_1, x_2) \left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right|$$

Box-Muller (Continued) Here we have $x_1 = e^{-\frac{1}{2}(y_1^2+y_2^2)}$ and $x_2 = \frac{1}{2\pi}\ \arctan\left(\frac{y_2}{y_1}\right)$

$$Hence \left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right| = \left(\frac{1}{\sqrt{2\pi}}\ e^{-y_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}\ e^{-y_2^2/2}\right)$$

Bivariate density is the product of two $N(0,1)$ densities, so we have generated two independent $N(0,1)$ deviates.

Generating Deviates Derived from N(0,1)

$$\chi_1^2 = [N(0,1)]^2$$

$\chi_d^2 = \sum_{i=1}^{d}\ [N_i(0,1)]^2$, where the $N_i(0,1)$ are independent

$N(\mu, \sigma^2) = \mu + \sigma\ N(0,1)$

$t_d = N(0,1)/\sqrt{x_d^2/d}$, where $N(0,1)$ and χ_d^2 are independent

$F_{d_1,d_2} = \chi_{d_1}^2/d_1/\chi_{d_2}^2/d_2$ where $\chi_{d_1}^2$ and $\chi_{d_2}^2$ are independent

Multivariate Normal

$N(0, I)$ (N-dimensional) – Just stack N $N(0,1)$'s

$N(\mu, \Sigma)$ (N-dimensional)

Let $PP' = \Sigma$ (P is the Cholesky factor of Σ)

Let $X \sim N(0, I)$. Then $PX \sim N(0, \Sigma)$

To sample from $N(\mu, \Sigma)$, take $\mu + PX$

7.2.3 Simple Accept-Reject

Accept-Reject

(Naive but Revealing Example)

We want to sample $x \sim f(x)$

Draw:

$$\nu_1 \sim U(\alpha,\ \beta)$$

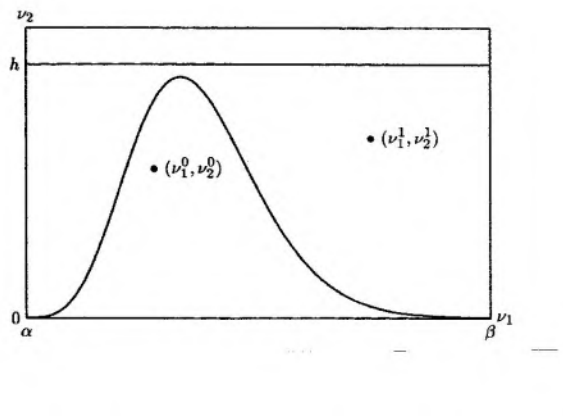


Figure 7.3: Naive Accept-Reject Method

$$\nu_2 \ \sim \ U(0, \ h)$$

If ν_1, ν_2 lies under the density $f(x)$, then take $x = \nu_1$

Otherwise reject and repeat

Graphical Representation of Naive Accept-Reject

Accept-Reject

General (Non-Naive) Case

We want to sample $x \sim f(x)$ but we only know how to sample $x \sim g(x)$.

Let M satisfy $\frac{f(x)}{g(x)} \leq M < \infty$, $\forall x$. Then:

1. Draw $x_0 \sim g(x)$
2. Take $x = x_0$ $w.p.$ $\frac{f(x_0)}{g(x_0)M}$; else go to 1.

(Allows for "blanket" functions $g(\cdot)$ more efficient than the uniform)

Note that accept-reject requires that we be able to evaluate $f(x)$ and $g(x)$ for any x.

Mixtures

On any draw i,

$$x \ \sim \ f_i(x), \ w.p. \ p_i$$

where

$$0 \leq p_i \leq 1, \ \forall \ i$$

$$\sum_{i=1}^{N} p_i = 1$$

For example, all of the f_i could be uniform, but with different location and scale.

7.3 SIMULATING EXACT AND APPROXIMATE REALIZATIONS OF TIME SERIES PROCESSES

Simulating Time Series Processes

VAR(1) simulation is key (state transition dynamics).

1. Nonparametric: Exact realization via Cholesky factorization of desired covariance matrix. One need only specify the autocovariances.
2. Parametric I: Exact realization via Cholesky factorization of covariance matrix corresponding to desired parametric model
3. Parametric II: Approximate realization via arbitrary startup value with early realization discarded
4. Parametric III: Exact realization via drawing startup values from unconditional density

7.4 MORE

Slice Sampling

Copulas and Sampling From a General Joint Density

7.5 ECONOMIC THEORY BY SIMULATION: "CALIBRATION"

7.6 ECONOMETRIC THEORY BY SIMULATION: MONTE CARLO AND VARIANCE REDUCTION

Monte Carlo

Key: Solve deterministic problems by simulating stochastic analogs, with the analytical unknowns reformulated as parameters to be estimated.

Many important discoveries made by Monte Carlo.

Also, numerous *mistakes avoided* by Monte Carlo!

The pieces:

(I) Experimental Design

(II) Simulation (including variance reduction techniques)

(III) Analysis: Response surfaces (which also reduce variance)

7.6.1 Experimental Design

(I) Experimental Design

- Data-Generating Process (DGP), $M(\theta)$
- Objective
  - e.g., MSE of an estimator:

$$E[(\theta - \hat{\theta})^2] \;=\; g(\theta, T)$$

  - e.g., Power function of a test:

$$\pi \;=\; g(\theta, T)$$

Experimental Design, Continued

- Selection of (θ, T) Configurations to Explore

  a. Do we need a "full design"? In general many values of θ and T need be explored. But if, e.g., $g(\theta, T) = g_1(\theta) + g_2(T)$, then only explore θ values for a single T, and T values for a single θ (i.e., there are no interactions).

  b. Is there parameter invariance ($g(\theta, T)$ unchanging in θ)? e.g., If $y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2\Omega(\alpha))$, then the exact finite-sample distributions of $\frac{\hat{\beta}_{MLE}-\beta}{\sigma}$ and $\frac{\hat{\sigma}^2_{MLE}}{\sigma^2}$ are invariant to true β, σ^2. So vary only α, leaving β and σ alone (e.g., set to 0 and 1). Be careful not to implicitly assume invariance regarding unexplored aspects of the design (e.g., structure of X variables above.)

Experimental Design, Continued

- Number of Monte Carlo Repetitions (N)

  e.g., MC computation of test size

  nominal size α_0, true size α, estimator $\hat{\alpha} = \frac{\#rej}{N} = \frac{\Sigma_{i=1}^{N} I(rej_i)}{N}$

$$Normal\ approximation: \;\; \hat{\alpha} \overset{a}{\sim} N\left(\alpha,\; \frac{\alpha(1-\alpha)}{N}\right)$$

$$P\left(\alpha \in \left[\hat{\alpha} \;\pm\; 1.96\sqrt{\frac{\alpha(1-\alpha)}{N}}\right]\right) \;=\; .95$$

  Suppose we want the 95% CI for α to be .01 in length.

Experimental Design, Continued

Strategy 1 (Use $\alpha = \alpha_0$; not conservative enough if $\alpha > \alpha_0$):

$$2 * 1.96 \sqrt{\frac{\alpha_0(1-\alpha_0)}{N}} \;=\; .01$$

If $\alpha_0 = .05$, N = 7299

Strategy 2 (Use $\alpha = \frac{1}{2} = argmax_\alpha[\alpha(1-\alpha)]$; conservative):

$$2 * 1.96 \sqrt{\frac{\frac{1}{2}\left(\frac{1}{2}\right)}{N}} \;=\; .01 \;\Rightarrow\; N = 38416$$

Strategy 3 (Use $\alpha = \hat{\alpha}$; the obvious strategy)

7.6.2 Simulation

(II) Simulation

Running example: Monte Carlo integration

Definite integral: $\theta = \int_0^1 m(x)dx$

Key insight:

$\theta = \int_0^1 m(x)dx = E(m(x))$

$x \sim U(0,\, 1)$

Notation:

$\theta = E[m(x)]$

$\sigma^2 = var(m(x))$

Direct Simulation:

Arbitrary Function, Uniform Density

Generate N $U(0, 1)$ deviates $x_i,\ i = 1, ..., N$

Form the N deviates $m_i = m(x_i),\ i = 1, ..., N$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} m_i$$

$$\sqrt{N}(\hat{\theta} - \theta) \;\xrightarrow{d}\; N(0,\ \sigma^2)$$

Direct Simulation General Case:

Arbitrary Function, Arbitrary Density

$$\theta = E(m(x)) = \int m(x)f(x)dx$$

– Indefinite integral, arbitrary function $m(\cdot)$, arbitrary density $f(x)$
Draw $x_i \sim f(\cdot)$, and then form $m_i(x_i)$,

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} m_i$$

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0,\ \sigma^2)$$

Direct Simulation Leading Case
Mean Function, Arbitrary Density
(e.g., Posterior Mean)

$$\theta = E(x) = \int x f(x) dx$$

– Indefinite integral, x has arbitrary density $f(x)$
Draw $x_i \sim f(\cdot)$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0,\ \sigma^2)$$

7.6.3 Variance Reduction: Importance Sampling, Antithetics, Control Variates and Common Random Numbers

Importance Sampling to Facilitate Sampling
Sampling from $f(\cdot)$ may be difficult. So change to:

$$\theta = \int x \frac{f(x)}{g(x)} g(x) dx$$

where the "importance sampling density" $g(\cdot)$ is easy to sample
Draw $x_i \sim g(\cdot)$, and then form $m_i = \frac{f(x_i)}{g(x_i)} x_i,\ i = 1, ..., N$

$$\hat{\theta}_* = \frac{1}{N}\sum_{i=1}^{N} m_i = \frac{1}{N}\sum_{i=1}^{N} \frac{f(x_i)}{g(x_i)} x_i = \sum_{i=1}^{N} w_i x_i$$

$$\sqrt{N}(\hat{\theta}_* - \theta) \xrightarrow{d} N(0,\ \sigma_*^2)$$

– Avg of $f(x)$ draws replaced by *weighted* avg of $g(x)$ draws
– Weight w_i reflects relative heights of $f(x_i)$ and $g(x_i)$

Importance Sampling

Consider the classic problem of calculating the mean $E(y)$ of r.v. y with marginal density:

$$f(y) = \int f(y/x)f(x)dx.$$

The standard solution is to form:

$$\widehat{E}(y) = \frac{1}{N}\sum_{i=1}^{N} f(y|x_i)$$

where the x_i are iid draws from $f(x)$.

But $f(x)$ might be hard to sample from! What to do?

Importance Sampling

Write

$$f(y) = \int f(y|x)\frac{f(x)}{I(x)}g(x)dx,$$

where the "importance sampler," $g(x)$, is easy to sample from.

Take

$$\widehat{E}(y) = \sum_{i=1}^{N} \frac{\frac{f(x_i)}{I(x_i)}}{\sum_{j=1}^{N}\frac{f(x_j)}{g(x_j)}} f(y|x_i) = \sum_{i=1}^{N} w_i f(y|x_i).$$

So importance sampling replaces a simple average of $f(y|x_i)$ based on initial draws from $f(x)$ with a weighted average of $f(y|x_i)$ based on initial draws from $g(x)$, where the weights w_i reflect the relative heights of $f(x_i)$ and $g(x_i)$.

Indirect Simulation

"Variance-Reduction Techniques"

("Swindles")

Importance Sampling to Achieve Variance Reduction

Again we use:

$$\theta = \int x\frac{f(x)}{g(x)}g(x)dx,$$

and again we arrive at

$$\sqrt{N}(\hat{\theta}_* - \theta) \quad \overset{d}{\rightarrow} \quad N(0,\ \sigma_*^2)$$

If $g(x)$ is chosen judiciously, $\sigma_*^2 \ll \sigma^2$

Key: Pick $g(x)$ s.t. $\frac{xf(x)}{g(x)}$ has small variance

Importance Sampling Example

Let $x \sim N(0,1)$, and estimate the mean of $I(x > 1.96)$:

$$\theta = E(I(x > 1.96)) = P(x > 1.96) = \int \underbrace{I(x > 1.96)}_{m(x)} \; \underbrace{\phi(x)}_{f(x)} \; dx$$

$$\hat{\theta} = \sum_{i=1}^{N} \frac{I(x_i > 1.96)}{N} \quad (with\ variance\ \sigma^2)$$

Use importance sampler:

$$g(x) = N(1.96,\, 1)$$

$$P(x > 1.96) = \int \frac{I(x > 1.96)\ \phi(x)}{g(x)}\ g(x)\ dx$$

$$\hat{\theta}_* = \frac{\sum_{i=1}^{N} \frac{I(x_i > 1.96)\ \phi(x_i)}{g(x_i)}}{N} \quad (with\ variance\ \sigma_*^2)$$

$$\frac{\sigma_*^2}{\sigma^2} \approx 0.06$$

Antithetic Variates

We average negatively correlated unbiased estimators of θ (Unbiasedness maintained, variance reduced)

The key: If $x \sim symmetric(\mu, v)$, then $x_i \pm \mu$ are equally likely

e.g., if $x \sim U(0,1)$, so too is $(1 - x)$

e.g., if $x \sim N(0, v)$, so too is $-x$

Consider for example the case of zero-mean symmetric $f(x)$

$$\theta = \int m(x) f(x) dx$$

$$Direct: \quad \hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} m_i,\ (\hat{\theta}\ is\ based\ on\ x_i, i = 1, ..., N)$$

$$Antithetic: \quad \hat{\theta}_* = \frac{1}{2}\hat{\theta}_{(x)} + \frac{1}{2}\hat{\theta}_{(-x)}$$

($\hat{\theta}_{(x)}$ is based on $x_i,\ i = 1, ..., N/2$, and

$\hat{\theta}_{(-x)}$ is based on $-x_i,\ i = 1, ..., N/2$)

Antithetic Variates, Cont'd

More concisely,

$$\hat{\theta}_* = \frac{2}{N}\sum_{i=1}^{N/2} k_i(x_i)$$

where:

$$k_i = \frac{1}{2}m(x_i) + \frac{1}{2}m(-x_i)$$

$$\sqrt{N}(\hat{\theta}_* - \theta) \overset{d}{\rightarrow} N(0, \sigma_*^2)$$

$$\sigma_*^2 = \frac{1}{4}var\left(m(x)\right) + \frac{1}{4}var\left(m(-x)\right) + \frac{1}{2}\underbrace{cov\left(m(x), m(-x)\right)}_{<0 \text{ for } m \text{ monotone incr.}}$$

Often $\sigma_*^2 \ll \sigma^2$

$$\theta = \int m(x)f(x)dx = \int g(x)f(x)dx + \int [m(x) - g(x)]f(x)dx$$

Control function $g(x)$ simple enough to integrate analytically and flexible enough to absorb most of the variation in $m(x)$.

We just find the mean of $m(x) - g(x)$, where $g(x)$ has known mean and is highly correlated with $m(x)$.

Control Variates

$$\hat{\theta} = \int g(x)dx + \frac{1}{N}\sum_{i=1}^{N}[m(x_i) - g(x_i)]$$

$$\sqrt{N}(\hat{\theta} - \theta) \overset{d}{\rightarrow} N(0, \sigma_*^2)$$

If $g(x)$ is chosen judiciously, $\sigma_*^2 \ll \sigma^2$.

Related method (conditioning): Find the mean of $E(z|w)$ rather than the mean of z. The two are of course the same (the mean conditional mean is the unconditional mean), but $var(E[z|w]) \leq var(z)$.

Control Variate Example

$$f(x) = \int_0^1 e^x dx$$

Control variate: $g(x) = 1 + 1.7x$

$$\Rightarrow \int_0^1 g(x)dx = \left(x + \frac{1.7}{2}x^2\right) \Bigg|_0^1 = 1.85$$

$$\hat{\theta}_{direct} = \frac{1}{N}\sum_{i=1}^{N} e^{x_i}$$

$$\hat{\theta}_{cv} = 1.85 + \frac{1}{N}\sum_{i=1}^{N}[e^{x_i} - (1 + 1.7x_i)]$$

$$\frac{var(\hat{\theta}_{direct})}{var(\hat{\theta}_{CV})} \approx 78$$

Common Random Numbers

We have discussed estimation of a single integral:

$$\int_0^1 f_1(x)dx$$

But interest often centers on *difference* (or ratio) of the two integrals:

$$\int_0^1 f_1(x)dx \; - \; \int_0^1 f_2(x)dx$$

The key: Evaluate each integral using the *same* random numbers.

Common Random Numbers in Estimator Comparisons

Two estimators $\hat{\theta}, \tilde{\theta}$; true parameter θ_0

Compare MSEs: $E(\hat{\theta} - \theta_0)^2, \; E(\tilde{\theta} - \theta_0)^2$

Expected difference: $E\left((\hat{\theta} - \theta_0)^2 \; - \; (\tilde{\theta} - \theta_0)^2\right)$

Estimate:

$$\frac{1}{N}\sum_{i=1}^{N} \left((\hat{\theta}_i - \theta_0)^2 \; - \; (\tilde{\theta}_i - \theta_0)^2\right)$$

Variance of estimate:

$$\frac{1}{N}var\left((\hat{\theta} - \theta_0)^2\right) + \frac{1}{N}var\left((\tilde{\theta} - \theta_0)^2\right) - \frac{2}{N}cov\left((\hat{\theta} - \theta_0)^2, (\tilde{\theta} - \theta_0)^2\right)$$

Extensions...

- Sequential importance sampling: Builds up improved proposal densities across draws

7.6.4 Response Surfaces

(III) Response surfaces

1. Direct Response Surfaces
2. Indirect Responses Surfaces:
   - Clear and informative graphical presentation
   - Variance reduction
   - Imposition of known asymptotic results (e.g., power $\to 1$ as $T \to \infty$)
   - Imposition of known features of functional form (e.g. power $\in$ [0,1])

Example: Assessing Finite-Sample Test Size

$$\alpha = P\left(s > s^* | T, H_0 \; true\right) = g(T)$$

(α is empirical size, s is test statistic, s^* is asymptotic c.v.)

$$\hat{\alpha} = \frac{rej}{N}$$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\alpha(1-\alpha)}{N}\right)$$

or

$$\hat{\alpha} = \alpha + \varepsilon = g(T) + \varepsilon$$

$$\varepsilon \sim N\left(0, \frac{g(T)(1-g(T))}{N}\right)$$

Note the heteroskedasticity: variance of ε changes with T.

Example: Assessing Finite-Sample Test Size

Enforce analytically known structure on $\hat{\alpha}$.

Common approach:

$$\hat{\alpha} = \alpha_0 + T^{-\frac{1}{2}}\left(c_0 + \sum_{i=1}^{p} c_i T^{\frac{-i}{2}}\right) + \varepsilon$$

α_0 is nominal size, which obtains as $T \to \infty$. Second term is the vanishing size distortion.

Response surface regression:

$$(\hat{\alpha} - \alpha_0) \to T^{-\frac{1}{2}}, T^{-1}, T^{-\frac{3}{2}}, ...$$

Disturbance will be approximately normal but heteroskedastic.
So use GLS or robust standard errors.

7.7 ESTIMATION BY SIMULATION: GMM, SMM AND INDIRECT INFERENCE

7.7.1 GMM

k-dimensional economic model parameter θ

$$\hat{\theta}_{GMM} = argmin_\theta \, d(\theta)' \, \Sigma d(\theta)$$

where

$$d(\theta) = \begin{pmatrix} m_1(\theta) - \hat{m}_1 \\ m_2(\theta) - \hat{m}_2 \\ \vdots \\ m_r(\theta) - \hat{m}_r \end{pmatrix}$$

The $m_i(\theta)$ are model moments and the $\hat{m}_i$ are data moments.
MM: $k = r$ and the $m_i(\theta)$ calculated analytically
GMM: $k < r$ and the $m_i(\theta)$ calculated analytically

- Inefficient relative to MLE, but useful when likelihood is not available

7.7.2 Simulated Method of Moments (SMM)

($k \leq r$ and the $m_i(\theta)$ calculated by simulation)

- Model moments for GMM may also be unavailable (i.e., analytically intractable)
- SMM: if you can simulate, you can consistently estimate
  - Simulation ability is a good test of model understanding
  - If you can't figure out how to simulate pseudo-data from a given probabilistic model, then you don't understand the model (or the model is ill-posed)
  - Assembling everything: If you understand a model you can simulate it, and if you can simulate it you can estimate it consistently. Eureka!
  - No need to work out what might be very complex likelihoods even if they are in principle "available."

- MLE efficiency lost may be a small price for SMM tractability gained.

SMM Under Misspecification
All econometric models are misspecified.
GMM/SMM has special appeal from that perspective.

- Under correct specification any consistent estimator (e.g., MLE or GMM/SMM) takes you to the right place asymptotically, and MLE has the extra benefit of efficiency.

- Under misspecification, consistency becomes an issue, quite apart from the secondary issue of efficiency. Best DGP approximation for one purpose may be very different from best for another.

- GMM/SMM is appealing in such situations, because it forces thought regarding which moments $M = \{m_1(\theta), ..., m_r(\theta)\}$ to match, and then by construction it is consistent for the M-optimal approximation.

SMM Under Misspecification, Continued

- In contrast, pseudo-MLE ties your hands. Gaussian pseudo-MLE, for example, is consistent for the KLIC-optimal approximation (1-step-ahead mean-squared prediction error).

- The bottom line: under misspecification MLE may not be consistent for what you want, whereas by construction GMM is consistent for what you want (once you *decide* what you want).

7.7.3 Indirect Inference

k-dimensional economic model parameter θ
$\delta > k$-dimensional auxiliary model parameter β

$$\hat{\theta}_{IE} = argmin_\theta \, d(\theta)' \, \Sigma d(\theta)$$

where

$$d(\theta) = \begin{pmatrix} \hat{\beta}_1(\theta) - \hat{\beta}_1 \\ \hat{\beta}_2(\theta) - \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_d(\theta) - \hat{\beta}_d \end{pmatrix}$$

$\hat{\beta}_i(\theta)$ are est. params. of aux. model fit to simulated model data

$\hat{\beta}_i$ are est. params. of aux. model fit to real data

– Consistent for true θ if economic model correctly specified
– Consistent for pseudo-true θ otherwise
– We introduced "Wald form"; also LR and LM forms
Ruge-Murcia (2010)

7.8 INFERENCE BY SIMULATION: BOOTSTRAP

7.8.1 i.i.d. Environments

Simplest (iid) Case

$$\{x_t\}_{t=1}^{T} \overset{iid}{\sim} (\mu, \sigma^2)$$

100α percent confidence interval for μ:

$$I = [\bar{x}_T - u_{(1+\alpha)/2}\frac{\sigma(x)}{\sqrt{T}}, \quad \bar{x}_T - u_{(1-\alpha)/2}\frac{\sigma(x)}{\sqrt{T}}]$$

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t, \ \sigma^2(x) = E(x-\mu)^2$$

$$u_\alpha \text{ solves } P\left(\frac{(\bar{x}_T - \mu)}{\frac{\sigma}{\sqrt{T}}} \leq u_\alpha\right) = \alpha$$

Exact interval, regardless of the underlying distribution.
Operational Version

$$I = [\bar{x}_T - \hat{u}_{(1+\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}, \bar{x}_T - \hat{u}_{(1-\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}]$$

$$\hat{\sigma}^2(x) = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x}_T)^2$$

$$\hat{u}_\alpha \text{ solves } P\left(\frac{(\bar{x}_T - \mu)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}} \leq \hat{u}_\alpha\right) = \alpha$$

Classic (Gaussian) example:

$$I = \bar{x}_T \pm t_{(1-\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}$$

Bootstrap approach: No need to assume Gaussian data.
"Percentile Bootstrap"

$$Root: \ S = \frac{(\bar{x}_T - \mu)}{\frac{\sigma}{\sqrt{T}}}$$

$$Root \ c.d.f.: \ H(z) = P\left(\frac{(\bar{x}_T - \mu)}{\frac{\sigma}{\sqrt{T}}} \leq z\right)$$

1. Draw $\{x_t^{(j)}\}_{t=1}^T$ with replacement from $\{x_t\}_{t=1}^T$
2. Compute $\frac{\bar{x}_T^{(j)} - \bar{x}_T}{\frac{\hat{\sigma}(x)}{\sqrt{T}}}$
3. Repeat many times and build up the sampling distribution of $\frac{\bar{x}_T^{(j)} - \bar{x}_T}{\frac{\hat{\sigma}(x)}{\sqrt{T}}}$ which is an approximation to the distribution of $\frac{\bar{x}_T - \mu}{\frac{\sigma}{\sqrt{T}}}$

"Russian doll principle"
Percentile Bootstrap, Continued
Bootstrap estimator of H(z):

$$\hat{H}(z) \ = \ P\left(\frac{(\bar{x}_T^{(j)} - \bar{x}_T)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}} \ \leq \ z\right)$$

Translates into bootstrap 100α percent CI:

$$\hat{I} \ = \ [\bar{x}_T - \hat{u}_{(1+\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}, \bar{x}_T - \hat{u}_{(1-\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}]$$

$$where \ P\left(\frac{(\bar{x}_T^{(j)} - \bar{x}_T)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}} \ \leq \ \hat{u}_\alpha\right) \ = \ \hat{H}(\hat{u}_\alpha) \ = \ \alpha$$

"Percentile-t" Bootstrap

$$S = \frac{(\bar{x}_T - \mu)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}}$$

$$H(z) = P\left(\frac{(\bar{x}_T - \mu)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}} \leq z\right)$$

$$\hat{H}(z) = P\left(\frac{(\bar{x}_T^{(j)} - \bar{x}_T)}{\frac{\hat{\sigma}(x^{(j)})}{\sqrt{T}}} \leq z\right)$$

$$\hat{I} = [\bar{x}_T - \hat{u}_{(1+\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}},\ \bar{x}_T - \hat{u}_{(1-\alpha)/2}\frac{\hat{\sigma}(x)}{\sqrt{T}}]$$

$$P\left(\frac{(\bar{x}_T^{(j)} - \bar{x}_T)}{\frac{\hat{\sigma}(x^{(j)})}{\sqrt{T}}} \ \leq\ \hat{u}_\alpha\right) = \alpha$$

Percentile-t Bootstrap, Continued

Key Insight:

Percentile: $\bar{x}_T^{(j)}$ changes across bootstrap replications

Percentile-t: both $\bar{x}_T^{(j)}$ and $\hat{\sigma}(x^{(j)})$ change across bootstrap replications

Effectively, the percentile method bootstraps the parameter, whereas the percentile-t bootstraps the t statistic

Key Bootstrap Property: Consistent Inference

Real-world root:

$$S \ \xrightarrow{d}\ D \quad (as\ T \to \infty)$$

Bootstrap-world root:

$$S^* \ \xrightarrow{d}\ D^* \quad (as\ T, N \to \infty)$$

Bootstrap consistent ("valid," "first-order valid") if $D = D^*$.

Holds under regularity conditions.

But Aren't There Simpler ways to do Consistent Inference?
Of Course. BUT:

1. Bootstrap idea extends mechanically to much more complicated models
2. Bootstrap can deliver higher-order refinements
   (e.g., percentile-t)
3. Monte Carlo indicates that bootstrap often does very well in finite samples (not unrelated to 2, but does not require 2)
4. Many variations and extensions of the basic bootstraps

7.8.2 Time-Series Environments

Stationary Time Series Case Before:

1. Use $S = \frac{(\bar{x}_T - \mu)}{\frac{\hat{\sigma}(x)}{\sqrt{T}}}$
2. Draw $\{x_t^{(j)}\}_{t=1}^T$ with replacement from $\{x_t\}_{t=1}^T$

Issues:

1. Inappropriate standardization of S for dynamic data. So replace $\hat{\sigma}(x)$ with $2\pi f_x^*(0)$, where $f_x^*(0)$ is a consistent estimator of the spectral density of x at frequency 0.
2. Inappropriate to draw $\{x_t^{(j)}\}_{t=1}^T$ with replacement for dynamic data. What to do?

Non-Parametric Time Series Bootstrap
(Overlapping Block Sampling)
Overlapping blocks of size b in the sample path:

$$\xi_t = (x_t, ..., x_{t+b-1}), t = 1, ..., T - b + 1$$

Draw k blocks (where $T = kb$) from $\{\xi_t\}_{t=1}^{T-b+1}$:

$$\xi_1^{(j)}, ..., \xi_k^{(j)}$$

Concatenate: $(x_1^{(j)}, ..., x_T^{(j)}) = (\xi_1^{(j)} ... \xi_k^{(j)})$
Consistent if $b \to \infty \; as \; T \to \infty \; with \; b/T \to 0$
AR(1) Parametric Time Series Bootstrap

$$x_t = c + \phi x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \; iid$$

1. Regress $x_t \to (c, x_{t-1})$ to get $\hat{c}$ and $\hat{\phi}$, and save residuals, $\{e_t\}_{t=1}^T$
2. Draw $\{\varepsilon_t^{(j)}\}_{t=1}^T$ with replacement from $\{e_t\}_{t=1}^T$
3. Draw $x_0^{(j)}$ from $\{x_t\}_{t=1}^T$
4. Generate $x_t^{(j)} = \hat{c} + \hat{\phi} x_{t-1}^{(j)} + \varepsilon_t^{(j)}, t = 1, ..., T$
5. Regress $x_t^{(j)} \to (c, \; x_{t-1}^{(j)})$ to get $\hat{c}^{(j)} \; and \; \hat{\phi}^{(j)}$, associated t-statistics, etc.
6. Repeat $j = 1, ..., R$, and build up the distributions of interest

General State-Space Parametric Time Series Bootstrap
Recall the prediction-error state space representation:

$$a_{t+1/t} = T a_{t/t-1} + T K_t v_t$$

$$y_t = Z a_{t/t-1} + v_t$$

1. Estimate system parameters θ. (We will soon see how to do this.)
2. At the estimated parameter values $\hat{\theta}$, run the Kalman filter to get the corresponding 1-step-ahead prediction errors $\hat{v}_t \sim (0, \hat{F}_t)$ and standardize them to $\hat{u}_t = \hat{\Omega}_t^{-1/2} \hat{v}_t \sim (0, I)$, where $\hat{\Omega}_t \hat{\Omega}_t' = \hat{F}_t$.

3. Draw $\{u_t^{(j)}\}_{t=1}^T$ with replacement from $\{\hat{u}_t\}_{t=1}^T$ and convert to $\{v_t^{(j)}\}_{t=1}^T = \{\hat{\Omega}_t u_t^{(j)}\}_{t=1}^T$.
4. Using the prediction-error draw $\{v_t^{(j)}\}_{t=1}^T$, simulate the model, obtaining $\{y_t^{(j)}\}_{t=1}^T$.
5. Estimate the model, obtaining $\hat{\theta}^{(j)}$ and related objects of interest.
6. Repeat $j = 1, ..., R$, simulating the distributions of interest.

Many Variations and Extensions...

- Stationary block bootstrap: Blocks of random (exponential) length
- Wild bootstrap: multiply bootstrap draws of shocks randomly by ± 1 to enforce symmetry
- Subsampling

7.9 OPTIMIZATION BY SIMULATION

Markov chains yet again.

7.9.1 Local

Using MCMC for MLE (and Other Extremum Estimators)

Chernozukov and Hong show how to compute extremum estimators as mean of pseudo-posterior distributions, which can be simulated by MCMC and estimated at the parametric rate $1/\sqrt{N}$, in contrast to the much slower nonparametric rates achievable (by any method) by the standard posterior mode extremum estimator.

7.9.2 Global

Summary of Local Optimization:

1. initial guess $\theta^{(0)}$
2. **while** stopping criteria not met **do**
3. select $\theta^{(c)} \in N(\theta^{(m)})$ (Classically: use gradient)
4. **if** $\Delta \equiv lnL(\theta^{(c)}) - lnL(\theta^{(m)}) > 0$ **then** $\theta^{(m+1)} = \theta^{(c)}$
5. end while

Simulated Annealing
(Illustrated Here for a Discrete Parameter Space)

Framework:

1. A set Θ, and a real-valued function lnL (satisfying regularity conditions) defined on Θ. Let $\Theta^* \subset \Theta$ be the set of global maxima of lnL
2. $\forall \theta^{(m)} \in \Theta$, a set $N(\theta^{(m)}) \subset \Theta - \theta^{(m)}$, the set of neighbors of $\theta^{(m)}$
3. A nonincreasing function, $T(m) : N \to (0, \infty)$ ("the cooling schedule"), where $T(m)$ is the "temperature" at iteration m
4. An initial guess, $\theta^{(0)} \in \Theta$

Simulated Annealing Algorithm

1. initial guess $\theta^{(0)}$
2. **while** stopping criteria not met **do**
3. select $\theta^{(c)} \in N(\theta^{(m)})$

4. **if** $\Delta > 0$ *or* $\exp(\Delta/T(m)) > U(0,1)$ **then** $\theta^{(m+1)} = \theta^{(c)}$
5. **end while**

Note the extremes:

$T = 0$ implies no randomization (like classical gradient-based)

$T = \infty$ implies complete randomization (like random search)

A (Heterogeneous) Markov Chain
If $\theta^{(c)} \notin N(\theta^{(m)})$ then
$P(\theta^{(m+1)} = \theta^{(c)} | \theta^{(m)}) = 0$
If $\theta^{(c)} \in N(\theta^{(m)})$ then
$P(\theta^{(m+1)} = \theta^{(c)} \,|\, \theta^{(m)}) \;=\; exp\ (\min[0,\ \Delta/T(m)])$
Convergence of a Global Optimizer
Definition. We say that the simulated annealing algorithm converges if
$\lim_{m\to\infty}\ P[\theta^{(m)} \in \Theta^*] \;=\; 1.$
Definition: We say that $\theta^{(m)}$ communicates with Θ^* at depth d if there exists a path in Θ (with each element of the path being a neighbor of the preceding element) that starts at $\theta^{(m)}$ and ends at some element of Θ^*, such that the smallest value of lnL along the path is $lnL(\theta^{(m)}) - d$.
Convergence of Simulated Annealing
Theorem: Let d^* be the smallest number such that every $\theta^{(m)} \in \Theta$ communicates with Θ^* at depth d^*. Then the simulated annealing algorithm converges if and only if, as $m \to \infty$,
$T(m) \;\to\; 0$
and
$\sum exp(-d^*/T(m)) \;\to\; \infty.$
Problems: How to choose T, and moreover we don't know d^*
Popular choice of cooling function: $T(m) \;=\; \frac{1}{\ln\ m}$
Regarding speed of convergence, little is known

7.9.3 Is a Local Optimum Global?

1. Try many startup values (sounds trivial but very important)
2. At the end of it all, use extreme value theory to assess the likelihood that the local optimum is global ("Veall's Method")

$\theta \in \Theta \subset R^k$
$lnL(\theta)$ is continuous
$lnL(\theta^*)$ is the unique finite global max of $lnL(\theta), \theta \in \Theta$
$H(\theta^*)$ exists and is nonsingular
$lnL(\hat{\theta})$ is a local max
Develop *statistical inference* for θ^*
Draw $\{\theta_i\}_{i=1}^N$ uniformly from Θ and form $\{lnL(\theta_i)\}_{i=1}^N$
lnL_1 first order statistic, lnL_2 second order statistic
$P[lnL(\theta^*) \in (lnL_1,\ lnL^\alpha)] \;=\; (1-\alpha),\ as\ N \to \infty$
where
$lnL^\alpha \;=\; lnL_1 \;+\; \frac{lnL_1 - lnL_2}{(1-\alpha)^{\frac{-2}{k}} - 1}.$

7.10 INTERVAL AND DENSITY FORECASTING BY SIMULATION

7.11 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Convex relaxation.

   Our approaches to global optimization involved attacking a nasty objective function with methods involving clever randomization. Alternatively, one can approximate the nasty objective with a friendly (convex) objective, which hopefully has the same global optimum. This is called "convex relaxation," and when the two optima coincide we say that the relaxation is "tight."

7.12 NOTES

Chapter Eight

Bayesian Time Series Posterior Analysis by Markov Chain Monte Carlo

8.1 BAYESIAN BASICS

8.2 COMPARATIVE ASPECTS OF BAYESIAN AND FREQUENTIST PARADIGMS

Overarching Paradigm ($T \to \infty$)

$$\sqrt{T}(\hat{\theta} - \theta) \sim N(0, \Sigma)$$

Shared by classical and Bayesian, but interpretations differ.

Classical: $\hat{\theta}$ random, θ fixed

Bayesian: $\hat{\theta}$ fixed, θ random

Classical: Characterize the distribution of the random data ($\hat{\theta}$) conditional on fixed "'true" θ. Focus on the likelihood max ($\hat{\theta}_{ML}$) and likelihood curvature in an ϵ-neighborhood of the max.

Bayesian: Characterize the distribution of the random θ conditional on fixed "true" data ($\hat{\theta}$). Examine the entire likelihood.

Bayesian Computational Mechanics

Data $y \equiv \{y_1, \ldots, y_T\}$

Bayes' Theorem:

$$f(\theta/y) = \frac{f(y/\theta)f(\theta)}{f(y)}$$

or

$f(\theta/y) = c\ f(y/\theta)f(\theta)$

where $c^{-1} = \int_\theta f(y/\theta)f(\theta)$

$f(\theta/y) \propto f(y/\theta)f(\theta)$

$p(\theta/y) \propto L(\theta/y)g(\theta)$

posterior $\propto$ likelihood $\cdot$ prior

Classical Paradigm ($T \to \infty$)

$$\sqrt{T}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} N\left(0, \left(\frac{I_{EX,H}(\theta)}{T}\right)^{-1}\right)$$

or more crudely

$$\sqrt{T}(\hat{\theta}_{ML} - \theta) \sim N(0, \Sigma)$$

(Enough said.)

Bayesian Paradigm ($T \to \infty$)

(Note that as $T \to \infty$, $p(\theta/y) \approx L(\theta/y)$,
so the likelihood below can be viewed as the posterior.)

Expand $lnL(\theta/y)$ around fixed $\hat{\theta}_{ML}$:
$lnL(\theta/y) \approx lnL(\hat{\theta}_{ML}/y) + S(\hat{\theta}_{ML}/y)'(\theta - \hat{\theta}_{ML})$
$-1/2(\theta - \hat{\theta}_{ML})' I_{OB,H}(\hat{\theta}_{ML}/y)(\theta - \hat{\theta}_{ML})$
But $S(\hat{\theta}_{ML}/y) \equiv 0$, so:
$lnL(\theta/y) \approx lnL(\hat{\theta}_{ML}/y) - 1/2(\theta - \hat{\theta}_{ML})' I_{OB,H}(\hat{\theta}_{ML}/y)(\theta - \hat{\theta}_{ML})$
Neglecting the expansion remainder, we then have:
$L(\theta/y) \propto exp(-1/2(\theta - \hat{\theta}_{ML})' I_{OB,H}(\hat{\theta}_{ML}/y)(\theta - \hat{\theta}_{ML}))$
or
$L(\theta/y) \propto N(\hat{\theta}_{ML}, I^{-1}_{OB,H}(\hat{\theta}_{ML}/y))$
Or, a posteriori, $\sqrt{T}(\theta - \hat{\theta}_{ML}) \sim N(0, \Sigma)$
Bayesian estimation and model comparison
Estimation:

Full posterior density
Highest posterior density intervals
Posterior mean, median, mode (depending on loss function)

Model comparison:

$$\underbrace{\frac{p(M_i|y)}{p(M_j|y)}}_{posterior\ odds} = \underbrace{\frac{p(y|M_i)}{p(y|M_j)}}_{Bayes\ factor} \bullet \underbrace{\frac{p(M_i)}{p(M_j)}}_{prior\ odds}$$

The Bayes factor is a ratio of *marginal likelihoods*, or *marginal data densities*.
– For comparison, simply report the posterior odds
– For selection, invoke 0-1 loss which implies selection of the model with highest posterior probability.
– Hence select the model with highest marginal likelihood if prior odds are 1:1.
Understanding the Marginal Likelihood
As a penalized log likelihood:
As $T \to \infty$, the marginal likelihood is approximately the maximized log likelihood minus $KlnT/2$. It's the SIC!
As a predictive likelihood:

$$P(y) = P(y_1, ..., y_T) = \prod_{t=1}^{T} P(y_t|y_{1:t-1})$$

$$\implies lnP(y) = \sum_{t=1}^{T} lnP(y_t|y_{1:t-1})$$

$$= \sum_{t=1}^{T} ln \int P(y_t|\theta, y_{1:t-1}) P(\theta|y_{1:t-1})\, d\theta$$

Bayesian model averaging:
Weight by posterior model probabilities:

$$P(y_{t+1}|y_{1:t}) = \pi_{it} P(y_{t+1}|y_{1:t}, M_i) + \pi_{jt} P(y_{t+1}|y_{1:t}, M_j)$$

As $T \to \infty$, the distinction between model averaging and selection vanishes, as one π goes to 0 and the other goes to 1.

If one of the models is true, then both model selection and model averaging are consistent for the true model. Otherwise they're consistent for the X-optimal approximation to the truth. Does X = KLIC?

8.3 MARKOV CHAIN MONTE CARLO

Metropolis-Hastings

We want to draw S values of θ from $p(\theta)$. Initialize chain at $\theta^{(0)}$ and burn it in.

1. Draw θ^* from proposal density $q(\theta; \theta^{(s-1)})$
2. Calculate the acceptance probability $\alpha(\theta^{(s-1)}, \theta^*)$
3. Set

$$\theta^s = \begin{cases} \theta^* \text{ w.p. } \alpha(\theta^{(s-1)}, \theta^*) & \text{“accept”} \\ \\ \theta^{(s-1)} \text{ w.p. } 1 - \alpha(\theta^{(s-1)}, \theta^*) & \text{“reject”} \end{cases}$$

4. Repeat 1-3, $s = 1, ..., S$

The question, of course, is what to use for step 2.

8.3.1 Metropolis-Hastings Independence Chain

Fixed proposal density:

$$q(\theta; \theta^{(s-1)}) = q^*(\theta)$$

Acceptance probability:

$$\alpha(\theta^{(s-1)}, \theta^*) = min \left[\frac{p(\theta = \theta^*)\ q^*\left(\theta = \theta^{(s-1)}\right)}{p(\theta = \theta^{(s-1)})\ q^*\left(\theta = \theta^*\right)},\ 1 \right]$$

8.3.2 Metropolis-Hastings Random Walk Chain

Random walk proposals:

$$\theta^* = \theta^{(s-1)} + \varepsilon$$

Acceptance probability reduces to:

$$\alpha(\theta^{(s-1)}, \theta^*) = min \left[\frac{p(\theta = \theta^*)}{p(\theta = \theta^{(s-1)})}, 1 \right]$$

8.3.3 More

Burn-in, Sampling, and Dependence

“total simulation” = “burn-in” + “sampling”

Questions:

How to assess convergence to steady state?

In the Markov chain case, why not do something like the following. Whenever time t is a multiple of m, use a distribution-free non-parametric (randomization) test for equality of distributions to test whether the unknown distribution f_1 of $x_t, ..., x_{t-(m/2)}$ equals the unknown distribution f_2 of $x_{t-(m/2)+1}, ..., x_{t-m}$. If, for example, we pick $m = 20,000$, then whenever time t is a multiple of 20,000 we would test equality of the distributions of $x_t, ..., x_{t-10000}$ and $x_{t-10001}, ..., x_{t-20000}$. We declare arrival at the steady state when the null is not rejected. Or something like that.

Of course the Markov chain is serially correlated, but who cares, as we're only trying to assess equality of unconditional distributions. That is, randomizations of $x_t, ..., x_{t-(m/2)}$ and of $x_{t-(m/2)+1}, ..., x_{t-m}$ destroy the serial correlation, but so what?

How to handle dependence in the sampled chain?

Better to run one long chain or many shorter parallel chains?

A Useful Property of Accept-Reject Algorithms
(e.g., Metropolis)

Metropolis requires knowing the density of interest only up to a constant, because the acceptance probability is governed by the RATIO $p(\theta = \theta^*)/p(\theta = \theta^{(s-1)})$. This will turn out to be important for Bayesian analysis.

Metropolis-Hastings (Discrete)

For desired π, we want to find P such that $\pi P = \pi$. It is sufficient to find P such that $\pi_i P_{ij} = \pi_j P_{ji}$. Suppose we've arrived at z_i. Use symmetric, irreducible transition matrix $Q = [Q_{ij}]$ to generate proposals. That is, draw proposal z_j using probabilities in i^{th} row of Q.

Move to z_j w.p. α_{ij}, where:

$$\alpha_{ij} = \begin{cases} 1, \text{ if } \frac{\pi_j}{\pi_i} \geq 1 \\ \\ \frac{\pi_j}{\pi_i} \text{ otherwise} \end{cases}$$

Equivalently, move to z_j w.p. α_{ij}, where:

$$\alpha_{ij} = min\left(\frac{\pi_j}{\pi_i},\ 1\right)$$

Metropolis-Hastings, Continued...

This defines a Markov chain P with:

$$P_{ij} = \begin{cases} \alpha_{ij} Q_{ij}, \text{ for } i \neq j \\ \\ 1 - \sum_{j \neq i} P_{ij}, \text{ for } i = j \end{cases}$$

Iterate this chain to convergence and start sampling from π.

Blocking strategies: Yu and Meng (2010)

Note that I have set this up to list bibliography at end. It gives a compiling error, but you can skip through it, and everything looks fine at the end.

Blocking MH algorithms: Ed Herbst Siddhartha Chib and Srikanth Ramamurthy. Tailored randomized block mcmc methods with application to dsge models. Journal of Econometrics, 155(1):1938, 2010. Vasco Curdia and Ricardo Reis. Correlated disturbances and u.s. business cycles. 2009. Nikolay Iskrev. Evaluating the information matrix in linearized dsge models. Economics Letters, 99:607610, 2008. Robert Kohn, Paolo Giordani, and Ingvar Strid. Adaptive hybrid metropolis-hastings samplers for dsge models. Working Paper, 2010. G. O. Roberts and S.K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. Journal of the Royal Statistical Society. Series B (Methodological), 59(2):291317, 1997.

8.3.4 Gibbs and Metropolis-Within-Gibbs

Bivariate Gibbs Sampling

We want to sample from $f(z) = f(z_1, z_2)$

Initialize $(j = 0)$ using z_2^0

Gibbs iteration $j = 1$:

a. Draw z_1^1 from $f(z_1|z_2^0)$

b. Draw z_2^1 from $f(z_2|z_1^1)$

Repeat $j = 2,\ 3,\ ....$

Theorem (Clifford-Hammersley): $lim_{j\to\infty} f(z^j) = f(z)$

Useful if/when conditionals are known and easy to sample from, but joint and marginals are not. (This happens a lot in Bayesian analysis.)

General Gibbs Sampling

We want to sample from $f(z) = f(z_1, z_2, ..., z_k)$

Initialize $(j = 0)$ using $z_2^0, z_3^0, ..., z_k^0$

Gibbs iteration $j = 1$:
a. Draw z_1^1 from $f(z_1|z_2^0, ..., z_k^0)$
b. Draw z_2^1 from $f(z_2|z_1^1, z_3^0, ..., z_k^0)$
c. Draw z_3^1 from $f(z_3|z_1^1, z_2^1, z_4^0, ..., z_k^0)$
...
k. Draw z_k^1 from $f(z_k|z_1^1, ..., z_{k-1}^1)$
Repeat $j = 2,\ 3, ...$
Again, $lim_{j\to\infty} f(z^j) = f(z)$
Metropolis Within Gibbs
Gibbs breaks a big draw into lots of little (conditional) steps. If you're lucky, those little steps are simple.
If/when a Gibbs step is difficult, i.e., it's not clear how to sample from the relevant conditional, it can be done by Metropolis.
("Metropolis within Gibbs")
Metropolis is more general but also more tedious, so only use it when you must.
Composition
We may want $(x_1, y_1), ..., (x_N, y_N) \sim iid$ from $f(x, y)$
Or we may want $y_1, ..., y_N \sim iid$ from $f(y)$
They may be hard to sample from directly.
But sometimes it's easy to:
Draw $x^* \sim f(x)$
Draw $y^* \sim f(y|x^*)$
Then:
$(x_1, y_1), ..., (x_N, y_N) \sim iid\ f(x, y)$
$(y_1, ..., y_N) \sim iid\ f(y)$

8.4 CONJUGATE BAYESIAN ANALYSIS OF LINEAR REGRESSION

Bayes for Gaussian Regression with Conjugate Priors
$y = X\beta + \varepsilon$
$\varepsilon \sim iid\ N(0, \sigma^2 I)$
Standard results:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2_{ML} = \frac{e'e}{T}$$

$$\hat{\beta}_{ML} \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$$

$$\frac{T\hat{\sigma}^2_{ML}}{\sigma^2} \sim \chi^2_{T-K}$$

Bayesian Inference for β/σ
Prior:
$\beta/\sigma^2 \sim N(\beta_0, \Sigma_0)$
$g(\beta/\sigma^2) \propto exp(-1/2(\beta - \beta_0)'\Sigma_0^{-1}(\beta - \beta_0))$
Likelihood:
$L(\beta/\sigma^2, y) \propto exp(\frac{-1}{2\sigma^2}(y - X\beta)'(y - X\beta))$
Posterior:
$p(\beta/\sigma^2, y) \propto exp(-1/2(\beta - \beta_0)'\Sigma_0^{-1}(\beta - \beta_0) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta))$
This is the kernel of a normal distribution (*Problem*):
$\beta/\sigma^2, y \sim N(\beta_1, \Sigma_1)$
where
$\beta_1 = \left(\Sigma_0^{-1} + \sigma^{-2}(X'X)\right)^{-1}(\Sigma_0^{-1}\beta_0 + \sigma^{-2}(X'X)\hat{\beta}_{ML})$

$\Sigma_1 = (\Sigma_0^{-1} + \sigma^{-2}(X'X))^{-1}$

Gamma and Inverse Gamma Refresher

$$z_t \overset{iid}{\sim} N\left(0, \frac{1}{\delta}\right), \; x = \sum_{t=1}^{v} z_t^2 \;\Rightarrow\; x \sim \Gamma\left(x; \frac{v}{2}, \frac{\delta}{2}\right)$$

(Note $\delta = 1 \Rightarrow x \sim \chi_v^2$, so χ^2 is a special case of Γ)

$$\Gamma\left(x; \frac{v}{2}, \frac{\delta}{2}\right) \propto x^{\frac{v}{2}-1} exp\left(\frac{-x\delta}{2}\right)$$

$$E(x) = \frac{v}{\delta}$$

$$var(x) = \frac{2v}{\delta^2}$$

$x \sim \Gamma^{-1}(\frac{v}{2}, \frac{\delta}{2})$ ("inverse gamma") $\Leftrightarrow \frac{1}{x} \sim \Gamma(\frac{v}{2}, \frac{\delta}{2})$

Bayesian Inference for σ^2/β

Prior:

$\frac{1}{\sigma^2}/\beta \sim \Gamma\left(\frac{v_0}{2}, \frac{\delta_0}{2}\right)$

$g\left(\frac{1}{\sigma^2}/\beta\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{v_0}{2}-1} exp\left(-\frac{\delta_0}{2\sigma^2}\right)$

(Independent of β, but write σ^2/β for completeness.)

$L\left(\frac{1}{\sigma^2}/\beta, y\right) \propto \left(\sigma^2\right)^{-T/2} exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$

(*Problem*: In contrast to $L(\beta/\sigma^2, y)$ earlier, we don't absorb the $(\sigma^2)^{-T/2}$ term into the constant of proportionality. Why?)

Hence (*Problem*):

$p\left(\frac{1}{\sigma^2}/\beta, y\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{v_1}{2}-1} exp\left(\frac{-\delta_1}{2\sigma^2}\right)$

or $\frac{1}{\sigma^2}/\beta, y \sim \Gamma\left(\frac{v_1}{2}, \frac{\delta_1}{2}\right)$

$v_1 = v_0 + T$

$\delta_1 = \delta_0 + (y - X\beta)'(y - X\beta)$

Bayesian Pros Thus Far

1. Feels sensible to focus on $p(\theta/y)$. Classical relative frequency in repeated samples replaced with subjective degree of belief conditional on the single sample actually obtained
2. Exact finite-sample full-density inference

Bayesian Cons Thus Far

1. From where does the prior come? How to elicit prior distributions?
2. How to do an "objective" analysis?
   (e.g. what is an "uninformative" prior? Uniform?)
   (Note, however, that priors can be desirable and helpful. See, for example, the cartoon at http://fxdiebold.blogspot.com/2014/04/more-from-xkcdcom.html)
3. We still don't have the marginal posteriors that we really want: $p(\beta, \sigma^2/y)$, $p(\beta/y)$.
   – Problematic in any event!

8.5 GIBBS FOR SAMPLING MARGINAL POSTERIORS

Markov Chain Monte Carlo Solves the Problem! 0. Initialize: $\sigma^2 = (\sigma^2)^{(0)}$

Gibbs sampler at generic iteration j:

$j1$. Draw $\beta^{(j)}$ from $p(\beta^{(j)}/(\sigma^2)^{(j-1)}, y)$ $\quad (N(\beta_1, \Sigma_1))$

$j2$. Draw $(\sigma^2)^{(j)}$ from $p(\sigma^2/\beta^{(j)}, y)$ $\quad \left(\Gamma^{-1}\left(\frac{v_1}{2}, \frac{\delta_1}{2}\right)\right)$

Iterate to convergence to steady state, and then estimate posterior moments of interest

8.6 GENERAL STATE SPACE: CARTER-KOHN MULTI-MOVE GIBBS

Bayesian Analysis of State-Space Models

$$\alpha_t = T\alpha_{t-1} + R\eta_t$$
$$y_t = Z\alpha_t + \varepsilon_t$$

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \overset{iid}{\sim} N \begin{pmatrix} Q & 0 \\ 0 & H \end{pmatrix}$$

Let $\tilde{\alpha}_T = (\alpha_1', \ldots, \alpha_T')'$, $\theta = (T', R', Z', Q', H')'$

The key: Treat $\tilde{\alpha}_T$ as a *parameter*, along with system matrices θ

Recall the State-Space Model in Density Form

$$\alpha_t|\alpha_{t-1} \sim N(T\alpha_{t-1},\ RQR')$$

$$y_t|\alpha_t \sim N(Z\alpha_t,\ H)$$

Recall the Kalman Filter in Density Form

Initialize at a_0, P_0

State prediction:

$\alpha_t|\tilde{y}_{t-1} \sim N(a_{t/t-1},\ P_{t/t-1})$

$a_{t/t-1} = Ta_{t-1}$

$P_{t/t-1} = TP_{t-1}T' + RQR'$

State update:

$\alpha_t|\tilde{y}_t \sim N(a_t,\ P_t)$

$a_t = a_{t/t-1} + K_t(y_t - Za_{t/t-1})$

$P_t = P_{t/t-1} - K_tZP_{t/t-1}$

Data prediction:

$y_t|\tilde{y}_{t-1} \sim N(Za_{t/t-1},\ F_t)$

where $\tilde{y}_t = (y_1', ..., y_t')'$

Carter-Kohn Multi-move Gibbs Sampler

Let $\tilde{y}_T = (y_1', \ldots, y_T')'$

0. Initialize $\theta^{(0)}$

Gibbs sampler at generic iteration j:

$j1$. Draw from posterior $\tilde{\alpha}_T^{(j)}/\theta^{(j-1)}, \tilde{y}_T$ ("hard")

$j2$. Draw from posterior $\theta^{(j)}/\tilde{\alpha}_T^{(j)}, \tilde{y}_T$ ("easy")

Iterate to convergence, and then estimate posterior moments of interest

Just two Gibbs draws: (1) $\tilde{\alpha}_T$ parameter, (2) θ parameter

Multimove Gibbs Sampler, Step 2 $(\theta^{(j)}|\tilde{\alpha}_T^{(j)}, \tilde{y}_T)$ ("easy")

Conditional upon draws $\tilde{\alpha}_T^{(j)}$, sampling $\theta^{(j)}$ becomes a multivariate regression problem.

We have already seen how to do univariate regression. We can easily extend to multivariate regression. The Gibbs sampler continues to work.

Multivariate Regression

$$\underbrace{Y}_{T\times n} = \underbrace{X}_{T\times k}\underbrace{B}_{k\times n} + \underbrace{E}_{T\times n},$$

where $E = [\epsilon_1, \epsilon_2, ..., \epsilon_T]'$, $\epsilon_t = [\epsilon_{1,t}, ...\epsilon_{n,t}]'$

$$\epsilon_t \overset{iid}{\sim} N(0, \Sigma)$$

Important differences compared to univariate regression:

a) B is a matrix rather than a vector.

b) Σ is matrix rather than a scalar.

Conjugate Priors for Multivariate Regression
$B|\Sigma$ multivariate normal prior:

$$vec(B)|\Sigma \sim N(B_0, \Sigma_0)$$

Inverse Wishart refresher (multivariate inverse gamma):

$$X \sim W^{-1}(n, \mathbf{V}) \leftrightarrow X^{-1} \sim W(n, \mathbf{V})$$

where

$$W(X; n, \mathbf{V}) \propto |X|^{\frac{n-p-1}{2}} \exp\left(-\frac{1}{2} tr(X\mathbf{V}^{-1})\right)$$

$\Sigma|B$ inverse Wishart prior:

$$p(\Sigma^{-1}|vec(B)) \propto |\Sigma^{-1}|^{\frac{n-p-1}{2}} \exp(-\frac{1}{2} tr(\Sigma^{-1}\mathbf{V}^{-1}))$$

Bayesian Inference for $B|\Sigma$
Prior:

$$p(vec(B)|\Sigma) \propto \exp\left(-\frac{1}{2} tr\left(vec(B - B_0)' V_0^{-1} vec(B - B_0)\right)\right)$$

Likelihood:

$$p(Y, X|B, \Sigma) \propto \exp\left(-\frac{1}{2} \sum_{t=1}^{T} (Y_t - B'X_t)' \Sigma^{-1} (Y_t - B'X_t)\right)$$

$$\propto \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1}(Y - XB)'(Y - XB)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2} tr\left(vec(B - \hat{B})' \left(\Sigma^{-1} \otimes X'X\right) vec(B - \hat{B})\right)\right)$$

Posterior:

$$p(vec(B)|\Sigma, Y) \propto$$

$$\exp\left(-\frac{1}{2}\left(tr\left(vec(B - \hat{B})' \left(\Sigma^{-1} \otimes X'X\right) vec(B - \hat{B}) + vec(B - B_0)' V_0^{-1} vec(B - B_0)\right)\right)\right)$$

This is the kernel of a multivariate normal distribution:

$$vec(B)|\Sigma, Y \sim N(B_1, V_1)$$

$vec(B_1) = V_1 \left[(\Sigma^{-1} \otimes X'X) vec(\hat{B}) + V_0^{-1} B_0\right], V_1 = \left[\Sigma^{-1} \otimes X'X + V_0^{-1}\right]^{-1}$
and $\hat{B} = (X'X)^{-1}(X'Y)$
Bayesian Inference for $\Sigma|B$
Prior:

$$p(\Sigma^{-1}|vec(B)) \propto |\Sigma^{-1}|^{\frac{n-p-1}{2}} \exp(-\frac{1}{2} tr(\Sigma^{-1}\mathbf{V}^{-1}))$$

Likelihood:

$$p(Y, X|B, \Sigma) \propto |\Sigma|^{-\frac{T}{2}} \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1}(Y - XB)'(Y - XB)\right)\right)$$

Posterior:

$$p(\Sigma^{-1}|vec(B), Y) \propto |\Sigma^{-1}|^{\frac{T+n-p-1}{2}} \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1}\left((Y - XB)'(Y - XB) + \mathbf{V}^{-1}\right)\right)\right)$$

This is the kernel of a Wishart distribution:

$$\Sigma^{-1}|vec(B), Y \sim W(T+n, ((Y-XB)'(Y-XB)+\mathbf{V}^{-1})^{-1})$$

Multimove Gibbs Sampler, Step 1 $(\tilde{\alpha}_T^{(j)}/\theta^{(j-1)}, \tilde{y}_T)$ ("hard")
For notational simplicity we write $p(\tilde{\alpha}_T/\tilde{y}_T)$, suppressing the dependence on θ.
$p(\tilde{\alpha}_T/\tilde{y}_T) = p(\alpha_T/\tilde{y}_T)p(\tilde{\alpha}_{T-1}/\alpha_T, \tilde{y}_T)$
$= p(\alpha_T/\tilde{y}_T)p(\alpha_{T-1}/\alpha_T, \tilde{y}_T)p(\tilde{\alpha}_{T-2}/\alpha_{T-1}, \alpha_T, \tilde{y}_T)$
$= \ldots$
$= p(\alpha_T/\tilde{y}_T)\Pi_{t=1}^{(T-1)}p(\alpha_t/\alpha_{t+1}, \tilde{y}_t)$
(*Problem*: Fill in the missing steps subsumed under "...")
So, to draw from $p(\tilde{\alpha}_T/\tilde{y}_T)$, we need to be able to draw from $p(\alpha_T/\tilde{y}_T)$ and $p(\alpha_t/\alpha_{t+1}, \tilde{y}_t), t = 1, \ldots, (T-1)$
Multimove Gibbs sampler, Continued
The key is to *work backward*:
Draw from $p(\alpha_T/\tilde{y}_T)$,
then from $p(\alpha_{T-1}/\alpha_T, \tilde{y}_{T-1})$,
then from $p(\alpha_{T-2}/\alpha_{T-1}, \tilde{y}_{T-2})$,
etc.
Time T draw is easy:
$p(\alpha_T/\tilde{y}_T)$ *is* $N(a_{T,T}, P_{T,T})$
(where the Kalman filter delivers $a_{T,T}$ and $P_{T,T}$)
Earlier-time draws are harder:
How to get $p(\alpha_t/\alpha_{t+1}, \tilde{y}_t),\ t = (T-1), ..., 1$?
Multimove Gibbs sampler, Continued
It can be shown that (*Problem*):
$p(\alpha_t/\alpha_{t+1}, \tilde{y}_t),\ t = (T-1), ..., 1,$ *is* $N(a_{t/t,\alpha_{t+1}}, P_{t/t,\alpha_{t+1}})$
where
$a_{t/t,\alpha_{t+1}} = E(\alpha_t/\tilde{y}_t, \alpha_{t+1}) = E(\alpha_t|a_t, \alpha_{t+1})$
$= a_t + P_tT'(TP_tT'+Q)^{-1}(\alpha_{t+1} - Ta_t)$
$P_{t/t,\alpha_{t+1}} = cov(\alpha_t/\tilde{y}_t, \alpha_{t+1}) = cov(\alpha_t|a_t, \alpha_{t+1})$
$= P_t - P_tT'(TP_tT'+Q)^{-1}TP_t$
*** Expanding $S(\hat{\theta}_{ML})$ around θ yields:
$S(\hat{\theta}_{ML}) \approx S(\theta) + S'(\theta)(\hat{\theta}_{ML} - \theta) = S(\theta) + H(\theta))(\hat{\theta}_{ML} - \theta)$.
Noting that $S(\hat{\theta}_{ML}) \equiv 0$ and taking expectations yields:
$0 \approx S(\theta) - I_{EX,H}(\theta)(\hat{\theta}_{ML} - \theta)$
or
$(\hat{\theta}_{ML} - \theta) \approx I_{EX,H}^{-1}(\theta)$.
Using $S(\theta) \overset{a}{\sim} N(0, I_{EX,H}(\theta))$ then implies:
$(\hat{\theta}_{ML} - \theta) \overset{a}{\sim} N(0, I_{EX,H}^{-1}(\theta))$
or
Case 3 β *and* σ^2
Joint prior $g(\beta, \frac{1}{\sigma^2}) = g(\beta/\frac{1}{\sigma^2})g(\frac{1}{\sigma^2})$
where $\beta/\frac{1}{\sigma^2} \sim N(\beta_0, \Sigma_0)$ and $\frac{1}{\sigma^2} \sim G(\frac{v_0}{2}, \frac{\delta_0}{2})$
HW Show that the joint posterior,
$p(\beta, \frac{1}{\sigma^2}/y) = g(\beta, \frac{1}{\sigma^2})L(\beta, \frac{1}{\sigma^2}/y)$
can be factored as $p(\beta/\frac{1}{\sigma^2}, y)p(\frac{1}{\sigma^2/y})$
where $\beta/\frac{1}{\sigma^2}, y \sim N(\beta_1, \Sigma_1)$
and $\frac{1}{\sigma^2}/y \sim G(\frac{v_1}{2}, \frac{\delta_1}{2})$,
and derive expressions for $\beta_1, \Sigma_1, v_1, \delta_1$

in terms of $\beta_0, \Sigma_0, \delta_0, x$, and y.

Moreover, the key marginal posterior
$P(\beta/y) = \int_0^\infty p(\beta, \frac{1}{\sigma^2}/y)d\sigma^2$ is multivariate t.

Implement the Bayesian methods via Gibbs sampling.

8.7 EXERCISES, PROBLEMS AND COMPLEMENTS

8.8 NOTES

Chapter Nine

Non-Stationarity: Integration, Cointegration and Long Memory

9.1 RANDOM WALKS AS THE I(1) BUILDING BLOCK: THE BEVERIDGE-NELSON DECOMPOSITION

Random Walks

Random walk:

$$y_t = y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)$$

Random walk with drift:

$$y_t = \delta + y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)$$

Properties of the Random Walk

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$$

(shocks perfectly persistent)

$$E(y_t) = y_0$$
$$var(y_t) = t\sigma^2$$
$$\lim_{t\to\infty} var(y_t) = \infty$$

Properties of the Random Walk with Drift

$$y_t = t\delta + y_0 + \sum_{i=1}^{t} \varepsilon_i$$

(shocks again perfectly persistent)

$$E(y_t) = y_0 + t\delta$$
$$var(y_t) = t\sigma^2$$
$$\lim_{t\to\infty} var(y_t) = \infty$$

The Random Walk as a Building Block

Generalization of random walk: $ARIMA(p, 1, q)$

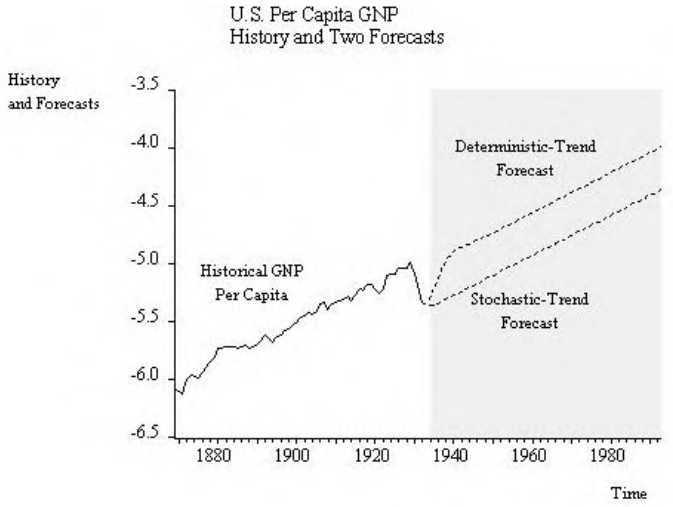
Beveridge-Nelson decomposition:

$y_t \sim ARIMA(p, 1, q) \Rightarrow y_t = x_t + z_t$

$x_t = random\ walk$

$z_t = covariance\ stationary$

– So shocks to $ARIMA(p, 1, q)$ are persistent, but not perfectly so.



– This was univariate BN. We will later derive multivariate BN.

Forecasting a Random Walk with Drift

$$x_t = b + x_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)$$

Optimal forecast:

$$x_{T+h,T} = bh + x_T$$

Forecast does not revert to trend

Forecasting a Linear Trend + Stationary AR(1)

$x_t = a + bt + y_t$

$$y_t = \phi y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)$$

Optimal forecast:

$$x_{T+h,T} = a + b(T+h) + \phi^h y_T$$

Forecast reverts to trend

9.2 STOCHASTIC VS. DETERMINISTIC TREND

Some Language...

"Random walk with drift" vs. "stat. $AR(1)$ around linear trend"

"unit root" vs. "stationary root"

"Difference stationary" vs. "trend stationary"

"Stochastic trend" vs. "deterministic trend"

"$I(1)$" vs. "$I(0)$"

Stochastic Trend vs. Deterministic Trend

9.3 UNIT ROOT DISTRIBUTIONS

Unit Root Distribution in the AR(1) Process

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

$$T(\hat{\phi}_{LS} - 1) \quad \overset{d}{\rightarrow} \quad DF$$

Superconsistent
Biased in finite samples ($E\hat{\phi} < \phi \ \forall \phi \in (0, 1]$)
"Hurwicz bias" "Dickey-Fuller bias"
"Nelson-Kang spurious periodicity"
Bigger as $T \to 0$, as $\phi \to 1$, and as intercept, trend included
Non-Gaussian (skewed left)
DF tabulated by Monte Carlo
Studentized Version

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s\sqrt{\frac{1}{\sum_{t=2}^{T} y_{t-1}^2}}}$$

Not t in finite samples
Not $N(0, 1)$ asymptotically
Again, tabulate by Monte Carlo
Nonzero Mean Under the Alternative

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + \varepsilon_t$$
$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t$$

where $\alpha = \mu(1 - \phi)$
Random walk null vs. mean-reverting alternative
Studentized statistic $\hat{\tau}_\mu$
Deterministic Trend Under the Alternative

$$(y_t - a - b\,t) = \phi(y_{t-1} - a - b\,(t-1)) + \varepsilon_t$$
$$y_t = \alpha + \beta t + \phi y_{t-1} + \varepsilon_t$$

where $\alpha = a(1 - \phi) + b\phi$ and $\beta = b(1 - \phi)$
$H_0 : \phi = 1$ (unit root)
$H_1 : \phi < 1$ (stationary root)
"Random walk with drift" vs. "stat. $AR(1)$ around linear trend"
"Difference stationary" vs. "trend stationary"
"Stochastic trend" vs. "deterministic trend"
"$I(1)$" vs. "$I(0)$"
Studentized statistic $\hat{\tau}_\tau$
Tabulating the Dickey-Fuller Distributions

1. Set T
2. Draw $TN(0, 1)$ variates
3. Construct y_t
4. Run three DF regressions (using common random numbers)

- $\hat{\tau}$: $y_t = \phi y_{t-1} + e_t$
- $\hat{\tau}_\mu$: $y_t = c + \phi y_{t-1} + e_t$
- $\hat{\tau}_\tau$: $y_t = c + \beta t + \phi y_{t-1} + e_t$

5. Repeat N times, yielding $\{\hat{\tau}^i,\ \hat{\tau}_\mu^i,\ \hat{\tau}_\tau^i\}_{i=1}^N$
6. Sort and compute fractiles
7. Fit response surfaces

$AR(p)$

$$y_t \ + \ \sum_{j=1}^{p} \phi_j y_{t-j} \ = \ \varepsilon_t$$

$$y_t \ = \ \rho_1 y_{t-1} \ + \ \sum_{j=2}^{p} \rho_j (y_{t-j+1} - y_{t-j}) \ + \ \varepsilon_t$$

where $p \geq 2$, $\rho_1 = -\sum_{j=1}^{p} \phi_j$, and $\rho_i = \sum_{j=i}^{p} \phi_j$, $i = 2, ..., p$
Studentized statistic $\hat{\tau}$
Allowing for Nonzero Mean Under the Alternative

$$(y_t - \mu) \ + \ \sum_{j=1}^{p} \phi_j \ (y_{t-j} - \mu) \ = \ \varepsilon_t$$

$$y_t \ = \ \alpha \ + \ \rho_1 y_{t-1} \ + \ \sum_{j=2}^{p} \rho_j \ (y_{t-j+1} - y_{t-j}) \ + \ \varepsilon_t$$

where $\alpha = \mu(1 + \sum_{j=1}^{p} \phi_j)$
Studentized statistic $\hat{\tau}_\mu$
Allowing for Trend Under the Alternative

$$(y_t \ - \ a \ - \ bt) \ + \ \sum_{j=1}^{p} \phi_j (y_{t-j} \ - \ a \ - \ b(t-j)) \ = \ \varepsilon_t$$

$$y_t \ = \ k_1 \ + \ k_2 \ t + \ \rho_1 y_{t-1} \ + \ \sum_{j=2}^{p} \rho_j (y_{t-j+1} - y_{t-j}) \ + \ \varepsilon_t$$

$$k_1 \ = \ a(1 + \sum_{i=1}^{p} \phi_i) \ - \ b \sum_{i=1}^{p} i\phi_i$$

$$k_2 \ = \ b \ (1 + \sum_{i=1}^{p} \phi_i)$$

Under the null hypothesis, $k_1 = -b \sum_{i=1}^{p} i\phi_i$ and $k_2 = 0$
Studentized statistic $\hat{\tau}_\tau$

9.4 UNIVARIATE AND MULTIVARIATE AUGMENTED DICKEY-FULLER REPRESENTATIONS

General $ARMA$ Representations
("Augmented Dickey-Fuller" (ADF))
Use one of the representations:

$$y_t \;=\; \rho_1 y_{t-1} \;+\; \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) \;+\; \varepsilon_t$$

$$y_t \;=\; \alpha \;+\; \rho_1 y_{t-1} \;+\; \sum_{j=2}^{k-1} \rho_j \;(y_{t-j+1} - y_{t-j}) \;+\; \varepsilon_t$$

$$y_t \;=\; k_1 \;+\; k_2 t \;+\; \rho_1 y_{t-1} \;+\; \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) \;+\; \varepsilon_t$$

Let $k \to \infty$ with $k/T \to 0$
"Trick Form" of ADF

$$(y_t - y_{t-1}) = \;(\rho_1 - 1) y_{t-1} \;+\; \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) \;+\; \varepsilon_t$$

- Unit root corresponds to $(\rho_1 - 1) = 0$
- Use standard automatically-computed t-stat
  (which of course does not have the t-distribution)

9.5 SPURIOUS REGRESSION

Multivariate Problem: Spurious Time-Series Regressions

Regress a persistent variable on an *unrelated* persistent variable:

$$y_t = \beta \; x_t + \varepsilon_t$$

(Canonical case: y, x independent driftless random walks)

$R^2 \;\xrightarrow{d}\; RV$ (not zero)

$\frac{t}{\sqrt{T}} \;\xrightarrow{d}\; RV$ (t diverges)

$\frac{\hat{\beta}}{\sqrt{T}} \;\xrightarrow{d}\; RV$ ($\hat{\beta}$ diverges)

When are I(1) Levels Regressions *Not* Spurious?
Answer: When the variables are cointegrated.

9.6 COINTEGRATION, ERROR-CORRECTION AND GRANGER'S REPRESENTATION THEOREM

Cointegration

Consider an N-dimensional variable x:
$x \;\sim\; CI\;(d,\, b)$ if

1. $x_i \sim I(d),\; i = 1, \ldots, N$
2. $\exists$ 1 or more linear combinations $z_t = \alpha' x_t$ s.t. $z_t \sim I(d-b), b > 0$

Leading Case

$x \sim CI(1,1)$ if
(1) $x_i \sim I(1),\ i = 1, \ldots, N$
(2) $\exists$ 1 or more linear combinations
$z_t = \alpha' x_t$ s.t. $z_t \sim I(0)$
Example

$$x_t = x_{t-1} + v_t,\ v_t \sim WN$$
$$y_t = x_{t-1} + \varepsilon_t,\ \varepsilon_t \sim\ WN,\ \varepsilon_t \perp v_{t-\tau},\ \forall t, \tau$$
$$\Rightarrow (y_t\ - x_t) = \varepsilon_t - v_t = I(0)$$

Cointegration and "Attractor Sets"
x_t is N-dimensional but does not wander randomly in $\mathbb{R}^N$
$\alpha' x_t$ is attracted to an $(N-R)$-dimensional subspace of $\mathbb{R}^N$
N: space dimension
R: number of cointegrating relationships
Attractor dimension $= N - R$
("number of underlying unit roots")
("number of common trends")
Example
3-dimensional $VAR(p)$, all variables $I(1)$
$R = 0 \Leftrightarrow$ no cointegration $\Leftrightarrow x$ wanders throughout $\mathbb{R}^3$
$R = 1 \Leftrightarrow$ 1 cointegrating vector $\Leftrightarrow x$ attracted to a 2-Dim hyperplane in $\mathbb{R}^3$ given by $\alpha' x = 0$
$R = 2 \Leftrightarrow$ 2 cointegrating vectors $\Leftrightarrow x$ attracted to a 1-Dim hyperplane (line) in $\mathbb{R}^3$ given by intersection of two 2-Dim hyperplanes, $\alpha_1' x = 0$ and $\alpha_2' x = 0$
$R = 3 \Leftrightarrow 3$ cointegrating vectors $\Leftrightarrow x$ attracted to a 0-Dim hyperplane (point) in $\mathbb{R}^3$ given by the intersection of three 2-Dim hyperplanes, $\alpha_1' x = 0$, $\alpha_2' x = 0$ and $\alpha_3' x = 0$
(Covariance stationary around $E(x)$)
Cointegration Motivation: Dynamic Factor Structure
Factor structure with $I(1)$ factors
$(N-R)$ $I(1)$ factors driving N variables
e.g., single-factor model:

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} f_t + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$
$$f_t = f_{t-1} + \eta_t$$

$R = (N-1)$ cointegrating combs: $(y_{2t} - y_{1t})$, ..., $(y_{Nt} - y_{1t})$
$(N-R) = N - (N-1) = 1$ common trend
Cointegration Motivation: Optimal Forecasting
$I(1)$ variables always co-integrated with their optimal forecasts
Example:
$x_t = x_{t-1} + \varepsilon_t$
$x_{t+h|t} = x_t$
$\Rightarrow x_{t+h} - x_{t+h|t} = \sum_{i=1}^{h} \varepsilon_{t+i}$
(finite MA, always covariance stationary)
Cointegration Motivation:
Long-Run Relation Augmented with Short-Run Dynamics

Simple AR Case (ECM):

$$\begin{aligned}\Delta y_t &= \alpha\ \Delta y_{t-1}\ +\ \beta\,\Delta x_{t-1}\ -\ \gamma(y_{t-1}\ -\ \delta x_{t-1})\ +\ u_t\\ &= \alpha\ \Delta y_{t-1}\ +\ \beta\,\Delta x_{t-1}\ -\ \gamma z_{t-1}\ +\ u_t\end{aligned}$$

General AR Case (VECM):
$A(L)\ \Delta x_t\ = -\gamma z_{t-1}\ +\ u_t$
where:
$A(L) = I - A_1 L - ... - A_p L^p$
$z_t = \alpha' x_t$
Multivariate ADF
Any VAR can be written as:

$$\Delta x_t\ \ =\ \ -\,\Pi x_{t-1}\ +\ \sum_{i=1}^{p-1} B_i \Delta x_{t-i}\ +\ u_t$$

Integration/Cointegration Status

- $Rank(\Pi) = 0$

  0 cointegrating vectors, N underlying unit roots
  (all variables appropriately specified in differences)

- $Rank(\Pi) = N$

  N cointegrating vectors, 0 unit roots
  (all variables appropriately specified in levels)

- $Rank(\Pi) = R \quad (0 < R < N)$

  R cointegrating vectors, $N - R$ unit roots

  New and important intermediate case
  (not possible in univariate)

Granger Representation Theorem

$$x_t\ \sim\ VECM\quad \Leftrightarrow\quad x_t\ \sim\ CI(1,1)$$

$VECM \Leftarrow$ Cointegration
We can always write
$\Delta x_t\ = \sum_{i=1}^{p-1}\ B_i \Delta x_{t-i}\ -\ \Pi\, x_{t-1}\ +\ u_t$
But under cointegration, $rank(\Pi) = R < N$, so

$$\underset{N\times N}{\Pi} = \underset{N\times R}{\gamma}\ \ \underset{R\times N}{\alpha'}$$

$\Rightarrow\ \Delta x_t\ = \sum_{i=1}^{p-1}\ B_i\ \Delta x_{t-i} - \gamma\alpha' x_{t-1} + u_t$
$= \sum_{i=1}^{p-1} B_i \Delta x_{t-i} - \gamma z_{t-1} + u_t$
$VECM \Rightarrow$ Cointegration

$$\Delta x_t\ \ =\ \ \sum_{i=1}^{p-1} B_i\ \Delta x_{t-i}\ \ -\ \ \gamma\ \alpha'\ x_{t-1}\ \ +\ \ u_t$$

Premultiply by α':

$$\alpha'\Delta x_t\ \ =\ \ \alpha'\ \sum_{i=1}^{p-1} B_i\ \ \Delta x_{t-i}\ \ -\ \ \underbrace{\alpha'\gamma}_{\text{full rank}}\ \ \alpha' x_{t-1}\ \ +\ \ \alpha'\ u_t$$

So equation balance requires that $\alpha' x_{t-1}$ be stationary.

Stationary-Nonstationary Decomposition

$$\underset{(N\times N)}{M'} \quad \underset{(N\times 1)}{x} = \begin{pmatrix} \alpha' \\ (R\times N) \\ \\ \delta \\ (N-R)\times N \end{pmatrix} x = \begin{pmatrix} CI\ combs \\ \\ com.\ trends \end{pmatrix}$$

(Rows of $\delta \perp$ to columns of γ)

Intuition Transforming the system by δ yields

$$\delta\ \Delta\ x_t\ =\ \sum_{i=1}^{p-1} \delta\ B_i\ \Delta\, x_{t-i}\ -\ \underbrace{\delta'\gamma}_{0 \text{ by orthogonality}}\ \alpha'\ x_{t-1}\ +\ \delta\ \mu_t$$

So δ isolates that part of the VECM
that is appropriately specified as a VAR in differences.

Note that if we *start* with M′x, then the observed series is
$(\mathrm{M}')^{-1}$ M′x, so nonstationarity is spread throughout the system.

Example

$$x_{1t}\ =\ x_{1t-1}\ +\ u_{1t}$$

$$x_{2t}\ =\ x_{1t-1}\ +\ u_{2t}$$

Levels form:

$$\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} L\right) \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

Dickey-Fuller form:

$$\begin{pmatrix} \Delta x_{1t} \\ \Delta x_{2t} \end{pmatrix} = - \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x_{1t-1} \\ x_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

Example, Continued

$$\Pi = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \end{pmatrix} = \gamma\alpha'$$

$$M'\ =\ \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}\ =\ \begin{pmatrix} \alpha' \\ \perp \gamma \end{pmatrix}$$

$$M' \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix}\ =\ \begin{pmatrix} u_{2t} - u_{1t} \\ x_{1t} \end{pmatrix}\ =\ \begin{pmatrix} x_{2t} - x_{1t} \\ x_{1t} \end{pmatrix}$$

9.7 FRACTIONAL INTEGRATION AND LONG MEMORY

Long Memory and Fractional Integration

"Integer-integrated" $ARIMA(p,d,q)$ $I(d)$:

$$(1-L)^d\Phi(L)y_t = \Theta(L)\varepsilon_t,\ d=0,1,...$$

Covariance stationary: $ARIMA(p,0,q)$

Random walk: $ARIMA(0,1,0)$ "pure unit root process"

"Fractionally-integrated" $ARFIMA(p,d,q)$ $I(d)$:

$$(1-L)^d\Phi(L)y_t = \Theta(L)\varepsilon_t,\ -\frac{1}{2} < d < \frac{1}{2}$$

Covariance stationary: $-\frac{1}{2} < d < \frac{1}{2}$ (focus on $0 < d < \frac{1}{2}$)

$ARFIMA(0,d,0)$: "pure fractionally-integrated process"

$$(1-L)^d = 1 - dL + \frac{d(d-1)}{2!}L^2 - \frac{d(d-1)(d-2)}{3!}L^3 + ...$$

Long Memory and Fractional Integration, Continued Time domain, $\tau \to \infty$

- $I(1):\ \rho(\tau) \propto const$
- $I(0):\ \rho(\tau) \propto r^\tau\ (0 < r < 1)$
- $I(d):\ \rho(\tau) \propto \tau^{2d-1}\ (0 < d < 1/2)$

9.7.1 Characterizing Integration Status

Frequency domain, $\omega \to 0$

- $I(1):\ f(\omega) \propto \omega^{-2}$
- $I(0):\ f(\omega) \propto const$
- $I(d):\ f(\omega) \propto \omega^{-2d}\ (0 < d < 1/2)$

Frequency-domain $I(d)$ behavior implies that for low frequencies,

$$ln\, f^*(\omega) = \beta_0 + \underbrace{\beta_1}_{-2d}\, ln\omega + \varepsilon_t$$

GPH estimator of d: Regress $ln\, f^*(\omega) \to const,\ ln\, \omega$

So take $\hat{d} = -\frac{1}{2}\hat{\beta}_1$. "GPH estimator"

9.8 EXERCISES, PROBLEMS AND COMPLEMENTS

1. Applied modeling.

Obtain a series of U.S. industrial production, monthly, 1947.01-present, not seasonally adjusted. Discard the last 20 observations. Now graph the series. Examine its trend and seasonal patterns in detail. Does a linear trend (fit to logs) seem appropriate? Do monthly seasonal dummies seem appropriate? To the log of the series, fit an ARMA model with linear trend and monthly seasonal dummies. Be sure to try a variety of the techniques we have covered (sample autocorrelation and partial autocorrelation functions, Bartlett standard errors, Box-Pierce test, standard error of the regression, adjusted R^2, etc.) in your attempt at selecting an adequate model. Once a model has been selected and estimated, use it to forecast the last 20 observations, and compare your forecast to the actual realized values. Contrast the results of the approach with those of the Box-Jenkins seasonal ARIMA approach, which involves taking first and seasonal differences as necessary to induce covariance stationarity, and then fitting a multiplicative seasonal ARMA model.

2. Aggregation.

   Granger (1980) shows that aggregation of a very large number of stationary ARMA time series results, under regularity conditions (generalized in Robinson, 1991), in a fractionally-integrated process. Thus, aggregation of short-memory processes results in a long-memory process. Discuss this result in light of theorems on aggregation of ARMA processes. In particular, recall that aggration of ARMA processes results in new ARMA processes, generally of higher order than the components.

3. Over-differencing and UCM's.

   The stochastic trend:

$$\begin{aligned} T_t &= T_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t, \end{aligned}$$

   has two unit roots. Discuss as regards "overdifferencing" in economic time series.

9.9 NOTES

Chapter Ten

Volatility Dynamics

10.1 VOLATILITY AND FINANCIAL ECONOMETRICS

10.2 GARCH

10.3 STOCHASTIC VOLATILITY

10.4 OBSERVATION-DRIVEN VS. PARAMETER-DRIVEN PROCESSES

Prologue: Reading

Much of what follows draws heavily upon:

- Andersen, T.G., Bollerslev, T., Christoffersen, P.F. and Diebold, F.X. (2012), "Financial Risk Measurement for Financial Risk Management," in G. Constantinedes, M. Harris and Rene Stulz (eds.), *Handbook of the Economics of Finance*, Elsevier.
- Andersen, T.G., Bollerslev, T. and Diebold, F.X. (2010), "Parametric and Nonparametric Volatility Measurement," in L.P. Hansen and Y. Ait-Sahalia (eds.), *Handbook of Financial Econometrics*. Amsterdam: North-Holland, 67-138.
- Andersen, T.G., Bollerslev, T., Christoffersen, P.F., and Diebold, F.X. (2006), "Volatility and Correlation Forecasting," in G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*. Amsterdam: North-Holland, 778-878.

Prologue

- Throughout: Desirability of *conditional* risk measurement
- Aggregation level
  - Portfolio-level (aggregated, univariate) Risk measurement
  - Asset-level (disaggregated, multivariate): Risk management
- Frequency of data observations
  - Low-frequency vs. high-frequency data
  - Parametric vs. nonparametric volatility measurement

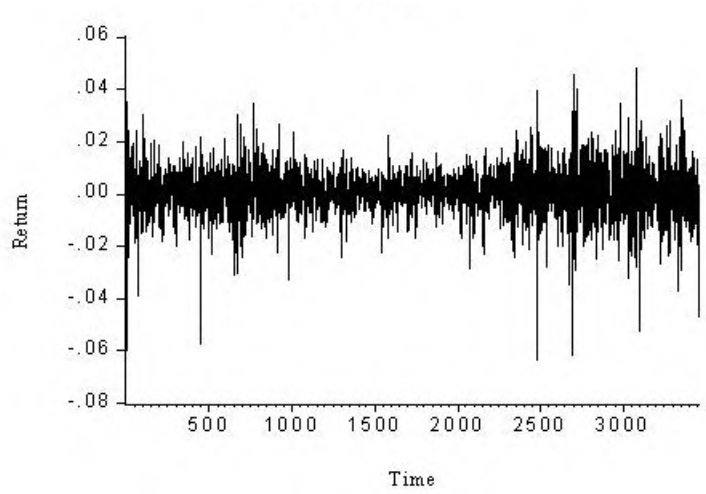


Figure 10.1: Time Series of Daily NYSE Returns

- Object measured and modeled
  - From conditional variances to conditional densities
- Dimensionality reduction in "big data" multivariate environments
  - From ad hoc statistical restrictions to factor structure

What's in the Data?

Returns

Key Fact 1: Returns are Approximately Serially Uncorrelated

Key Fact 2: Returns are not Gaussian

Key Fact 3: Returns are Conditionally Heteroskedastic I

Key Fact 3: Returns are Conditionally Heteroskedastic II

Why Care About Volatility Dynamics?

Everything Changes when Volatility is Dynamic

- Risk management
- Portfolio allocation
- Asset pricing
- Hedging
- Trading

Risk Management

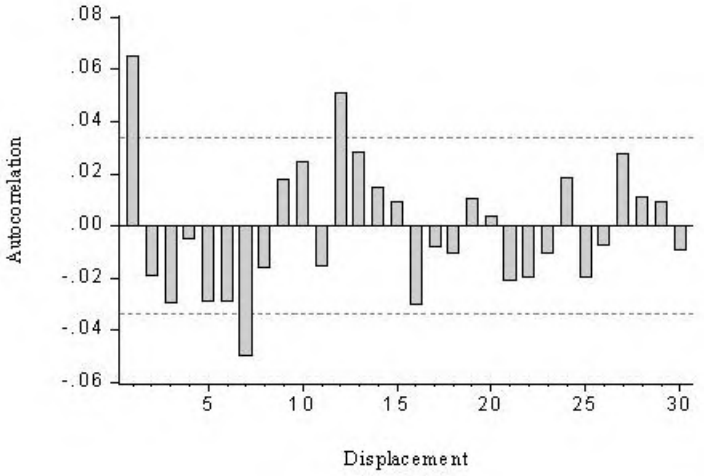


Figure 10.2: Correlogram of Daily NYSE Returns.

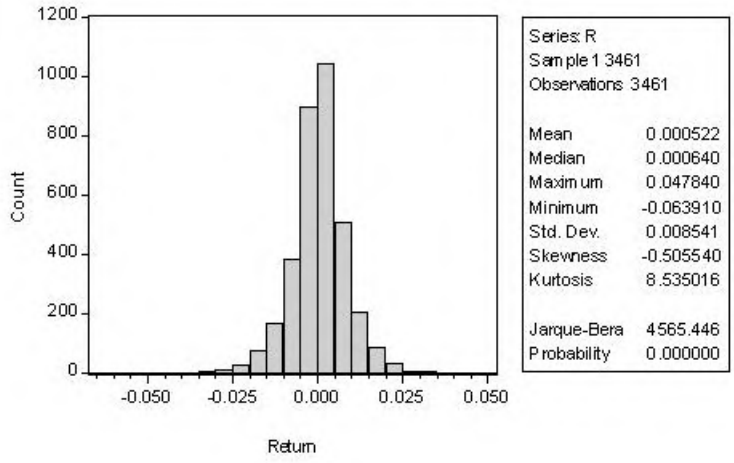


Figure 10.3: Histogram and Statistics for Daily NYSE Returns.

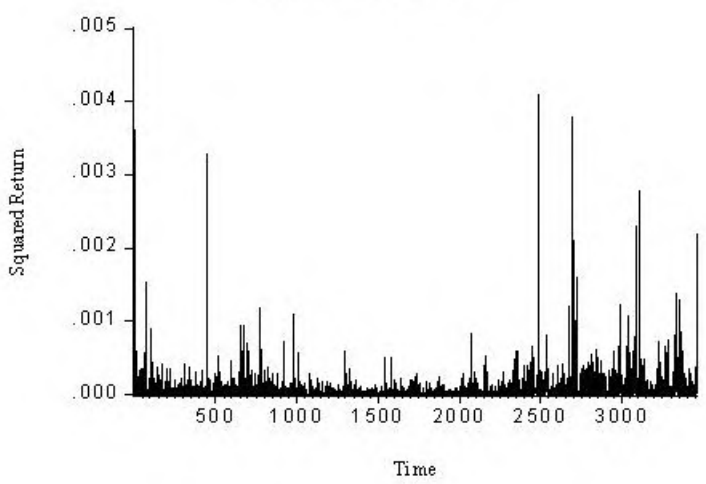


Figure 10.4: Time Series of Daily Squared NYSE Returns.

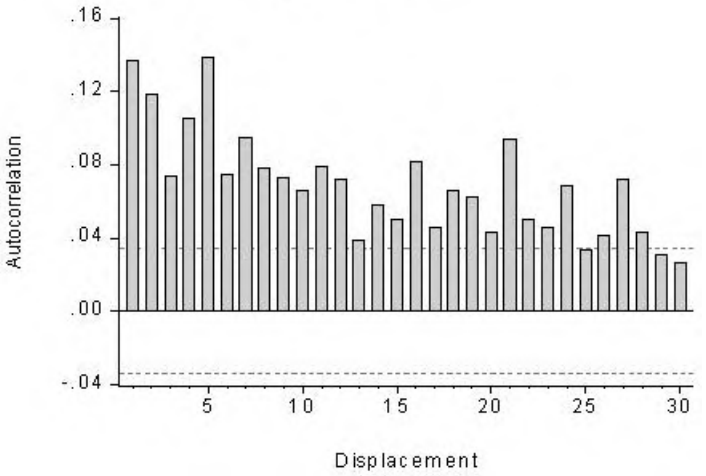


Figure 10.5: Correlogram of Daily Squared NYSE Returns.

Individual asset returns:

$$r \ \sim \ (\mu, \ \Sigma)$$

Portfolio returns:

$$r_p \ = \ \lambda' r \ \sim \ (\lambda'\mu, \ \lambda'\Sigma\lambda)$$

If Σ varies, we need to track time-varying portfolio risk, $\lambda'\Sigma_t\lambda$
Portfolio Allocation
Optimal portfolio shares w^* solve:

$$min_w \ w'\Sigma w$$

$$s.t. \ w'\mu = \mu_p$$

Importantly, $w^* = f(\Sigma)$
If Σ varies, we have $w_t^* = f(\Sigma_t)$
Asset Pricing I: Sharpe Ratios
Standard Sharpe:

$$\frac{E(r_{it} - r_{ft})}{\sigma}$$

Conditional Sharpe:

$$\frac{E(r_{it} - r_{ft})}{\sigma_t}$$

Asset Pricing II: CAPM Standard CAPM:

$$(r_{it} - r_{ft}) = \alpha + \beta(r_{mt} - r_{ft})$$

$$\beta \ = \ \frac{cov((r_{it} - r_{ft}), \ (r_{mt} - r_{ft}))}{var(r_{mt} - r_{ft})}$$

Conditional CAPM:

$$\beta_t \ = \ \frac{cov_t((r_{it} - r_{ft}), \ (r_{mt} - r_{ft}))}{var_t(r_{mt} - r_{ft})}$$

Asset Pricing III: Derivatives
Black-Scholes:

$$C = N(d_1)S - N(d_2)Ke^{-r\tau}$$

$$d_1 = \frac{ln(S/K) + (r + \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$$

$$d_2 = \frac{ln(S/K) + (r - \sigma^2/2)\tau}{\sigma\sqrt{\tau}}$$

$$P_C \;=\; BS(\sigma,\; ...)$$

(Standard Black-Scholes options pricing)

Completely different when σ varies!

Hedging

- Standard delta hedging

$$\Delta H_t \;=\; \delta\, \Delta S_t \;+\; u_t$$

$$\delta \;=\; \frac{cov(\Delta H_t,\; \Delta S_t)}{var(\Delta S_t)}$$

- Dynamic hedging

$$\Delta H_t \;=\; \delta_t\, \Delta S_t \;+\; u_t$$

$$\delta_t \;=\; \frac{cov_t(\Delta H_t,\; \Delta S_t)}{var_t(\Delta S_t)}$$

Trading

- Standard case: no way to trade on fixed volatility
- Time-varying volatility I: Options straddles, strangles, etc. Take position according to whether $P_C >< f(\sigma_{t+h,t}, \ldots)$
  (indirect)
- Time-varying volatility II: Volatility swaps

  Effectively futures contracts written on underlying
  "realized volatility"

  (direct)

Some Warm-Up

Unconditional Volatility Measures

Variance: $\sigma^2 \;=\; E(r_t - \mu)^2$ (or standard deviation: σ)

Mean Absolute Deviation: $MAD \;=\; E|r_t - \mu|$

Interquartile Range: $IQR \;=\; 75\% \;-\; 25\%$

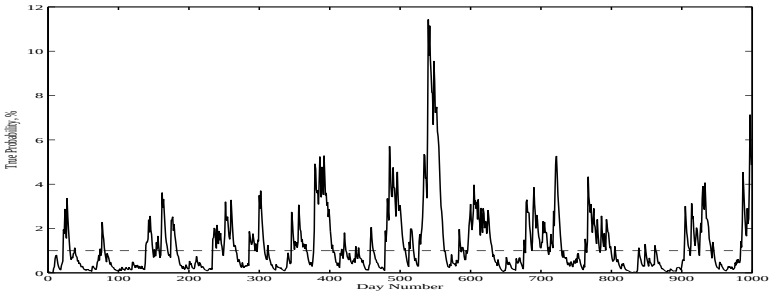


Figure 10.6: True Exceedance Probabilities of Nominal 1% HS-VaR When Volatility is Persistent. We simulate returns from a realistically-calibrated dynamic volatility model, after which we compute 1-day 1% HS-VaR using a rolling window of 500 observations. We plot the daily series of true conditional exceedance probabilities, which we infer from the model. For visual reference we include a horizontal line at the desired 1% probability level.

p% Value at Risk (VaR^p)): x s.t. $P(r_t < x) = p$

Outlier probability: $P|r_t - \mu| > 5\sigma$ (for example)

Tail index: γ s.t. $P(r_t > r) = k\ r^{-\gamma}$

Kurtosis: $K = E(r-\mu)^4/\sigma^4$

Dangers of a Largely Unconditional Perspective (HS-VaR)

Dangers of an Unconditional Perspective, Take II

The unconditional HS-VaR perspective encourages incorrect rules of thumb, like scaling by $\sqrt{h}$ to convert 1-day vol into h-day vol.

Conditional VaR

Conditional VaR ($VaR^p_{T+1|T}$) solves:

$$p = P_T(r_{T+1} \leq -VaR^p_{T+1|T}) = \int_{-\infty}^{-VaR^p_{T+1|T}} f_T(r_{T+1}) dr_{T+1}$$

($f_T(r_{T+1})$ is density of r_{T+1} conditional on time-T information)

But VaR of any Flavor has Issues

- VaR is silent regarding expected loss when VaR is exceeded (fails to assess the entire distributional tail)
- VaR fails to capture beneficial effects of portfolio diversification

Conditionally expected shortfall:

$$ES^p_{T+1|T} = p^{-1} \int_0^p VaR^\gamma_{T+1|T}\, d\gamma$$

- ES assesses the entire distributional tail
- ES captures the beneficial effects of portfolio diversification

Exponential Smoothing and RiskMetrics

$$\sigma_t^2 \;=\; \lambda\,\sigma_{t-1}^2 \;+\; (1-\lambda)\,r_{t-1}^2$$

$$\sigma_t^2 \;=\; \sum_{j=0}^{\infty} \varphi_j\, r_{t-1-j}^2$$

$$\varphi_j = (1-\lambda)\,\lambda^j$$

(Many initializations possible: r_1^2, sample variance, etc.)

$$\text{RM-VaR}_{T+1|T}^{p} = \sigma_{T+1}\Phi_p^{-1}$$

- Random walk for variance
- Random walk plus noise model for squared returns
- Volatility forecast at any horizon is current smoothed value
- But flat volatility term structure is not realistic

Rigorous Modeling I

Conditional Univariate Volatility Dynamics from "Daily" Data

Conditional Return Distributions

$f(r_t)$ *vs.* $f(r_t|\Omega_{t-1})$

Key 1: $E(r_t|\Omega_{t-1})$

Are returns conditional mean independent? Arguably yes.

Returns are (arguably) approximately serially uncorrelated, and (arguably) approximately free of additional non-linear conditional mean dependence.

Conditional Return Distributions, Continued Key 2: $var(r_t|\Omega_{t-1}) \;=\; E((r_t-\mu)^2|\Omega_{t-1})$

Are returns conditional variance independent? No way!

Squared returns serially correlated, often with very slow decay.

The Standard Model

(Linearly Indeterministic Process with iid Innovations)

$$y_t \;=\; \sum_{i=0}^{\infty} b_i\;\varepsilon_{t-i}$$

$$\varepsilon \sim iid\ (0,\ \sigma_\varepsilon^2) \quad \sum_{i=0}^{\infty} b_i^2 < \infty \quad b_0 = 1$$

Uncond. mean: $E(y_t) = 0$ (constant)

Uncond. variance: $E(y_t - E(y_t))^2 = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} b_i^2$ (constant)

Cond. mean: $E(y_t \mid \Omega_{t-1}) = \sum_{i=1}^{\infty} b_i\ \varepsilon_{t-i}$ (varies)

Cond. variance: $E([y_t - E(y_t \mid \Omega_{t-1})]^2 \mid \Omega_{t-1}) = \sigma_\varepsilon^2$ (constant)

The Standard Model, Continued

k-Step-Ahead Least Squares Forecasting

$$E(y_{t+k} \mid \Omega_t) = \sum_{i=0}^{\infty} b_{k+i}\ \varepsilon_{t-i}$$

Associated prediction error:

$$y_{t+k} - E(y_{t+k} \mid \Omega_t) = \sum_{i=0}^{k-1} b_i\ \varepsilon_{t+k-i}$$

Conditional prediction error variance:

$$E([y_{t+k} - E(y_{t+k} \mid \Omega_t)]^2 \mid \Omega_t) = \sigma_\varepsilon^2 \sum_{i=0}^{k-1} b_i^2$$

Key: Depends only on k, not on Ω_t

ARCH(1) Process

$$r_t|\Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha r_{t-1}^2$$

$$E(r_t) = 0$$

$$E(r_t - E(r_t))^2 = \frac{\omega}{(1-\alpha)}$$

$$E(r_t|\Omega_{t-1}) = 0$$

$$E([r_t - E(r_t|\Omega_{t-1})]^2|\Omega_{t-1}) = \omega + \alpha r_{t-1}^2$$

GARCH(1,1) Process

"Generalized ARCH"

$$r_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$$

$$E(r_t) = 0$$

$$E(r_t - E(r_t))^2 = \frac{\omega}{(1-\alpha-\beta)}$$

$$E(r_t|\Omega_{t-1}) = 0$$

$$E([r_t - E(r_t \mid \Omega_{t-1})]^2 \mid \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$$

Conditionally-Gaussian GARCH-Based 1-Day VaR

$$\text{GARCH-VaR}_{T+1|T}^{p} \equiv \sigma_{T+1|T}\,\Phi_p^{-1}$$

– Consistent with fat tails of unconditional return distribution
– Can be extended to allow for fat-tailed conditional distribution
Unified Theoretical Framework

- Volatility dynamics (of course, by construction)
- Conditional symmetry translates into unconditional symmetry
- Volatility clustering produces unconditional leptokurtosis

Tractable Empirical Framework

$$L(\theta; r_1, \ldots, r_T) \approx f(r_T|\Omega_{T-1}; \theta)\, f(r_{T-1}|\Omega_{T-2}; \theta) \ldots f(r_{p+1}|\Omega_p; \theta)$$

If the conditional densities are Gaussian,

$$f(r_t|\Omega_{t-1}; \theta) = \frac{1}{\sqrt{2\pi}} h_t(\theta)^{-1/2} \exp\left(-\frac{1}{2}\frac{r_t^2}{h_t(\theta)}\right)$$

$$\ln L(\theta; r_{p+1}, \ldots, r_T) \approx -\frac{T-p}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=p+1}^{T} \ln h_t(\theta) - \frac{1}{2}\sum_{t=p+1}^{T} \frac{r_t^2}{h_t(\theta)}$$

The Squared Return as a Noisy Volatility Proxy
Note that we can write:
$r_t^2 = h_t + \nu_t$
Thus r_t^2 is a *noisy* indicator of h_t
Various approaches handle the noise in various ways.

GARCH(1,1) and Exponential Smoothing

Exponential smoothing recursion:

$$\bar{r}_t^2 \;=\; \gamma\, r_t^2 \;+\; (1-\gamma)\,\bar{r}_{t-1}^2$$

Back substitution yields:

$$\bar{r}_t^2 \;=\; \sum w_j \; r_{t-j}^2$$

where $w_j \;=\; \gamma\,(1-\gamma)^j$

But in GARCH(1,1) we have:

$$h_t \;=\; \omega \;+\; \alpha r_{t-1}^2 \;+\; \beta h_{t-1}$$

$$h_t \;=\; \frac{\omega}{1-\beta} \;+\; \alpha \sum \beta^{j-1} r_{t-j}^2$$

Variance Targeting

Sample unconditional variance:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} r_t^2$$

Implied unconditional GARCH(1,1) variance:

$$\sigma^2 = \frac{\omega}{1-\alpha-\beta}$$

We can constrain $\sigma^2 = \hat{\sigma}^2$ by constraining:

$$\omega = (1-\alpha-\beta)\hat{\sigma}^2$$

– Saves a degree of freedom and ensures reasonableness

ARMA Representation in Squares

r_t^2 has the ARMA(1,1) representation:

$$r_t^2 = \omega + (\alpha+\beta) r_{t-1}^2 - \beta\nu_{t-1} + \nu_t,$$

where $\nu_t = r_t^2 - h_t$.

Variations on the GARCH Theme

Regression with GARCH Disturbances

$$y_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t|\Omega_{t-1} \;\sim\; N(0, h_t)$$

- Regression with GARCH Disturbances
- Incorporating Exogenous Variables
- Asymmetric Response and the Leverage Effect:
- Fat-Tailed Conditional Densities
- Time-Varying Risk Premia

Incorporating Exogenous Variables

$$h_t \;=\; \omega \;+\; \alpha\, r_{t-1}^2 \;+\; \beta\, h_{t-1} \;+\; \gamma' z_t$$

γ is a parameter vector
z is a set of positive exogenous variables.
Asymmetric Response and the Leverage Effect I: TARCH
Standard GARCH: $h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$
TARCH: $h_t = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 D_{t-1} + \beta h_{t-1}$
$D_t = \begin{cases} 1 \; if \; r_t < 0 \\ 0 \; otherwise \end{cases}$
positive return (good news): α effect on volatility

negative return (bad news): $\alpha + \gamma$ effect on volatility

$\gamma \neq 0$: Asymetric news response
$\gamma > 0$: "Leverage effect"
Asymmetric Response II: E-GARCH

$$\ln(h_t) \;=\; \omega \;+\; \alpha \left| \frac{r_{t-1}}{h_{t-1}^{1/2}} \right| \;+\; \gamma \frac{r_{t-1}}{h_{t-1}^{1/2}} \;+\; \beta \ln(h_{t-1})$$

- Log specification ensures that the conditional variance is positive.
- Volatility driven by both size and sign of shocks
- Leverage effect when $\gamma < 0$

Fat-Tailed Conditional Densities: t-GARCH
If r is conditionally Gaussian, then $\frac{r_t}{\sqrt{h_t}} \;\sim\; N(0,1)$
But often with high-frequency data, $\frac{r_t}{\sqrt{h_t}} \;\sim\; fat\ tailed$
So take:
$r_t \;=\; h_t^{1/2}\, z_t$

Dependent Variable: R
Method: ML - ARCH (Marquardt)

Sample: 1 3461
Included observations: 3461
Convergence achieved after 19 iterations
Variance backcast: ON

Coefficient	Std. Error	z-Statistic	Prob.	
C	0.000640	0.000127	5.036942	0.0000
Variance Equation				
C	1.06E-06	1.49E-07	7.136840	0.0000
ARCH(1)	0.067410	0.004955	13.60315	0.0000
GARCH(1)	0.919714	0.006122	150.2195	0.0000

Figure 10.7: GARCH(1,1) Estimation, Daily NYSE Returns.

$$z_t \overset{iid}{\sim} \frac{t_d}{std(t_d)}$$

Time-Varying Risk Premia: GARCH-M
Standard GARCH regression model:

$$y_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t|\Omega_{t-1} \sim N(0, h_t)$$

GARCH-M model is a special case:

$$y_t = x_t'\beta + \gamma h_t + \varepsilon_t$$

$$\varepsilon_t|\Omega_{t-1} \sim N(0, h_t)$$

A GARCH(1,1) Example
A GARCH(1,1) Example
A GARCH(1,1) Example
A GARCH(1,1) Example
After Exploring Lots of Possible Extensions...

Rigorous Modeling II

Conditional Univariate Volatility Dynamics from High-Frequency Data

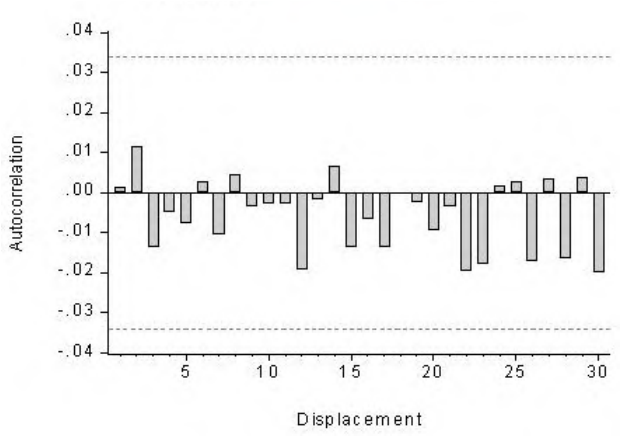


Figure 10.8: Correlogram of Squared Standardized GARCH(1,1) Residuals, Daily NYSE Returns.

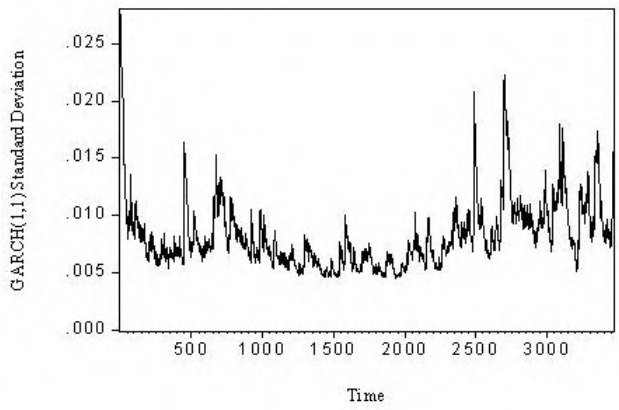


Figure 10.9: Estimated Conditional Standard Deviation, Daily NYSE Returns.

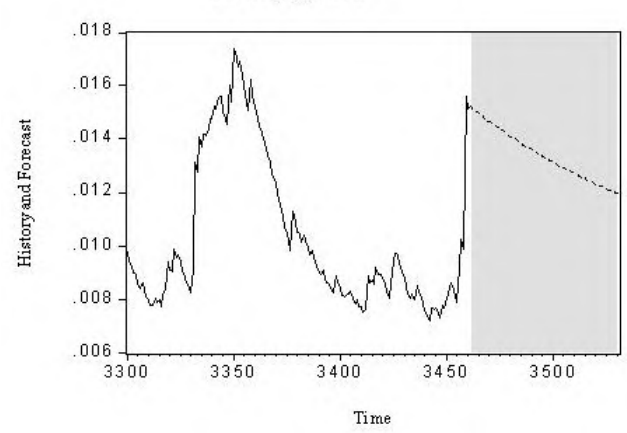


Figure 10.10: Conditional Standard Deviation, History and Forecast, Daily NYSE Returns.

Dependent Variable: R
Method: ML - ARCH (Marquardt) - Student's t distribution
Date: 04/10/12 Time: 13:48
Sample (adjusted): 2 3461
Included observations: 3460 after adjustments
Convergence achieved after 19 iterations
Presample variance: backcast (parameter = 0.7)
GARCH = C(4) + C(5)*RESID(-1)^2 + C(6)*RESID(-1)^2*(RESID(-1)<0)
+ C(7)*GARCH(-1)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
@SQRT(GARCH)	0.083360	0.053138	1.568753	0.1167
C	1.28E-05	0.000372	0.034443	0.9725
R(-1)	0.073763	0.017611	4.188535	0.0000
Variance Equation				
C	1.03E-06	2.23E-07	4.628790	0.0000
RESID(-1)^2	0.014945	0.009765	1.530473	0.1259
RESID(-1)^2*(RESID(-1)<0)	0.094014	0.014945	6.290700	0.0000
GARCH(-1)	0.922745	0.009129	101.0741	0.0000
T-DIST. DOF	5.531579	0.478432	11.56188	0.0000

Figure 10.11: AR(1) Returns with Threshold t-GARCH(1,1)-in Mean.

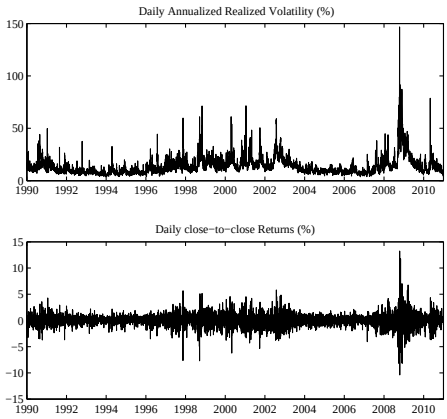


Figure 10.12: S&P500 Daily Returns and Volatilities (Percent). The top panel shows daily S&P500 returns, and the bottom panel shows daily S&P500 realized volatility. We compute realized volatility as the square root of $AvgRV$, where $AvgRV$ is the average of five daily RVs each computed from 5-minute squared returns on a 1-minute grid of S&P500 futures prices.

Intraday Data and Realized Volatility

$$dp(t) = \mu(t)dt + \sigma(t)dW(t)$$

$$RV_t(\Delta) \equiv \sum_{j=1}^{N(\Delta)} \left(p_{t-1+j\Delta} - p_{t-1+(j-1)\Delta}\right)^2$$

$$RV_t(\Delta) \;\rightarrow\; IV_t = \int_{t-1}^{t} \sigma^2(\tau)\, d\tau$$

Microstructure Noise

– State space signal extraction

– AvgRV

– Realized kernel

– Many others

RV is Persistent

RV is Reasonably Approximated as Log-Normal

RV is Long-Memory

Exact and Approximate Long Memory

Exact long memory:

$$(1-L)^d RV_t = \beta_0 + \nu_t$$

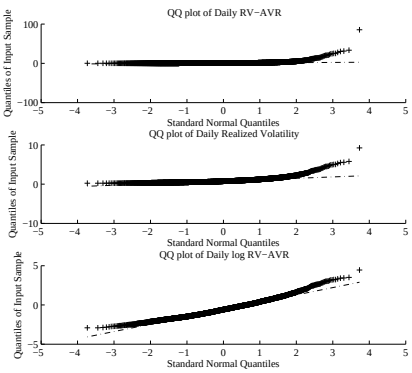


Figure 10.13: S&P500: QQ Plots for Realized Volatility and Log Realized Volatility. The top panel plots the quantiles of daily realized volatility against the corresponding normal quantiles. The bottom panel plots the quantiles of the natural logarithm of daily realized volatility against the corresponding normal quantiles. We compute realized volatility as the square root of $AvgRV$, where $AvgRV$ is the average of five daily RVs each computed from 5-minute squared returns on a 1-minute grid of S&P500 futures prices.

"Corsi model" (HAR):

$$RV_t \;=\; \beta_0 \;+\; \beta_1\, RV_{t-1} \;+\; \beta_2\, RV_{t-5:t-1} \;+\; \beta_3 RV_{t-21:t-1} \;+\; \nu_t$$

Even better:

$$\log RV_t \;=\; \beta_0 + \beta_1 \log RV_{t-1} + \beta_2 \log RV_{t-5:t-1} + \beta_3 \log RV_{t-21:t-1} + \nu_t$$

– Ensures positivity and promotes normality

RV-VaR

$$RV - VaR^p_{T+1|T} \;=\; \widehat{RV}_{T+1|T}\, \Phi_p^{-1},$$

GARCH-RV

$$\sigma_t^2 \;=\; \omega \;+\; \beta\, \sigma_{t-1}^2 \;+\; \gamma\, RV_{t-1}$$

- Fine for 1-step
- Multi-step requires "closing the system" with an RV equation
  - "Realized GARCH"
  - "HEAVY"

Separating Jumps

$$QV_t \;=\; IV_t \;+\; JV_t$$

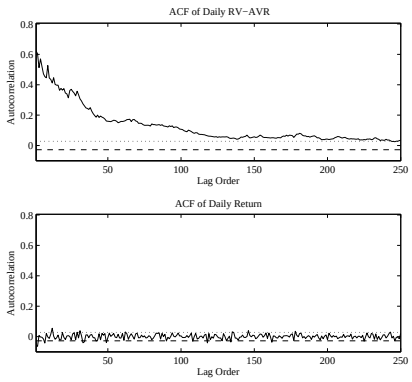


Figure 10.14: S&P500: Sample Autocorrelations of Daily Realized Variance and Daily Return. The top panel shows realized variance autocorrelations, and the bottom panel shows return autocorrelations, for displacements from 1 through 250 days. Horizontal lines denote 95% Bartlett bands. Realized variance is $AvgRV$, the average of five daily RVs each computed from 5-minute squared returns on a 1-minute grid of S&P500 futures prices.

where

$$JV_t \;=\; \sum_{j=1}^{\mathcal{J}_t} J_{t,j}^2$$

e.g., we might want to explore:

$$\begin{aligned} RV_t \;&= \beta_0 \,+\, \beta_1\, IV_{t-1} \,+\, \beta_2\, IV_{t-5:t-1} \,+\, \beta_3\, IV_{t-21:t-1} \\ &+\; \alpha_1\, JV_{t-1} \,+\, \alpha_2\, JV_{t-5:t-1} \,+\, \alpha_3\, JV_{t-21:t-1} \,+\, \nu_t \end{aligned}$$

But How to Separate Jumps?

- Truncation:

$$TV_t\,(\Delta) \;=\; \sum_{j=1}^{N(\Delta)} \Delta p_{t-1+j\Delta}^2\, I\left(\,\Delta p_{t-1+j\Delta} \,<\, \mathcal{T}\,\right)$$

- Bi-Power Variation:

$$BPV_t\,(\Delta) \;=\; \frac{\pi}{2}\,\frac{N\,(\Delta)}{N\,(\Delta)-1} \sum_{j=1}^{N(\Delta)-1} |\Delta p_{t-1+j\Delta}|\;|\Delta p_{t-1+(j+1)\Delta}|$$

- Minimum:

$$MinRV_t\,(\Delta) \;=\; \frac{\pi}{\pi-2}\left(\frac{N\,(\Delta)}{N\,(\Delta)-1}\right) \sum_{j=1}^{N(\Delta)-1} \min\left\{|\Delta p_{t-1+j\Delta}|, |\Delta p_{t-1+(j+1)\Delta}|\right\}^2$$

Rigorous Modeling III

Conditional Asset-Level (Multivariate) Volatility Dynam-

ics from "Daily" Data

Multivariate

Univariate volatility models useful for portfolio-level risk measurement (VaR, ES, etc.)

But what about risk *management* questions:

- Portfolio risk change under a certain scenario involving price movements of set of assets or asset classes?
- Portfolio risk change if certain correlations increase suddenly
- Portfolio risk change if I double my holdings of Intel?
- How do optimal portfolio shares change if the covariance matrix moves in a certain way?

Similarly, what about almost any other question in asset pricing, hedging, trading? Almost all involve correlation.

Basic Framework and Issues I

$N \times 1$ return vector R_t

$N \times N$ covariance matrix Ω_t

- $\frac{N(N+1)}{2}$ distinct elements
- Structure needed for pd or even psd
- Huge number of parameters even for moderate N
- And N may be *not* be moderate!

Basic Framework and Issues II

Univariate:

$$r_t = \sigma_t z_t$$

$$z_t \sim i.i.d.(0,1)$$

Multivariate:

$$R_t = \Omega_t^{1/2} Z_t$$

$$Z_t \sim i.i.d.(0,\mathcal{I})$$

where $\Omega_t^{1/2}$ is a "square-root" (e.g., Cholesky factor) of Ω_t

Ad Hoc Exponential Smoothing (RM)

$$\Omega_t \;=\; \lambda\,\Omega_{t-1} \;+\; (1-\lambda)\,R_{t-1}R_{t-1}'$$

- Assumes that the dynamics of all the variances and covariances are driven by a single scalar parameter λ (identical smoothness)
- Guarantees that the smoothed covariance matrices are pd so long as Ω_0 is pd
- Common strategy is to set Ω_0 equal to the sample covariance matrix $\frac{1}{T}\sum_{t=1}^{T} R_t R_t'$ (which is pd if $T > N$)
- But covariance matrix forecasts inherit the implausible scaling properties of the univariate RM forecasts and will in general be suboptimal

Multivariate GARCH(1,1)

$$vech\,(\Omega_t) \;=\; vech\,(C) \;+\; B\,vech\,(\Omega_{t-1}) \;+\; A\,vech\,(R_{t-1}R_{t-1}')$$

- *vech* operator converts the upper triangle of a symmetric matrix into a $\frac{1}{2}N(N+1)\times 1$ column vector
- A and B matrices are both of dimension $\frac{1}{2}N(N+1)\times\frac{1}{2}N(N+1)$
- Even in this "parsimonious" GARCH(1,1) there are $O(N^4)$ parameters
  - More than 50 *million* parameters for $N = 100$!

Encouraging Parsimony: Diagonal GARCH(1,1)
Diagonal GARCH constrains A and B matrices to be diagonal.

$$vech\,(\Omega_t) \;=\; vech\,(C) \;+\; (I\beta)\,vech\,(\Omega_{t-1}) \;+\; (I\alpha)\,vech\,(R_{t-1}R_{t-1}')$$

– Still $O(N^2)$ parameters.
Encouraging Parsimony: Scalar GARCH(1,1)
Scalar GARCH constrains A and B matrices to be scalar:

$$vech\,(\Omega_t) \;=\; vech\,(C) \;+\; (I\beta)\,vech\,(\Omega_{t-1}) \;+\; (I\alpha)\,vech\,(R_{t-1}R_{t-1}')$$

– Mirrors RM, but with the important difference that the Ω_t forecasts now revert to $\Omega = (1-\alpha-\beta)^{-1}C$
– Fewer parameters than diagonal, but still $O(N)^2$
(because of C)
Encouraging Parsimony: Covariance Targeting
Recall variance targeting:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} r_t^2, \quad \sigma^2 = \frac{\omega}{1-\alpha-\beta} \quad \Longrightarrow \quad take\ \omega = (1-\alpha-\beta)\hat{\sigma}^2$$

Covariance targeting is the obvious multivariate generalization:

$$vech(C) \;=\; (\mathcal{I} - A - B)\; vech\,(\,\frac{1}{T}\,\sum_{t=1}^{T} R_t R_t'\,)$$

– Encourages both parsimony and reasonableness

Constant Conditional Correlation (CCC) Model

[Key is to recognize that correlation matrix
is the covariance matrix of standardized returns]

Two-step estimation:

- Estimate N appropriate univariate GARCH models
- Calculate standardized return vector, $\hat{e}_t = R_t \hat{D}_t^{-1}$
- Estimate correlation matrix Γ (assumed constant) as $\frac{1}{T}\sum_{t=1}^{T} \hat{e}_t \hat{e}_t'$

– Quite flexible as the N models can differ across returns

Dynamic Conditional Correlation (DCC) Model

Two-step estimation:

- Estimate N appropriate univariate GARCH models
- Calculate standardized return vector, $\hat{e}_t = R_t \hat{D}_t^{-1}$
- Estimate correlation matrix Γ_t (assumed to have scalar GARCH(1,1)-style dynamics) as following

$$vech(\Gamma_t) = vech(C) + (I\beta)vech(\Gamma_{t-1}) + (I\alpha)vech(e_{t-1}e_{t-1}')$$

  – "Correlation targeting" is helpful

DECO

- Time-varying correlations assumed identical across all pairs of assets, which implies:

$$\Gamma_t \;=\; (1-\rho_t)\,\mathcal{I} \;+\; \rho_t\,\mathcal{J},$$

  where $\mathcal{J}$ is an $N \times N$ matrix of ones

- Analytical inverse facilitates estimation:

$$\Gamma_t^{-1} \;=\; \frac{1}{(1-\rho_t)}\,\left[\mathcal{I} \;-\; \frac{\rho_t}{1+(N-1)\rho_t}\,\mathcal{J}\right]$$

- Assume GARCH(1,1)-style conditional correlation structure:

$$\rho_t \;=\; \omega_\rho \;+\; \alpha_\rho\, u_t \;+\; \beta_\rho\, \rho_{t-1}$$

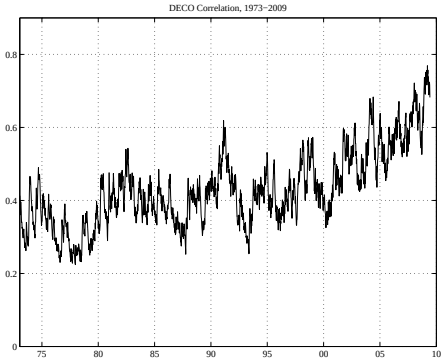


Figure 10.15: Time-Varying International Equity Correlations. The figure shows the estimated equicorrelations from a DECO model for the aggregate equity index returns for 16 different developed markets from 1973 through 2009.

- Updating rule is naturally given by the average conditional correlation of the standardized returns,

$$u_t \;=\; \frac{2\sum_{i=1}^{N}\sum_{j>i}^{N} e_{i,t}\, e_{j,t}}{N\sum_{i=1}^{N} e_{i,t}^2}$$

- Three parameters, ω_ρ, α_ρ and β_ρ, to be estimated.

DECO Example
Factor Structure

$$R_t = \lambda F_t + \nu_t$$

where

$$F_t = \Omega_{Ft}^{1/2} Z_t$$

$$Z_t \sim i.i.d.(0,\ \mathcal{I})$$

$$\nu_t \sim i.i.d.(0,\ \Omega_\nu)$$

$$\Longrightarrow\ \Omega_t \;=\; \lambda\,\Omega_{Ft}\,\lambda' \;+\; \Omega_{\nu t}$$

One-Factor Case with Everything Orthogonal

$$R_t = \lambda f_t + \nu_t$$

where

$$f_t = \sigma_{ft} z_t$$

$$z_t \sim i.i.d.(0,\ 1)$$

$$\nu_t \sim i.i.d.(0,\ \sigma_\nu^2)$$

$$\Longrightarrow\ \Omega_t\ =\ \sigma_{ft}^2 \lambda\lambda'\ +\ \Omega_\nu$$

$$\sigma_{it}^2 = \sigma_{ft}^2 \lambda_i^2 + \sigma_{\nu i}^2$$

$$\sigma_{ijt}^2 = \sigma_{ft}^2 \lambda_i \lambda_j$$

Rigorous Modeling IV

Conditional Asset-Level (Multivariate) Volatility Dynamics from High-Frequency Data

Realized Covariance

$$dP(t)\ =\ M(t)\,dt\ +\ \Omega(t)^{1/2}\,dW(t)$$

$$RCov_t(\Delta)\ \equiv\ \sum_{j=1}^{N(\Delta)} R_{t-1+j\Delta,\Delta}\,R'_{t-1+j\Delta,\Delta}$$

$$RCov_t(\Delta)\ \rightarrow\ ICov_t\ =\ \int_{t-1}^{t} \Omega\,(\tau)\,d\tau$$

– p.d. so long as $N(\Delta) > N$; else use regularization methods

Asynchronous Trading and the Epps Effect

– Epps effect biases covariance estimates downward

– Can overcome Epps by lowering sampling frequency to accommodate least-frequently-traded asset, but that wastes data

– Opposite extreme: Calculate each pairwise realized covariance matrix using appropriate sampling; then assemble and regularize

Regularization (Shrinkage)

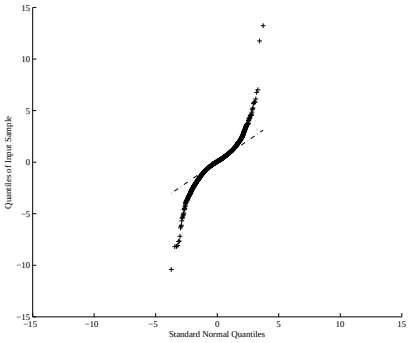


Figure 10.16: QQ Plot of S&P500 Returns. We show quantiles of daily S&P500 returns from January 2, 1990 to December 31, 2010, against the corresponding quantiles from a standard normal distribution.

$$\hat{\Omega}_t^S \;=\; \kappa\, RCov_t(\Delta) \;+\; (1-\kappa)\,\Upsilon_t$$

– Υ_t is p.d. and $0 < \kappa < 1$
  – $\Upsilon_t = I$ (naive benchmark)
  – $\Upsilon_t = \Omega$ (unconditional covariance matrix)
  – $\Upsilon_t = \sigma_f^2 \lambda\lambda' + \Omega_\nu$ (one-factor market model)

Multivariate GARCH-RV

$$vech\,(\Omega_t) \;=\; vech\,(C) \;+\; B\,vech\,(\Omega_{t-1}) \;+\; A\,vech(\hat{\Omega}_{t-1})$$

- Fine for 1-step
- Multi-step requires "closing the system" with an RV equation
  – Noureldin et al. (2011), multivariate HEAVY

Rigorous Modeling V

Distributions

Modeling Entire Return Distributions:
Returns are not Unconditionally Gaussian

Modeling Entire Return Distributions:
Returns are Often not Conditionally Gaussian

Modeling Entire Return Distributions: Issues

- Gaussian QQ plots effectively show calibration of Gaussian VaR at different levels
- Gaussian unconditional VaR is terrible
- Gaussian conditional VaR is somewhat better but left tail remains bad

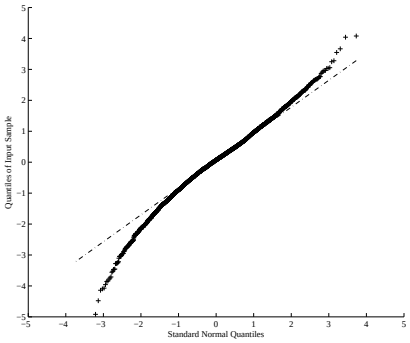


Figure 10.17: QQ Plot of S&P500 Returns Standardized by NGARCH Volatilities. We show quantiles of daily S&P500 returns standardized by the dynamic volatility from a NGARCH model against the corresponding quantiles of a standard normal distribution. The sample period is January 2, 1990 through December 31, 2010. The units on each axis are standard deviations.

- Gaussian conditional expected shortfall, which integrates over the left tail, would be terrible
- So we want more accurate assessment of things like $VaR^p_{T+1|T}$ than those obtained under Gaussian assumptions

  –Doing so for all values of $p \in [0, 1]$ requires estimating the entire conditional return distribution

  – More generally, best-practice risk measurement is about tracking the entire conditional return distribution

Observation-Driven Density Forecasting

Using $r = \sigma\,\varepsilon$ and GARCH

Assume:

$$r_{T+1} = \sigma_{T+1/T}\varepsilon_{T+1}$$

$$\varepsilon_{T+1} \sim iid(0, 1)$$

Multiply ε_{T+1} draws by $\sigma_{T+1/T}$ (fixed across draws, from a GARCH model) to build up the conditional density of r_{T+1}.

- ε_{T+1} simulated from standard normal
- ε_{T+1} simulated from standard t
- ε_{T+1} simulated from kernel density fit to $\frac{r_{T+1}}{\sigma_{T+1/T}}$
- ε_{T+1} simulated from any density that can be simulated

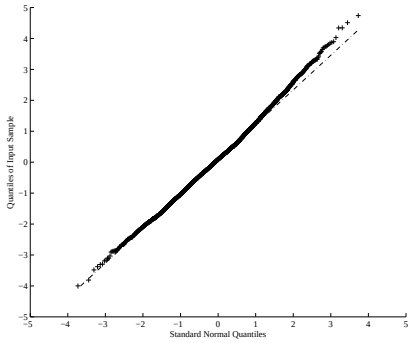


Figure 10.18: QQ Plot of S&P500 Returns Standardized by Realized Volatilities. We show quantiles of daily S&P500 returns standardized by $AvgRV$ against the corresponding quantiles of a standard normal distribution. The sample period is January 2, 1990 through December 31, 2010. The units on each axis are standard deviations.

Parameter-Driven Density Forecasting

Using $r = \sigma\,\varepsilon$ and SV

Assume:

$$r_{T+1} = \sigma_{T+1}\varepsilon_{T+1}$$

$$\varepsilon_{T+1} \sim iid(0,1)$$

Multiply ε_{T+1} draws by σ_{T+1} draws (from a simulated SV model) to build up the conditional density of r_{T+1}.

– Again, ε_{T+1} simulated from any density deemed relevant

Modeling Entire Return Distributions:

Returns Standardized by RV *are* Approximately Gaussian

A Special Parameter-Driven Density Forecasting Approach

Using $r = \sigma\,\varepsilon$ and RV

(Log-Normal / Normal Mixture)

Assume:

$$r_{T+1} = \sigma_{T+1}\varepsilon_{T+1}$$

$$\varepsilon_{T+1} \sim iid(0,1)$$

Multiply ε_{T+1} draws from $N(0,1)$ by σ_{T+1} draws (from a simulated RV model fit to log realized standard deviation) to build up the conditional density of r_{T+1}.

Pitfalls of the "$r = \sigma\,\varepsilon$" Approach

In the conditionally *Gaussian* case we can write with no loss of generality:

$$r_{T+1} = \sigma_{T+1/T}\varepsilon_{T+1}$$

$$\varepsilon_{T+1} \sim iidN(0,1)$$

But in the conditionally non-Gaussian case there *is* potential loss of generality in writing:

$$r_{T+1} = \sigma_{T+1/T}\varepsilon_{T+1}$$

$$\varepsilon_{T+1} \sim iid(0,1),$$

because there may be time variation in conditional moments other than $\sigma_{T+1/T}$, and using $\varepsilon_{T+1} \sim iid(0,1)$ assumes that away

Multivariate Return Distributions

– If reliable realized covariances are available, one could do a multivariate analog of the earlier lognormal/normal mixture model. But the literature thus far has focused primarily on conditional distributions for "daily" data.

Return version:

$$Z_t \;=\; \Omega_t^{-1/2}\,R_t, \qquad Z_t \sim i.i.d., \quad E_{t-1}(Z_t) = 0 \quad Var_{t-1}(Z_t) = \mathcal{I}$$

Standardized return version (as in DCC):

$$e_t \;=\; D_t^{-1}\,R_t, \quad E_{t-1}(e_t) \;=\; 0, \quad Var_{t-1}(e_t) \;=\; \Gamma_t$$

where D_t denotes the diagonal matrix of conditional standard deviations for each of the assets, and Γ_t refers to the potentially time-varying conditional correlation matrix.

Leading Examples

Multivariate normal:

$$f(e_t) \;=\; C\,(\Gamma_t)\,\exp\left(-\tfrac{1}{2}\,e_t'\,\Gamma_t^{-1}\,e_t\right)$$

Multivariate t:

$$f(e_t) \;=\; C\,(d,\Gamma_t)\left(1 + \frac{e_t'\Gamma_t^{-1}e_t}{(d-2)}\right)^{-(d+N)/2}$$

Multivariate asymmetric t:

$$f\left(e_t\right) = \frac{C\left(d, \dot{\Gamma}_t\right) K_{\frac{d+N}{2}}\left(\sqrt{\left(d + \left(e_t - \dot{\mu}\right)' \dot{\Gamma}_t^{-1}\left(e_t - \dot{\mu}\right)\right) \xi' \dot{\Gamma}_t^{-1} \xi}\right) \exp\left(\left(e_t - \dot{\mu}\right)' \dot{\Gamma}_t^{-1} \xi\right)}{\left(1 + \frac{\left(e_t - \dot{\mu}\right)' \dot{\Gamma}_t^{-1}\left(e_t - \dot{\mu}\right)}{d}\right)^{\frac{(d+N)}{2}} \left(\sqrt{\left(d + \left(e_t - \dot{\mu}\right)' \dot{\Gamma}_t^{-1}\left(e_t - \dot{\mu}\right)\right) \xi' \dot{\Gamma}_t^{-1} \xi}\right)^{-\frac{(d+N)}{2}}}$$

– More flexible than symmetric t but requires estimation of N asymmetry parameters simultaneously with the other parameters, which is challenging in high dimensions.

Copula methods sometimes provide a simpler two-step approach.

Copula Methods

Sklar's Theorem:

$$F(e) \;=\; G(\, F_1(e_1), ..., F_N(e_N)\,) \;\equiv\; G(\, u_1, ..., u_N\,) \;\equiv\; G(u)$$

$$f(e) \;=\; \frac{\partial^N G(F_1(e_1), ..., F_N(e_N))}{\partial e_1 ... \partial e_N} \;=\; g\left(u\right) \times \prod_{i=1}^{N} f_i(e_i)$$

$$\Longrightarrow \quad \log L \;=\; \sum_{t=1}^{T} \log g(u_t) \;+\; \sum_{t=1}^{T} \sum_{i=1}^{N} \log f_i(e_{i,t})$$

Standard Copulas

Normal:

$$g(u_t; \Gamma_t^*) = |\Gamma_t^*|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \Phi^{-1}(u_t)' (\Gamma_t^{*-1} - I) \Phi^{-1}(u_t)\right\}$$

where $\Phi^{-1}(u_t)$ refers to the $N \times 1$ vector of standard inverse univariate normals, and the correlation matrix Γ_t^* pertains to the $N \times 1$ vector e_t^* with typical element,

$$e_{i,t}^* \;=\; \Phi^{-1}(u_{i,t}\,) \;=\; \Phi^{-1}(F_i\left(e_{i,t}\right)).$$

– Often does not allow for sufficient dependence between tail events.

– t copula

– Asymmetric t copula

Asymmetric Tail Correlations

Multivariate Distribution Simulation (General Case)

Simulate using:

$$R_t = \hat{\Omega}_t^{1/2} Z_t$$

$$Z_t \sim i.i.d.(0, \mathcal{I})$$

– Z_t may be drawn from parametrically-(Gaussian, t, ...) or nonparametrically-fitted

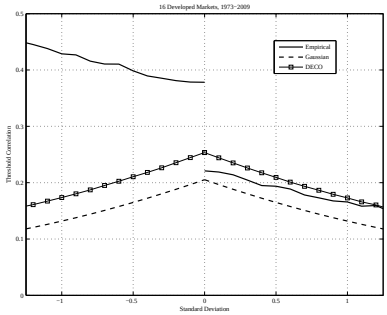


Figure 10.19: Average Threshold Correlations for Sixteen Developed Equity Markets. The solid line shows the average empirical threshold correlation for GARCH residuals across sixteen developed equity markets. The dashed line shows the threshold correlations implied by a multivariate standard normal distribution with constant correlation. The line with square markers shows the threshold correlations from a DECO model estimated on the GARCH residuals from the 16 equity markets. The figure is based on weekly returns from 1973 to 2009.

distributions, or with replacement from the empirical distribution.

Multivariate Distribution Simulation (Factor Case)

Simulate using:

$$F_t = \hat{\Omega}_{F,t}^{1/2} Z_{F,t}$$

$$R_t \; = \; \hat{\lambda} \, F_t + \nu_t$$

– $Z_{F,t}$ and ν_t may be drawn from parametrically- or nonparametrically-fitted distributions, or with replacement from the empirical distribution.

Rigorous Modeling VI

Risk, Return and Macroeconomic Fundamentals

We Want to Understand the Financial / Real Connections

Statistical vs. "scientific" models

Returns $\leftrightarrow$ Fundamentals

$r \leftrightarrow f$

Disconnect?

"excess volatility," "disconnect," "conundrum," ...

μ_r, σ_r, σ_f, μ_f

Links are complex:

$\mu_r \leftrightarrow \sigma_r \leftrightarrow \sigma_f \leftrightarrow \mu_f$

Volatilities as intermediaries?

For Example...

	Mean Recession Volatility Increase	Standard Error	Sample Period
Aggregate Returns	43.5%	3.8%	63Q1-09Q3
Firm-Level Returns	28.6%	6.7%	69Q1-09Q2

Table 10.1: Stock Return Volatility During Recessions. Aggregate stock-return volatility is quarterly realized standard deviation based on daily return data. Firm-level stock-return volatility is the cross-sectional inter-quartile range of quarterly returns.

	Mean Recession Volatility Increase	Standard Error	Sample Period
Aggregate Growth	37.5%	7.3%	62Q1-09Q2
Firm-Level Growth	23.1%	3.5%	67Q1-08Q3

Table 10.2: Real Growth Volatility During Recessions. Aggregate real-growth volatility is quarterly conditional standard deviation. Firm-level real-growth volatility is the cross-sectional inter-quartile range of quarterly real sales growth.

GARCH σ_r usually has no μ_f or σ_f:

$$\sigma_{r,t}^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{r,t-1}^2.$$

One might want to entertain something like:

$$\sigma_{r,t}^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{r,t-1}^2 + \delta_1 \mu_{f,t-1} + \delta_2 \sigma_{f,t-1}.$$

$\mu_f \leftrightarrow \sigma_r$

Return Volatility is Higher in Recessions

Schwert's (1989) "failure": Very hard to link market risk to expected fundamentals (leverage, corporate profitability, etc.).

Actually a great success:

Key observation of robustly higher return volatility in recessions!

– Earlier: Officer (1973)

– Later: Hamilton and Lin (1996), Bloom et al. (2009)

Extends to business cycle effects in credit spreads via the Merton model

$\mu_f \leftrightarrow \sigma_r$, Continued

Bloom et al. (2009) Results

$\mu_f \leftrightarrow \sigma_f$

Fundamental Volatility is Higher in Recessions

More Bloom, Floetotto and Jaimovich (2009) Results

$\sigma_f \leftrightarrow \sigma_r$

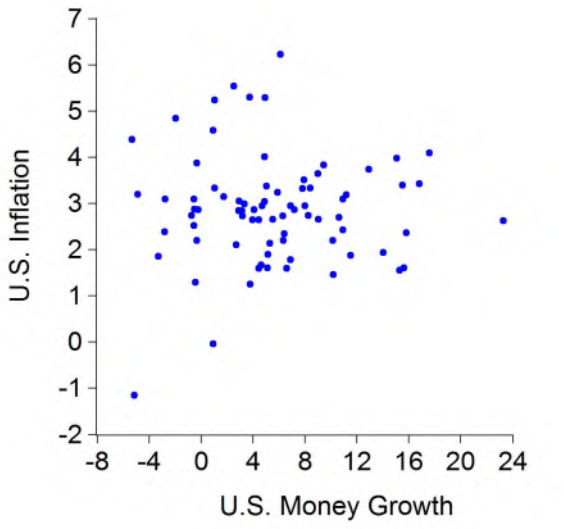
Return Vol is Positively Related to Fundamental Vol

Follows immediately from relationships already documented

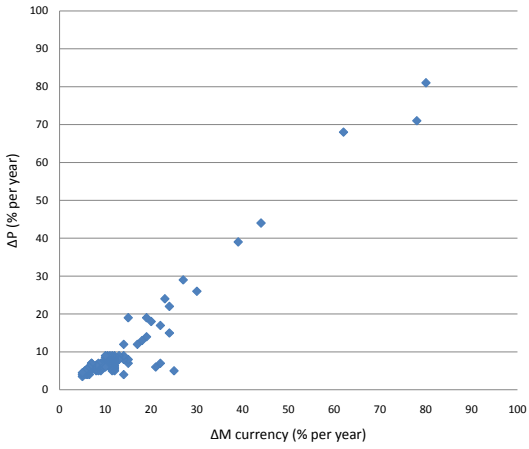
Moreover, direct explorations provide direct evidence:

– Engle et al. (2006) time series

– Diebold and Yilmaz (2010) cross section
– Engle and Rangel (2008) panel
Can be extended to fundamental determinants of correlations (Engle and Rangle, 2011)
[Aside: Inflation and its Fundamental (U.S. Time Series)]

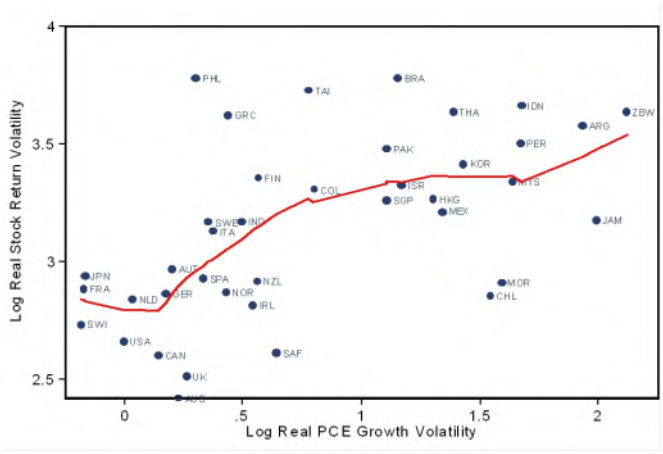


Weak inflation / money growth link
[Inflation and its Fundamental (Barro's Cross Section)]
Strong inflation / money growth link



Back to $\sigma_f \leftrightarrow \sigma_r$: Cross-Section Evidence

Real Stock Return Volatility and Real PCE Growth Volatility, 1983-2002



Now Consider Relationships Involving the Equity Premium

?? μ_r ??

$\mu_r \leftrightarrow \sigma_r$

"Risk-Return Tradeoffs" (or Lack Thereof)

Studied at least since Markowitz

ARCH-M characterization:

$$R_t = \beta_0 + \beta_1 X_t + \beta_2 \sigma_t + \varepsilon_t$$

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$$

– But subtleties emerge...

$\mu_r \leftrightarrow \mu_f$

Odd Fama-French (1989):

$$r_{t+1} = \beta_0 + \beta_1 dp_t + \beta_2 term_t + \beta_3 def_t + \epsilon_{t+1}$$

Less Odd Lettau-Ludvigson (2001):

$$r_{t+1} = \beta_0 + \beta_1 dp_t + \beta_2 term_t + \beta_3 def_t + \beta_4 cay_t + \epsilon_{t+1}$$

Natural Campbell-Diebold (2009):

$r_{t+1} = \beta_0 + \beta_1 dp_t + \beta_2 term_t + \beta_3 def_t + \beta_4 cay_t + \beta_5 g_t^e + \epsilon_{t+1}$

– Also Goetzman et al. (2009) parallel cross-sectional analysis

Expected Business Conditions are Crucially Important!

$\mu_r \leftrightarrow \sigma_f$

Bansal and Yaron (2004)

(and many others recently)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
g_t^e	-0.22	–	–	-0.21	-0.20	–	-0.20
	(0.08)	–	–	(0.09)	(0.09)	–	(0.10)
DP_t	–	–	0.25	0.17	–	0.19	0.12
			(0.10)	(0.10)		(0.12)	(0.11)
DEF_t	–	–	-0.11	-0.01	–	-0.10	0.00
			(0.07)	(0.09)		(0.08)	(0.09)
$TERM_t$	–	–	0.15	0.17	–	0.09	0.11
			(0.07)	(0.07)		(0.09)	(0.09)
CAY_t	–	0.24	–	–	0.22	0.17	0.15
		(0.07)			(0.08)	(0.11)	(0.10)

So, Good News:

We're Learning More and More About the Links

– We've Come a Long Way Since Markowitz:

$\mu_r \leftrightarrow \sigma_r$

The Key Lesson

The business cycle is of central importance for both μ_r and σ_r

– Highlights the importance of high-frequency business cycle monitoring. We need to interact high-frequency real activity with high-frequency financial market activity

e.g., Aruoba-Diebold-Scotti real-time framework at Federal Reserve Bank of Philadelphia

Conclusions

- Reliable risk measurement requires *conditional* models that allow for time-varying volatility.
- Risk measurement may be done using univariate volatility models. Many important recent developments.
- High-frequency return data contain a wealth of volatility information.
- Other tasks require multivariate models. Many important recent developments, especially for N large. Factor structure is often useful.
- The business cycle emerges as a key macroeconomic fundamental driving risk.
- New developments in high-frequency macro monitoring yield high-frequency real activity data to match high-frequency financial market data.

****

Models for non-negative variables (from Minchul)

Introduction **Motivation:** Why do we need dynamic models for positive values?

- Volatility: Time-varying conditional variances

- Duration: Intertrade duration, Unemployment spell
- Count: Defaults of U.S. corporations

Autoregressive Gamma processes (ARG)

- Gourieroux and Jasiak (2006)
- Monfort, Pegoraro, Renne and Roussellet (2014)

Alternative model

- ACD (autoregressive conditional duration) by Engle and Russell (1998)
- Its extension through Dynamic conditional score models
  - Harvey (2013)
  - Creal, Koopman, and Lucas (2013)

Autoregressive Gamma Processes

Autoregressive Gamma Processes (ARG): Definition **Definition:** Y_t follows the autoregressive gamma process if

Y_t conditional on Y_{t-1} follows the non-central gamma distribution with

- degree of freedom parameter: δ
- non-centrality parameter: βY_{t-1}
- scale parameter: c

Very exotic ... but we can guess

- Gamma distribution –¿ Maybe it takes positive values
- Conditional dynamics through non-centrality parameter

ARG Processes: State space representation If Y_t follows ARG, then

Measurement:

$$Y_t \left| Z_t \sim Gamma(\delta + Z_t, c) \right.$$

Transition:

$$Z_t \left| Y_{t-1} \sim Poisson(\beta Y_{t-1}) \right.$$

- Y_t takes positive real number
- Z_t takes positive integer
- Dynamics through Z_t

Conditional moments **Measurement:**

$$Y_t \mid Z_t \sim Gamma(\delta + Z_t, c)$$

Transition:

$$Z_t \mid Y_{t-1} \sim Poisson(\beta Y_{t-1})$$

Conditional moments:

$$\begin{aligned} E(Y_t|Y_{t-1}) &= \rho Y_{t-1} + c\delta \\ V(Y_t|Y_{t-1}) &= 2\rho c Y_{t-1} + c^2\delta \\ Corr(Y_t, Y_{t-h}) &= \rho^h \end{aligned}$$

where $\rho = \beta c > 0$.

The process is stationary when $\rho < 1$.
Conditional over-dispersion The conditional over-dispersion exists if and only if

$$V(Y_t|Y_{t-1}) > E(Y_t|Y_{t-1})^2$$

When $\delta < 1$,

- The stationary ARG process features marginal over-dispersion.
- The process may feature either conditional under- or over-dispersion, depending on the value of Y_{t-1}.

Remark: ACD (autoregressive conditional duration) model assumes the path-indepednet over-dispersion.
Continuous time limit of ARG(1) The stationary ARG process is a discretized version of the CIR process.

$$dY_t = a(b - Y_t)dt + \sigma\sqrt{Y_t}dW_t$$

where

$$\begin{aligned} a &= -\log\rho \\ b &= \frac{c\delta}{1-\rho} \\ \sigma^2 &= \frac{-2\log\rho}{1-\rho}c \end{aligned}$$

- This process is non-negative almost surely.
  - Originally model for interest rates.
  - Also used for volatility dynamics.

Figure 10.20: Simulated data, $\rho = 0.5$

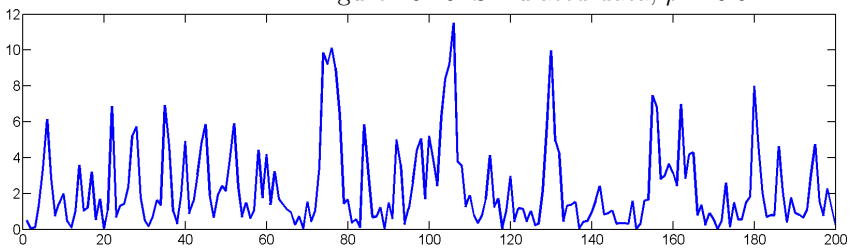
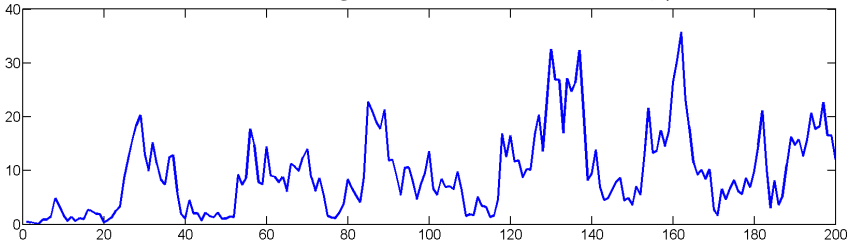


Figure 10.21: Simulated data, $\rho = 0.9$



Long memory Case 1) Let $\rho = 1$, then

- Y_t is a stationary Markov process.
- An autocorrelation function with a hyperbolic rate of decay.

Case 2) Stochastic autoregressive coefficient. Let $\rho \sim \pi$. Then

$$Corr(Y_t, Y_{t-h}|\delta, c, \pi) = E_\pi(\rho^h)$$

The autocorrelation function features hyperbolic decay when the distribution π assigns sufficiently large probabilities to values close to one.

Figures 1

Application in the original paper

Measurement:

$$Y_t \left| Z_t \sim Gamma(\delta + Z_t, c) \right.$$

Transition:

$$Z_t \left| Y_{t-1} \sim Poisson(\beta Y_{t-1}) \right.$$

- Y_t: Interquote durations of the Dayton Mining stock traded on the Toronto Stock Exchange in October 1998.
- Estimation based on QMLE

Extension Creal (2013) considers the following non-linear state space

Measurement

$$y_t \sim p(y_t|h_t, x_t; \theta)$$

where x_t is an exogenous regressor.

Transition

$$h_t \sim Gamma(\delta + z_t, c)$$
$$z_t \sim Poisson(\rho h_{t-1})$$

- When $y_t = h_t$, the process becomes ARG.
- Various applications are fall into this form.

Example 1: Stochastic volatility models **Measurement**

$$y_t = \mu + x_t\beta + \sqrt{h_t}e_t, \quad e_t \sim N(0,1)$$

Transition

$$h_t \sim Gamma(\delta + z_t, c)$$
$$z_t \sim Poisson(\rho h_{t-1})$$

Example 2: Stochastic duration and intensity models **Measurement**

$$y_t \sim Gamma\left(\alpha,\ h_t \exp(x_t\beta)\right)$$

Transition

$$h_t \sim Gamma(\delta + z_t, c)$$
$$z_t \sim Poisson(\rho h_{t-1})$$

Example 3: Stochastic count models **Measurement**

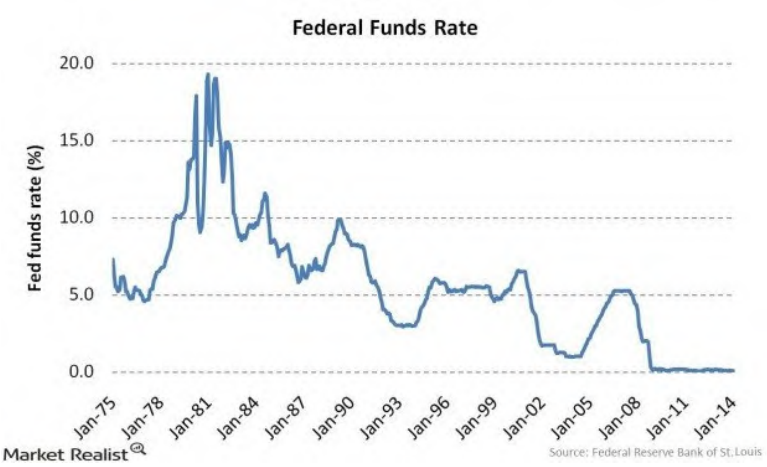
$$y_t \sim Poisson\left(h_t \exp(x_t\beta)\right)$$

Transition

$$h_t \sim Gamma(\delta + z_t, c)$$
$$z_t \sim Poisson(\rho h_{t-1})$$

Recent extension: ARG-zero processes 1 Monfort, Pegoraro, Renne, Roussellet (2014) extend ARG process to take account for zero-lower bound spells,

Recent extension: ARG-zero processes 2 Monfort, Pegoraro, Renne, Roussellet (2014) extend ARG process to take account for zero-lower bound spells,



If Y_t follows ARG, then

$$Y_t \,|Z_t \sim Gamma(\delta + Z_t, c)$$
$$Z_t \,|Y_{t-1} \sim Poisson(\beta Y_{t-1})$$

If Y_t follows ARG-zero, then

$$Y_t \,|Z_t \sim Gamma(Z_t, c)$$
$$Z_t \,|Y_{t-1} \sim Poisson(\alpha + \beta Y_{t-1})$$

Two modifications

- $\delta = 0$: As $\delta \to 0$, $Gamma(\delta, c)$ converges to dirac delta function.
- α is related with a probability of escaping from the zero lower bound.

Characterization **Probability density** for ARG-zero is

$$p(Y_t|Y_{t-1}; \alpha, \beta, c) = \sum_{z=1}^{\infty} g(Y_t, Y_{t-1}, \alpha, \beta, c, z) 1_{\{Y_t > 0\}} + \exp(-\alpha - \beta Y_{t-1}) 1_{\{Y_t = 0\}}$$

- Second term is consequence of $\delta \to 0$.
- If $\alpha = 0$, $Y_t = 0$ becomes an absorbing state.

Conditional moments

$$E[Y_t|Y_{t-1}] = \alpha c + \rho Y_{t-1}$$

and

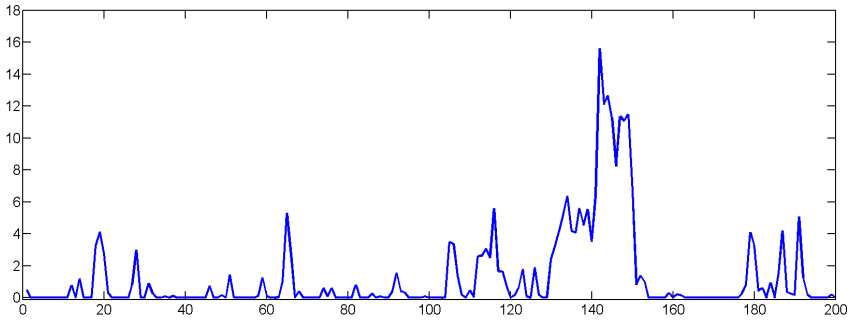
$$V(Y_t|Y_{t-1}) = 2c^2\alpha + 2c\rho Y_{t-1}$$

where $\rho = \beta c$.

Figure: ARG-zero

ACD and DCS

Figure 10.22: Simulated data



Autoregressive conditional duration model (ACD) Y_t follows the autoregressive conditional duration model if

$$y_t = \mu_t \, e_t, \quad E[e_t] = 1$$
$$\mu_t = w + \alpha \mu_{t-1} + \beta y_{t-1}$$

- Because of its multiplicative form, it is classified as the multiplicative error model (MEM).
- Conditional moments

$$E[y_t | y_{1:t-1}] = \mu_t$$
$$V(y_t | y_{1:t-1}) = k_0 \mu_t^2$$

- Conditional over-dispersion is path-independent

$$\frac{V(y_t | y_{1:t-1})}{E[y_t | y_{1:t-1}]^2} = k_0$$

  Recall that ARG process can have path-dependent over-dispersion.

Dynamic conditional score (DCS) model Dynamic conditional score model (or Generalized Autoregressive Score model) is a general class of observation-driven model.

- Observation-driven model is a time-varying parameter model where time-varying parameter is a function of histories of observable. For example, GARCH, ACD, ...
- DCS (GAS) model encompasses GARCH, ACD, and other observation-driven models.

The idea is very simple and pragmatic

- Give me a conditional likelihood and time-varying parameters, I will give you a law of motion for time-varying parameters.

Convenient and general modelling strategy. I will describe it within the MEM class of model.
DCS Example: ACD 1 Recall

$$y_t = \mu_t \, e_t, \quad E[e_t] = 1$$
$$\mu_t = w + \alpha \mu_{t-1} + \beta y_{t-1}$$

Instead, we apply DCS principle: "Give me conditional likelihood and time-varying parameters, then I will give you a law motion"

$$y_t = \mu_t \, e_t, \quad e_t \sim Gamma(\kappa, 1/\kappa)$$

DCS Example: ACD 2

$$y_t = \mu_t \, e_t, \quad e_t \sim Gamma(\kappa, 1/\kappa)$$

Then DCS specifies a law of motion for μ_t as follows:

$$\mu_t = w + \alpha \mu_{t-1} + \beta s_{t-1}$$

where (w, α, β) are additional parameters and s_t is a scaled score,

$$s_t = E_{t-1}\left[\frac{\partial \log p(y_t|\mu_t, y_{1:t}; \kappa)}{\partial \mu_t} \frac{\partial \log p(y_t|\mu_t, y_{1:t}; \kappa)}{\partial \mu_t}'\right]^{-1} \frac{\partial \log p(y_t|\mu_t, y_{1:t}; \kappa)}{\partial \mu_t}$$

In this case, it happens to be

$$\mu_t = w + \alpha \mu_{t-1} + \beta y_{t-1}$$

which is ACD.

However, a law of motion will be different with different choice of distribution – Generalized Gamma, Log-Logistic, Burr, Pareto, and many other distributions

10.5 EXERCISES, PROBLEMS AND COMPLEMENTS

10.6 NOTES

Chapter Eleven

Non-Linear Non-Gaussian State Space and Optimal Filtering

11.1 VARIETIES OF NON-LINEAR NON-GAUSSIAN MODELS

11.2 MARKOV CHAINS TO THE RESCUE (AGAIN): THE PARTICLE FILTER

11.3 PARTICLE FILTERING FOR ESTIMATION: DOUCET'S THEOREM

11.4 KEY APPLICATION I: STOCHASTIC VOLATILITY (REVISITED)

11.5 KEY APPLICATION II: CREDIT-RISK AND THE DEFAULT OPTION

11.6 KEY APPLICATION III: DYNAMIC STOCHASTIC GENERAL EQUILIBRIUM (DSGE) MACROECONOMIC MODELS

11.7 A PARTIAL "SOLUTION": THE EXTENDED KALMAN FILTER

Familiar Linear / Gaussian State Space

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

$$y_t = Z\alpha_t + \varepsilon_t$$

$$\eta_t \sim N(0, Q), \varepsilon_t \sim N(0, H)$$

Linear / Non-Gaussian

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

$$y_t = Z\alpha_t + \varepsilon_t$$

$$\eta_t \sim D^{\eta}, \varepsilon_t \sim D^{\varepsilon}$$

Non-Linear / Gaussian

$$\alpha_t = Q(\alpha_{t-1}, \eta_t)$$

$$y_t = G(\alpha_t, \varepsilon_t)$$

$$\eta_t \sim N(0, Q), \varepsilon_t \sim N(0, H)$$

Non-Linear / Gaussian II
(Linear / Gaussian with time-varying system matrices)

$$\alpha_t = T_t \alpha_{t-1} + R_t \eta_t$$

$$y_t = Z_t \alpha_t + \varepsilon_t$$

$$\eta_t \sim N^{\eta}, \quad \varepsilon_t \sim N^{\varepsilon}$$

"Conditionally Gaussian"
White's theorem
Non-Linear / Non-Gaussian

$$\alpha_t = Q(\alpha_{t-1}, \eta_t)$$

$$y_t = G(\alpha_t, \varepsilon_t)$$

$$\eta_t \sim D^{\eta}, \quad \varepsilon_t \sim D^{\varepsilon}$$

(DSGE macroeconomic models are of this form)
Non-Linear / Non-Gaussian, Specialized

$$\alpha_t = Q(\alpha_{t-1}) + \eta_t$$

$$y_t = G(\alpha_t) + \varepsilon_t$$

$$\eta_t \sim D^\eta, \quad \varepsilon_t \sim D^\varepsilon$$

Non-Linear / Non-Gaussian, Generalized

$$\alpha_t = Q_t(\alpha_{t-1}, \eta_t)$$

$$y_t = G_t(\alpha_t, \varepsilon_t)$$

$$\eta_t \sim D_t^\eta, \quad \varepsilon_t \sim D_t^\varepsilon$$

Credit Risk Model (Nonlinear / Non-Gaussian)

Asset value of the firm V_t:

$$V_t = \mu V_{t-1}\eta_t', \eta_t' \sim lognormal$$

Firm issues liability D: Zero-coupon bond, matures at T, pays D

Equity value of the firm S_t:

$$S_t = max(V_t - D, 0)$$

From the call option structure of S_t, Black-Scholes gives:

$S_t = BS(V_t)\varepsilon_t'$

$\varepsilon_t' \sim lognormal$

(ε_t' captures BS misspecification, etc.)

Credit Risk Model (Nonlinear / Gaussian Form)

Taking logs makes it Gaussian, but it's intrinsically non-linear:

$ln\, V_t = ln\, \mu + V_{t-1} + \eta_t$

$ln\, S_t = ln\, BS(V_t) + \varepsilon_t$

$\eta_t \sim N, \varepsilon_t \sim N$

Regime Switching Model (Nonlinear / Gaussian)

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix} = \begin{pmatrix} \phi & 0 \\ 0 & \gamma \end{pmatrix} \begin{pmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{pmatrix} + \begin{pmatrix} \eta_{1t} \\ \eta_{2t} \end{pmatrix}$$

$$y_t = \mu_0 + \delta\, I(\alpha_{2t} > 0) + (1, 0) \begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix}$$

$$\eta_{1t} \sim N^{\eta_1} \quad \eta_{2t} \sim N^{\eta_2} \quad \eta_{1t} \perp \eta_{2t}$$

Extensions to:

– Richer α_1 dynamics (governing the observed y)

– Richer α_2 dynamics (governing the latent regime)

– Richer η_t distribution (e.g., η_{2t} asymmetric)

– More than two states

– Switching also on dynamic parameters, volatilities, etc.

– Multivariate

Stochastic Volatility Model (Nonlinear/Gaussian Form)

$$h_t = \omega + \beta h_{t-1} + \eta_t \quad (transition)$$
$$r_t = \sqrt{e^{h_t}}\varepsilon_t \quad (measurement)$$

$\eta_t \sim N(0, \sigma_\eta^2), \quad \varepsilon_t \sim N(0,1)$

Stochastic Volatility Model (Linear/Non-Gaussian Form)

$$h_t = \omega + \beta h_{t-1} + \eta_t \quad (transition)$$
$$2ln|r_t| = h_t + 2ln|\varepsilon_t| \quad (measurement)$$

or

$$h_t = \omega + \beta h_{t-1} + \eta_t$$
$$y_t = h_t + u_t$$

$\eta_t \sim N(0, \sigma_\eta^2), \quad u_t \sim D^u$

– A "signal plus (non-Gaussian) noise"

components model for volatility

Realized and Integrated Volatility

$$IV_t = \phi IV_{t-1} + \eta_t$$

$$RV_t = IV_t + \varepsilon_t$$

ε represents the fact that RV is based on less than an infinite sampling frequency.

Microstructure Noise Model

**Hasbrouck

(Non-linear / non-Gaussian)

A Distributional Statement of the Kalman Filter ****

Multivariate Stochastic Volatility with Factor Structure

***

Approaches to the General Filtering Problem Kitagawa (1987), numerical integration (linear / non-Gaussian) More recently, Monte Carlo integration

Extended Kalman Filter (Non-Linear / Gaussian)

$$\alpha_t = Q(\alpha_{t-1}, \eta_t)$$
$$y_t = G(\alpha_t, \varepsilon_t)$$
$$\eta_t \sim N, \varepsilon_t \sim N$$

Take first-order Taylor expansions of:

Q around a_{t-1}

G around $a_{t,t-1}$

Use Kalman filter on the approximated system

Unscented Kalman Filter (Non-Linear / Gaussian)

Bayes Analysis of SSMs: Carlin-Polson-Stoffer 1992 *JASA*

"single-move" Gibbs sampler

(Many parts of the Gibbs iteration: the parameter vector, and then *each observation* of the state vector, period-by-period)

Multi-move Gibbs sampler can handle non-Gaussian (via mixtures of normals), but not nonlinear.

Single-move can handle nonlinear and non-Gaussian.

Expanding $S(\hat{\theta}_{ML})$ around θ yields:
$S(\hat{\theta}_{ML}) \approx S(\theta) + S'(\theta)(\hat{\theta}_{ML} - \theta) = S(\theta) + H(\theta))(\hat{\theta}_{ML} - \theta)$.
Noting that $S(\hat{\theta}_{ML}) \equiv 0$ and taking expectations yields:
$0 \approx S(\theta) - I_{EX,H}(\theta)(\hat{\theta}_{ML} - \theta)$
or
$(\hat{\theta}_{ML} - \theta) \approx I^{-1}_{EX,H}(\theta)$.
Using $S(\theta) \overset{a}{\sim} N(0, I_{EX,H}(\theta))$ then implies:
$(\hat{\theta}_{ML} - \theta) \overset{a}{\sim} N(0, I^{-1}_{EX,H}(\theta))$
or
***

Case 3 *β and σ^2*
Joint prior $g(\beta, \frac{1}{\sigma^2}) = g(\beta/\frac{1}{\sigma^2})g(\frac{1}{\sigma^2})$
where $\beta/\frac{1}{\sigma^2} \sim N(\beta_0, \Sigma_0)$ and $\frac{1}{\sigma^2} \sim G(\frac{v_0}{2}, \frac{\delta_0}{2})$
HW Show that the joint posterior,
$p(\beta, \frac{1}{\sigma^2}/y) = g(\beta, \frac{1}{\sigma^2})L(\beta, \frac{1}{\sigma^2}/y)$
can be factored as $p(\beta/\frac{1}{\sigma^2}, y)p(\frac{1}{\sigma^2/y})$
where $\beta/\frac{1}{\sigma^2}, y \sim N(\beta_1, \Sigma_1)$
and $\frac{1}{\sigma^2}/y \sim G(\frac{v_1}{2}, \frac{\delta_1}{2})$,
and derive expressions for $\beta_1, \Sigma_1, v_1, \delta_1$
in terms of $\beta_0, \Sigma_0, \delta_0, x$, and y.

Moreover, the key marginal posterior
$P(\beta/y) = \int_0^\infty p(\beta, \frac{1}{\sigma^2}/y)d\sigma^2$ is multivariate t.

Implement the Bayesian methods via Gibbs sampling.

Linear Quadratic Business Cycle Model

Hansen and Sargent

Linear Gaussian state space system

Parameter-Driven vs. Observation-Driven Models

Parameter-driven: Time-varying parameters measurable w.r.t. latent variables

Observation-driven: Time-varying parameters measurable w.r.t. observable variables

Parameter-driven models are mathematically appealing but hard to estimate. Observation-driven models are less mathematically appealing but easy to estimate. State-space models, in general, are parameter-driven. Stochastic volatility models are parameter-driven, while ARCH models are observation-driven.

Regime Switching

We have emphasized dynamic linear models, which are tremendously important in practice. They're called linear because y_t is a simple linear function of past y's or past ε 's. In some forecasting situations, however, good statistical characterization of dynamics may require some notion of regime switching, as between "good" and "bad" states, which is a type of nonlinear model.

Models incorporating regime switching have a long tradition in business-cycle analysis, in which expansion is the good state, and contraction (recession) is the bad state. This idea is also manifest in the great interest in the popular press, for example, in identifying and forecasting turning points in economic activity. It is only within a regime-switching framework that the concept of a turning point has intrinsic meaning; turning points are naturally and immediately defined as the times separating expansions and contractions.

Observable Regime Indicators

Threshold models are squarely in line with the regime-switching tradition. The following threshold model,

for example, has three regimes, two thresholds, and a d-period delay regulating the switches:

$$y_t \;=\; \left\{ \begin{array}{c} c^{(u)} \;+\; \phi^{(u)} y_{t-1} \;+\; \varepsilon_t^{(u)},\; \theta^{(u)} < y_{t-d} \\ c^{(m)} \;+\; \phi^{(m)} y_{t-1} \;+\; \varepsilon_t^{(m)},\; \theta^{(l)} < y_{t-d} < \theta^{(u)} \\ c^{(l)} \;+\; \phi^{(l)} y_{t-1} \;+\; \varepsilon_t^{(l)},\; \theta^{(l)} > y_{t-d}\;. \end{array} \right\}.$$

The superscripts indicate "upper," "middle," and "lower" regimes, and the regime operative at any time t depends on the observable past history of y – in particular, on the value of y_{t-d} .

Latent Markovian Regimes

Although observable threshold models are of interest, models with *latent* states as opposed to observed states may be more appropriate in many business, economic and financial contexts. In such a setup, time-series dynamics are governed by a finite-dimensional parameter vector that switches (potentially each period) depending upon which of two unobservable states is realized, with state transitions governed by a first-order Markov process. To make matters concrete, let's take a simple example. Let $\{s_t\}_{t=1}^T$ be the (latent) sample path of two-state first-order autoregressive process, taking just the two values 0 or 1, with transition probability matrix given by

$$M \;=\; \left(\begin{array}{cc} p_{00} & 1-p_{00} \\ 1-p_{11} & p_{11} \end{array} \right)\;.$$

The ij-th element of M gives the probability of moving from state i (at time $t-1$) to state j (at time t). Note that there are only two free parameters, the staying probabilities, p_{00} and p_{11}. Let $\{y_t\}_{t=1}^T$ be the sample path of an observed time series that depends on $\{s_t\}_{t=1}^T$ such that the density of y_t conditional upon $\{s_t\}$ is

$$f(y_t|s_t;\;\theta) \;= \frac{1}{\sqrt{2\pi}\;\sigma}\;\exp\;\left(\frac{-(y_t-\mu_{s_t})^2}{2\sigma^2} \right)\;.$$

Thus, y_t is Gaussian white noise with a potentially switching mean. The two means around which y_t moves are of particular interest and may, for example, correspond to episodes of differing growth rates ("booms" and "recessions", "bull" and "bear" markets, etc.).

Appendices

Appendix A

A "Library" of Useful Books

Ait-Sahalia, Y. and Hansen, L.P. eds. (2010), Handbook of Financial Econometrics. Amsterdam: North-Holland.

Ait-Sahalia, Y. and Jacod, J. (2014), High-Frequency Financial Econometrics, Princeton University Press.

Beran, J., Feng, Y., Ghosh, S. and Kulik, R. (2013), Long-Memory Processes: Probabilistic Properties and Statistical Methods, Springer.

Box, G.E.P. and Jenkins, G.W. (1970), Time Series Analysis, Forecasting and Control, Prentice-Hall.

Davidson, R. and MacKinnon, J. (1993), Estimation and Inference in Econometrics, Oxford University Press.

Diebold, F.X. (1998), Elements of Forecasting, South-Western.

Douc, R., Moulines, E. and Stoffer, D.S. (2014), Nonlinear Time Series: Theory, Methods, and Applications with R Examples, Chapman and Hall.

Durbin, J. and Koopman, S.J. (2001), Time Series Analysis by State Space Methods, Oxford University Press.

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap, Chapman and Hall.

Elliott, G., Granger, C.W.J. and Timmermann, A., eds. (2006), Handbook of Economic Forecasting, Volume 1, North-Holland.

Elliott, G., Granger, C.W.J. and Timmermann, A., eds. (2013), Handbook of Economic Forecasting, Volume 2, North-Holland.

Engle, R.F. and McFadden, D., eds. (1995), Handbook of Econometrics, Volume 4, North-Holland.

Geweke, J. (2010), Complete and Incomplete Econometric Models, Princeton University Press.

Geweke, J., Koop, G. and van Dijk, H., eds. (2011), The Oxford Handbook of Bayesian Econometrics, Oxford University Press.

Granger, C.W.J. and Newbold, P. (1977), Forecasting Economic Time Series, Academic Press.

Granger, C.W.J. and Tersvirta, Y. (1996), Modeling Nonlinear Economic Relationships, Oxford University Press.

Hall, P. (1992), The Bootstrap and Edgeworth Expansion, Springer Verlag.

Hammersley, J.M. and Handscomb, D.C. (1964), Monte Carlo Methods, Chapman and Hall.

Hansen, L.P. and Sargent, T.J. (2013), Recursive Models of Dynamic Linear Economies, Princeton University Press.

Harvey, A.C. (1989), Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge University Press.

Harvey, A.C. (1993.), Time Series Models, MIT Press.

Harvey, A.C. (2013), Dynamic Models for Volatility and Heavy Tails, Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag.

Herbst, E. and Schorfheide, F. (2015), Bayesian Estimation of DSGE Models, Manuscript.

Kim, C.-J. and Nelson, C.R. (1999), State-Space Models with Regime Switching, MIT Press.

Koop, G. (2004), Bayesian Econometrics, John Wiley.

Nerlove, M., Grether, D.M., Carvalho, J.L. (1979), Analysis of Economic Time Series: A Synthesis, Academic Press.

Priestley, M. (1981), Spectral analysis and Time Series, Academic Press.

Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis, Chapman and Hall.

Whittle, P. (1963), Prediction and Regulation by Linear Least Squares Methods, University of Minnesota Press.

Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics, John Wiley and Sons.

Appendix B

Elements of Continuous-Time Processes

B.1 DIFFUSIONS

A key general reference is Karatzas and Shreve (1991).

Continuous-time white noise with finite variance is hard to define. (Why? See, for example, Priestley, 1980, pp. 156-158). Continuous-time analogs of random walks are easier to define, so they wind up playing a more crucial role as building blocks for continuous-time processes.

A *diffusion* is a process with Markovian structure and a continuous (but non-differentiable) sample path. Some important special cases:

• Standard Brownian motion

$$dx = dW,$$

where

$$W(t) = \int_0^t \varepsilon(u)du$$

(that is, it is an *additive process*). W(t) is the continuous-time analog of a discrete-time driftless Gaussian random walk. Intuitively, the normality arises from central-limit considerations stemming from the additive nature of the process. A key property of Brownian motion is its independent Gaussian increments,

$$(W(t) - W(s)) \underset{\sim}{\overset{iid}{}} N(0, t-s), \forall 0 \leq s \leq t \leq \infty$$

Brownian motion is fundamental, because processes with richer dynamics are built up from it, via location and scale shifts. "W" stands for "Wiener process." Standard Brownian motion is the simplest example of the slightly more general Wiener process.

• Wiener process. Standard Brownian motion shifted and scaled.

$$dx = \alpha\ dt + \sigma\ dW.$$

Figure: Gaussian random walk with drift, optimal point and interval forecasts

Wiener process arises as the continuous limit of a discrete-time binomial tree. Discrete periods Δt . Each period the process moves up by Δh w.p. p, and down by Δh w.p. 1-p. If we take limits as $\Delta t \to 0$ and adjust Δh and p apropriately (as they depend on Δt), we obtain the Wiener process. Useful for simplified derivatives pricing, as in Cox, Ross and Rubinstein (1979, *JFE*).

• Wiener process subject to reflecting barriers

$$dx = \alpha\ dt + \sigma\ dW.$$

s.t. $|x| < c$
Figure: Gaussian random walk with drift subject to reflecting barriers.
Stationary distribution exists. Symmetry depends on whether drift exists, and if so, on the sign of the drift.
Process arises as the continuous limit of a discrete-time binomial tree subject to reflecting barriers.
Discrete periods Δt . Each period the process moves up by Δh w.p. p, and down by Δh w.p. 1-p, except that if it tries to move to c or -c, it is prohibited from doing so.

• Ito process

$$dx = \alpha(x,t)\ dt + \sigma(x,t)\ dW$$

An important generalization of a Wiener process.

• Geometric Brownian motion

$$dx = \alpha\ x\ dt + \sigma\ x\ dW.$$

Simple and important Ito process.
Figure: Exp of a logarithmic Gaussian random walk with drift, optimal point and interval forecasts

• Orstein-Uhlenbeck process

$$dx = (\alpha + \beta x)dt + \sigma dW.$$

Simple and important Ito process. Reverts to a mean of -α/β. Priestley (1980) shows how it arises as one passes to continuous time when starting from a discrete-time AR(1) process.
Figure: Gaussian AR(1) with nonzero mean, optimal point and interval forecasts

• The square-root process of Cox, Ingersoll and Ross (1985, *Econometrica*):

$$dx = (\alpha + \beta x)dt + \sigma\sqrt{x}dW.$$

This is an important example of an Ito process – this time heteroskedastic, as the variance depends on the level.

• Generalized CIR process (Chan *et al.*, *JOF* 1992; Kroner *et al.*) [1]

$$dr = (a + \beta r)dt + \psi r^{\gamma} dW$$

Discrete approximation:

$$\triangle r_t = a + \beta r_{t-1} + \varepsilon_t \Rightarrow r_t = a + b r_{t-1} + \varepsilon_t$$
$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$
$$h_t = \psi^2 r_{t-1}^{2\gamma},$$

where

$$b \equiv (1 + \beta)$$

.

More precisely, let $r_{1,t}$ denote a "1-aggregated" series at time t. Then

$$r_{1,t} = a + b r_{1,t-1} + \varepsilon_{1,t}$$
$$\varepsilon_{1,t} | \Omega_{t-1} \sim N(0, \psi^2 r_{1,t-1}^{2\gamma}).$$

[1] We change the notation from x to r, in keeping with the fact that the models are commonly used for interest rates.

Back substitution gives:

$$r_{1,t} = a\sum_{i=0}^{h-1} b^i + b^h r_{1,t-h} + \sum_{i=0}^{h-1} b^i \varepsilon_{1,t-i}.$$

Thus the h-aggregated series follows:

$$r_{h,t} = a\sum_{i=0}^{h-1} b^i + b^h r_{h,t-1} + \varepsilon_{h,t}$$

$$\varepsilon_{h,t}|\Omega_{t-1} \sim N\left(0, \sum_{i=0}^{h-1} b^{2i} var(\varepsilon_{1,t-i})\right),$$

where

$$var(\varepsilon_{1,t-i}) = \psi^2 r_{1,t-i-1}^{2\gamma}.$$

Note that, although the "discretization interval" must be set by the investigator, and is therefore subject to discretion, from that point on the parameter estimates are (asymptotically) invariant to the data recording interval.

• Diffusion limit of GARCH (Nelson, 1990; Drost-Werker, 1996)

$$dr_t = \sigma_t dW_{pt}$$
$$d\sigma_t^2 = \theta(\omega - \sigma_t^2) + \sqrt{2\lambda\theta}\sigma_t^2 dW_{st}$$
$$\omega > 0, \theta > 0, 0 < \lambda < 1, W_{pt} indepof W_{st}$$

Drost and Werker (1996) show that approximate discretizations are available, which follow weak-GARCH processes and are therefore closed under temporal aggregation. They provide formulae for the continuous-time coefficients in terms of the discrete weak-GARCH coefficients at any aggregation level. Makes for a tidy framework bridging continuous and discrete time. See also Andersen and Bollerslev (1998).

• A Poisson Jump Diffusion

It can be shown that *any* additive process can be written as the sum of a Wiener process and a "jump" process. "Jump diffusion." In many aplications the jump part is absent. Here we consider the opposite case, pure jump diffusion driven by Poisson jumps,

$$dx = \alpha(x,t)\ dt + \sigma(x,t)\ dP$$

where $dP = \begin{cases} 0, w.p.1 - \lambda dt \\ \\ u, w.p.\lambda dt \end{cases}$

and u, is a jumpsize, which can itself be a random variable.

This is an example of a non-standard Brownian motion, which is to say that the increments are not Gaussian.

Ito's Lemma

Let F(x, t) be a function of a diffusion, at least twice differentiable in x and once in t. Then:

$$dF = \left[\frac{\partial F}{\partial t} + \alpha(x,t)\frac{\partial F}{\partial x} + \frac{1}{2}\ \sigma^2(x,t)\frac{\partial^2 F}{\partial x^2}\right] dt + \sigma(x,t)\frac{\partial F}{\partial x} dW$$

Ito's Lemma is central because we often need to characterize the diffusion followed by a function of an underlying diffusion, as in derivatives pricing.

As an example, suppose that x follows a geometric Brownian motion. Then a simple application of Ito's Lemma reveals that lnx follows the simple Wiener process:
$dF = (\alpha - \frac{1}{2}\sigma^2)\ dt + \sigma\ dW$,
where F = lnx.
Hence lnx is the continuous time limit of a logarithmic Gaussian random walk.

B.2 JUMPS

B.3 QUADRATIC VARIATION, BI-POWER VARIATION, AND MORE

B.4 INTEGRATED AND REALIZED VOLATILITY

B.5 REALIZED COVARIANCE MATRIX MODELING IN BIG DATA MULTIVARIATE ENVIRONMENTS

B.6 EXERCISES, PROBLEMS AND COMPLEMENTS

1. A key problem in nonparametric estimation.
   Estimate the drift function $f(t)$ in:

$$dy_t = f(t)dt + \frac{1}{\sqrt{N}}dW, \quad t \in [0,1]$$

   consistently $(N \to \infty)$.

B.7 NOTES

Appendix C

Seemingly Unrelated Regression

$y_{it} = X'_{it}\beta^i + \varepsilon_{it}$

$cov(\varepsilon_{it},\, \varepsilon_{jt}) = \sigma_{ij},\ \Sigma = [\sigma_{ij}]$

$i = 1, ..., N;\ t = 1, ...T$

Matrix form: $y^i = X^i\beta^i + \varepsilon^i,\ i = 1, ..., N$

Stacked version:

$$\begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix} = \begin{pmatrix} X^1 & & & \\ & X^2 & & 0 \\ & & \ddots & \\ 0 & & & X^N \end{pmatrix} \begin{pmatrix} \beta^1 \\ \vdots \\ \beta^N \end{pmatrix} + \begin{pmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^N \end{pmatrix}$$

$y = X\beta + \varepsilon$

$cov(\varepsilon) = \Sigma \otimes I \equiv \Omega$

$$\hat{\beta}_{SUR} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}y$$

Bibliography

Aldrich, E.M., F. Fernndez-Villaverde, A.R. Gallant, and J.F. Rubio-Ramrez (2011), "Tapping the Supercomputer Under Your Desk: Solving Dynamic Equilibrium Models with Graphics Processors," *Journal of Economic Dynamics and Control*, 35, 386–393.

Aruoba, S.B., F.X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song (2013), "Improving GDP Measurement: A Measurement Error Perspective," Working Paper, University of Maryland, Federal Reserve Board, and University of Pennsylvania.

Nerlove, M., D.M. Grether, and J.L. Carvalho (1979), *Analysis of Economic Time Series: A Synthesis*. New York: Academic Press. Second Edition.

Ruge-Murcia, Francisco J. (2010), "Estimating Nonlinear DSGE Models by the Simulated Method of Moments," Manuscript, University of Montreal.

Yu, Yaming and Xiao-Li Meng (2010), "To Center or Not to Center: That is Not the Question An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency," Manuscript, Harvard University.